Interprétation des essais cliniques pour la pratique médicale

Michel Cucherat

- *Accueil*

- *Contexte général*
  - *Introduction*
  - *Pourquoi vouloir des preuves scientifiques de l'efficacité d'un traitement*
  - *Les critères scientifiques de démonstration de l'efficacité des traitements*

- *Lecture critique*
  - *Les principes de la mise en pratique des résultats des essais thérapeutiques*
  - *Démonstration de l'efficacité*
  - *Evaluation de la validité interne*
  - *Evaluation de la pertinence clinique*

- *Méthodologie des essais thérapeutiques*

  - *Principes généraux de la méthodologie*
    - *Les biais*
    - *Groupe contrôle*
    - *Randomisation*
    - *Double aveugle*
    - *Au total*
  - *Etude empirique des biais*
  - *Plan factoriel*
  - *Cross over*
  - *Autres plan d'expérience*
  - *Le contrôle du biais d'attrition : Analyse en intention de traiter et remplacement des données manquantes*
  - *Flux de patients*

- *Concepts statistiques*
  - *Le test statistique*
  - *Multiplicité des comparaisons statistiques*
  - *Le critère de jugement principal*
  - *Analyses en sous groupes*
  - *L'intervalle de confiance*
  - *Les analyses intermédiaires*
  - *Ajustement statistique*
  - *Puissance et calcul du nombre de sujets nécessaires*
  - *Test unilateral - bilateral*
  - *Parallèle entre test statistique et test diagnostique*
  - *Courbes de survie*

- *Méta-analyse*
  - *Principes généraux*
  - *Apports*
  - *Biais de publication*
  - *Hétérogénéité, modèle aléatoire et méta-analyse en sous groupe*
  - *Méta-régression*
  - *Méta-analyse sur données individuelles*
  - *Lecture critique*

- *Essai de non infériorité*
  - *Principe général*
  - *Placebo putatif*
  - *Lecture critique d'un essai de non infériorité*

- *Pertinence clinique*
  - *Indices d'efficacités pour critères binaires*
    - *Risque relatif*
    - *Odds ratio*
    - *Différence des risque et NNT*
    - *Comparaison des indices*
    - *Autres indices*
  - *Indices d'efficacités pour critères continus*
  - *Pertinence des critères de jugement*
    - *Critère clinique intermédiaire de substitution*
    - *Les différents types de critères*
      - *Les événements cliniques*
      - *Critère composite*
      - *Les scores et les échelles*
      - *Les durées*
      - *Les mesures quantitatives*
  - *Pertinence des populations incluses*
  - *Balance bénéfice risque*

Interprétation des essais cliniques pour la pratique médicale

Michel Cucherat
Faculté de médecine Laennec - Lyon

Les résultats des essais thérapeutiques et des méta-analyses prennent une place de plus en plus importante dans la médecine actuelle. Ils présentent de répondre de façon fiable aux questions thérapeutiques se posant dans la prise en charge des malades. Ils permettent de valider les soins proposés aux patients et donnent les preuves de leur efficacité.

Tirer les conséquences pour la pratique d'un résultat d'essais thérapeutique ou de méta-analyse n'est pas un processus trivial, naturellement maîtrisé par tout un chacun. Ce processus fait appel à des concepts de statistique et de méthodologie qui ne sont enseignés que de façon récente durant les études médicales. De plus il existe une activité de recherche intense dans ce domaine avec une évolution rapide des idées et des outils. Il existe donc un réel besoin d'un document de synthèse actualisé présentant l'ensemble des concepts et de la démarche qui permet de tirer les conséquences pour la pratique des données brutes présentes dans les publications d'essais cliniques et de méta-analyse.

Ce document électronique a pour objectif de combler ce manque en présentant de manière didactique les concepts impliqués dans la mise en œuvre dans la pratique médicale des résultats des essais thérapeutiques. Il détaille le processus qui partant des résultats des essais cliniques permet la construction de référentiels thérapeutiques ou de recommandations pour la pratique en appliquant les critères scientifiques de l'évaluation des traitements.

Il s'adresse principalement aux cliniciens, aux enseignants-chercheurs, aux formateurs et aux décideurs de santé publique impliqués dans la réalisation des stratégie thérapeutique et des recommandations pour la pratique qui souhaitent faire une analyse critique des publications d'essais thérapeutiques et de méta-analyses, dans le but de décider s'il y a lieu de modifier les pratiques en fonction des résultats de la recherche thérapeutique.

CONTEXTE GENERAL

Introduction

Les résultats des essais thérapeutiques et des méta-analyses prennent une place de plus en plus importante dans la médecine actuelle. Ils présentent de répondre de façon fiable aux questions thérapeutiques se posant dans la prise en charge des malades. Ils permettent de valider les soins proposés aux patients et donnent les preuves de leur efficacité. Ainsi ils permettent de répondre à un impératif éthique de l'exercice médical qui est de proposer au patient le meilleur traitement en fonction des données acquises de la science comme le précise l'article 32 du code de déontologie (« …, le médecin s'engage à assurer personnellement au patient des soins consciencieux, dévoués et fondés sur les données acquises de la science, … ».).
De ce fait l'essai thérapeutique est un outil, mis à la disposition des praticiens pour faire progresser leur pratique thérapeutique. Il est donc nécessaire, pour tout médecin, qui veut rester maître de ses choix thérapeutiques, de connaître la démarche nécessaire à leur mise en pratique. Les essais apportent aussi le substratum rationnel des recommandations de pratiques en leur donnant des fondements basés sur des critères scientifiques.

Tirer les conséquences pour la pratique d'un résultat d'essais thérapeutique ou de méta-analyse n'est pas un processus trivial, naturellement maîtrisé par tout un chacun. Ce processus fait appel à des concepts de statistique et de méthodologie qui ne sont enseignés que de façon récente durant les études médicales. De plus il existe une activité de recherche intense dans ce domaine avec une évolution rapide des idées et des outils. Il existe donc un réel besoin d'un document de synthèse actualisé présentant l'ensemble des concepts et de la démarche qui permet de tirer les conséquences pour la pratique des données brutes présentes dans les publications d'essais cliniques et de méta-analyse.
Par exemple, tous les guides de construction de recommandations pour la pratique (comme celui de l'ancienne ANAES) ou les grilles d'évaluation des recommandations [1] insiste sur la nécessité que les énoncées de recommandations s'appuient principalement les résultats de l'évaluation scientifique. Mais il n'existe très peu de guide pour ce processus de dérivation d'énoncée de recommandation à partir des résultats des essais cliniques et des méta-analyses. Le texte le plus aboutit est un document australien du National Health and Medical Research Council (l'équivalent de notre Haute Autorité de Santé), intitulé « How to use the evidence: assessment and application of scientific evidence » (http://www.nhmrc.gov.au/publications/synopses/cp69syn.htm).
Il n'existe rien de substantiel en langue Française. Cette étape est évoquée en quelques lignes (cf. §II.3.1) dans le document « les recommandations pour la pratique clinique - base méthodologique pour leur réalisation en France » de l'ANAES (http://www.has-sante.fr/anaes/Publications.nsf/wEdition/RA_APEH-3YJ9MJ).
Ce document électronique a pour objectif de combler ce manque en présentant de manière didactique les concepts impliqués dans la mise en œuvre dans la pratique médicale des résultats des essais thérapeutiques. Il détaille le processus qui partant des résultats des essais cliniques permet la construction de référentiels thérapeutiques ou de recommandations pour la pratique en appliquant les critères scientifiques de l'évaluation des traitements.

Il s'adresse principalement aux cliniciens, aux enseignants-chercheurs, aux formateurs et aux décideurs de santé publique impliqués dans la réalisation des stratégie thérapeutique et des recommandations pour la pratique qui souhaitent faire une analyse critique des publications d'essais thérapeutiques et de méta-analyses, dans le but de décider s'il y a lieu de modifier les pratiques en fonction des résultats de la recherche thérapeutique.

Bibliographie

1. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Qual Saf Health Care 2003;12(1):18-23. *PMID: 12571340*.

Pourquoi vouloir des preuves scientifiques de l'efficacité d'un traitement

Introduction

L'inscription de l'activité de soin dans le paradigme scientifique vise à limiter l'influence de des déterminants arbitraires dans le choix des traitements : c'est la ***confrontation à la réalité*** *qui* ***corrobore ou infirme les hypothèses thérapeutiques****.*

L'obtention de preuve est ainsi essentielle pour éviter la situation inéquitable où un bénéfice (commercial ou autre) serait obtenu par le promoteur du traitement sans qu'il soit assuré que le patient ou la société en tire un quelconque bénéfice.

Dans ce cadre, l'utilisation en pratique courante d'un traitement nécessite de disposer des preuves rigoureuses et impartiales qu'il permet effectivement d'atteindre l'objectif thérapeutique pour lequel son utilisation est envisagée (c'est-à-dire le bénéfice clinique que l'on souhaite apporté au patient : guérison, éviter le décès prématuré, etc.). L'obtention de ces preuves est indispensable pour que la pratique médicale thérapeutique s'inscrive dans le paradigme scientifique. Mais qu'est ce que la science ?

Qu'est ce que la science ?

Un des principes de la méthode scientifique est de ne prendre en compte que des faits vérifiés.

Une partie des considérations épistémologiques modernes a été proposée par le philosophe Karl Popper (voir encadré). L'exposé de ces principes et de leur justification permet de mieux comprendre certains principes de la méthode des essais cliniques. L'exposé épistémologique que nous faisons ici est basique. L'objectif n'est pas de faire un panorama exhaustif des considérations modernes de la philosophie des sciences, mais simplement de montrer qu'un certain nombre de principes méthodologiques sont simplement la transcription au domaine des essais cliniques de fondements scientifiques plus généraux [1Ê, 2].

Un fait scientifique est une hypothèse qui a été corroborée par sa confrontation à la réalité dans une expérience spécifique

Les faits scientifiques sont des faits qui ont été vérifiés dans une confrontation avec la réalité. L'intelligence humaine est capable d'échafauder de nombreuses théories spéculatives pour expliquer les phénomènes issus de l'observation de notre univers. Même si ces théories sont extrêmement attrayantes pour l'esprit humain, elles peuvent ne pas être exactes. Ce sont des théories métaphysiques. La démarche scientifique évite d'être spéculative en ne se fondant que sur des faits réels, vérifiés. La science procède par déduction, par test d'hypothèse. Une théorie n'est retenue que si elle a fait l'objet de test et si elle a été suffisamment corroborée par l'expérience.

Une petite introduction à la philosophie de Karl Popper

Les idées du philosophe Karl Popper en épistémologie ont fortement marqué la science du milieu du 20ème siècle. Dans son livre La logique de la découverte scientifique Karl Popper ébauche une théorie des sciences [3]. Qu'est une théorie scientifique ? Quels sont les critères qui permettent de définir la science ?

Il récuse l'induction comme démarche scientifique. L'induction consiste à avancer une loi générale à partir de l'observation d'un nombre plus ou moins grand de cas particuliers. Dès le 18ème siècle, le philosophe anglais David Hume avait attiré l'attention sur les faiblesses de ce raisonnement.

La conclusion inductive n'est pas logiquement contraignante. La conclusion peut être fausse alors que les prémisses sont parfaitement exactes. Ce n'est pas parce que nous n'avons observé que des cygnes blancs que tous les cygnes sont blancs. Des prémisses exactes (nous n'avons observé que des cygnes blancs) n'induisent pas forcément une conclusion exacte. La conclusion fort générale « tous les cygnes sont blancs » peut être fausse bien que les prémisses fussent parfaitement exactes « nous ,

n'avons effectivement observé jusqu'à présent que des cygnes blancs ». Mais il se peut très bien que, quelque part, existent des cygnes noirs, que nous n'avons pas encore eu l'occasion d'observer.

L'intelligence humaine peut concevoir une infinité de conjectures de toute nature, de théories. Pour Popper c'est le domaine de la métaphysique, qui repose sur des suppositions librement crées par l'esprit, en particulier par induction. L'objectif de la science est de rechercher des théories vraies (la vérité scientifique) c'est-à-dire des théories qui correspondent à la structure de la réalité.

Ainsi, une fois énoncées, les théories spéculatives doivent être confrontées rigoureusement et impitoyablement à l'observation et à l'expérience. Il faut éliminer les théories incapables de résister aux tests de l'observation ou de l'expérience et les remplacer par d'autres conjectures spéculatives. La science progresse par essais et erreurs, par conjectures et réfutations. Seules les théories les mieux adaptées survivent. On ne s'autorisera jamais à dire d'une théorie qu'elle est vraie, mais seulement qu'elle est corroborée et donc crédible car elle n'a jamais été jusqu'à présent contredite par l'expérience.

Cette démarche de confrontation à la réalité est en fait un processus de déduction logique qui, contrairement à l'induction, est toujours logiquement vrai. Il repose sur le *modus tollens* : si p est déductible de t et si p est faux alors t aussi est faux. Une conclusion fausse, c'est-à-dire l'observation d'un fait en contradiction avec la conclusion, implique obligatoirement et sans exception l'inexactitude des prémisses. Si nous faisons l'hypothèse que tous les cygnes sont blancs, l'observation d'un seul cygne noir entraîne la fausseté de l'hypothèse : l'hypothèse est alors réfutée.

Pour Popper le critère permettant de caractériser un énoncé scientifique, qu'il appelle critère de démarcation (entre science et métaphysique) est sa réfutabilité (« *falsifiability* »). La réfutabilité est la possibilité de soumettre l'énoncé à une épreuve logique de réfutation déductive. S'il est impossible de concevoir une expérience pouvant amener à la réfutation de l'énoncé, celui-ci n'appartient pas au domaine de la science mais seulement à celui des possibilités de conception abstraite de l'esprit humain, à la métaphysique.

Le constat que seule la déduction est logiquement satisfaisante débouche sur une asymétrie cruciale. On ne peut jamais prouver qu'une théorie est vraie. On peut seulement prouver qu'une théorie est fausse car il suffit pour cela d'une seule observation qui la contredise. Lorsqu'une théorie se soumet à un test de réfutation sans être réfutée, un scientifique considérera qu'elle est partiellement confirmée et lui accordera une crédibilité plus grande. Les non réfutations ne font que corroborer une théorie sans jamais la démontrer formellement. Mais plus une théorie a été soumise à réfutation sans être réfutée, plus elle devient crédible.

Karl Popper n'est pas le seul philosophe ou penseur des sciences qui soit arrivé à la conclusion de la nécessité de l'approche hypothético-déductive en sciences. Par exemple, le statisticien Pearson arriva par des voies différentes au même principe. Popper est cependant celui qui décrivit avec le plus de détails la méthode hypothético-déductive et qui donna les raisons logiques les plus rigoureuses pour l'adopter.

Application à la thérapeutique

Pour s'inscrire dans le paradigme scientifique, l'énoncé qu'un traitement apporte un bénéfice doit être basé sur les résultats de la confrontation de cette hypothèse à la réalité par une expérience, et non pas n'être qu'une affirmation conceptuelle issue d'un raisonnement inductif.

Les sciences biologiques fondamentales et l'épidémiologie nous permettent de connaître de mieux en mieux les mécanismes des maladies et les mécanismes d'action des traitements. Ces connaissances fondamentales permettent d'imaginer qu'un certain traitement apportera un bénéfice clinique dans une certaine maladie. Mais tant que cette hypothèse n'a pas été vérifiée, elle reste théorique et spéculative. Ce n'est pas parce qu'elle se fonde sur des faits scientifiques prouvés que la conclusion générique : « le traitement apporte un bénéfice clinique », est elle aussi prouvée. Nous verrons par la suite plusieurs exemples illustrant ce point (dans la section « Comment obtenir des preuves fiables »).

L'évolution actuelle de la médecine vers le paradigme scientifique conduit naturellement à exiger des traitements qu'ils apportent la preuve de leur efficacité clinique.

> Pour certains, ce principe, suivant lequel l'utilisation d'un traitement dans une pathologie doit trouver sa justification au niveau de faits scientifiques, est le manifeste de la médecine factuelle (« *evidence based medicine* ») [4].

Justification par l'exemple

Il existe de nombreux exemples de traitements qui ont diffusé dans la pratique médicale sans attendre les preuves de leur efficacité clinique et qui se sont avérés sans effet, voir délétères, quand les résultats des essais cliniques appropriés furent disponibles. Le Tableau 1 donne la liste des exemples les plus marquants.

Tous ces exemples représentent autant de justifications empiriques de l'impérative nécessité de disposer des résultats des essais cliniques avant de recommander l'utilisation d'un nouveau traitement dans une situation donnée.

Tableau 1 – Exemple de traitement couramment utilisé sans preuves de leur efficacité clinique et qui se sont avérés inefficace ou délétères lorsque les résultats des essais furent disponibles

Traitement	**Idée préconçue de l'intérêt du traitement**	**Résultats des essais thérapeutiques**
Anti arythmique de classe 1c en post infarctus	Prévention de la mort subite en post infarctus	Augmentation de la mortalité [5]
Traitement hormonal substitutif de la ménopause	Prévention des pathologies coronariennes chez la femme ménopausée	Augmentation de la fréquence des pathologies coronariennes [6]

Pourquoi rechercher les preuves

Pour la pratique médicale, la finalité de l'interprétation des résultats d'essais thérapeutiques est d'évaluer de manière critique un résultat avant de le mettre en application. C'est pour cette raison que l'interprétation est parfois appelée lecture critique. Pour un praticien, il s'agit de répondre à la question « le bénéfice apporté par ce traitement est-il suffisamment établi et cliniquement pertinent pour justifier son utilisation » [7].

Qui est concerné ?

Avant d'aller plus loin et d'aborder le vif du sujet, il est licite de se poser la question « Qui est concerné par la lecture critique et l'interprétation des essais thérapeutiques ? »

L'importance grandissante que prennent les résultats des essais cliniques dans les décisions thérapeutiques suggère que tout médecin prescripteur est concerné par l'interprétation de leurs résultats. En effet, ces informations sont maintenant à la base des arguments apportés aux médecins pour qu'ils réactualisent leur pratique thérapeutique. Pour garder une indépendance intellectuelle, le médecin doit alors être capable d'analyser ces arguments, de les interpréter pour adapter sa pratique en toute connaissance de cause.

Une autre réponse possible serait de dire que la lecture critique et l'interprétation des essais thérapeutiques est une affaire de spécialistes. Le médecin praticien n'a pas les compétences pour le faire et c'est à d'autres de l'effectuer pour lui. Le résultat de ce travail lui est alors transmis sous la forme de sources secondaires, comme des guides de pratique. Malheureusement cette proposition n'est pas entièrement satisfaisante à l'heure actuelle, en particulier à cause de la difficulté de maintenir les recommandations à jour au fur et à mesure de la publication de nouveaux résultats scientifiques.

Suite à la publication des résultats d'un essai thérapeutique, de nombreux documents d'interprétation, de synthèse, de recommandation pour la pratique sont diffusés à destination des prescripteurs, sous des formes variées : publicité, articles de synthèse publiés dans des revues professionnelles, recommandations officielles ou de sociétés savantes, avis d'expert, etc…

Nous allons voir que ces sources d'informations secondaires sont susceptibles de distordre la réalité en raison de conflits d'intérêts ou d'un manque de compétences méthodologiques qui conduisent à des interprétations partisanes des faits scientifiques disponibles.

Ainsi, pour éviter de se faire abuser, et pour séparer le bon grain de l'ivraie, le médecin doit posséder les compétences nécessaires à l'interprétation correcte des faits produits par les essais. Cela est une condition nécessaire au maintient de son indépendance dans ces choix thérapeutiques.

Tableau 5 – Sources secondaires proposées aux médecins pour prendre connaissance des évolutions thérapeutiques

• Revues de la littérature (publiées ou présentées dans des congrès) • Article de synthèse • Éditorial • Publicité de l'industrie pharmaceutique • Recommandation officielle (référence médicale, RMO) • Formation médicale continue	• Recommandations de sociétés savantes • Avis d'expert (spécialiste, hospitalo-universitaire) • Vidal ou Résumé des caractéristiques du produit • Presse et média grand public

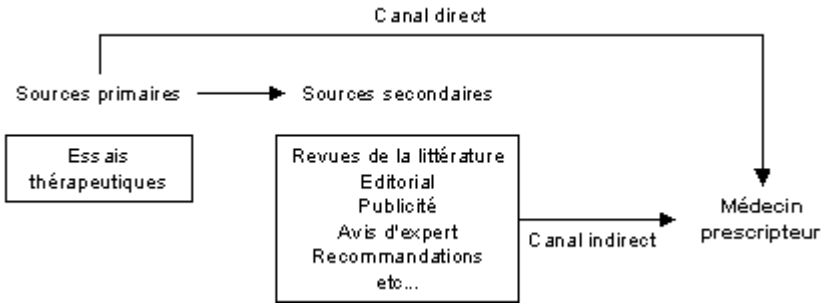


Figure 1 – Les différentes sources d'informations sur les évolutions thérapeutiques auxquelles le médecin prescripteur a accès. Le canal direct nécessite de faire la lecture critique et l'interprétation des résultats. Les sources secondaires se proposent de le faire pour les médecins mais exposent au risque de distorsion du message véhiculé.

Conflits d'intérêts

Les commentaires des leaders d'opinion sont potentiellement influencés par les conflits d'intérêts de leur auteur.

Un travail a comparé les positions prises par différents leaders d'opinion à propos d'une récente polémique concernant les antagonistes calciques en fonction des liens que ces personnes avaient avec des fabricants de ces médicaments [8]. Une méta-analyse et une étude cas-témoin ont suggéré un surcroît d'infarctus du myocarde induit par les antagonistes calciques utilisés dans l'hypertension artérielle ou l'ischémie coronarienne. Ce résultat a donné lieu à la publication de nombreux commentaires les uns le soutenant, les autres le rejetant. Il s'avère cependant que les personnes contestant l'existence d'effets délétères des antagonistes calciques présentent plus fréquemment un conflit d'intérêt (96%) que les personnes ayant un avis neutre (60%) ou défendant le résultat de la méta-analyse (37%). Cette tendance s'avère hautement statistiquement significative ($p<0,001$). Ce travail révèle donc qu'un commentaire de résultats scientifiques est potentiellement soumis aux éventuels conflits d'intérêts de son auteur et nécessite lui aussi une lecture critique.

Retard de prise en compte dans les sources secondaires

Il s'avère que les traités de thérapeutiques n'intègrent qu'avec retard les résultats des essais thérapeutiques [9]. Ce phénomène a été mis en évidence par Antman en comparant les résultats de méta-analyses cumulatives aux recommandations des livres de thérapeutiques [10].

Par exemple, pour la fibrinolyse à la phase aiguë de l'infarctus du myocarde des preuves de son efficacité étaient disponibles en méta-analyse dès 1982 et directement à partir d'un essai de grande taille (GISSI) en 1986. Cependant, en 1987, seulement 5 traités de thérapeutiques sur 10 recommandaient l'utilisation de la fibrinolyse, 4/10 réservaient son usage à des situations particulières, et un n'en parlait pas.

Une situation identique est détectée pour l'injection de lidocaïne en prévention des fibrillations ventriculaires. Aucun des 9 essais réalisés entre 1970 et 1990 n'a mis en évidence de bénéfice. Cependant, en 1989, seulement 10 traités sur 24 excluent ce produit du traitement de la phase aiguë de l'infarctus, alors que 5/24 continuent à le recommander.

Bien d'autres exemples similaires ont été rapportés. L'actualisation des pratiques thérapeutiques, qui ne se baserait que sur les traités de thérapeutiques, n'intégreraient qu'avec retard les dernières données de l'évaluation thérapeutique. Ce retard entraîne une perte de chance pour les patients qui durant cette période ne sont pas traités avec le traitement le plus efficace. Une utilisation directe des résultats des essais pour l'actualisation des pratiques thérapeutiques évite ce retard et limite ses conséquences. En fait, pour la majorité des clinicien l'utilisation est semis-directe (l'article princeps leur est signalé par des revues secondaires spécialisées).

Recommandations

La qualité des recommandations est très variable. Une publication du Lancet a analysé la qualité des recommandations ou des guides de pratique issus des sociétés savantes [11]. La qualité était évaluée en utilisant trois critères : la description des professionnels impliqués dans leur édification, la stratégie utilisée pour identifier les données factuelles primaires et la gradation des recommandations en fonction du niveau de preuve des faits sur lesquels elles s'appuient.

Sur un total de 431 guides de pratique publiés, seulement 5% d'entre elles présentaient tous les critères de qualité ; 67% ne rapportent pas le type de professionnels impliqués, 88% ne donnent pas d'informations sur la méthode employée pour rechercher les sources primaires et 82% ne présentent pas de gradations des recommandations en fonction du niveau de preuve.

La production de recommandations discordantes n'est pas exceptionnelle. L'analyse en 1998 de 20 recommandations disponibles en Grande Bretagne sur le traitement anticoagulant dans la fibrillation ventriculaire révèle de fortes variations dans les conseils donnés [12]. Des différences importantes sont notées au niveau des traitements par age ou des cibles d'INR. Il est donc bon que le médecin praticien ait une certaine visibilité des données sources à travers les recommandations pour qu'il puisse se faire sa propre opinion et garder une certaine maîtrise de la phase de décision.

Comités de lecture

De même il n'est pas possible de se fier aux comités de lecture des revues pour filtrer les résultats pertinents des autres [13-15]. Il a été montré que la publication dans une revue médicale de grande notoriété n'était pas un gage absolu de qualité et de pertinence des résultats [16, 17]. Même si les revues majeures comme le New England Journal Medicine, The lancet, *le* BMJ *ou le* JAMA *publient une majorité d'articles de très grande valeur, elles ne peuvent pas rapporter la totalité des essais décisifs. Bons nombres d'entre eux paraissent aux côtés de résultats d'intérêt moindre dans des revues où le crible est moins performant [18, 19].*

Interprétation globale

Pendant longtemps l'analyse critique n'était envisagée qu'au niveau d'un seul essai. Les essais pris isolément étaient analysés, puis, les éventuelles discordances notées entre les résultats étaient expliquées par une analyse comparative discursive. Il était alors facile de trouver a posteriori des explications ayant trait aux patients ou aux contextes de soins. Nous verrons par la suite les problèmes soulevés par cette approche, en particulier par la non prise en compte de la fluctuation aléatoire des résultats et par la nature exploratoire (post-hoc) de l'explication des différences.

En fait, l'approche doit être plus globale, envisageant simultanément l'ensemble des essais concernant la même question clinique. Le problème thérapeutique (c'est-à-dire la triade traitement - pathologie - type de patients) est à mettre au centre de la démarche qui analyse et interprète toutes les données disponibles, documentant ainsi, avec précision et rigueur, l'efficacité clinique du traitement dans cette situation. La méta-analyse formalise cette approche.

Il s'avère aussi que l'analyse ne doit pas être limitée à un seul traitement mais aussi inclure les traitements concurrents. Cette démarche aboutit alors à la production d'un tableau de bord présentant pour les différentes thérapeutiques concurrentes les résultats comparatifs de leur évaluation.

Intégration des données factuelles dans la pratique médicale

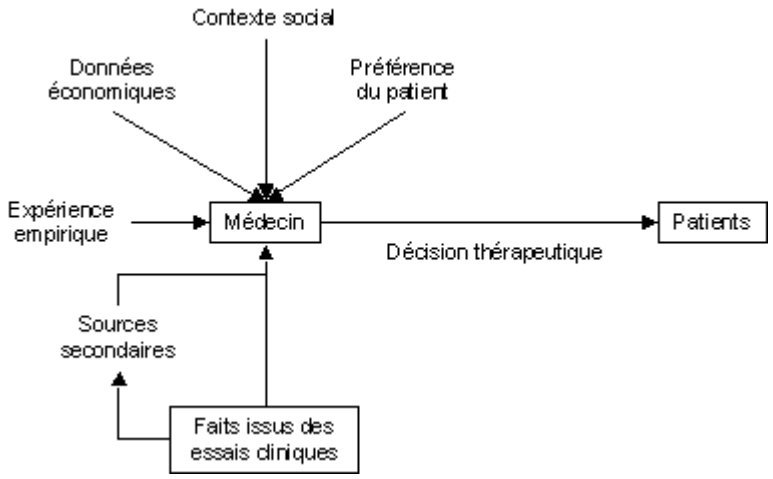
Bien que prépondérants, les résultats des essais cliniques ne constituent pas les seules informations à prendre en compte dans l'élaboration de recommandations pour la pratique ou de guides de décision :

- *les données de pharmacovigilances,*
- *les traitements alternatifs disponibles,*
- *les choix collectifs de politique de santé, le contexte de soins,*
- *les choix sociétaux,*
- *la fréquence de la maladie et l'importance du problème de santé publique qu'elle engendre,*
- *les coûts,*

Au niveau du colloque singulier entre le médecin et le patient, il est évident que d'autres éléments interviennent à coté des résultats des essais, comme :

- *l'applicabilité des résultats au patient (le patient est-il similaire à ceux qui ont été étudiés dans les essais),*
- *l'attente du patient vis à vis de sa prise en charge médicale (objectifs thérapeutiques personnels),*
- *les préférences du patient (et/ou de ses proches),*
- *son profil psycho-affectif (et/ou de ses proches),*
- *le contexte social et d'accès aux soins du patient.*

Les résultats des essais représentent, ainsi, qu'une partie des multiples données que doit intégrer « l'art médical » exercé par le médecin dans sa pratique quotidienne.



En pratique, pour être facilement applicable cette démarche nécessite d'avoir accès facilement aux résultats des essais et à leur méta-analyse. A l'heure actuelle ce scénario est encore légèrement futuriste, mais de nombreuses initiatives sont en train de se mettre en place pour permettre aux praticiens d'accéder facilement et directement à cette « information thérapeutique » représentée par les résultats des essais thérapeutiques

Références

1. *Cucherat M. Karl Popper et la recherche clinique. Rev Prat 2000;50:1286-89.* PMID:
2. *Boissel JP. Note on the epistemology of clinical pharmacology: comparison with the approach of Karl Popper. Therapie 1999;54(1):67-73.* PMID: 10216427.
3. *Popper KR. La logique de la découverte scientifique. Paris: Payot; 1973.*
4. *Evidence-based medicine working group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992;268:2420-2425.* PMID:

5. *Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. N Engl J Med 1991;324(12):781-8.* PMID: 1900101.

6. *Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. Jama 2002;288(3):321-33.* PMID: 12117397.

7. *Bouvenot G, Eschwege E. Essais thérapeutiques. Principes d'interprétation. Rev Prat 1991;41:1853-7.* PMID:

8. *Stelfox HT, Chua G, O'Rourke K, Detsky AS. Conflict of interest in the debate over calcium-channel antagonists. NEJM 1998;338:101-106.* PMID:

9. *Nony P, Cucherat M, Boissel JP. Implication of evidence-based medicine in prescription guidelines taught to French medical students: current status in the cardiovascular field. Clin Pharmacol Ther 1999;66:173-84.* PMID:

10. *Antman EM, Lau J, Kulpelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analysis of randomized control trials and recommendations of clinical experts: treatment for myocardial infraction. JAMA 1992;268:240-248.* PMID:

11. *Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. Lancet 2000;355:103-106.* PMID:

12. *Thomson R, McElroy H, Sudlow M. Guidelines on anticoagulant treatment in atrial fibrillation in great Britain: variation in content and implications for treatment. BMJ 1998;316:509-13.* PMID:

13. *Horton R. The less acceptable face of bias. Lancet 2000;356:959-960.* PMID:

14. *Altman DG. The scandal of poor medical research. BMJ 1994;308:283-84.* PMID:

15. *Adetugbo K, Williams H. How well are randomized controlled trials reported in the dermatology literature? Arch Dermatol 2000;136:381-5.* PMID:

16. *Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. NEJM 1987;317:426-12.* PMID:

17. *DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. NEJM 1982;306:1332-7.* PMID:

18. *Talley NJ, Owen BK, Boyce P, Paterson K. Psychological treatments for irritable bowel syndrome: a critique of controlled treatment trials. Am J Gastroenterol 1996;91(2):277-83.* PMID:

19. *Veldhuyzen van Zanten SJ, Cleary C, Talley NJ, Peterson TC, Nyren O, Bradley LA, et al. Drug treatment of functional dyspepsia: a systematic analysis of trial methodology with recommendations for design of future trials. Am J Gastroenterol 1996;91(4):660-73.* PMID:

Les critères scientifiques de démonstration de l'efficacité des traitements

Qu'est ce qu'une preuve fiable ?

La nécessité de disposer de preuves de l'efficacité clinique d'un traitement étant posée, apparaît alors le problème de définir ce qu'est une preuve fiable du bénéfice clinique apporté par un traitement. Cette analyse permettra d'en déduire les critères scientifiques de l'efficacité clinique des traitements.

Les arguments avancés comme des preuves doivent être d'un haut niveau de fiabilité car de nombreux facteurs favorisent l'apparition et la sélection des faux positifs. Les intérêts en jeux font que dès qu'un « argument » favorable est disponible, même s'il s'agit d'un artefact, et donc d'un résultat erroné, cet « argument » sera retenu et avancé par les promoteurs du traitement. Si aucune précaution n'est prise, des « arguments » seront disponibles même pour des traitements sans efficacité. Les intérêts en jeu conduisent à la mise en avant des arguments faussement en faveur de l'efficacité et à la mise sous silence des arguments contre l'efficacité même si ces derniers sont le reflet de la réalité. Devant cette pression favorisant l'émergence des arguments faussement positifs, il est ***impératif*** *de n'accepter que les arguments qui constituent des* ***preuves quasiment irréfutables du bénéfice*** *apporté par un traitement. Ne peuvent donc être retenues que les preuves les plus fiables, celles qui ont un risque infime d'être erronées (biaisées).*

> *Un argument positif doit toujours être considéré comme douteux car on sait que, si un argument faussement positif apparaît par hasard, il sera utilisé et mis en avant. Dans ce contexte, seul les arguments d'un haut niveau de fiabilité sont probants. Ceux moins robustes et potentiellement sujet aux erreurs (biais) ne sont pas convaincants car rien ne permet d'exclure qu'ils sont positifs à tort.*

Essai explicatif – essai pragmatique

Les essais explicatifs sont des essais thérapeutiques entrepris pour tester des ***hypothèses cognitives sans finalité thérapeutique directe****. Ce sont, par exemple, des essais réalisés pour connaître les mécanismes d'actions des traitements, avant d'envisager leur utilisation en thérapeutique ou pour expliquer a posteriori leur efficacité. L'objectif de ces essais est seulement d'enrichir nos connaissances fondamentales, connaissances qui dans un second temps peuvent déboucher sur la réalisation d'essais pragmatiques à la recherche des preuves de l'intérêt de l'utilisation d'un traitement [1].*

Les essais explicatifs (« explanatory trial *») impliquent en général le recueil de nombreuses d'informations sur un petit nombre de patients. Les essais pragmatiques («* pragmatic trial *») sont conduits sur un grand nombre de patients en focalisant le recueil d'information sur les critères correspondant à l'objectif thérapeutique. Les principes d'interprétation qui sont présentés dans cet ouvrage ne concernent que les essais pragmatiques (ou les essais qui ont une finalité pragmatique) et non pas les essais explicatifs dont l'analyse repose en partie sur d'autres critères.*

Les essais pragmatiques peuvent intégrer une composante explicative, tandis que le contraire n'est pas possible.

Classiquement on distingue l'efficacité (« efficacy *») du bénéfice clinique («* effectiveness *»). L'efficacité peut être définie comme l'effet du traitement sur les processus biologiques, son aptitude à modifier l'état ou le devenir de phénomènes biologiques. Efficacité clinique a plus trait aux conséquences du traitement sur l'état global du patient et intègre plus volontiers les deux dimensions efficacité et sécurité. Le bénéfice clinique ou l'efficacité clinique sont synonymes. Le bénéfice clinique est en rapport direct avec l'objectif thérapeutique envisagé. Lorsque le traitement s'accompagne d'effets délétères ou d'effets indésirables sérieux, l'effet obtenu au niveau de l'objectif thérapeutique doit être pondéré par ces effets négatifs. Le mieux est de disposer dans ce cas-là d'un critère qui intègre à la fois les effets positifs et les effets négatifs.*

En fait, cette distinction est celle qui se retrouve entre critères de jugement cliniques et critères intermédiaires, et entre essai explicatif et essai pragmatique [2]. L'efficacité proprement dite se mesure plutôt avec les critères intermédiaires dans des essais de nature explicative. Le bénéfice clinique repose sur les critères cliniques et sa mesure fait appel à des essais pragmatiques.

Les conceptions ont évolué depuis ces dernières années. L'accent est de plus en plus mis sur l'évaluation de l'efficacité clinique des traitements et sur la nécessité de réaliser des essais pragmatiques pour guider la pratique médicale [3].

Comment obtenir des preuves fiables

L'impérative nécessité de disposer de preuves solides de l'efficacité d'un traitement étant posée, surgit la question de savoir comment les obtenir. Cette quête est semée d'embûches, et il convient de se protéger contre plusieurs phénomènes pouvant conduire à des conclusions erronées, c'est-à-dire à de fausses preuves. Si ces pièges ne sont pas évités, ou le sont de façon insuffisante, il existe un risque d'avancer des arguments erronés comme preuve de l'existence d'une efficacité qui en réalité n'existe pas. Ces pièges sont les nombreux et divers :

- les limites du raisonnement théorique basées sur la physiopathologie,
- la variabilité biologique et les risques statistiques,
- les facteurs de confusion et les biais,
- la sélection des arguments en fonction des résultats,
- le biais de publication.

S'il n'est pas protégé contre ces sources potentielles d'erreur, un argument en faveur de l'efficacité ne présente pas un niveau de fiabilité suffisante pour constituer une preuve [4]. Il est susceptible d'être biaisé par de nombreux phénomènes.

Quelles sont donc les qualités que doivent posséder les arguments en faveur de l'efficacité pour être acceptés comme des preuves suffisamment fiables ?

Limites des raisonnements théoriques

L'utilisation d'un traitement pour une pathologie donnée peut être justifiée de plusieurs manières. Une première approche, traditionnelle, consiste à justifier le traitement par son mécanisme d'action. A partir de la connaissance de la physiopathologie de la maladie et des actions pharmacologiques du traitement, il est possible d'élaborer un raisonnement théorique mécaniste qui laisse présager un bénéfice thérapeutique.

Malgré l'importance de nos connaissances fondamentales, celles-ci restent parcellaires et les raisonnements théoriques spéculatifs. Toutes ces raisons font que les déductions faites à ce niveau ne peuvent garantir l'exactitude du raisonnement et de ses conclusions (le traitement est efficace). Il existe de nombreux exemples où ce type de raisonnement a été pris en défaut et où les prédictions du modèle mécaniste n'ont pas été confirmées dans un essai clinique. Le plus fréquemment un traitement prédit comme efficace s'est avéré sans effet, mais avec parfois, chose plus gênante, la révélation d'un effet délétère.

Ce type de raisonnement n'est évidemment pas strictement théorique. Il se fonde sur des résultats expérimentaux documentant chacune de ces étapes. C'est la combinaison des arguments qui est spéculative comme l'est, par voie de conséquence, l'effet induit. À l'expérience, cette approche apparaît d'une fiabilité imparfaite.

L'essai clinique est une confrontation à la réalité des hypothèses du raisonnement théorique. Son résultat, lorsque l'essai est correctement conçu et réalisé, mesure en conditions réelles le bénéfice apporté par le traitement. Il permet de démontrer que l'utilisation d'un traitement s'accompagne bien d'un bénéfice réel.

Avec certaines limites, ce procédé s'apparente au processus de réfutation et corroboration Poppérienne. Le raisonnement sur les mécanismes d'action est de nature inductive. A partir d'un énoncé singulier : « le traitement a tel effet pharmacologique », il induit un fait plus global : « le traitement apporte un bénéfice clinique ». Des épistémologues, comme Hume et Popper, ont montré les limites de l'induction. Par contre, la vérification de l'hypothèse réalisée par l'essai thérapeutique est de nature déductive et permet de manière fiable de rejeter ou de corroborer l'hypothèse de l'efficacité.

Par exemple, les antiarrythmiques de classe 1 ont été prescrits après infarctus du myocarde en cas d'extrasystolie ventriculaire pour prévenir la mort subite. La valeur péjorative des extra systoles ventriculaires

fréquentes est bien documentée. De même, les antiarrythmiques de classe 1 ont montré qu'ils diminuaient fortement la fréquence des extra systoles. De ce fait, il semblait logique de penser que l'utilisation de ces traitements devait diminuer la mortalité par mort subite. Ces traitements furent utilisés en pratique pendant de nombreuses années, leur usage étant uniquement justifié par ce raisonnement théorique. Ce n'est quand 1991 que l'essai CAST a évalué quels étaient les résultats réellement produits sur la mortalité (Tableau 1) [5].

Les résultats de cet essai furent à l'opposé de ce qui était attendu. Au lieu de confirmer la réduction de mortalité pressentie, il mettait en évidence un doublement de celle-ci. On a pu calculer que la pratique d'utilisation non fondées sur des preuves cliniques de ces médicaments auraient entraîné aux États-Unis au moins autant de décès que les guerres de Corée et du Vietnam [6].

Tableau 1 – Résultats de l'essai CAST de prévention de la mort subite après infarctus par les antiarrythmiques de classe 1.

	DC / n	mortalité
groupe antiarrythmiques	39 / 432	9%
groupe contrôle	18 / 423	4%
RR=2,13, p=0,0004		

Une explication fut rapidement avancée faisant intervenir les effets proarrythmogènes et inotropes négatifs des antiarrythmiques de classe 1. Ces effets pour lesquels il été possible de proposer un mécanisme physiopathologique n'avaient pas été pris en compte dans l'élaboration du modèle thérapeutique théorique alors qu'ils étaient connus. On touche là une limitation des modèles thérapeutiques discursifs : la difficulté de prendre en compte des effets délétères à cause de l'absence d'intégration des aspects quantitatifs de la connaissance.

De très nombreux autres exemples existent comme ceux : des inotropes positifs dans l'insuffisance cardiaque (cf. encadré ci dessous), du traitement hormonal substitutif de la ménopause [7], de la vitamine E (cf. ci-dessous), etc.

En fait tous les médicaments dont le développement a été arrêté lors des essais cliniques de phases 3 constituent un exemple où les hypothèses physiopathologiques n'ont pas été confirmées. Dans d'autre cas, les essais thérapeutiques ont mis en évidence un bénéfice clinique apporté par des traitements pour lesquels les raisonnements théoriques n'en prédisaient pas. C'est par exemple le cas des bêta-bloquants dans l'insuffisance cardiaque (cf. encadré) ou des statines dans la prévention des AVC.

Les connaissances sur les mécanismes d'actions sont indispensables pour la recherche de nouveaux traitements, mais elles ne peuvent pas être utilisées comme preuves ultimes de l'efficacité.

Malgré ses limites, l'étude des mécanismes d'action est indispensable pour imaginer de nouveaux traitements. Cette approche est obligatoire pour rationaliser l'évaluation des traitements. En effet, la connaissance des mécanismes fondamentaux de la physiopathologie et de la pharmacologie est une voie sans égale pour la génération de nouvelles hypothèses thérapeutiques. Sans connaissance fondamentale, comment trouver de nouveaux antiagrégants plaquettaires ? Comment avoir l'idée d'utiliser des fibrinolytiques dans l'infarctus du myocarde ?

Il faut cependant être conscient de ce que dans de nombreux cas, soit les essais n'ont pas confirmé la théorie pourtant séduisante et paraissant robuste, soit après démonstration clinique de l'efficacité du traitement, un autre mécanisme d'action a été trouvé (par exemple les bêta-bloqueurs dans le post infarctus du myocarde). Ces deux points illustrent le fait que, quel que soit le haut degré de connaissances que l'on puisse atteindre sur la physiopathologie des maladies et sur les mécanismes d'action des traitements, la complexité biologique peut mettre en échec une approche purement théorique.

Exemples de mécanismes non confirmés

Vitamine E et prévention cardiovasculaire

Une alimentation riche en vitamine E est associée, dans les études d'observation, avec une faible mortalité cardiovasculaire. Au niveau physiopathologique, les propriétés antioxydantes de la vitamine E et ses actions sur le métabolisme des lipides expliqueraient cette observation. Cependant, malgré 5 essais regroupant 56 505

patients, il n'a pas été possible de mettre en évidence de bénéfice clinique de la vitamine E. Aucune modification n'est observée ni sur la mortalité totale (RR=1.00 ; IC95%=[0.95 ;1.05]), ni sur la fréquence des événements cardiovasculaire mortels ou non mortels (RR=0.98 ; IC95%=[0.93 ;1.02]).

Statines et AVC

Épidémiologiquement, aucune relation entre cholestérolémie et risque d'accidents vasculaires cérébraux mortels n'a été trouvée même dans la méta-analyses des études de cohortes [8]. Cependant, la méta-analyse des essais de statines montrent une réduction statistiquement significative de 31% (odds ratio=0.69, IC95%=0.57;0.83) de la fréquence des accidents vasculaires cérébraux mortels ??? [9].

Bêtabloquant dans l'insuffisance cardiaque

Les bêtabloquants du fait de leur propriété inotrope négative ont été pendant longtemps contre-indiqués dans l'insuffisance cardiaque. L'observation assez fréquente de décompensation cardiaque consécutive à l'administration de bêta-bloquant confirme en pratique cette induction physiopathologique. A l'opposé de ces arguments théoriques et empiriques existait aussi d'autres arguments physiopathologiques faisant penser que les bêta-bloquant pourraient être bénéfiques en diminuant l'hypersensibilité aux catécholamines adrénergiques existante dans l'insuffisance cardiaque. Ainsi au niveau des mécanismes physiopathologiques deux théories radicalement opposées s'affrontaient. Elles ont pu cependant être départagées par des essais de mortalité qui ont montrés une réduction substantielle de mortalité apportée par les bêtabloquants [10, 11].

Insuffisance cardiaque et agents inotropes inhibiteurs de la phophodiestérase

D'après leurs effets pharmacologiques favorables sur des critères physiopathologiques d'ordre hémodynamique, deux produits, l'emoxinome et la vesnarinone, sont apparus comme des traitements qui devaient augmenter la survie des patients insuffisants cardiaques. Ce raisonnement théorique n'a cependant pas été confirmé lors des essais de mortalité où une surmortalité a été observée. De plus dans cet exemple, des critères assez proches de critères cliniques, comme la tolérance à l'exercice ou la qualité de vie, étaient aussi influencés de façon favorable. La seule prise en considération de ces critères qui sont déjà des critères cliniques n'aurait pas permis de mettre en évidence cette surmortalité.

Dans un essai comparant deux doses de vesnarinone, 30 et 60 mg par jour, à un placebo chez des insuffisants cardiaques sévères de classe 3 ou 4 de la NYHA, un surcroît de décès à court et long terme à été observé avec la dose de 60 mg [12]. La mortalité a été de 18,9% dans le groupe placebo et de 22,9% dans le groupe vesnarinone (risque relatif = 1,21). Dans ce même groupe, la qualité de vie (appréciée par le « Minnesota Living with Heart Failure questionnaire ») était améliorée significativement à 8 semaines et à 16 semaines, mais pas à 26 semaines. Les mêmes tendances ont été observées dans le groupe 30 mg sans qu'elles s'avèrent statistiquement significatives.

Un résultat similaire a été observé avec un essai de l'enoximone [13]. Dans huit essais contre placebo de petite taille (entre 10 et 100 patients) et de courte durée (3 à 16 semaines), l'enoximone a été associée à des améliorations statistiquement significatives de la fraction d'éjection, de la symptomatologie ou du stade NYHA, mais aussi à une amélioration de critères cliniques intermédiaires comme la tolérance à l'exercice, la durée de marche [14]. Cependant, dans un essai contre placebo incluant 151 patients souffrant d'une insuffisance cardiaque sévère, un excès de décès a été observé dans le groupe enoximone par rapport au placebo (27 vs 18 décès, p<0,05). La qualité de vie était pourtant significativement améliorée à 2 semaines ainsi qu'un score de mobilité physique après 3 mois. Ce résultat n'est pas isolé et se retrouve dans d'autres essais [15].

Cet exemple illustre une fois de plus le danger qu'il pourrait y avoir à se contenter de la démonstration d'un effet sur un critère intermédiaire physiologique, comme la fonction ventriculaire, à la place d'un essai sur le critère clinique ultime. Il montre aussi que la mise en évidence d'un bénéfice sur des critères cliniques intermédiaires, comme la tolérance à l'effort ou la durée de l'effort, n'est pas non plus suffisante.

La nature probabiliste des phénomènes étudiée comme limites de l'expérience personnelle

L'expérience clinique possède des limites. Dans bon nombre de pathologies, la faible fréquence de survenue des événements et la petitesse des bénéfices attendus font qu'il est impossible de juger à partir de quelques cas de l'efficacité d'un traitement

Tout médecin a déjà fait la constatation que des patients ayant des caractéristiques identiques peuvent avoir des évolutions très différentes. Par exemple, deux patients similaires avec une hypertension artérielle auront

des devenirs différents : l'un présentant un accident vasculaire cérébral très rapidement, l'autre pas. Cette variabilité non réductible, des phénomènes biologiques ne permet pas de raisonner dans un champ déterministe mais seulement en probabilité : une élévation de la pression artérielle ***ne détermine pas*** *la survenue d'un AVC, elle en* ***augmente*** *seulement sa* ***probabilité****.*

Le recours aux probabilités est nécessaire chaque fois que nos connaissances ne nous permettent pas de prédire avec une quasi certitude l'évolution du phénomène étudié comme la survenue d'un événement, la durée d'une maladie, etc. Pour pouvoir se passer des probabilités, il conviendrait d'être dans une situation où l'événement surviendrait systématiquement chez 100% des patients pris en considération. Dans cette situation hypothétique un traitement qui empêcherait ne serait-ce qu'un événement pourrait être considéré comme efficace. Il suffit cependant d'une incertitude sur le diagnostic pour que ces conditions ne soient plus réunies. Il convient alors de travailler sur des groupes.

La nécessité de raisonner en probabilité empêche de pouvoir conclure à partir de l'observation d'un seul individu. Il convient de travailler sur des groupes de patients pour pouvoir mesurer les probabilités et les variations de probabilités avec suffisamment de précision. Chez un hypertendu, la probabilité annuelle d'accidents cardiovasculaires est de 4%. Il est évident que l'observation d'un seul sujet qui présente ou pas un événement ne permet pas de savoir si le traitement a modifié ce risque.

> *La relative rareté des événements et la petitesse des bénéfices attendus font qu'il est impossible de juger de l'efficacité d'un traitement à partir de quelques cas isolés. En effet, est-il possible qu'un médecin perçoive l'effet de la pravastatine, qui réduit la fréquence des événements coronariens de 2,26% au bout de 5 ans (la fréquence des événements coronariens mortels et non mortels passe de 7,5% à 5,3% à 5 ans), à partir de l'observation de quelques dizaines de patients de sa clientèle qu'il aura mis sous ce traitement ?*

La plupart des traitements ne font que modifier la probabilité d'événement sans l'annuler. La mise en évidence de leur efficacité ne peut donc se faire qu'à partir de groupes de patients. L'expérience empirique, que peut avoir un praticien sur une série limitée de ses patients, ne lui permet pas d'appréhender des modifications de probabilité.

L'observation de l'évolution satisfaisante de quelques patients, ne démontre pas l'efficacité. En l'absence de traitement des évolutions favorables sont aussi observées. De plus l'esprit humain a tendance à oublier les mauvaises expériences et à ne mettre en avant que les bonnes. Tous ces phénomènes concourent à fausser favorablement nos impressions subjectives sur l'efficacité jugée par l'expérience personnelle.

Bien que les impressions issues de l'expérience personnelle ne puissent pas être acceptées comme preuve de l'efficacité, elles ont de réelles conséquences psychologiques sur les prescripteurs : la réticence que peut avoir un médecin à utiliser un traitement après avoir fait l'expérience d'un événement indésirable gravissime est bien compréhensible.

Dans tous les cas où le lien de causalité est irréfutable ou quasiment irréfutable parce que la même action est toujours suivie du même effet, le recours aux probabilités est inutile. Il existe, ainsi, en médecine des situations très fortement déterministes où l'outil probabiliste n'est pas nécessaire. Montrer que le sondage vésical soulage sur le champ le patient en rétention aiguë ne nécessite pas de faire appel au raisonnement en termes de probabilité. Il en est de même pour l'effet narcotique d'un produit anesthésique. Par contre l'évaluation de la sécurité de ces deux anesthésiques retombe dans le champ du probable et nécessite l'outil statistique.

Les études d'observation ne contrôlent pas tous les biais

Plusieurs types d'étude de recherche clinique peuvent être envisagés a priori pour réaliser la confrontation à la réalité des hypothèses thérapeutiques. Il s'avère cependant qu'elles ne sont pas toutes équivalentes et qu'elles ne permettent pas toutes d'obtenir des preuves fiables [16] Quels sont les types d'études qui réalisent une confrontation à la réalité suffisamment fiable des hypothèses thérapeutiques ?

> *Toutes les études de recherches cliniques ne permettent pas d'obtenir des preuves fiables de l'efficacité des traitements.*

La simple observation n'est pas suffisante car elle ne permet pas de prendre en compte les ***facteurs de confusion*** *(«* confounding factor *») et ses résultats sont potentiellement biaisés.*

Dans un cadre observationnel, c'est-à-dire dans la vie de tous les jours, les patients qui reçoivent un traitement ne sont pas comparables à ceux qui ne le reçoivent pas. Ils diffèrent sur de nombreux points :

ils sont plus sévèrement malades que les autres ce qui impliquera une plus morbidité chez eux
ils sont un accès aux soins plus facile et seront donc en meilleure santé
ils ont accès à un contexte de soin plus développés
ils ont présenté un échec thérapeutique lors de l'administration des autres traitements
ils ont une mauvaises tolérances aux autres traitements

Ainsi les études d'observation (« observational study *») épidémiologiques ne donnent pas suffisamment de garanties pour fournir les preuves recherchées (le domaine visé par les études d'observation n'est pas l'évaluation des thérapeutiques mais l'étude des déterminants des pathologies. Un exemple particulièrement démonstratif de ces limites est apporté par le traitement substitutif de la ménopause (cf. l'étude de cas si rapportant) ???*

Les études d'observation permettent à ce niveau d'analyser des phénomènes inaccessibles à l'essai randomisé). Par contre elles sont idéales pour générer de nouvelles hypothèses. Pour la recherche d'une efficacité thérapeutique les limites des études d'observation sont énumérées dans le Tableau 2.

Tableau 2 – Les limites des études d'observation pour la recherche d'une efficacité thérapeutique.

Type d'étude	**Limites**
Série de cas (« *case study* »)	Pas de prise en considération des facteurs de confusion.
Étude écologique (« *ecological study* »)	Prise en compte insuffisante des facteurs de confusion : pas de prise en compte des différences géographiques (génomiques, environnementales, etc.).
Étude longitudinale (« *longitudinal study* »)	Prise en compte insuffisante des facteurs de confusion : pas de prise en compte de l'évolution séculaire de la maladie, de l'évolution de ses déterminants, de sa prise en charge, de sa prévention.
Étude cas-témoins (« *case-control study* »)	Biais d'indication. Biais de mémorisation (les sujets atteints se remémorent plus facilement les traitements qu'ils ont pris). En plus d'une prise en compte insuffisante des facteurs de confusion.
Étude de cohortes (« *cohort study* »)	Biais d'indication (les patients reçoivent ou ne reçoivent pas le traitement étudié en fonction de la gravité de leur maladie). En plus d'une prise en compte insuffisante des facteurs de confusion.

L'épidémiologie est un outil indispensable pour étudier les conséquences et surveiller l'utilisation d'un traitement dans une population (pharmaco-épidémiologie).

Les études d'observations sont des outils conçus pour étudier les déterminants des maladies [17] (objectif pour lesquels l'essai randomisé est totalement inadapté) et elles ne sont pas adaptées à la démonstration de l'efficacité d'un traitement.

Par contre, les études d'observation contribuent fréquemment à l'évaluation des effets indésirables des traitements, en particulier lorsque ceux-ci sont rares ou nécessitent de longue période d'exposition pour survenir mais à conditions qu'ils soient nettement distincts des événements survenant spontanément avec la maladie considérée.

La méthodologie des essais thérapeutiques a été développée pour éliminer les facteurs de confusion et les biais. Ainsi les limites de l'observation sont levées par l'essai thérapeutique.

Exemples

Beta-carotène

Un exemple des limitations des études d'observation pour la recherche des effets d'un traitement est donné par l'évaluation des effets du bêta-carotène en prévention. Les données épidémiologiques et biologiques suggèrent que le bêta-carotène, grâce à ses propriétés antioxydantes, est protecteur contre les cancers et les maladies cardiovasculaires.

Les études de cohortes montrent que les sujets consommant le plus de bêta-carotène ont une mortalité cardiovascualire (CV) réduite par rapport à ceux qui en consomment le moins (réduction relative du risque de 31%, IC 95% = [41% ; 20%], P<0.0001) [18]. L'hypothèse issue des études d'observation que le bêta-carotène pouvait réduire la mortalité CV a été testée dans 4 essais thérapeutiques [19-21] (

Figure 1). *Leurs résultats vont à l'encontre de ceux des études épidémiologiques et concluent à une augmentation de mortalité (augmentation relative du risque de décès cardiovasculaire de 12%, IC 95% = [4% ; 22%], P=0.005) [22].*

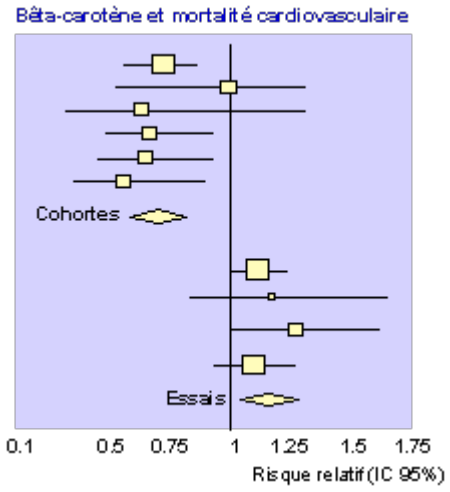


Figure 1 – Comparaison des résultats des études d'observation et des essais randomisés documentant l'efficacité de la vitamine E dans la prévention cardiovasculaire.

Plusieurs hypothèses sont disponibles pour expliquer ces discordances : type de sujets différents, doses différentes, mais surtout, et avant tout, le fait que l'effet observé dans les études épidémiologiques peut être dû à un facteur de confusion [18]. Par exemple, il est possible que la consommation de forte quantité de bêta-carotène soit simplement un marqueur d'une plus forte préoccupation vis à vis de la santé. Les personnes qui apportent le plus de soins à leur santé font d'avantage attention à leur alimentation et ont une alimentation plus diversifiée et plus riche en vitamines. De ce fait, les sujets consommant le plus de bêta-carotène pourraient avoir moins de conduites à risque (tabagisme, alcoolisme), se soumettraient plus à des actions de prévention, auraient une plus grande facilité de recours aux soins, etc. et, finalement, auraient donc un meilleur pronostic cardiovasculaire que ceux consommant moins de bêta-carotène qui prêteraient moins d'attention à leur santé. Même si en fait le bêta-carotène augmente quelque peu leur mortalité, celle-ci se maintient

inférieure à celle des sujets contrôles. Dans les essais randomisés, où tous ces facteurs comportementaux et liés à la catégorie socioprofessionnelle sont également répartis dans les deux groupes, le vrai effet du bêta-carotène peut apparaître. Cette explication n'a pas la prétention d'être « l'Explication » de ces discordances observées entre études d'observation et études randomisées. Elle n'est qu'une possibilité parmi d'autres. Mais elle montre comment une étude d'observation peut être faussée par un facteur de confusion. Les conséquences des facteurs de confusion connus peuvent être corrigées par un ajustement, mais pas ceux des facteurs de confusion non connus ou non enregistrés.

En pratique, si les prescriptions avaient suivi les résultats des études d'observation, sans attendre les résultats des essais, la pratique qui en aurait découlé, aurait eu un résultat inverse à celui recherché et auraient induit des décès.

L'histoire du bêta-carotène est d'ailleurs exemplaire. Les études d'observation ont généré une hypothèse thérapeutique qui a ensuite été testée dans des essais thérapeutiques. Finalement cette piste c'est avérée infructueuse. Dans d'autres cas, les études d'observation ont permis de découvrir des voies thérapeutiques qui se sont, ensuite, avérées très efficaces dans les essais randomisés.

Neuro stimulation et antalgie post chirurgicale

La neuro-stimulation transcutanée donne un autre exemple de l'inadéquation des études d'observation à la démonstration de l'efficacité des traitements. La neuro-stimulation transcutanée est proposée pour le traitement, en autre, des douleurs post-opératoires. La comparaison des résultats obtenus avec ce traitement dans les essais randomisés et dans les études non randomisés (

Figure 2) met en évidence le très grand risque de faux positifs des comparaisons non randomisées [23]. Les études non randomisées sont dans leur très grande majorité en faveur de l'efficacité de la neuro-stimulation avec 17 études positives sur 19 (89%). Cependant, cette efficacité n'est retrouvée que dans 2 essais randomisés parmi 17 (12%).

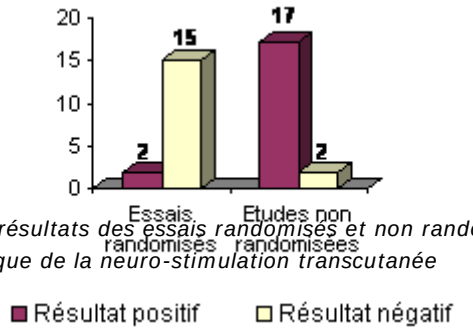


Figure 2 – Comparaison des résultats des essais randomisés et non randomisés dans le cadre de l'évaluation de l'effet antalgique de la neuro-stimulation transcutanée

Ces exemples ne sont pas des cas isolés. Une surestimation de l'effet par les études non randomisée a été retrouvée dans de nombreux domaines [16, 24Ê, 25Ê, 26].

Nécessité de confirmation des résultats

Le résultat d'un seul essai répondant aux critères que nous venons de voir ne constitue cependant pas encore une preuve suffisamment fiable. En effet, malgré une méthodologie irréprochable, le résultat d'un essai unique peut encore être inexact du fait de la présence d'une erreur statistique ou d'un artefact. La nature aléatoire des phénomènes considérés implique un risque d'erreur dans la conclusion sur l'existence de l'effet du traitement qui ne peut pas être éliminé, même s'il est contrôlable grâce au test statistique. Un résultat statistiquement significatif au seuil α=5% laisse la possibilité d'une conclusion fausse dans 5% des cas. De plus, la valeur prédictive positive d'un résultat significatif est variable et dépend de la probabilité à priori de l'hypothèse testée et de la puissance de l'essai (cf. chapitre Tests statistiques*).*

Le résultat d'un essai peut être dû à un artefact. Même avec une méthodologie apparemment irréprochable, il n'est pas possible d'exclure avec certitude l'existence d'un biais ou l'existence de circonstances exceptionnelles conduisant à un résultat positif à tort. Seule la confirmation du résultat par au moins un autre essai permet d'éliminer ces deux possibilités. Déjà en 1983, Zelen attirait l'attention sur le fait que de nombreux éditeurs de journaux étaient opposés à publier des articles qui étaient des confirmations de résultats précédents [27]. Les essais dont le but est de confirmer un résultat antérieur ne sont pas considérés comme aussi excitants ou innovateurs que le premier report d'une avancée thérapeutique. Ce propos est à modérer actuellement. Il n'est plus exceptionnel que des revues comme The Lancet ou le New England Journal of Medicine publie simultanément dans un même numéro deux essais sur la même question (par exemple le dépistage du cancer colique par recherche de sang occulte dans les selles [28, 29] ; ou la comparaison des antibiothérapies orale versus intraveineuse dans les neutropénies des chimiothérapies anticancéreuses [30, 31]).

Il faut cependant noter qu'il existe une réticence franche dans certains domaines à dupliquer les essais. Même au niveau réglementaire, un seul essai est parfois accepté comme preuve de l'efficacité d'un traitement, surtout lorsque la démonstration demande des effectifs importants ou des durées de suivi prolongées. C'est pourtant dans ces situations où les preuves doivent être les plus solides.

Exemple

Dans le cancer colorectal métastasé, un premier essai [32] montrait la supériorité de la triple association irinotecan, fluoro-uracile et leucovorine par rapport au traitement standard fluorouracile + leucovorine en termes de régression tumorale, de survie sans progression, et même de survie totale. À la suite de cet essai, cette trithérapie a été homologuée par la FDA comme traitement de première intention dans le cancer colorectal métastasé et a été utilisée de façon standard par de nombreux cancérologues. Cependant, ces résultats extrêmement favorables, n'ont pas été retrouvés dans deux essais ultérieurs, financés par le National Cancer Institute, qui ont été rapidement arrêtés en raison d'une surmortalité [33].

Cet exemple, illustre que, même en l'absence de tout biais détectable un résultat, obtenu pourtant sur un critère « robuste » comme la mortalité, peut être spécieux et non reproductible.

Multiplication des résultats

La nécessité de vérifier les résultats va entraîner dans de nombreux cas une multiplication des résultats à prendre en compte avant de se déterminer sur l'efficacité du traitement. Cette multiplicité de l'information entraîne à son tour de nouvelles difficultés : comment réaliser une synthèse de résultats qui peuvent être apparemment discordants car soumis aux risques d'erreurs statistiques a et β. Ces difficultés sont résolues par la méta-analyse.

Sélection arbitraire et biais d'information

Un autre écueil à éviter dans la recherche des preuves est celui de la sélection arbitraire des essais utilisés comme argument en fonction de leurs résultats. Assez fréquemment, les synthèses réalisées sous forme de revue générale ne retiennent que les résultats positifs, donnant ainsi une impression en faveur de l'efficacité plus favorable que ce qu'elle aurait dû être après prise en compte de tous les résultats, positifs ou négatifs.

Pour éviter ce problème, les preuves d'efficacité doivent être issues d'une synthèse non arbitraire de tous les résultats d'essais, qu'ils soient en faveur ou non de l'efficacité

Un exemple de sélection arbitraire des résultats couramment avancés pour justifier l'efficacité d'un traitement est donné par le travail de Ranskov [34] réalisé sur les essais d'hypocholestérolémiants dans la prévention du risque cardiovasculaire.

En 1992, les résultats de 24 essais étaient disponibles, 14 étaient favorables à l'efficacité en prévention des hypocholestérolémiants et 10 ne l'étaient pas (résultats non statistiquement significatifs ou effet délétère). Ranskov a étudié la fréquence de citations de ces essais dans la littérature au travers des articles de synthèse, des éditoriaux, des articles de recommandation pour la pratique (Tableau 3 et Tableau 4).

Tableau 3 – Fréquence de citations des essais en fonction de leur résultat.

	Nombre moyen de citations par an
Résultats favorable (n=14)	40
Résultats non favorables (n=10)	7,4

Tableau 4 - Fréquence de citations dans les années suivant la publication de deux essais publiés dans le JAMA

Essai	1ère	2ème	3ème	4ème
LRC (favorable)	109	121	202	180
Miettinen (non favorable)	6	5	3	0

Il apparaît que la fréquence de citation dépend grandement du résultat. Les essais négatifs disparaissent presque totalement de la mémoire collective.

Pour constituer une preuve fiable, une série de résultats d'essais doit être issue d'un processus de collecte non arbitraire, sélectionnant les essais sur leur caractère non-biaisé et non sur leurs résultats. Les résultats en faveur du traitement ainsi que ceux en défaveur doivent être pris en considération.

Biais de publication

Les essais thérapeutiques ont d'autant plus de chance d'être publiés que leurs résultats s'avèrent positifs, c'est-à-dire statistiquement significatifs. Il existe ainsi une publication sélective des résultats positifs au détriment des résultats négatifs. Cela ne veut pas dire que ces derniers ne soient jamais publiés mais ils le sont plus difficilement et seulement pour une partie d'entre eux. Les raisons de cette censure sont multiples et peuvent provenir soit des comités de lecture des journaux, soit des firmes finançant l'étude, mais aussi d'une autocensure que s'infligent spontanément les investigateurs.

De ce fait la littérature biomédicale ne reflète pas exactement la réalité et en donne un aperçu exagérément optimiste, en taisant les résultats en défaveur de l'efficacité des traitements. C'est le biais de publication (« publication biais *») (cf. chapitre* Méta-analyse*).*

La prise en considération de tous les essais entrepris avec un traitement, publiés et non publiés, est indispensable avant de conclure à l'efficacité du traitement et doit comporter une recherche poussée des essais non publiés. Ces synthèses exhaustives sont réalisées sous forme de méta-analyse (cf. chapitre méta-analyse).

Exemple

Les conséquences potentiellement dommageables du biais de publication sont parfaitement illustrées par l'exemple des antiarythmiques de classe 1 en post infarctus. Leur nocivité n'a été mise en évidence qu'en 1991 (par l'essai CAST) alors que dès 1980 un essai de petite taille avait observé une forte augmentation de mortalité avec une molécule de cette classe, le lorcainide [35]. Cependant cet essai n'a pas été publié. Bien qu'il soit impossible de réécrire l'histoire, il est raisonnable de penser que la publication de cet essai aurait peut-être accéléré la mise en place de l'essai de confirmation qu'a été CAST. Ici la non publication d'un résultat non concluant a certainement retardé la mise en évidence d'un effet délétère avec comme conséquence de nombreuses morts prématurées.

Résumé

Les différents points que nous venons de voir conduisent à conclure que l'obtention de preuves fiables de l'efficacité d'un traitement nécessite :

- une vérification directe, sur critères cliniques, que le traitement permet d'atteindre l'objectif thérapeutique pour lequel il est pressenti
- que cette vérification s'effectue sans biais par le moyen d'essais thérapeutiques randomisés correctement conçus et réalisés
- que plusieurs résultats concordants soient disponibles pour éliminer un **résultat artefactuel** dû, soit au risque d'erreur statistique à, soit à une étude biaisée
- que l'ensemble des essais conduits, publiés et non publiés, quels que soient leurs résultats, soit disponible afin de pouvoir en faire une synthèse loyale pesant les résultats positifs et négatifs

la synthèse des essais en prenant en compte l'inflation du risque alpha et le risque bêta par la technique de la méta-analyse

Bibliographie

1. *Roland M, Torgerson DJ. Understanding controlled trials: what are pragmatic trials? BMJ 1998;316:285.* PMID:

2. *Eschwege E, Bouvenot G. Essais explicatifs ou pragmatiques. Le dualisme. Rev Med Int 1994;15:357-61.* PMID:

3. *Vray M, Bouvenot G. Il faut faire des essais pragmatiques. Presse Med 1995;24.* PMID:

4. *Cucherat M, Dürr F. Contexte de la médecine factuelle. Med Hyg 2000;58:850-55.* PMID:

5. *Echt DS, Liebson PR, Mitchell LB, Peters RW. Mortality and morbidity in patients receiving encainide, flecainide or placebo. The Cardiac Arrhythmia Suppression Trial. NEJM 1991;324:781-8.* PMID:

6. *Moore TJ. Deadly medicine. New York: Simon & Schuster; 1995.*

7. *Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. Jama 2002;288(3):321-33.* PMID: 12117397.

8. *Cholesterol, diastolic blood pressure, and stroke: 13,000 strokes in 450,000 people in 45 prospective cohorts. Lancet 1995;346:1647-1652.* PMID:

9. *Blauw GJ, Lagaay AM, Smelt AHM, Westendorp RGP. Stroke, statins, and Cholesterol. A meta-analysis of randomized, placebo-controlled, bouble-blind trials with HMG-CoA reductase inhibitors. Stroke 1997;28:946-950.* PMID:

10. *Lechat P, Packer M, Chalon S, Cucherat M, Arab T, Boissel JP. Clinical effects of beta-adrenergic blockade in chronic heart failure: a meta-analysis of double-blind, placebo-controlled, randomized trials. Circulation 1998;98(12):1184-91.* PMID:

11. *The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial. Lancet 1999;353(9146):9-13.* PMID:

12. *Cohn JNG, S.O. A dose-dependant increase in mortality with vesnarinone among patient with severe heart failure. Vesnarinone Trial Investigators. NEJM 1998;339:1810-6.* PMID:

13. *Cowley AJS, A.M. Treatment of severe heart failure: quantity or quality of life? A trial of enoximone. Enoximone Investigators. Br Heart J 1994;72:226-30.* PMID:

14. *Vernon MWH, R.C.;Brogden,R.N. Enoximone. A review of its phamacological properties and therapeutic potential. Drugs 1991;42:997-1017.* PMID:

15. *Uretsky BF, Jessup M, Konstam MA, Dec GW, Leier CV, Benotti J, et al. Multicenter trial of oral enoximone in patients with moderate to moderately severe congestive heart failure. Lack of benefit compared with placebo. Enoximone Multicenter Trial Group. Circulation 1990;82(3):774-80.* PMID:

16. *MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies. Lancet 2001;357:455-62.* PMID:

17. *Jenicek M. Epidemiology. The logic of modern medicine. Montreal: EPIMED; 1995.*

18. *Jha P, Flather M, Lonn E, Farkouh M, Yusuf S. The antioxidant vitamins and cardiovascular disease. Ann Intern Med 1995;123:860-72.* PMID:

19. *Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. NEJM 1994;330:1029-35.* PMID:

20. *Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, et al. Effects of a combination of beta carotene and vitamin A on lung cancer and cardiovascular disease. NEJM 1996;334:1150-5.* PMID:

21. *Hennekens CH, Buring JE, Manson JC, Stampfer M, Rosner B, Cook NR, et al. Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. NEJM 1996;334:1145-9.* PMID:

22. *Egger M, Davey Smith G. Misleading meta-analysis. BMJ 1995;310:752-754.* PMID:

23. *Carroll D, Tramèr M, McQuay H, Nye B, Moore A. Randomization is important in studies with pain outcomes: systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain. British Journal of Anaesthesia 1996;77:798-803.* PMID:

24. *Kunz R, Owman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317:1185-90.* PMID:

25. *Diehl LF, Perry DJ. A comparison of randomized concurrent control groups with matched historical control groups: are historical controls valid? J Clin Oncol 1986;4:1114-20.* PMID:

26. *Gleave ME EM, Fradet Y, Davis I, Venner P, Saad F, Klotz LH, Moore MJ, Paton V, Bajamonde A. Interferon gamma-1b compared with placebo in metastatic renal-cell carcinoma. Canadian Urologic Oncology Group. NEJM 1998;338(18):1265-71.* PMID:

27. *Zelen M. Guidelines for publishing papers on cancer clinical trials. J Clin Oncology 1983;1:164-169.* PMID:

28. *Hardcastle JD, Chamberlain JO, Robinson MH, Moss SM, Amar SS, Balfour TW, et al. Randomised controlled trial of faecal-occult-blood screening for colorectal cancer. Lancet 1996;348:1472-7.* PMID:

29. *Kronborg O, Fenger C, Olsen J, Jorgensen OD, Sondergaard O. Randomised study of screening for colorectal cancer with faecal-occult-blood test. Lancet 1996;348(9040):1467-71.* PMID:

30. *Freifeld A, Marchigiani D, Walsh T, Chanock S, Lewis L, Hiemenz J, et al. A double-blind comparison of empirical oral and intravenous antibiotic therapy for low-risk febrile patients with neutropenia during cancer chemotherapy. NEJM 1999;341(5):305-11.* PMID:

31. *Kern WV, Cometta A, De Bock R, Langenaeken J, Paesmans M, Gaya H. Oral versus intravenous empirical antimicrobial therapy for fever in patients with granulocytopenia who are receiving cancer chemotherapy. International Antimicrobial Therapy Cooperative Group of the European Organization for Research and Treatment of Cancer. NEJM 1999;341(5):312-8.* PMID:

32. *Saltz LB, Cox JV, et al. Irinotecan plus fluorouracil and leucovoin for metastatic colorectal cancer. NEJM 2000;343:905-14.* PMID:

33. *Sargent DJ, Niedzwiecki D, O'Connell MJ, Schilsky RL. Recommendation for caution with irinotecan, fluorouracil, and leucovorin for colorectal cancer. NEJM 2001.* PMID:

34. *Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. BMJ 1992;305:15-9.* PMID:

35. *Cowley AJ, Skene A, Stainer K, Hampton JR. The effect of lorcainide on arrhythmias and survival in patients with acute myocardial infarction: an example of publication bias. Int J Cardiol 1993;40(2):161-6.* PMID:

LECTURE CRITIQUE

Les principes de la mise en pratique des résultats des essais thérapeutiques

Contexte général

L'évolution de la médecine l'a conduit à chercher dans la science le moteur de son progrès. La médecine est, et doit rester avant tout, une « pratique », mais qui va chercher dans l'approche scientifique les fondements de son exercice. Ainsi, au-delà des connaissances fondamentales biologique et physiopathologique que nous fournie la science, la démarche scientifique est aussi présente à la fin du processus de développement des nouvelles approches thérapeutiques, au moment de la confirmation de l'intérêt clinique du traitement pour le patient. Cette base scientifique permet de rendre rationnel la pratique thérapeutique afin d'obtenir une efficacité optimale et certaine. Elle apporte les ***preuves*** *assurant que les soins proposés aux patients sont cliniquement efficace. La nécessité de cette recherche de preuve est détaillée dans un autre chapitre (Les critères scientifiques de démonstration de l'efficacité des traitements).*

De plus, cette approche évite des pratiques dont la motivation serait autre que la recherche d'un résultat thérapeutique optimum ou qui ne prendraient pas le soin de vérifier que les traitements proposés aux patients sont en mesure de répondre à leur attente ou à leur espoir. Ainsi, en dehors du paradigme scientifique, le choix des traitements est potentiellement de l'ordre du rapport de force ; du mercantilisme, de la sociologie, de l'esthétique, de l'idéologie, de la croyance, de l'ésotérisme, etc. (1) Négliger la recherche des preuves scientifiques de l'efficacité des traitements a pour conséquence de retourner à une thérapeutique fondée sur l'intuition, qui est devenue insuffisante au regard des exigences de performance et de moyen du public consommateur de soins.

De ce fait, les connaissances thérapeutiques (énoncées par les règles de l'art, les recommandations, les guides de pratiques, les protocoles de traitement, etc.) doivent reposées sur les résultats de l'évaluation scientifique de l'efficacité des traitements par les essais thérapeutiques. Ce principe est renforcé par le constat par le passé, à plusieurs reprises, que certaines pratiques de soins courantes mais dont les bénéfices n'avait jamais été prouvées, se sont avérées inefficaces voire dangereuses lorsque les essais ont été finalement réalisés (voir chapitre Pourquoi vouloir des preuves scientifiques de l'efficacité d'un traitement*).*

Les grandes lignes de la démarche

L'intégration des résultats des essais cliniques à la pratique médicale nécessite :

- *d'identifier les nouveaux essais thérapeutiques (dans le cadre d'une veille scientifique) ou de rechercher les essais correspondant à une question thérapeutique*
- *de faire la synthèse des résultats disponibles concernant la question d'intérêt*
- *de vérifier le niveau de démonstration d'un résultat : le bénéfice clinique est il formellement démontré par les essais disponibles ou seulement suggéré ?*
- *de déterminer sa pertinence clinique (le bénéfice est il médicalement intéressant et extrapolable aux patients vus en pratique)*
- *de déterminer la place exacte du nouveau traitement dans l'arsenal thérapeutique*
- *d'intégrer les données factuelles au cas par cas dans la décision thérapeutique au moment de l'acte de soins en tenant compte des aspects humains, sociologiques, psychologiques et autres.*

Les trois premières étapes de ce travail sont en général effectuées en aval de la décision thérapeutique concernant un patient (consultation ou au pied du lit) lors, par exemple, de l'élaboration de recommandations pour la pratique ou lors de la formation médicale continue ou initiale.

Interprétation et mise en œuvre des résultats des essais cliniques

Pour la pratique médicale, l'objectif de l'interprétation des résultats des essais thérapeutiques est d'évaluer de manière critique un résultat avant de le mettre, éventuellement, en application. Pour cette raison, l'interprétation est parfois appelée « lecture critique ». Pour un médecin, il s'agit de répondre à l'interrogation « le bénéfice apporté par ce traitement est-il suffisamment établi et cliniquement pertinent pour justifier son utilisation » (2).

La lecture critique et l'interprétation des essais consistent à évaluer s'il existe des preuves que le traitement apportera en pratique un bénéfice suffisant et pertinent.

En d'autres termes, il s'agit d'évaluer si les données disponibles fournissent une preuve suffisamment fiable que le traitement permettra d'atteindre de manière suffisamment satisfaisante le bénéfice thérapeutique recherché. Nous verrons par la suite que la recherche de preuves est indispensable car tout concourt à la production de faux positifs.

Composantes de l'interprétation

L'interprétation d'un résultat d'essais cliniques porte sur trois axes différents :

- *la **validité interne** du résultat ou « est-ce que le résultat est réel ? est-il non biaisé ? »*
- *a **pertinence clinique** et sa **représentativité** ou « ce résultat représente-t-il un bénéfice intéressant en pratique ? et est-il extrapolable à mes patients ? »*
- *sa **cohérence externe** ou « est-ce que ce résultat est concordant avec les autres résultats disponibles et les autres connaissances sur le sujet (physiopathologiques, pharmacologiques, épidémiologiques) ? »*

L'interprétation d'un résultat consiste donc à analyser s'il est licite de répondre par l'affirmative à ces trois groupes de questions ou si, au contraire, il existe des points conduisant à émettre des réserves sur la réalité et/ou la pertinence du résultat considéré. Dans ce cas un nouvel essai devra lever ces réserves.

Les éléments clés de la lecture critique sont :

- *la **réalité du résultat** : fiabilité de la méthodologie de l'étude, réalité statistique du résultat, valeur épistémologique,*
- *la **reproductibilité** du résultat : le résultat est-il confirmé par d'autres et est-il cohérent avec les connaissances disponibles,*
- *la **taille** de l'effet et la **précision** de son estimation,*
- *la **balance bénéfice – risque**,*
- *la **pertinence** vis à vis des questions cliniques se posant en pratique,*
- *la **représentativité** du résultat et son **extrapolabilité**.*

D'autres approches sont possibles, qui représentent des déclinaisons différentes des mêmes dimensions. Comme celle qui met le traitement au centre et qui consiste à se poser trois types de questions quelque peu différentes des précédentes :

- *existe-t-il un réel effet du traitement ? Dans quelle mesure l'effet observé peut il être dû à des biais ou au hasard ?*
- *la taille de l'effet est-elle cliniquement importante ?*
- *ce résultat est-il pertinent pour la pratique médicale (répond-t-il à une question se posant réellement en pratique) et est-il généralisable (a-t-il été obtenu sur des patients et dans des contextes de soins similaires à ceux de ma pratique quotidienne, sinon puis-je extrapoler) ?*

Avant d'appliquer un résultat en pratique, il convient de s'assurer :

Validité interne	• que le résultat est très probablement réel, c'est-à-dire qu'il est statistiquement significatif et qu'il est épistémologiquement valide, • que le résultat est sûr (exempt de biais), c'est-à-dire que le plan d'expérience choisi évite les biais et que l'étude a été correctement réalisée,
Cohérence externe	• que le résultat est confirmé par les autres résultats du domaine et qu'il est cohérent avec les connaissances disponibles (biologiques, épidémiologiques),
Pertinence clinique	• que le critère de jugement est pertinent cliniquement et correspond à un objectif thérapeutique opportun pour la maladie, • que le résultat et la balance bénéfice-risque, sont cliniquement pertinents : c'est-à-dire qu'il est de taille suffisante pour être intéressant en pratique, et que la balance bénéfice-risque est acceptable, • qu'il a été obtenu sur des patients comparables à ceux vus en pratique, • que le traitement a été utilisé dans un contexte de soins similaires à celui de la pratique quotidienne.

L'analyse critique et l'interprétation d'un essai thérapeutique requièrent donc de s'intéresser aux éléments suivants :

- *la méthode de l'essai,*
- *la qualité de réalisation de l'essai,*
- *la réalité statistique du résultat,*
- *l'intensité de l'efficacité,*
- *la précision avec laquelle cette taille est connue,*
- *les effets indésirables observés dans l'essai et ceux connus par ailleurs afin d'étudier la balance bénéfice-risque,*
- *la nature du critère de jugement,*
- *la nature du traitement de référence : placebo, traitement actif validé, non validé,*
- *le type de patients recrutés : stade de la maladie, méthode diagnostique,*
- *le contexte de soins dans lequel s'est déroulé l'essai : diagnostic, traitements concomitants, procédure de surveillance, etc.,*
- *les autres essais du domaine concernant le même traitement ou des traitements concurrents.*

Le but de ce chapitre était d'introduire les grands principes de l'interprétation d'un résultat d'essai clinique. Ces points seront explicités en détail dans un chapitre ultérieur. Mais auparavant, les concepts nécessaires à cette interprétation vont être présentés.

Intégration des données factuelles dans la pratique médicale

Bien que prépondérants, les résultats des essais cliniques ne constituent pas les seules informations à prendre en compte dans l'élaboration de recommandations pour la pratique ou de guides de décision. Les autres points entrant en ligne de compte sont :

- *les données de pharmacovigilance,*
- *l'efficacité et la sécurité des traitements alternatifs disponibles,*
- *les choix collectifs de politique de santé, le contexte de soins,*
- *les choix sociétaux,*
- *la fréquence de la maladie et l'importance du problème de santé publique qu'elle engendre,*
- *les coûts,*

Au niveau du colloque singulier entre le médecin et le patient, il est évident que d'autres éléments interviennent à coté des résultats des essais, comme :

- *l'applicabilité des résultats au patient (le patient est-il similaire à ceux qui ont été étudiés dans les essais),*
- *l'attente du patient vis à vis de sa prise en charge médicale (objectifs thérapeutiques personnels),*
- *les préférences du patient (et/ou de ses proches),*
- *son profil psycho-affectif (et/ou de ses proches),*
- *l'offre de soins disponible*
- *le contexte social et d'accès aux soins du patient.*

Les résultats des essais représentent, ainsi, qu'une partie des multiples données que doit intégrer « l'art médical » exercé par le médecin dans sa pratique quotidienne (Figure 1).

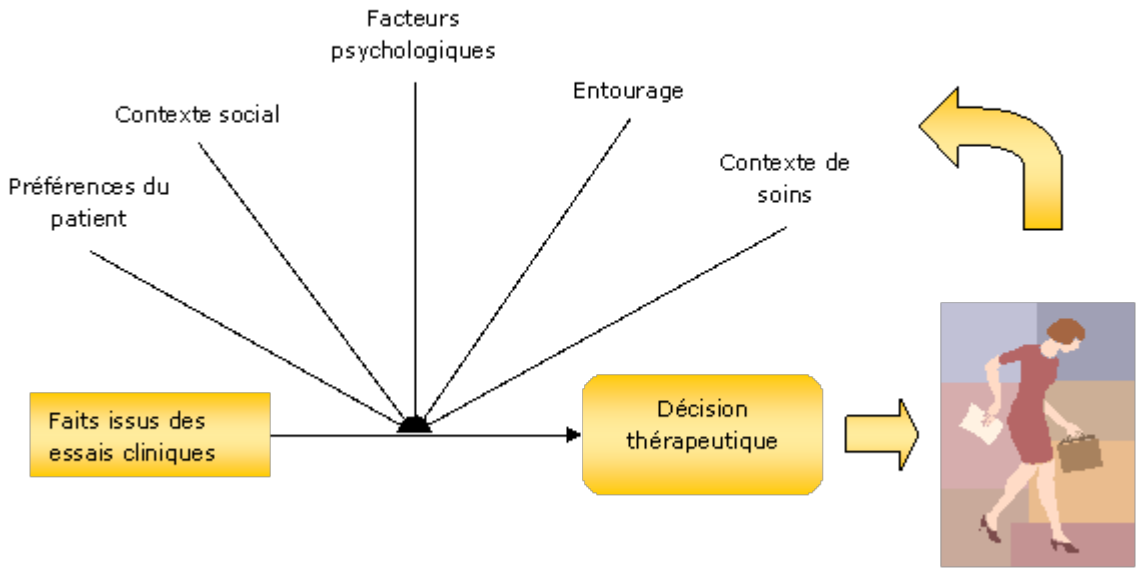


Figure 1 – Les différentes composantes de la décision thérapeutique

En pratique, pour être facilement applicable cette démarche nécessite d'avoir accès facilement aux résultats des essais et à leur méta-analyse. A l'heure actuelle ce scénario est encore légèrement futuriste, mais de nombreuses initiatives sont en train de se mettre en place pour permettre aux praticiens d'accéder facilement et directement à cette « information thérapeutique » représentée par les résultats des essais thérapeutiques (3).

Ce schéma conduit inévitablement à des décisions qui dans certains cas particulier ne sont pas en pleine conformité avec les résultats de la recherche thérapeutique. Dans ces cas, les écarts sont justifiés par les éléments propres aux patients et/ou au contexte que nous venons de voir. Cependant ces situations sont exceptionnelles et ne doivent pas être systématiques.

Les écarts entre pratique et résultats de la recherche thérapeutique

Plusieurs études ont cherchées à mesurer dans différents domaines de la thérapeutique le taux de prescriptions en cohérence avec les essais randomisés et les méta-analyses. Une récente revue systématique de ces travaux empirique en dénombre 15 de bonne qualité.(4) Globalement, dans ces études 50% des prescriptions de médecine interne s'avèrent fondées sur des preuves issues d'essais randomisés ou de méta-analyses.(5) Dans notre étude, le taux de prescriptions de traitements ayant démontrés leur bénéfice clinique est du même ordre de grandeur que ceux relevés dans ces autres études. Par exemple, Koyama et al. trouvent que 49% des soins primaires sont supportés par des essais randomisés dans leur service de médecine générale au Japon. (6) Des taux bien plus important sont cependant possible, montrant la possibilité de pratiques presque entièrement basées sur les preuves.(7) Dans un contexte de médecine libérale, Gill et coll. rapportent un taux de 81% d'interventions fondées sur des preuves. (8)

Liens internet

- *Seven alternatives to evidence based medicine.* BMJ *1999;319:1618-1618* *http://bmj.bmjjournals.com/cgi/content/full/319/7225/1618*

Bibliographie

1. *Isaacs D, Fitzgerald D. Seven alternatives to the evidence based medicine. BMJ 1999;319:1618-1618.*
2. *Bouvenot G, Eschwege E. Essais thérapeutiques. Principes d'interprétation. Rev Prat 1991;41:1853-7.*
3. *Boissel JP, Bossard N, Cucherat M, Haugh MC, Fardeheb M, Rachet F, et al. L'information thérapeutique. Paris: Masson; 2000.*
4. *Matzen P. [How evidence-based is medicine? A systematic literature review]. Ugeskr Laeger 2003;165(14):1431-5.*
5. *Nordin-Johansson A, Asplund K. Randomized controlled trials and consensus as a basis for interventions in internal medicine. J Intern Med 2000;247(1):94-104.*
6. *Koyama H, Matsui K, Goto M, Sekimoto M, Maeda K, Morimoto T, et al. In-patient interventions supported by results of randomized controlled trials in Japan. Int J Qual Health Care 2002;14(2):119-25.*
7. *Geddes JR, Game D, Jenkins NE, Peterson LA, Pottinger GR, Sackett DL. What proportion of primary psychiatric interventions are based on evidence from randomised controlled trials? Qual Health Care 1996;5(4):215-7.*
8. *Gill P, Dowell AC, Neal RD, Smith N, Heywood P, Wilson AE. Evidence based general practice: a retrospective study of interventions in one training practice. Bmj 1996;312(7034):819-21.*

Démonstration de l'efficacité

Introduction

La méta-analyse occupe une place centrale dans le processus d'établissement des preuves de l'efficacité clinique d'un traitement[1, 2]. Elle représente le dernier stade de la démonstration de l'efficacité et permet une formalisation de la représentation des connaissances [3]. Ce chapitre explicite la notion de « démonstration de l'efficacité » et décrit la place qu'occupe la méta-analyse dans cette démarche de formalisation.

Différents stades existent dans la démonstration de l'efficacité d'un traitement, créant ainsi une hiérarchie qui peut s'apparenter à une échelle de niveau de preuve. Il est ainsi possible de délimiter des situations où l'efficacité peut être considérée comme formellement démontrée, de celles où les données sont nettement insuffisantes et de celles, intermédiaires, où il convient de discuter avant de prendre une décision thérapeutique.

Méta-analyse concluante et au moins un essai concluant par lui-même

La situation où la méta-analyse est concluante avec au moins *un essai concluant par lui-même est la plus probante et permet de considérer que l'efficacité est formellement démontrée.*

La situation la plus probante est celle où la méta-analyse est concluante et contient au moins un essai correctement conçu et réalisé et concluant par lui-même (cf. encart). Cet essai montre un bénéfice statistiquement significatif sur son critère de jugement principal qui est un critère clinique pertinent. La méta-analyse regroupant cet essai et les autres du domaine est elle aussi concluante. Elle confirme le résultat de l'essai et valide la cohérence externe du domaine. Son apport est indispensable car la force de conviction d'un seul résultat isolé est limitée.

Critères de définition d'un essai concluant

- *essai contrôlé randomisé,*
- *correctement conçu, sans biais potentiel,*
- *correctement réalisé, sans biais apparent,*
- *analyse en intention de traiter,*
- *utilisant comme critère de jugement principal un critère clinique pertinent,*
- *résultat statistiquement significatif sur le critère de jugement principal*

En regroupant tous les essais réalisés, qu'ils soient en faveur ou contre l'existence de l'efficacité, la méta-analyse fait le bilan de l'existant, et vérifie que l'essai concluant, avancé pour justifier l'efficacité, n'est pas une aberration, due au hasard ou à un biais.

Dans le cas où l'essai concluant est le seul essai réalisé, le processus méta-analytique met ce fait en évidence et attire l'attention sur l'absence de vérification de ce résultat. À l'opposé, il peut exister non pas un seul mais plusieurs essais concluant par eux-mêmes. Le degré de conviction de la preuve de l'efficacité s'en trouve augmentée.
La méta-analyse permet aussi d'explorer la possibilité d'un biais de publication qui ferait que l'essai concluant est le produit d'un processus de sélection par les résultats et ne représente pas la réalité. Pour récuser l'existence d'un biais de publication, la méta-analyse utilise la recherche exhaustive des essais publiés et non publiés, le graphique en « entonnoir » (« funnel plot *») et le calcul de la robustesse du résultat [4].*

Elle permet aussi de rechercher la variabilité de l'efficacité du traitement entre les essais et de détecter les situations où une éventuelle interaction biologique pourrait être suspectée (analyse en sous-groupe, méta-

régression). La recherche de l'hétérogénéité est importante pour s'assurer que le résultat de l'essai concluant n'est pas dû à une situation où le traitement développe une efficacité particulièrement importante, non retrouvée dans les autres situations. En pratique, les difficultés rencontrées sont les suivantes. :

- Il est difficile de conclure formellement à l'absence d'un biais de publication en raison de la faible puissance de sa recherche dans la plupart des cas.
- La recherche d'une hétérogénéité est souvent de faible puissance statistique. Il est rarement possible d'exclure formellement une variation de l'effet même après l'utilisation de méthodes statistiques complexes. La conclusion est qu'il n'a pas été possible de mettre en évidence une hétérogénéité. Dans ce cas, selon le paradoxe de Stein, l'estimation globale est la meilleure estimation de l'effet du traitement pour chaque type de patients [5, 6].

Classiquement, il a été dit qu'il était nécessaire de disposer de deux essais significatifs pour conclure à l'efficacité. L'existence d'un deuxième essai permet une vérification externe du premier résultat et diminue le risque d'erreur alpha globale. Ce principe semblait être exigé au niveau des instances réglementaires bien qu'il soit impossible de trouver une trace écrite officielle de cela. Cette règle tend à ne plus être respectée dans plusieurs domaines où les essais nécessitent un très grand nombre de sujets. C'est par exemple le cas avec les essais de mortalité à la phase aiguë de l'infarctus qui regroupent plusieurs milliers de patients. La duplication de ce type d'essai est financièrement très difficile. Le principe, selon lequel tout résultat expérimental doit être vérifié par au moins une autre expérience, conduit à un certain nombre de difficultés en évaluation des traitements. En effet, il est parfois difficile, une fois qu'un essai correctement conçu et réalisé, a conclu à l'efficacité du traitement, d'en refaire un uniquement dans un but de vérification. Cela pose des problèmes éthiques et financiers. Cependant, dans certains cas, de nouveaux essais sont encore nécessaires, même après un essai concluant pour convaincre une opinion médicale réticente. Cela fut le cas avec la fibrinolyse à la phase aiguë de l'infarctus du myocarde. L'essai ISIS2 a continué à se dérouler après que l'essai GISSI 1 a montré une réduction de mortalité. Les craintes du corps médical étaient telles que la continuation de l'essai a permis d'habituer les médecins à ce traitement. En réalité, d'autres essais sont très souvent disponibles aux côtés d'un essai concluant car ils ont été entrepris simultanément ou avant lui. L'essai suffisamment puissant intervenant en fin de développement, après la réalisation d'essais de plus petite taille documentant plutôt des critères intermédiaires mais pour lesquels les critères cliniques sont aussi disponibles. D'autres possibilités conduisent à la coexistence de plusieurs essais dans le même domaine : réalisation d'essais dans des populations différentes, ou avec différentes molécules de la même classe pharmacologique. Le regroupement de ces essais permet une certaine vérification des résultats, mais pose le problème de l'effet de classe ou de l'hétérogénéité clinique des populations.

Méta-analyse concluante sans aucun essai concluant

La situation où seule la méta-analyse est concluante sans qu'il existe un essai concluant par lui même est moins convaincante.

Une position raisonnable consiste à admettre que le résultat de la méta-analyse n'est pas suffisant pour démontrer formellement l'efficacité et qu'il est nécessaire de le confirmer par un essai thérapeutique suffisamment puissant (sauf si de nouveaux essais s'avèrent totalement impossibles).

Cette attitude prudente trouve une justification dans les faits observés, par exemple, avec les dérivés nitrés utilisés à la phase aiguë de l'infarctus du myocarde. Une méta-analyse de petits essais laissait prévoir une réduction possible de la mortalité [7]. Des essais de grandes tailles ont été lancés pour la confirmer (ESPRIM [8], ISIS 4[9], et GISSI 3[10]). Ces trois essais n'ont pas confirmé la réduction de la mortalité. Un autre exemple est celui du magnésium dans la même pathologie (cf. chapitre La méta-analyse*). Ces exemples étayent donc le principe énoncé ci-dessus, même si dans la majorité des cas les essais de grandes tailles ont confirmé les méta-analyses.*

Autres situations

Une situation où il existe un essai concluant, mais où la méta-analyse n'est pas concluante révèle le plus souvent une hétérogénéité qui demande à être expliquée avant de pouvoir conclure (cf. section sur la validité externe). Le résultat de l'essai concluant n'est pas concordant avec les résultats des autres essais.

Il aussi possible de rencontrer des cas où le résultat de la méta-analyse est non significatif mais sans qu'il y est d'hétérogénéité. Cette situation qui pourrait paraître paradoxale traduit simplement le fait que l'essai concluant est probablement la manifestation du risque α de 5 % et que son résultat favorable soit le fait du hasard

Tableau 1 – Récapitulatif des différents niveaux possibles de démonstration de l'effet.

Démonstration de l'effet	◆	Méta-analyse concluante sans hétérogénéité et existence d'un essai concluant.
Effet suggéré mais non démontré	◆	Méta-analyse montrant un effet statistiquement significatif mais absence d'essai concluant.
	◆	Essai concluant mais la méta-analyse ne retrouve pas l'effet de manière statistiquement significative.
	◆	Essai concluant mais méta-analyse hétérogène.
Preuves insuffisantes	◆	Absence d'effet statistiquement significatif dans la méta-analyse et absence d'essai concluant.

Conclusion à l'absence d'efficacité

Ne pas mettre en évidence l'efficacité n'autorise pas à conclure à l'absence de l'efficacité.

La démonstration formelle de l'absence d'efficacité est difficile à obtenir et fait appel à une méthodologie spécifique (cf. essai d'équivalence). Ne pas mettre en évidence l'efficacité n'autorise pas à conclure à l'absence de l'efficacité. Il peut simplement s'agir d'un manque de puissance et de l'impossibilité de démontrer l'absence d'effet au moyen de l'instrument utilisé. Ainsi un essai non significatif ou une méta-analyse non significative ne permettent pas de conclure, formellement, à l'inefficacité du traitement.

Principe

Au point de vue statistique la démonstration de l'efficacité nulle est impossible à obtenir. Tout au plus il est possible de démontrer que l'efficacité est insuffisante dans une démarche type essais d'équivalence ou de non infériorité.

Le principe de cette démarche est le suivant. L'efficacité d'un traitement est jugée insuffisante lorsqu'il est fortement probable qu'elle soit inférieure à la plus petite efficacité intéressante *dans le domaine.*

*Le raisonnement se base sur l'analyse des intervalles de confiance. Lorsque la borne la plus favorable de l'intervalle de confiance de l'effet du traitement est en dessous du seuil d'effet minimal intéressant, il est possible de conclure que l'efficacité du traitement est insuffisante. Cette conclusion se fait avec un risque statistique de première espèce alpha contrôlé (*α/2 *pour un intervalle à* (1-α)100% *).*

La conclusion formelle à l'absence d'efficacité nécessiterait un essai de non-infériorité spécialement réalisé pour tester l'hypothèse d'absence d'efficacité. Cette éventualité n'est pas envisageable en pratique. Ainsi, il n'est jamais possible de conclure en toute rigueur à l'absence d'effet.

La ***Erreur ! Source du renvoi introuvable.*** *représente deux types de résultats non significatifs d'interprétation différente.*

- Le traitement n'entraîne pas de modification relative de la mortalité RRR=0% avec un IC95% de [4% ;-4%]. Ce résultat n'est pas significatif (p=0,95). Au mieux, il pourrait exister une réduction très faible de 4% qui ne présente aucun intérêt en pratique. Bien qu'en toute rigueur il ne soit pas possible de conclure à l'absence d'efficacité, l'interprétation de l'intervalle de confiance conduit à conclure que (très probablement) ce traitement serait d'aucune utilité en pratique. Étant donné la précision du résultat, il est licite de conclure à l'absence d'intérêt de ce traitement : même si celui-ci a une efficacité non nulle, la taille de l'effet serait trop petite pour être intéressante.
- Le traitement entraîne une réduction relative non significative de 20% (IC à 95% de [39%,-8%]) (p=0,16). Il apparaît clairement que ce résultat non significatif n'autorise pas à conclure à l'absence d'effet. En effet, ce résultat est compatible avec une réduction relative de 39%, effet de taille conséquente. De plus l'intervalle est en très grande partie du côté favorable ce qui renforce la

possibilité de l'existence de l'effet. En conclusion, il est possible que le traitement soit efficace et que cette efficacité soit suffisamment importante pour être intéressante. Ce résultat encourage à réaliser un nouvel essai de puissance appropriée.

Seuil

La détermination de l'efficacité minimale intéressante est un point délicat. Ce choix est arbitraire et conditionne la conclusion, mais le problème est moins aigu que dans le cadre de la démonstration de l'équivalence car le choix erroné ne se traduit pas par les mêmes pertes de chance pour le patient. Pour l'équivalence, la décision est d'utiliser un traitement à la place d'un autre et le risque couru est que le traitement finalement recommandé soit nettement moins efficace que le précédent.

La conclusion d'une efficacité insuffisante n'entraînera pas de perte de chance directe pour les patients. Un bénéfice minimal intéressant trop petit ne permet pas de conclure à tort à une efficacité insuffisante. Mais le traitement ne sera pas utilisé pour autant en pratique car il n'a pas démontré son efficacité. Au pire de nouveaux essais seront entrepris pour rien.

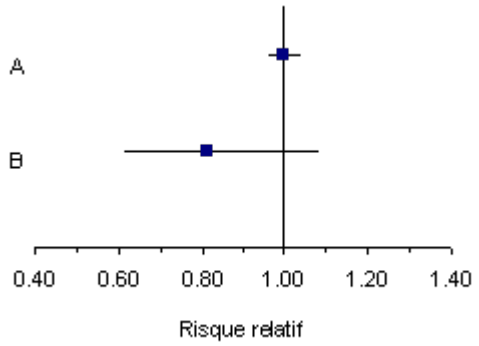


Figure 1 – Deux types de résultats non significatifs d'interprétation différente. Dans le cas A, l'intervalle de confiance est étroit. En toute rigueur il n'est pas possible d'exclure que l'effet du traitement soit non nul (partie de l'intervalle inférieure à 1), mais même dans la situation la plus favorable l'effet serait de très petit taille et sans intérêt en pratique. Il est donc raisonnable de conclure à l'absence d'efficacité. En B, l'intervalle est très large et il est compatible avec des efficacités très importantes. Il n'est pas possible de conclure à l'absence d'effet et il convient donc de réaliser un nouvel essai de puissance suffisante pour préciser l'effet du traitement.

Un bénéfice minimal intéressant trop grand peut conduire à rejeter un traitement dont l'efficacité est faible mais quand même intéressante. Dans cette situation, le développement est arrêté prématurément à tort. Le problème n'est pas de nature statistique. C'est le choix du bénéfice minimum intéressant qui est déterminant.

Intérêt de la méta-analyse dans cette situation

La méta-analyse est irremplaçable, voire indispensable, pour assoir la conclusion d'une efficacité insuffisante. Elle fournit la puissance statistique maximale et donc l'intervalle de confiance le plus petit possible.

Conclusion

Au total, la conclusion d'efficacité insuffisante *est possible quand :*

- il existe au moins un essai suffisamment puissant et correctement conçu et réalisé avec comme critère principal un critère clinique pertinent (et un calcul de puissance pour ce critère) non concluant,
- la méta-analyse incluant cet essai conduit à un intervalle de confiance qui exclut une efficacité du traitement au moins égale à l'efficacité minimale intéressante,
- il n'existe pas d'hétérogénéité.

En l'absence d'un essai suffisamment puissant avec un critère principal clinique, une conclusion d'efficacité insuffisante à partir uniquement d'une méta-analyse est moins probante. Les essais de la méta-analyse étaient peut être biaisés vers zéro en raison par exemple d'un recueil des critères cliniques peu fiable. En effet, l'analyse des critères cliniques ne faisait pas partie des objectifs des essais sélectionnés. Les procédures mises en œuvre ne permettent pas alors de garantir la qualité des données.

Cependant, en pratique, si la méta-analyse conduit à un intervalle de confiance très petit il sera difficile d'argumenter la réalisation d'un essai puissant. Même si aucune conclusion formelle d'efficacité insuffisante ne pourra être prononcée, le résultat sera le même : le traitement ne sera pas utilisé en pratique, et que l'on puisse affirmer ou non l'efficacité insuffisante sera sans grand intérêt.

Dans les comparaisons contre placebo, le recours à l'essai de non-infériorité pour montrer l'absence d'effet du traitement étudié n'est pas concevable. Il n'est pas réaliste d'investir dans un essai dont l'objectif serait de montrer qu'un traitement n'a pas d'efficacité ! La conclusion à l'absence d'effet d'un traitement ne peut donc être prise que dans le cadre de l'analyse d'essais négatifs. L'utilisation de la méta-analyse et des intervalles de confiances sont d'une grande aide dans cette démarche comme le montre l'exemple suivant de la vitamine E.

Exemple

La vitamine E a été envisagée dans la prévention des événements cardiovasculaires en raison de ses propriétés antioxydantes.

Cinq essais de bonne qualité méthodologique (randomisation imprévisible, explicitation du critère de jugement et du calcul de l'effectif, très faible taux de perdus de vue, analyse en intention de traiter) sont disponibles : ATBC bras vitamine E, HOPE, GISSI prevention, CHAOS, PPP. Ces essais ont comparé au placebo ou à l'absence de traitement des doses variables de vitamine E.

Dans la méta-analyse de ces essais, aucune modification n'est observée ni sur la mortalité cardio-vasculaire (RR=0,99, IC95%=[0,92 ;1,06]) ni sur la mortalité totale (RR=1,00 ; IC95%=[0,95 ;1,05]), ni sur la fréquence des événements cardiovasculaires mortels ou non mortels (infarctus, AVC) (RR=0,98 ; IC95%=[0,93 ;1,02]).

La précision élevée de ces résultats autorise à conclure à l'absence d'efficacité. En effet, au mieux, la vitamine E apporterait des bénéfices très minimes : réduction de 8% de la mortalité cardiovasculaire, de 5 % de la mortalité totale, etc. Ces valeurs sont sans intérêt clinique. Il est donc possible d'écarter définitivement ce traitement dans cette indication.

Absence de données

Certaines situations, au demeurant fort nombreuses, n'ont fait l'objet d'aucun essai thérapeutique sur critères cliniques. Aucun argument suffisamment consistant permet de justifier l'efficacité de ces traitements. Les essais disponibles sont souvent des essais de petite taille et utilisant un critère intermédiaire. Ces essais n'ont même pas recueilli les événements cliniques, si bien qu'il est impossible de réaliser une méta-analyse sur le critère pertinent. L'efficacité réelle de ces traitements est inconnue. Elle est tout au plus suspectée à partir d'éléments physiopathologiques et pharmacologiques, mais reste non démontrée tant que les essais appropriés ne sont pas réalisés.

Bibliographie

1. *Cucherat M. La méta-analyse des essais thérapeutiques [Thèse]. Lyon: Université Lyon-1; 2000.*
2. *Boissel JP, Cucherat M, Gueyffier F. Place de la méta-analyse dans la définition de la population cible d'une thérapeutique. Thérapie 1997;52:19-28.* PMID:
3. *Cucherat M. Représentation et gestion de la connaissance en médecine factuelle. Med Hyg 2000;58:1427-30.* PMID:
4. *Cucherat M, Boissel JP, Leizorovicz A, Haugh M. EasyMA: a program for the meta-analysis of clinical trials. Computer Methods and Programs in Biomedecine 1997;53:187-190.* PMID:
5. *Robbins H. An emperical Bayes approach to statistics. In: Proceeding of the 3rd Berkeley Symposium on Mathematical Statistics and Probability; 1955; 1955. p. 157-164.*
6. *Casella G. An introduction to emperical Bayes data analysis. American Statistician 1985;39:83-87.* PMID:
7. *Woods KL. Mega-trials and management of acute myocardial infarction. Lancet 1995;346:611-614.* PMID:

8. *European Study of Prevention of Infarct with Molsidomine (ESPRIM) Group. The ESPRIM trial: short term treatment of acute myocardial infarction with molsidomine. Lancet 1994;344:91-97.* PMID:

9. *ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58050 patients with suspected myocardial infarction. Lancet 1995;345:669-685.* PMID:

10. *Gruppo Italiano per lo Studio dell Streptochinasi nell'Infarto Miocardico (GISSI). GISSI 3: effects of lisinopril and transdermal glyceryl trinitrate singly and together on 6-week mortality and ventricular function after myocardial infraction. Lancet 1994;343:1115-1122.* PMID:

Validité interne

Introduction

Trois composantes contribuent à la validité interne et à la fiabilité du résultat :

- *la réalité statistique du résultat,*
- *l'absence de biais*
- *et la validité méthodologique.*

L'évaluation de la validité interne d'un résultat a ainsi pour but d'éliminer la possibilité qu'un résultat positif soit en fait produit par un biais ou soit le reflet du hasard.

Réalité statistique du résultat

Rappel

Lorsque le test statistique est significatif, il existe moins de 5% de chance que seul le hasard soit à l'origine du résultat observé. Il n'est jamais possible d'éliminer avec certitude le risque d'erreur statistique. En cas de valeur de p supérieure à 5%, il n'est en général pas raisonnable de considérer que le résultat observé est réel, le risque d'erreur statistique est trop important.

*Pour conclure à la réalité statistique d'un résultat, il convient aussi d'écarter une situation d'inflation du risque alpha produite par un mécanisme de répétition ou de multiplicité des tests statistiques (*cf. chapitre sur les tests statistiques*) : absence de critère de jugement principal fixé avant l'obtention des résultats, analyses en sous groupes, recherche de l'effet répétée au cours du temps, analyses intermédiaires non protégées.*

Points à vérifier

Les différents points à vérifier pour s'assurer de la réalité statistique de l'effet du traitement sont les suivants :

- *Le résultat est-il statistiquement significatif à un seuil inférieur ou égal à 5% ?*
- *Peut-on considérer que le risque alpha a été parfaitement contrôlé pour le résultat avancé (absence d'inflation) ?*
- *Le test statistique utilisé est-il adapté ? Ce point peut paraître technique et nécessitant des compétences statistiques avancées. En fait, il n'en est rien. Sauf cas exceptionnel, les tests statistiques nécessaires pour mettre en évidence un effet dans un essai clinique sont les tests de bases (chi-2, test-t, ANOVA, analyse de covariance).*

Situation à fort risque de problème

- *Le résultat avancé est issu d'une analyse en sous-groupes. Il existe un fort risque d'inflation du risque alpha. De plus, ce résultat n'est pas issu d'une démarche hypothético-déductive, ce qui limite sa valeur méthodologique.*
- *Le résultat est obtenu sur un critère de jugement qui n'a pas été clairement défini a priori comme étant le critère de jugement principal.*
- *Des analyses intermédiaires sont réalisées sans protection contre l'inflation du risque alpha.*
- *Il y a une répétition de la recherche de l'effet au cours du temps.*
- *Il y a absence du calcul préalable de l'effectif nécessaire. L'analyse porte alors sur un nombre arbitraire de sujets. Le choix du moment de l'analyse est peut-être conditionné par les résultats obtenus.*
- *Un ajustement est effectué sur des variables non prévues a priori. Un ajustement post-hoc sur des variables trouvées déséquilibrées entre les groupes entraîne un biais. Les variables d'ajustement doivent être prédéfinies et non pas déterminées en fonction des résultats.*
- *Pour les résultats négatifs, un résultat non significatif ne garantit pas l'absence d'effet. Le risque d'erreur bêta est incontrôlé.*
- *Des mesures multiples du critère de jugement chez le même patient sont analysées comme si elles étaient indépendantes, ou mesurées chez des sujets différents (le nombre d'unités statistiques sur lesquelles se base l'analyse statistique est supérieur au nombre de patients) [17,39].*
- *Résultat obtenu au niveau d'un sous groupe.*

Valeur méthodologique du résultat

Le résultat avancé doit correspondre directement à l'hypothèse formulée a priori, et dont le test était l'objet spécifique de l'essai, afin de respecter le principe de la méthode expérimentale. Cette condition est indispensable pour garantir la valeur méthodologique (« épistémologique ») du résultat.

Il convient, tout particulièrement, d'éliminer la possibilité que l'hypothèse ait pu être formulée après la prise de connaissance des résultats de l'essai. Dans ce cas, « l'expérience » ne peut que confirmer l'hypothèse puisque celle-ci a été formulée à partir de ces résultats. Cette situation tautologique enlève toute valeur au résultat.

L'hypothèse de l'essai doit avoir été formulée avant la réalisation de l'étude et non pas après.

Afin de pouvoir éliminer une génération post-hoc de l'hypothèse, l'introduction doit justifier de manière prospective l'hypothèse de l'essai, ses objectifs cliniques et les analyses en sous-groupes prévues. L'introduction doit démontrer que l'hypothèse testée découle naturellement des connaissances et des données disponibles avant le début de l'essai et que celui-ci a été spécifiquement entrepris pour la tester.

Un résultat non issu d'une démarche hypothético-déductive est de nature inductif. Il suggère alors un effet, mais ne peut le démontrer.

Tout changement post-hoc (ou définition post-hoc) de l'hypothèse testée, du critère de jugement, de la population cible supprime sa valeur déductive à un résultat et le transforme en un résultat inductif, exploratoire. Ce type de résultat suggère alors un effet, mais ne peut le démontrer.

Dans un essai d'un traitement topique des piqûres de moustiques où le critère de jugement principal est la durée du prurit, si une réduction significative de la mortalité est observée, celle-ci sera très vraisemblablement mise sur le compte du hasard et personne ne pensera à avancer comme effet de ce traitement une réduction de la mortalité. Cet effet est biologiquement peu plausible et un tel résultat ne sera pas considéré. Tout au plus il pourra mettre « la puce à l'oreille » et éventuellement faire découvrir une propriété insoupçonnée de ce traitement. Mais avant d'en arriver à une conclusion définitive, d'autres essais avec comme objectif la survie seront entrepris.

Par contre dans un essai s'adressant à un traitement de l'insuffisance cardiaque, avec comme critère de jugement la fréquence des hospitalisations, une réduction de mortalité pourrait éventuellement être considérée différemment en raison d'une plus forte plausibilité biologique. Pourtant la situation est identique à celle de l'exemple précédent. Le résultat observé sur la mortalité peut très bien provenir du hasard, des conditions particulières de l'essai, etc. Comme avec le traitement des piqûres de moustiques, ce résultat ne constitue pas une démonstration, même s'il semble concevable et plausible. Il suggère seulement un effet à ce niveau et permet de générer de nouvelles hypothèses à démontrer dans un essai de confirmation ayant comme critère principal la mortalité.

Absence de biais

Il convient de vérifier que la méthode utilisée évite la survenue des biais et que la réalisation de l'essai a été correcte.

L'analyse critique doit pouvoir éliminer la possibilité de l'existence d'un biais. Les situations propices à l'apparition des différents biais sont à rechercher, soit au niveau d'un défaut méthodologique, soit au niveau d'un défaut de réalisation.

Les différents biais pouvant affecter un essai thérapeutique vont être passés en revue. Pour chacun d'entre eux, un bref rappel de son mécanisme est effectué, puis les points à vérifier pour s'assurer que le résultat en est exempt sont listés. Pour terminer, une liste des situations à fort risque de biais est proposée.

Biais de confusion

Rappel

Le biais de confusion est le biais entraîné par l'absence de prise en considération des facteurs de confusion. Pour l'éviter l'essai doit être comparatif et doit comporter un groupe contrôle contemporain comme référence.

Questions à se poser pour vérifier les précautions prises pour éviter le biais

- *Existe-t-il un groupe contrôle ?*
- *L'effet du traitement est-il déterminé par rapport à ce groupe contrôle ?*

Situations à fort risque de biais

Dans les situations suivantes, le risque de biais de confusion est fort et remet en cause la validité interne du résultat obtenu.

- *Malgré la présence d'un groupe contrôle, l'effet est mesuré par une comparaison avant – après dans le groupe traité.*

Biais de sélection

Rappel

Le biais de sélection survient lorsque les deux groupes de l'essai ne sont pas comparables ce qui conditionne une différence dans le critère de jugement en dehors de tout effet traitement.

La randomisation a pour but d'éviter le biais de sélection qui survient lorsque les patients des deux groupes ne sont pas comparables. Il convient cependant de vérifier que la randomisation qui a été employée a bien permis d'atteindre ce but. Pour cela elle doit être décrite avec suffisamment de détails.

Questions à se poser pour vérifier les précautions prises pour éviter le biais

- *La méthode de randomisation garantit-elle l'imprévisibilité du traitement alloué à un patient ?*
  En effet, il est particulièrement important qu'un investigateur ne puisse pas connaître ou prédire le groupe auquel sera alloué le prochain patient [156,205]. À ce titre une « pseudo randomisation » basée sur la date de naissance du patient ou le jour de la consultation est inacceptable. L'utilisation d'enveloppe scellée n'est pas optimale, surtout pour les essais en ouvert (cf. exemple de l'essai CAPP page 63). Seules les procédures centralisées (téléphone, fax, informatique) donnent suffisamment de garantie.
- *Les groupes issus de la randomisation sont-ils comparables ?*
  Pour juger de cela (cf. chapitre Comparaison des groupes*), il convient de vérifier que les principaux facteurs pronostiques ou facteurs de risques du critère de jugement sont rapportés.*

Situations à fort risque de biais

Dans les situations suivantes, le risque de biais de confusion est fort et remet en cause la validité interne du résultat obtenu.

- *Le groupe contrôle n'est pas constitué de patients contemporains, mais de témoins historiques ou de témoins géographiques (en fait, il n'y a pas eu de randomisation).*
- *Le processus de randomisation était prévisible. Il était possible pour les investigateurs de sélectionner les patients dans les groupes de l'essai.*

Biais liés à l'absence ou un défaut de double insu

Rappel

L'absence, ou une mauvaise réalisation, du double insu est susceptible d'entraîner différents biais : biais de suivi, biais d'évaluation. Dans certaines situations, la réalisation d'un double insu n'est pas possible pour des raisons éthiques ou pratiques (cf. Tableau 1). Dans ce cas, les essais ne peuvent être réalisés qu'en simple insu ou en ouvert. Les points spécifiques à cette situation seront abordés dans une section suivante.

Questions à se poser pour vérifier les précautions prises pour éviter les biais

- *Le traitement du groupe contrôle est-il indiscernable du traitement du groupe traité ? Les deux groupes doivent recevoir un traitement qui a la même forme (gélule, perfusion IV, etc.), la même apparence (couleur, volume, conditionnement, étiquetage,), le même goût, etc…*
- *En cas de différence entre les traitements comparés (voie d'administration, forme galénique, etc. différentes), une technique de double placebo a-t-elle été employée ?*
- *Le code du traitement figurait-il sur les boîtes de traitements (par exemple code A, B)*

Biais de suivi

Rappel

Un biais de suivi survient lorsque les deux groupes ne sont pas suivis de la même manière au cours de l'essai. La comparabilité initiale est alors détruite et une différence peut apparaître en dehors de tout effet traitement. Le double aveugle est un élément central pour empêcher l'apparition de ce biais. À côté de l'évaluation de la qualité du double aveugle, d'autres points spécifiques du biais de suivi sont à prendre en considération.

Des points d'analyse spécifiques de l'essai en ouvert sont exposés dans la section suivante.

Questions à se poser pour vérifier les précautions prises pour éviter le biais de suivi

- *Est-ce que les arrêts de traitements, les déviations aux protocoles et les traitements concomitants ont été recueillis et sont convenablement documentés ?*
  Ces informations sont nécessaires pour répondre aux questions suivantes.
- *Le recours aux traitements concomitants a-t-il été aussi fréquent dans tous les groupes ? Une différence dans les traitements concomitants peut faire disparaître l'effet du traitement étudié, ou, à l'inverse, faire apparaître une fausse différence.*
  Une différence dans les traitements concomitants peut aussi être le reflet de l'effet du traitement étudié. Avec un traitement efficace, la fréquence de recours aux traitements de seconde ligne est réduite. Par exemple, un traitement doté d'un effet antalgique puissant entraîne une diminution de l'utilisation des antalgiques de seconde ligne prévus dans le protocole.
  À l'inverse un traitement ayant une mauvaise tolérance entraîne une augmentation de consommation des traitements prescrits en raison de cette mauvaise tolérance. Par exemple, des antiémétiques avec une chimiothérapie anticancéreuse.
- *Les taux de déviation au protocole sont-ils similaires dans les deux groupes ?*
- *Les taux d'arrêt du traitement de l'étude sont-ils similaires dans les deux groupes ?*
  En sachant que les différences observées peuvent être dues à une différence de tolérance des produits et non pas à une situation potentiellement biaisée.

Biais d'évaluation

Rappel

Le biais d'évaluation (aussi appelé biais de mesure) survient quand la mesure du critère de jugement n'est pas réalisée de la même manière dans les deux groupes. Le double insu limite le risque de biais d'évaluation.

Questions à se poser pour vérifier les précautions prises pour éviter le biais d'évaluation

- *L'évaluation du critère de jugement est-elle faite de la même façon quel que soit le traitement reçu ?*
- *Le traitement est-il susceptible d'influencer la mesure du critère de jugement ?*
- *Dans un essai en ouvert, la mesure du critère de jugement est-elle subjective ? La connaissance du traitement reçu par le patient peut influencer la mesure du critère de jugement. Avec ce type de critère, si le double aveugle est impossible (par exemple psychothérapie), l'évaluation des patients doit se faire, en insu du traitement reçu, par un évaluateur indépendant des médecins ayant en charge les patients (triple aveugle).*

Recherche des biais dans l'essai en ouvert

Rappel

Dans certaines situations, la réalisation d'un double insu n'est pas possible pour des raisons éthiques ou pratiques. Dans ce cas, les essais ne peuvent être réalisés qu'en simple insu ou en ouvert. La méthodologie employée n'empêchant pas la survenue d'un biais, il convient d'analyser soigneusement les marqueurs permettant de juger que le suivi et l'évaluation des critères de jugement se sont effectués de manière identique dans les deux groupes.

Seules quelques situations très particulières empêchent la réalisation d'un double insu (cf. Tableau 1). En dehors de ces situations, l'absence de double insu n'est ni satisfaisante, ni justifiable.

Questions à se poser pour vérifier les précautions prises pour éviter le biais

- *Le critère de jugement est-il un critère « dur », dont l'évaluation ne peut pas être influencée subjectivement par l'investigateur ?*
  Le décès est le critère le plus sûr dans un essai en ouvert car il ne demande aucune interprétation. Par contre, l'utilisation d'événements cliniques est moins robuste. Dans certains cas, le diagnostic de

survenue de l'événement clinique peut être subjectif et influencé par la connaissance du traitement du patient.

Tableau 1 – Liste des situations où l'absence de double insu est « acceptable ».

Un des traitements comparés est une intervention chirurgicale ou invasive (radiologie interventionnelle comme une angioplastie).
Un des traitements comparés nécessite un appareillage lourd dont il est impossible de faire un simulacre comme la radiothérapie.
Un des traitements comparés s'accompagne d'effet indésirable ou d'une toxicité évocatrice qui laisse deviner la nature du traitement dans presque tous les cas : chute de cheveux dans des chimiothérapies anticancéreuses.
Les traitements comparés sont des stratégies de prise en charge : traitement à domicile versus traitement hospitalier.
Un des traitements comparés concerne une prise en charge améliorée : stroke unit, kinésithérapie, aide à domicile, etc.
Le traitement factice risque d'avoir un effet : faux massage, placebo de chewing-gum pour l'arrêt du tabac, etc.
Un des traitements comparés délivre son action de façon évidente et non dissimulable. Il est donc impossible d'en faire un simulacre sans effet : (chirurgie,) dans une certaine mesure kinésithérapie, cure thermale, physiothérapie (chaleur), etc.
D'une manière générique, toutes les situations où la réalisation d'un traitement « placebo » ayant la même apparence que le traitement étudié s'avère trop compliqué à réaliser ou illusoire, par exemple, quand l'action du traitement est directement visible (comme la chirurgie, le recours à une aide humaine, etc.).

- *En cas d'utilisation d'événements cliniques comme critère de jugement, l'adjudication s'est-elle effectuée de manière centralisée, indépendante et en insu de la connaissance du traitement ?*
- *L'essai est réalisé en ouvert alors que sa réalisation en double insu était éthiquement et pratiquement possible.*
  La justification de l'absence d'aveugle pour des raisons pratiques, principalement de coûts, ne doit pas être acceptée trop facilement. L'expérience montre que, même avec des critères de jugement « durs » (mortalité), il existe une surestimation de l'effet dans les essais en ouvert par rapport aux essais en double aveugle (cf. exemple de l'amiodarone ci-dessous). Les situations où il est impossible de réaliser un double insu sont rares. Par exemple, la nécessité d'une adaptation posologique en fonction d'un paramètre biologique n'est pas un obstacle insurmontable à la réalisation d'un double aveugle. Une procédure d'ajustement centralisé peut être mise en place.

Exemple

La méta-analyse des essais comparant l'amiodarone au placebo ou à l'absence d'antiarythmiques dans l'insuffisance cardiaque congestive ou en post infarctus précoce montre une réduction significative de la mortalité totale et d'origine arythmique. Cependant, l'analyse restreinte aux essais en double aveugle contre placebo conduit à des résultats non significatifs, qui s'avèrent hétérogènes par rapport aux résultats des essais en ouvert, sans placebo, qui lui est significatif.

Type d'essais	Mortalité totale Risque relatif (95%)
Essais contre placebo	0,96 (0,84 ;1,10)
Essais contre pas d'antiarythmiques	0,64 (0,50 ;0,82)
Tous les essais	0,87 (0,78 ;0,99)

Cet exemple est l'un des cas où il a été mis en évidence que les essais réalisés en ouvert, sans recourir au placebo, avaient une certaine propension à surestimer l'efficacité et à perturber l'interprétation des résultats [15].

Biais d'attrition

Rappel

Le biais d'attrition survient quand des patients randomisés sont écartés de l'analyse. Tous les patients randomisés doivent être inclus dans l'analyse. Les patients inclus mais non analysés correspondent soit à des perdus de vue, soit à des données manquantes, ce qui a pour conséquence dans les deux cas de rendre le critère de jugement principal manquant.

Questions à se poser pour vérifier les précautions prises pour éviter le biais

- *Le nombre de patients analysés est-il égal au nombre de patients randomisés ?*
- *Qu'elle est la robustesse du résultat vis-à-vis de l'hypothèse du biais maximum ?*
- *Est-ce qu'une méthode de remplacement des données manquantes a été utilisée ? Dans ce cas, le nombre de patients analysés correspond au nombre de patients randomisés même si de nombreuses valeurs étaient manquantes. Ces méthodes nécessitent des hypothèses sur la nature des données manquantes. Même si elles sont pour la plupart conservatrices, leur utilisation ne doit pas faire oublier le problème initial et le risque de biais.*

Autres biais liés à la destruction de la comparabilité des groupes

Rappel

Différentes situations peuvent conduire à une destruction de la comparabilité initiale des groupes, comme, par exemple, une analyse en « per-protocole » où les patients inclus à tort, traités par erreur avec un mauvais traitement, ayant arrêté le traitement de l'étude ou ayant reçu des traitements concomitants sont exclus de l'analyse. Ces exclusions secondaires sont susceptibles de biaiser le résultat, principalement en détruisant la comparabilité initiale des groupes et du fait que les exclusions sont potentiellement liés à l'effet du traitement. Pour éviter ce biais, l'analyse doit être réalisée en intention de traiter. Le diagramme de flux de patients des recommandations « CONSORT » permet de juger de la population soumise à l'analyse.

Questions à se poser pour vérifier les précautions prises pour éviter le biais

Afin de vérifier l'absence d'un éventuel biais, il convient de se poser les questions suivantes.

- *L'analyse a-t-elle été faite en intention de traiter ?*
  C'est-à-dire tous les patients inclus dans l'essai ont-ils été analysés dans le groupe dans lequel ils ont été randomisés, quel que soit le traitement qu'ils ont reçu ?
- *Les patients randomisés mais non traités sont retenus pour l'analyse.*
- *Les patients alloués à un groupe mais traités par erreur avec le traitement d'un autre groupe sont analysés dans leur groupe d'origine.*

Biais des essais de non-infériorité

Rappel

Les biais spécifiques affectent l'essai de non-infériorité, en particulier, tout ce qui concourt à faire disparaître l'effet des traitements étudiés. La situation est inversée par rapport à l'essai de supériorité où ces situations n'entraînent pas de biais mais simplement une perte de puissance.

Questions à se poser pour vérifier les précautions prises pour éviter les biais dans les essais de non infériorité

- *Le traitement de référence a-t-il développé sa pleine efficacité ?*
  Les conditions d'administration du traitement de référence (dose utilisée, schéma d'administration, observance des patients) doivent garantir l'obtention de l'efficacité optimale du traitement de référence. Si ce n'est pas le cas, un nouveau traitement, en réalité, inférieur au traitement de référence, apparaîtrait comme non-inférieur.
- *Les patients inclus sont-ils similaires aux patients chez lesquels le traitement de référence a été validé ?*
- *Les patients inclus présentent-ils un risque suffisamment élevé pour permettre à l'effet du traitement de se manifester. La fréquence du critère de jugement doit être proche de celle qui est attendue et qui a été utilisée dans le calcul du nombre de sujets.*
- *L'analyse en intention de traiter donne-t-elle les mêmes résultats que l'analyse en per-protocole ?*
  Dans l'essai de non-infériorité, l'analyse per-protocole est la plus sensible et la moins biaisée. Cependant, elle ne reflète pas la vraie vie. L'analyse en intention de traiter est plus représentative de la pratique courante, mais elle est conservatrice et a tendance à faire disparaître les différences. Il convient donc de considérer simultanément ces deux analyses pour avoir à la fois une vue non biaisée et représentative de la réalité.

Situations à fort risque de biais

Dans les situations suivantes, le risque de biais dans l'essai de non-infériorité est fort et remet en cause la validité interne du résultat obtenu.

- *La mesure du critère de jugement est peu sensible et/ou peu spécifique. La mauvaise performance diagnostique de cette mesure tend à égaliser les résultats des deux groupes, et peut gommer une différence en défaveur du traitement étudié.*
- *De nombreux patients sont exclus de l'analyse per-protocole.*
- *Il existe un fort taux d'écarts au protocole.*
- *Le taux de données manquantes était élevé et des techniques de remplacements ont été utilisées. Ces techniques sont conservatrices et elles sont susceptibles de faire disparaître une réelle différence entre les traitements.*

Évaluation de la pertinence clinique

Introduction

L'évaluation de la pertinence clinique (« clinical relevance ») permet de s'assurer que le bénéfice apporté par le traitement est suffisamment important et concerne un critère cliniquement pertinent, que la balance bénéfice / effets indésirables est acceptable et que ce résultat est informatif pour les situations de la pratique médicale courante (résultat extrapolable).

Tableau 76 – Principales composantes de l'estimation de la pertinence clinique

L'estimation de la taille de l'effet doit être suffisamment précise pour pouvoir raisonnablement éliminer la possibilité que l'effet puisse être trop petit pour avoir un intérêt en pratique.
L'effet du traitement étudié doit être déterminé par rapport à un comparateur adapté : placebo ou traitement de référence validé.
Les patients de l'essai doivent être représentatifs des patients vus en pratique médicale courante afin d'assurer l'informativité et l'extrapolabilité du résultat : même définition de la maladie, pas de sélection excessive sur le sexe, l'âge, les comorbidités, etc.. En particulier, ils ne doivent pas avoir été sur-sélectionnés.

Pertinence clinique de l'objectif

Introduction

L'objectif de l'essai (« aims » ou « objective ») doit être clairement formulé et correspondre à une hypothèse testable. L'hypothèse doit pouvoir être justifiée à partir des connaissances disponibles au moment de la planification de l'essai (induction prospective de l'hypothèse et non pas justification rétrospective de l'hypothèse).

Pour être parfaitement défini, l'objectif doit préciser :

- le traitement testé,
- le traitement contrôle (placebo ou traitement actif),
- s'il s'agit d'une recherche de supériorité ou d'équivalence,
- le critère de jugement principal (et le moment de sa mesure),
- les patients concernés : maladie et éventuellement caractéristiques particulières.

« Montrer que l'alteplase entraîne une réduction supplémentaire de la mortalité à court terme par rapport à la streptokinase à la phase aiguë de l'infarctus » est un objectif correctement formulé. « Évaluer le ramipril dans l'hypertension » n'est pas un objectif énoncé avec suffisamment de précision.
L'objectif de l'essai doit correspondre à une question pertinente, représentant un problème réel de thérapeutique (jugé par l'expérience du lecteur et non pas uniquement à partir du rationnel de l'essai), pour lequel il n'y a pas encore de solution satisfaisante (jugée par l'étude de la méta-analyse centrée sur la question afin de connaître et de juger les résultats obtenus par les traitements concurrents).

Il existe actuellement des tentatives de création de toutes pièces de pseudos questions médicales pour donner matière, de façon tout à fait artificielle, à l'utilisation d'un traitement particulier. Le British Medical Journal retrace l'histoire de l'invention d'un syndrome de trouble sexuel chez la femmes [1], inventé de toute pièce lors d'une réunion sponsorisée par les laboratoires Pfizer. Cette création était motivée par l'intention de donner au sidenafil une indication chez la femme. L'histoire a été jusqu'à la réalisation d'essais thérapeutiques qui ce sont révélés négatifs, contrecarrant ainsi les plans initiaux.

Liste de contrôle 1 – Point à vérifier pour établir la pertinence clinique de l'objectif
• Pertinence de la question thérapeutique. • Pertinence du critère de jugement principal. • Pertinence du traitement testé. • Pertinence du traitement contrôle. • Quelle est l'hypothèse de l'essai : supériorité ou non-infériorité ?

Traitement de comparaison adapté

Les éléments de pertinence du traitement de comparaison sont exposés dans le chapitre : Comparateur. Nous ne reprendrons ici que les grandes lignes.

***Un essai contre placebo**, alors qu'il existe un traitement ayant déjà montré son efficacité par rapport au placebo, ne peut pas justifier l'utilisation du nouveau traitement en remplacement du traitement de référence. En effet, la supériorité du nouveau traitement par rapport au traitement de référence est inconnue. Il n'est pas exclu que le nouveau traitement soit, en fait, moins efficace que ce dernier.*

Faute de mieux, il est possible de comparer l'efficacité de ces deux traitements dans une comparaison indirecte qui permet d'estimer l'efficacité du nouveau traitement par rapport au traitement de référence à partir de son efficacité par rapport au placebo et de celle du traitement de référence par rapport au placebo (cf. section : Comparaison indirecte du chapitre : Comparateur,). Mais l'estimation de l'efficacité par cette comparaison indirecte n'est qu'une extrapolation sans garantie de fiabilité. Tout au plus, le nouveau traitement pourra servir d'alternative en cas d'intolérance ou d'effet secondaire avec le traitement de référence, mais il ne peut pas faire l'objet d'une utilisation systématique.

Cette règle est à nuancer, par exemple, dans les cas où la démonstration de l'efficacité du traitement de référence est peu solide. La comparaison d'un nouveau traitement au placebo, et non pas à ce traitement, est justifiée.

*A contrario, si l'essai est réalisé **contre un traitement actif**, sans que ce dernier ait été évalué, le résultat de l'essai sera peu informatif vis-à-vis du nouveau traitement. L'efficacité du traitement servant à la comparaison étant inconnue, il n'est pas possible d'exclure que ce traitement soit en fait délétère. Même si le nouveau traitement s'avère supérieur, il n'est pas possible d'en déduire qu'il est efficace et qu'il se serait révélé supérieur au placebo. Cependant dans certains cas où des traitements sont standards du fait de l'usage, ce type d'essai montre que les nouveaux traitements permettent de faire mieux (ou au pire moins mal) que les pratiques standards de soins de cette maladie.*

Dans les essais de supériorité contre traitement de référence, ainsi que dans les essais de non-infériorité, il convient aussi de se méfier d'une utilisation non optimale du comparateur entraînant une diminution de son efficacité. Dans ce cas un nouveau traitement moins efficace que le traitement de référence pourrait donner l'impression du contraire. Ce genre de situation se détecte en recherchant si le traitement de référence est utilisé dans les conditions où il a révélé son efficacité optimale (dose, schéma d'administration, types de patients, etc.) et s'il a été effectivement utilisé de cette façon (dose moyenne, dose moyenne rapportée au poids, durée moyenne, fréquence des arrêts de traitement). Le recours aux publications de l'essai de validation du traitement de référence est indispensable.

Liste de contrôle 2 – Point à vérifier pour établir la pertinence clinique du traitement de comparaison.
• En cas de comparaison au placebo, il n'existe pas de traitements de même mécanisme d'action ayant démontré leur efficacité. • En cas de comparaison à un traitement actif, l'efficacité du comparateur est établie. • Le traitement de référence actif a été administré de façon propice au développement de son efficacité optimale.

Traitements concomitants

La liste des traitements interdits par le protocole doit aussi être analysée ainsi que les taux d'utilisation effective des différents traitements concomitants. Elle peut révéler que le nouveau traitement n'a montré une efficacité que chez des patients sous-utilisant les traitements concomitants et ne bénéficiant pas de tous les traitements validés. Dans ce cas, la réelle efficacité du nouveau traitement utilisé conjointement avec les autres traitements est inconnue, de même que le risque d'interaction statistique et/ou pharmacologique.

Liste de contrôle 3 – Point à vérifier pour établir la pertinence clinique des traitements concomitants.
• Le traitement étudié a été évalué conjointement avec les autres traitements utilisés pour la pathologie considérée et avec des taux d'utilisation de ces traitements conformes à ceux habituellement rencontrés en pratique.

Pertinence de la taille de l'effet

L'estimation de la taille de l'effet (« size of effect *») doit être suffisamment précise pour pouvoir raisonnablement éliminer le fait que l'effet puisse être petit et donc sans intérêt en pratique. (voir le chapitre :* Intervalle de confiance*).*

La difficulté réside dans la fixation d'une valeur pour le « plus petit bénéfice intéressant en pratique ». Cette détermination pose des problèmes similaires à celui de la fixation du « seuil » dans les essais de non-infériorité (cf. chapitre *Essai d'équivalence*). La détermination de cette valeur, qui ne peut être qu'arbitraire, doit prendre en compte de nombreux facteurs :

la gravité de la pathologie,
les autres possibilités de traitement,
les autres intérêts du traitement (tolérance, coût, facilité d'administration, satisfaction des patients),
la fréquence de la maladie : un petit bénéfice peut correspondre, si la maladie est fréquente, à un nombre substantiel d'événements sérieux évités au niveau de la population toute entière,
les choix politiques de santé publique.

Des problèmes de même nature apparaissent dans les comparaisons indirectes : quelle différence d'efficacité entre deux traitements actifs représente vraiment un gain pertinent ?

Pertinence des critères de jugement

La pertinence du critère de jugement peut être remise en cause dans de nombreuses situations :

- *Le critère reflète un mécanisme biologique ou pharmacologique non directement lié à l'objectif thérapeutique.*
- *Dans un critère composite, la composante qui a le plus de poids (qui est la plus fréquente et/ou qui est le plus modifiée par le traitement) a une faible pertinence clinique (critère non clinique, ou critère clinique secondaire).*
- *Les critères continus posent assez souvent des problèmes au niveau de l'interprétation de leur pertinence clinique ou de la pertinence des effets observés.*

Liste de contrôle 1 – Point à vérifier pour établir la pertinence clinique du critère de jugement principal.
• Le critère de jugement est un critère clinique (ou un critère de substitution validé). • Le critère correspond à un objectif thérapeutique pertinent pour le patient. • Si le critère est composite, chaque composante a la même pertinence clinique. • Le moment de mesure du critère est pertinent.

Praticabilite des traitements

Une partie de la pertinence clinique d'un résultat dépend de la praticabilité du traitement étudié. Une utilisation complexe, une mauvaise tolérance, un coût élevé ou la nécessité d'investissement important, sont autant de facteurs qui peuvent limiter la pertinence d'un résultat, même en cas d'efficacité substantielle.

Par exemple, à la phase de l'infarctus, l'alteplase en perfusion accélérée a montré sa supériorité par rapport à la streptokinase [2]. Cette perfusion accélérée consiste à administrer une dose de 15mg en bolus suivie par une perfusion de 0,75mg/kg durant 30 minutes (sans excéder 50mg) puis 0,5mg/kg durant les 60 minutes suivantes (sans dépasser 35mg). La relative complexité de ce mode d'administration a motivé la recherche de mode d'administration plus simple (en double bolus) [3].

La documentation des changements de posologie qui ont eu lieu dans l'essai permet d'apprécier quelle fut la praticabilité du traitement tel qu'il était prévu au protocole. De nombreuses diminutions de posologie peuvent laisser supposer une posologie initiale inadéquate. De multiples abandons du traitement révèlent un traitement soit trop lourd soit mal supporté. Dans ces situations, le traitement sera difficilement utilisable en pratique même s'il montre un bénéfice malgré ces arrêts de traitement.

L'analyse des effets secondaires documente aussi la tolérance du traitement et sa praticabilité. Un traitement mal supporté peut ne pas présenter d'avantage par rapport à un autre mieux toléré, même s'il possède une efficacité plus importante (cf. justification de l'essai d'équivalence).

La réalisation d'essais pragmatiques est particulièrement nécessaire avec les traitements mal supportés. En effet, dans le cadre très protocolisé d'un essai, un traitement est moins facilement arrêté pour mauvaise tolérance que dans un contexte plus libre : les patients sont plus sollicités pour continuer et/ou des mesures d'accompagnement sont plus facilement prises. Dans ce contexte, le bénéfice mesuré surestimera celui qui sera effectivement obtenu dans la vraie vie où le traitement sera plus facilement arrêté.

Liste de contrôle 4 – Point à vérifier pour établir la pertinence clinique de la praticabilité du traitement
• Quelle est la praticabilité du traitement étudié ? • Le traitement a été arrêté en cas de mauvaise tolérance de manière similaire à ce qui se passe en pratique.

Bibliographie

1. *Moynihan R. The making of a disease: female sexual dysfunction. Bmj 2003;326(7379):45-7.* PMID: 12511464.

2. *The Global Use of Strategies to Open Occluded Coronary (GUSTO 3) Investigators. A comparison of reteplase with alteplase for acute myocardial infraction. NEJM 1997;337:1118-23.* PMID:

3. *The Continuous Infusion versus Double-bolus Administration of Alteplase (COBALT) Investigators. A comparison of continuous infusion of alteplase with double-bolus administration for acute myocardial infarction. NEJM 1997;337:1124-30.* PMID:

METHODOLOGIE DES ESSAIS THERAPEUTIQUES

Les principes de base de l'essai thérapeutique

Powerpoint

La mise en évidence du bénéfice clinique apporté par un traitement est un parcours plein d'embûches : les biais, qui font courir le risque de conclure à tort à l'efficacité du traitement. Les principes méthodologiques ont été développés pour éviter ces pièges et conduisent à la méthode de l'essai contrôlé randomisé en double insu.

Les principes méthodologiques de l'essai évitent la survenue des biais

Un critère de jugement (« *endpoint* » ou « *outcome* ») permettant une quantification va être utilisé pour mettre en évidence « numériquement » l'effet d'un traitement (« *treatment effect* »). Les critères de jugement sont des variables dont la valeur numérique est susceptible de changer sous l'effet du traitement. Ils permettront donc par leur modification de mettre en évidence un effet du traitement étudié. Ces critères peuvent être de nature très variée : durée d'une maladie, fréquence de survenue d'un événement, taux de mortalité, score clinique, etc.

Le but de l'essai thérapeutique est d'établir qu'il existe une relation causale entre une modification de cette variable et l'administration d'un traitement. Cependant l'attribution d'une modification du critère de jugement à un effet du traitement ne peut être faite que sous certaines conditions. En effet, il existe de nombreux phénomènes qui interfèrent avec l'éventuelle relation existant entre le traitement et le critère de jugement. En particulier ces phénomènes sont susceptibles d'entraîner une modification du critère de jugement, même si le traitement est, en réalité, sans effet ; pouvant ainsi faire croire à un effet du traitement. Ces phénomènes ayant des effets pouvant être confondus avec l'effet du traitement sont appelés facteurs de confusion, ou facteurs de méprise car ils conduisent l'investigateur à se méprendre sur la cause des modifications observées au niveau du critère de jugement. Les principes méthodologiques ont pour objectifs de contrôler ces facteurs de confusion et d'éviter leur influence nuisible.

L'essai prospectif contrôlé randomisé en double aveugle analysé en intention de traiter est à l'abri des biais

La suite de ce chapitre décrit les différents obstacles qui surviennent dans la recherche de preuves fiables de l'efficacité d'un traitement et les principes méthodologiques qui permettent de les contourner. L'encadré suivant les présente en quelques lignes.

Les grands principes de l'essai thérapeutique

Pour mettre en évidence de façon fiable (sans biais) l'effet d'un traitement, un essai doit être :

- **Prospectif** : les données sont recueillies spécialement pour répondre à la question posée dans l'essai.
- **Comparatif** : l'effet du traitement est déterminé par rapport à un groupe contrôle qui prend en compte les facteurs de confusion.
- **Randomisé** : la randomisation (allocation aléatoire des traitements) produit des groupes de patients initialement comparables en moyenne, soumis de la même manière aux facteurs de confusion et qui ne diffèreront que par le traitement qui leur sera appliqué dans l'essai. De ce fait une différence observée à la fin proviendra du traitement étudié.
- En **double aveugle** : afin d'assurer que les groupes comparés sont suivis de la même façon et qu'ils ne différeront durant le suivi que par les traitements appliqués.
- **Sans donnée manquante** et analysé **en intention de traiter** : le devenir de chaque patient inclus dans l'essai est pris en considération dans l'analyse et les patients ne doivent pas être changés de groupe.

Les biais

Le terme biais (« *bias* ») est omniprésent dans le discours se rapportant aux essais cliniques et il est souvent mal employé. Une partie de la confusion existant dans l'emploi de ce terme vient qu'il a plusieurs significations en fonction du contexte ou du domaine dans lequel il est employé.

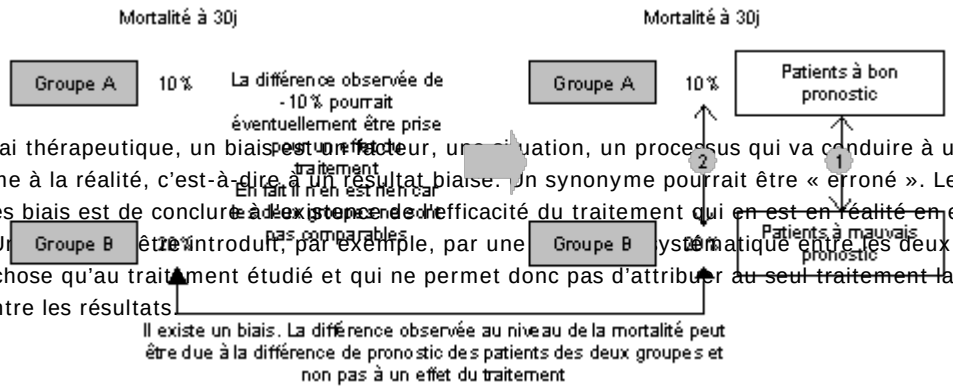


Figure 1 – Lorsque l'effet d'un traitement est recherché en comparant deux groupes de patients, l'un recevant le traitement étudié et l'autre un autre traitement, la comparaison est potentiellement biaisée quand, par exemple, il existe une différence entre les deux groupes au niveau d'un facteur autre que le traitement ① qui est susceptible d'entraîner une différence dans le résultat, même en l'absence d'effet du traitement ②.

Définition

Pour un essai thérapeutique, un biais est un facteur, une situation, un processus qui va conduire à un résultat non conforme à la réalité, c'est-à-dire à un résultat biaisé. Un synonyme pourrait être « erroné ». Le risque que fait courir les biais est de conclure à l'existence de l'efficacité du traitement qui en est en réalité en est dépourvu. Un biais peut être introduit, par exemple, par une différence systématique entre les deux groupes, dû à autre chose qu'au traitement étudié et qui ne permet donc pas d'attribuer au seul traitement la différence observée entre les résultats.

Il y a biais quand la différence observée à la fin de l'essai entre les deux groupes est due à un autre facteur que le traitement étudié

En statistique, le biais est une erreur systématique entre une estimation et la véritable valeur du paramètre estimé. Ainsi, le biais est une erreur qui se reproduit à l'identique (systématique) et qui, contrairement aux erreurs aléatoires, ne se compense pas en moyenne. Un estimateur biaisé est un estimateur qui donne des estimations qui comportent toujours la même erreur. Ainsi l'acception

courante du terme biais dans le domaine des essais est un cas particulier de la définition statistique du biais.

Le problème posé par les résultats biaisés est la difficulté de leur détection. Conclure qu'un résultat est faux nécessite de connaître la réalité, la vraie valeur de l'effet traitement, qui par essence est inconnue. Comme **il est impossible de savoir, a posteriori, si un résultat est biaisé** ou pas, des « astuces » ont été inventées pour **rendre les biais impossibles**. Ce sont les principes méthodologiques, qui, s'ils sont correctement suivis empêchent la survenue des biais. On dit qu'il y a contrôle des biais (dans le sens où l'on empêche leur survenue) et que le résultat est à l'abri des biais. Ainsi ne se pose plus la question de s'avoir si le résultat est biaisé ou non mais seulement la question : « est ce que la méthodologie utilisée empêchent les biais » ou « qu'elles sont les causes résiduelles de biais potentiel ».

En lecture critique, on dira que le résultat d'un essai est potentiellement biaisé quand la méthode utilisée ne permet pas de contrôler la survenue d'un biais. Devant un résultat, il est impossible de dire si celui-ci est effectivement biaisé ou non (car on ne connait pas la réalité), mais il est possible de dire si celui-ci est à l'abris des biais ou non. Par défaut de langage, on dit souvent qu'un résultat est biaisé à la place de dire qu'il n'est pas à l'abri des biais.

Dans le langage courant le terme biais est utilisé à la fois pour désigner la cause et l'erreur du résultat. On appelle ainsi biais les causes de biais, c'est-à-dire les conditions qui peuvent conduire à un résultat biaisé. On parle ainsi d'un biais de sélection quand les deux groupes d'un essai ne sont pas strictement comparables.

Contrairement aux fluctuations aléatoires (appelées aussi erreurs aléatoires « *random error* ») dont les conséquences sont diminuées par l'augmentation du nombre de sujets, l'erreur induite par un biais reste constante quel que soit le nombre de sujets.

Biais et plan d'expérience

Les principes méthodologiques ont été conçus pour éviter la survenue des biais.

Le but de la méthodologie est de construire des expériences qui ne sont pas exposées au risque de biais. Pour cela, le plan d'expérience (« *design* ») est construit de telle façon qu'il ne laisse pas la possibilité à un biais de venir fausser les résultats obtenus. On dit qu'il y a **contrôle des biais**. La plupart des principes méthodologiques trouvent leur justification dans le fait qu'ils évitent l'apparition d'un biais.

Tous les plans d'expériences ne permettent cependant pas de garantir de la même manière l'absence de biais. Par exemple, une étude d'observation ne contrôle pas le biais de sélection et ne garantit pas que les deux groupes seront strictement comparables. Il existe une hiérarchisation des plans d'expériences quant à leur aptitude à assurer l'absence de biais dans la recherche d'un effet thérapeutique. L'essai thérapeutique est celui qui contrôle le mieux les biais, mais il n'est pas toujours.

Biais et représentativité

Le biais est à distinguer du **manque de représentativité** qui entache un résultat obtenu sur des sujets sélectionnés et peu représentatifs de la situation étudiée. Le résultat obtenu n'est pas biaisé, mais il n'est pas extrapolable à l'ensemble des patients. Le résultat n'est pas biaisé car il estime correctement l'effet du traitement chez ces types particuliers de patients, mais cette estimation n'est pas celle recherchée (estimation de l'effet du traitement chez les patients standards). Le problème de la représentativité d'un résultat sera abordé ultérieurement dans cet ouvrage.

Groupe contrôle

Les limites de l'observation ponctuelle

L'administration d'un nouveau traitement antiviral à un patient souffrant d'un « rhume banal » a été suivi de la guérison complète du patient. Bien évidemment, ce type d'argument n'est pas acceptable comme preuve de l'efficacité du traitement car le rhume est une maladie qui guérit spontanément. L'observation de la survenue de la guérison après l'administration du traitement ne peut pas être attribuée avec certitude au traitement car elle serait aussi survenue spontanément. L'évolution spontanée de la maladie peut ainsi être confondue avec l'effet du traitement et conduit à se méprendre sur l'origine de la guérison observée. L'évolution naturelle est un facteur de confusion, c'est-à-dire un facteur qui produit un effet similaire à celui que l'on attend du traitement et pouvant ainsi être mis sur le compte du traitement même si celui-ci est dépourvu d'efficacité.

De même que l'observation d'un épistaxis après la prise du traitement ne peut pas être mis sur le compte d'un effet indésirable du traitement, ce genre de phénomène émaillant spontanément l'évolution des rhinites virales.

Ces propos peuvent paraître triviaux. Il faut cependant remarquer que la confiance en l'anecdote est assez fréquente dans le grand public et parfois s'immisce dans le monde médical. La revue Prescrire (février 1999, page 118) relate quelques exemples de cet ordre comme : « Essayez, faites vous une idée ; cela vaut mieux que toutes les études du monde ! ».

Les réserves mises sur ces observations portant sur l'évolution d de patients ne veulent pas dire que les observations sont toujours erronées en efficace, il entraîne une guérison accélérée du rhume.

Le problème n'est pas que l'observation ne reflète jamais la réalité, mais qu'il est impossible de faire la part des choses entre un effet du traitement étudié et l'effet des facteurs de confusion.

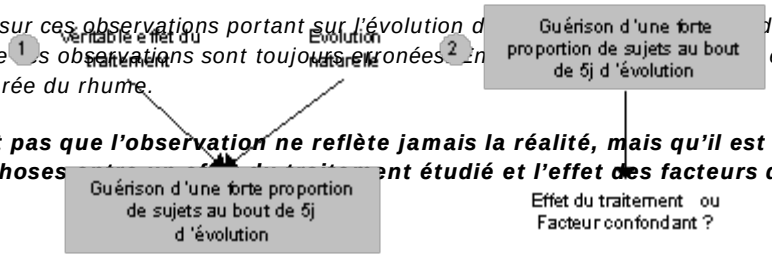


*Figure 1 – **(1)** Observer, qu'après prise d'eau minérale pendant 5 jours, 78% des sujets présentant un rhume banal n'ont plus de symptômes, n'apporte pas la preuve de l'efficacité de l'eau minérale. Il est évident dans cet exemple que l'effet observé correspond à l'évolution naturelle du rhume. Néanmoins cette évolution naturelle se traduit par le même résultat qu'un effet bénéfique du traitement, c'est-à-dire une guérison chez un certain nombre de patients. Comme cette évolution naturelle peut-être confondue avec un effet du traitement, elle est appelée facteur confondant.*
***(2)** L'évolution des maladies durant la prise d'un traitement ne permet pas de démontrer l'efficacité d'un traitement. Les résultats observés peuvent être dus à des facteurs confondants, indépendamment de tout effet bénéfique du traitement.*

Les facteurs de confusion

Les facteurs de confusion produisent des effets qui peuvent être pris pour ceux d'un traitement qui en fait est dépourvu d'efficacité.

Plusieurs facteurs produisent des effets pouvant être confondus avec un effet du traitement étudié. Les principaux facteurs de confusion (« confounding factor *») sont :*

- *l'évolution naturelle de la maladie,*
- *l'effet placebo,*
- *la régression à la moyenne,*
- *l'effet de traitements concomitants,*

mais bien d'autres existent.

Nous ne détaillerons pas l'évolution naturelle qui peut aussi bien faire croire à l'efficacité d'un traitement (lorsqu'elle consiste en une amélioration spontanée), qu'à l'inefficacité en cas d'aggravation spontanée de la maladie. Le biais de confusion induit par le recours à des traitements concomitants est lui aussi évident. Par contre, les autres facteurs de confusion nécessitent un peu plus d'explications.

L'effet placebo

La prise en charge médicale d'un patient est susceptible d'entraîner une modification de son état en dehors de toute administration d'un traitement actif. Cette amélioration apportée par des facteurs intangibles est appelée effet placebo. L'effet placebo peut être expliqué par des effets de conditionnement, d'auto et d'hétéro suggestion. Ainsi l'administration d'un traitement sans aucune activité biologique peut, par effet placebo, améliorer l'état des patients et faire croire à une efficacité du traitement.

L'effet placebo apparaît dans une expérience menée chez des volontaires qui étaient soumis à une épreuve stressante, et chez qui, était mesuré l'intensité de la réaction anxieuse (cf. Figure 2). Avant cette épreuve, les sujets prenaient des gélules qui contenaient soit un anxiolytique soit un placebo. De plus ils recevaient une information sur la nature du produit qu'ils prenaient, mais cette information ne correspondait pas forcément à la réalité. Ce plan expérimental crée quatre groupes de sujets :

- *ceux qui reçoivent le placebo et à qui l'on dit qu'ils prennent le placebo,*
- *ceux qui reçoivent l'anxiolytique mais à qui l'on dit qu'ils prennent le placebo,*
- *ceux qui reçoivent le placebo mais à qui l'on dit qu'ils prennent l'anxiolytique,*
- *ceux qui reçoivent l'anxiolytique et à qui l'on dit qu'ils prennent l'anxiolytique.*

Les sujets qui prennent l'anxiolytique, en croyant que c'est le placebo, manifestent un effet anxiolytique moindre que l'effet maximum obtenu chez les sujets qui prennent l'anxiolytique en le sachant. La réduction d'effet entraînée par "l'autosuggestion" est appelée l'effet nocebo.

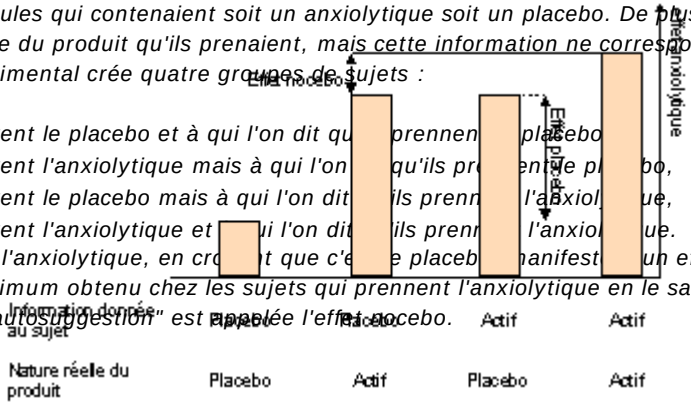


Figure 2 – Mise en évidence expérimentale de l'effet placebo et de l'effet nocebo.

Les sujets qui prennent le placebo, et qui donc ne bénéficient d'aucun effet anxiolytique exogène, mais à qui on fait croire qu'ils prennent l'anxiolytique, ont cependant une diminution de leur anxiété plus importante que ceux du groupe placebo-placebo. La différence entre ces deux groupes est l'effet apporté par le fait de « croire » que l'on prend un traitement actif.

Les sujets qui reçoivent le placebo, et qui savent qu'ils le prennent, ont l'effet anxiolytique le plus petit.

Les sujets qui reçoivent l'anxiolytique et à qui on fait croire qu'ils prennent l'anxiolytique ont l'anxiolyse la plus importante. Ils ont le bénéfice maximal : celui de la molécule et celui de « l'autosuggestion ».

En conclusion, le simple fait de prendre en charge un sujet peut produire un effet favorable. C'est l'effet placebo. Ainsi un traitement sans efficacité peut, par cet effet placebo, produire un apparent effet favorable qui est alors confondu avec un effet propre du traitement.

Exemple

L'effet placebo est parfois extrêmement important dans certaine pathologie, comme par exemple le colon irritable. Dans les groupes placebo de 45 essais randomisés, une amélioration globale des symptômes est observée chez 16% à 71% des patients (en moyenne 40%)(1).

La régression à la moyenne

La régression à la moyenne (« regression toward the mean ») est un phénomène purement statistique qui survient lorsque l'on s'intéresse à un groupe de sujets sélectionnés comme ayant une valeur du paramètre d'intérêt supérieure ou inférieure à un seuil. Comme, par exemple, un groupe de sujets hypertendus sélectionnés d'après une valeur de pression artérielle diastolique (PAD) supérieure à 90 mmHg. La régression à la moyenne va entraîner une diminution de la PAD moyenne du groupe au cours du temps, sans qu'il y ait de véritable évolution de la pression artérielle des sujets. Ce phénomène, purement statistique, lié à la grande variabilité intra-individuelle de la pression artérielle et aux erreurs de mesure pourrait donc faire croire à l'effet d'un traitement sur la pression artérielle.

> La variabilité intra individuelle est la variabilité des mesures réalisée successivement chez un même sujet. Il y a variabilité intra individuelle lorsque les valeurs varient d'une mesure à l'autre chez le même

individu. Cette variabilité est à distinguer de la variabilité inter-individuelle qui traduit des variations de valeur entre individus (d'un individu à l'autre).

Chez un même sujet, la pression artérielle est variable d'un moment à l'autre sous l'effet de nombreux paramètres physiologiques comme le stress, l'effort physique, etc. Cependant ces valeurs oscillent autour d'une valeur moyenne caractéristique du sujet : sa vraie valeur de PAD. Du fait de ces variations, la mesure de la PAD d'un individu, à un moment donné, peut révéler une valeur supérieure à la vraie valeur ou bien inférieure. Mais la moyenne d'un grand nombre de mesures est égale à la vraie valeur du sujet.

Au moment de la sélection, du fait de la variabilité de la PAD, certains sujets auront une valeur au dessus du seuil bien que leur vraie valeur soit en dessous. De même, l'expérimentateur pourra surestimer la valeur de la PAD du fait d'une erreur de mesure. Ces patients sont inclus à tort. En réalité leur PAD est en dessous de la valeur minimale recherchée, mais, au moment, de la sélection ils avaient par hasard une valeur située au-dessus.

Lorsque, quelque temps plus tard, la PAD est à nouveau mesurée, ces patients sélectionnés à tort auront une mesure probablement plus proche de leur vraie valeur, donc inférieure au seuil. Ces valeurs vont tirer la moyenne du groupe vers le bas.

Ainsi, ce phénomène de régression à la moyenne, lié uniquement à des sujets sélectionnés à tort dans le groupe du fait de la variabilité des mesures, va entraîner une diminution de la moyenne du groupe sans aucune modification de la vraie valeur des sujets.

Ce phénomène peut être réduit en sélectionnant les sujets, non plus sur une mesure, mais sur la moyenne de plusieurs. Cette valeur moyenne étant plus proche de la vraie valeur des sujets, il y aura moins de sujets sélectionnés à tort.

La prise en compte des facteurs de confusion

Le problème que pose la prise en compte des facteurs de confusion est de pouvoir faire la part des choses, dans le changement observé après l'administration d'un traitement, entre ce qui est dû aux facteurs de confusion et ce qui est dû au véritable effet du traitement.

La base de la prise en compte des facteurs de confusion est l'utilisation d'un groupe de référence appelé groupe contrôle.

Principes

La solution consiste à travailler comparativement à un groupe contrôle (« control group *»), qui ne reçoit pas le traitement étudié mais qui va subir les mêmes influences en provenance des facteurs de confusion que le groupe traité (avec le traitement étudié). Ce groupe contrôle constitue une référence, représentant ce qui se passe en l'absence d'effet thérapeutique (ce groupe n'est pas traité). L'effet propre d'un traitement pourra donc être déterminé en comparant l'évolution d'un groupe de patients traités à cette référence. Ce qui est observé dans ce groupe contrôle quantifie les effets des différents facteurs confondants. La part propre à l'effet du traitement étudié est obtenue en retranchant de ce que l'on observe sous traitement à ce qui est observé dans le groupe contrôle.*

Du fait de l'existence d'un groupe contrôle, l'essai est dit contrôlé *(«* controlled trial *»). C'est un essai* comparatif *qui compare un groupe traité à un groupe contrôle non traité afin de déduire l'effet propre d'un traitement.*

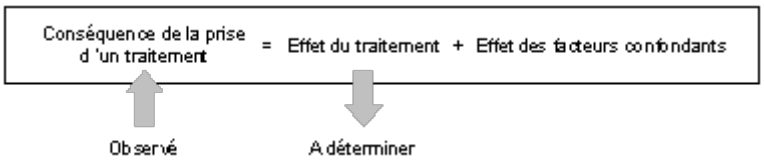


Figure 5 – La détermination de l'effet propre du traitement nécessite de pouvoir enlever des effets observés après l'administration d'un traitement ceux qui sont liés aux facteurs de confusion. Si les facteurs de confusion expliquent 100% de ce qui a été observé, le résultat de cette soustraction sera zéro, témoignant de l'absence d'effet du traitement. Si le résultat n'est pas nul, il s'agira de l'effet propre du traitement

L'absence de groupe contrôle induit un biais, appelé biais de confusion engendré par l'incapacité du plan d'expérience à prendre en compte l'effet des facteurs de confusion.

Pour annihiler complètement les effets des facteurs de confusions une autre condition est nécessaire en plus du groupe contrôle : celle que les 2 groupes subissent exactement la même influence des facteurs de confusion tout au long de l'essai. Ce sera l'objectif des autres principes méthodologiques (randomisation, aveugle, prévention de l'attrition).

Contrôle complet des biais engendrés par les facteurs de confusion = *Mesure de l'effet du traitement par rapport à un groupe contrôle*

Même influence des facteurs de confusion sur les 2 groupes

Application

Comment montrer qu'une nouvelle molécule a un effet hypocholestérolémiant ? La prise pendant 4 semaines du traitement étudié est associée avec une baisse du taux de cholestérol. Se pose alors la question de savoir si cette baisse est due à un effet propre du traitement ou aux facteurs de confusion. L'observation de l'évolution spontanée de la cholestérolémie dans un groupe de patients ne recevant pas le traitement étudié (Figure 6) permet d'appréhender l'effet des facteurs de confusion. Il apparaît alors qu'il existe un effet propre du traitement car, même s'il s'avère que spontanément le taux de cholestérol baisse sous l'action des facteurs de confusion, cette baisse n'est pas aussi importante que celle qui a été observée avec le traitement testé. L'effet propre du traitement est la différence existant entre ces deux groupes : C'est la part de la baisse de taux cholestérol observée avec le traitement qui ne peut pas être expliquée par l'effet des facteurs confondants.

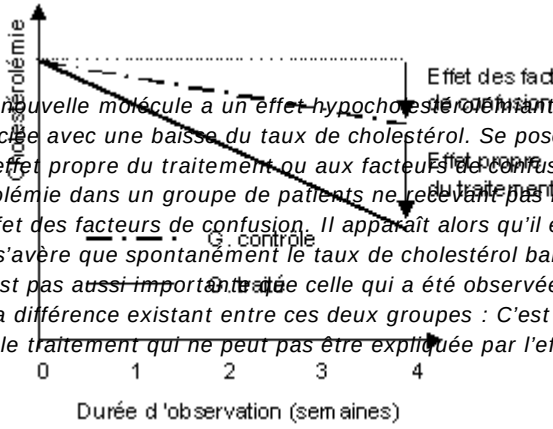


Figure 6 – Détermination de l'effet hypocholestérolémiant propre à un traitement par comparaison à un groupe contrôle.

L'effet observé dans le groupe placebo n'est pas seulement l'effet placebo. Il est égal à l'effet placebo ajouté aux effets des autres facteurs confondants.

Exemple

Un essai a comparé l'effet sur la réduction pondérale d'un nouveau traitement à un placebo chez 322 diabétiques de type 2. Le critère de jugement était la proportion de sujets ayant obtenu une perte de poids de plus de 5%. Dans le groupe placebo, 13,2% des patients ont présenté une telle perte de poids. Du fait de la fréquence importante de ces bonnes évolutions spontanées, la détermination de l'effet de ce traitement par rapport à un groupe contrôle est indispensable.

Nature du traitement administré dans le groupe contrôle

La nature du traitement que reçoit le groupe contrôle dépend de la question posée. Si cette question est d'évaluer l'efficacité « absolue » du traitement, le groupe contrôle ne doit recevoir aucun traitement actif. Par contre si la question est de savoir si le traitement étudié est plus efficace que les traitements déjà disponibles, le groupe contrôle doit recevoir le traitement standard (de référence).

Un des intérêts du placebo donné au groupe contrôle est de permettre la prise en compte de l'effet placebo en mettant les patients du groupe contrôle dans une situation identique à celle d'une prise en charge par un traitement réputé actif.

Dans le premier cas, le groupe contrôle doit être pris en charge de la même façon que s'il était traité, pour pouvoir prendre en compte l'effet placebo. En effet, nous avons vu que la simple prise en charge médicale, même sans administration de substance active, est susceptible d'entraîner une amélioration de l'état de santé par l'effet placebo. Si le traitement étudié est sans efficacité, les patients vont cependant bénéficier de cet effet placebo. Pour que les patients du groupe contrôle subissent eux aussi l'effet placebo, il est nécessaire qu'ils aient l'impression d'être pris en charge médicalement de la même façon. Pour cela, il convient de leur donner un traitement identique au traitement testé, mais qui n'a aucun effet. Ce traitement factice est appelé placebo.

Dans ce contexte, le placebo est un traitement identique au traitement étudié, s'administrant de la même façon, mais dénué d'activité biologique. Le placebo de médicament ne contient aucune substance active (en général du lactose) mais se présente sous la même galénique que le traitement étudié et sera administré de la même façon.

Le placebo a aussi un autre intérêt, celui de permettre le double insu et de faire que les deux groupes soient suivis de la même façon durant l'essai (cf. infra). Dans le cas d'un essai contre traitement de référence, les patients du groupe contrôle sont pris en charge médicalement. Il est donc inutile de donner un placebo sauf cas particulier (cf. infra double placebo).

Une mauvaise utilisation du groupe contrôle

Il arrive parfois que l'effet du traitement soit recherché de manière inadaptée par une comparaison avant-après effectuée dans le groupe recevant le traitement étudié. Le groupe contrôle est alors utilisé pour monter qu'il n'y a pas de différence sans le traitement. Cette procédure ne corrige pas la différence avant-après observée dans le groupe traité de l'effet des facteurs de confusion.

Exemple

Dans un essai de traitement de l'hyperuricémie par un uricosurique, l'uricémie dans le groupe expérimental est passée de 71 mg/l à 55 mg/l entre avant et après une période d'administration de l'allopurinol de 3 semaines ($p<0,05$). Dans le groupe contrôle, l'uricémie mesurée avant et après la prise d'un placebo durant la même durée est passée de 68 mg/l à 60 mg/l (différence statistiquement non significative). Cependant, la différence significative dans le groupe traité ne peut pas être avancée comme preuve de l'effet du traitement. Cet effet doit-être mesuré par la différence entre les deux groupes des changements d'uricémie entre avant et après la période de l'essai, soit : (55-71) - (60-68) = (-16) - (-8) = -8 ($p>0,05$). Le traitement abaisse de 8 mg/l l'uricémie de plus que le placebo, mais cette différence n'est pas statistiquement significative. Il n'est pas possible de conclure à l'efficacité de cet uricosurique.

La randomisation

L'utilisation d'un groupe contrôle est donc indispensable pour prendre en compte les facteurs de confusion et éviter le biais de confusion, mais cela n'est pas suffisant pour éviter tous les biais. D'autres précautions doivent être prises.

Le biais de sélection

Nous avons vu que l'effet du traitement se déduisait de la différence existant au niveau du critère de jugement à la fin de l'essai entre le groupe recevant le traitement étudié et le groupe contrôle. Une condition évidente pour que la différence entre les deux groupes représente uniquement l'effet du traitement est que les deux groupes soient comparables avant application du traitement. S'ils ne le sont pas, la différence initiale (en quelque sorte constitutionnelle) va se retrouver à la fin de l'essai (

Figure 7). *Dans ce cas, une différence constatée à la fin de l'essai pourrait n'être que la répercussion de la différence initiale et non pas l'effet du traitement.*

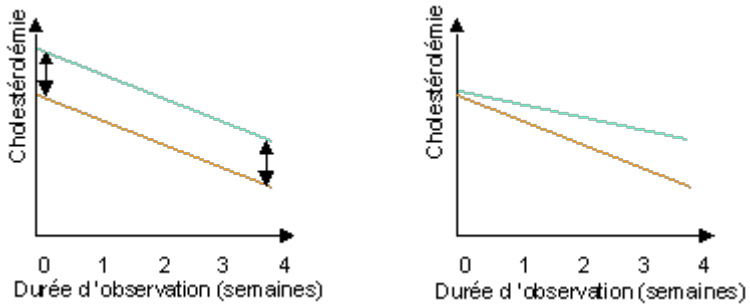


Figure 7 – Si les deux groupes n'ont pas initialement la même cholestérolémie moyenne, la différence observée après traitement peut être simplement la répercussion de la différence initiale, et ne pas être due à l'effet du traitement. Pour pouvoir, à coup sûr, mettre la différence finale sur le compte de l'effet du traitement, il est impératif qu'il n'y ait pas de différence initiale entre les deux groupes. C'est-à-dire que les deux groupes soient strictement identiques au début de l'essai

Le biais introduit par l'utilisation de groupes dissemblables initialement est appelé biais de sélection. Il y a risque de biais de sélection chaque fois qu'il existe une différence initiale dans la répartition, entre les groupes comparés, d'un facteur qui conditionne la valeur du critère de jugement (cf. aussi la Figure 1).

Dans l'exemple de l'hypocholestérolémiant, on pourrait imaginer écarter la possibilité d'un biais en corrigeant la différence finale de la différence initiale. Par exemple, s'il existe en début d'essai une différence de 0,3 mmol/L entre les deux groupes et après traitement une différence de 0,5 mmol/l, la différence 0,5 – 0,3 pourrait être avancée comme effet propre du traitement. Cette correction donne l'effet traitement réel que si la cholestérolémie évolue parallèlement dans les deux groupes sous l'effet des facteurs de confusion. Rien ne le garantit car les patients de ces deux groupes, en plus de différer par leur valeur initiale, peuvent aussi différer au niveau de la façon dont agiront les facteurs de confusion.

Les scénarios imaginables sont nombreux. Par exemple le niveau initial pourrait influencer l'évolution naturelle de la cholestérolémie ou l'importance de l'effet placebo. La régression à la moyenne est d'autant plus importante que la moyenne initiale du groupe est élevée. À partir du moment où les deux groupes ne sont pas initialement comparables, plus rien ne garantit que ces deux groupes auraient évolués de façon parallèle. La correction du résultat final de la différence initiale ne supprime donc pas toutes les causes de biais liées à une dissimilitude initiale des groupes.

L'utilisation de deux groupes non comparables perturbe la mesure de l'effet traitement et introduit potentiellement un biais.

Au maximum l'utilisation de deux groupes non comparables initialement est susceptible de faire croire à un effet du traitement qui n'existe pas ou, à l'opposé, de faire disparaître un réel effet.

Les mauvais groupes contrôles

Contrôles historiques

Un groupe de patients recevant un nouveau traitement peut être comparé au groupe des patients qui, par exemple, durant l'année précédant l'introduction du nouveau traitement ont été traités avec l'ancien traitement.

Ce groupe contrôle, appelé contrôle historique, ne remplit pas les conditions pour être un groupe acceptable :

- *rien ne garantit que les anciens patients soient comparables aux nouveaux, le recrutement du service a pu évoluer au cours du temps,*
- *les autres traitements concomitants ont aussi évolués. Un meilleur résultat obtenu avec les nouveaux patients signifie peut-être tout simplement que la prise en charge des patients s'est améliorée, sans que le nouveau traitement soit meilleur que le précédent.*

Une caractéristique importante d'un groupe contrôle est qu'il doit être constitué de patients contemporains aux patients inclus dans le groupe traité.

Contrôle géographique

Les contrôles géographiques sont constitués de patients traités dans un service hospitalier A par le traitement standard, auxquels on compare les patients du service B traités par le nouveau traitement. Cette comparaison est potentiellement biaisée car rien ne garantit que les patients recrutés dans ces deux services soient similaires et pris en charge de la même manière.

Ces deux types de groupes contrôles sont de mauvais groupes contrôles. Ils peuvent différer du groupe traité par de nombreux points autres que le traitement étudié. Ils ne garantissent pas l'absence de biais de sélection

Comment assurer la comparabilité initiale des deux groupes ?

Pour éviter un biais de sélection, la nature du traitement que reçoit un patient ne doit dépendre d'aucun facteur susceptible d'influencer le résultat. Ainsi, l'allocation du traitement ne doit dépendre ni de la gravité de la maladie, ni des caractéristiques du patient, ni du contexte des soins donnés au patient. Seule une attribution au hasard peut garantir la totale indépendance de la nature du traitement donné vis-à-vis de ces facteurs. Autrement, si un traitement est donné plus spécifiquement à des patients d'un certain type, les groupes ne seront plus comparables et il y aura biais. Pour obtenir cette indépendance les traitements sont attribués aléatoirement aux patients. Ainsi, chaque patient, quelles que soient ses caractéristiques, a la même probabilité de recevoir un traitement ou l'autre. Sur un grand nombre de patients, cette allocation aléatoire assure l'équilibre de toutes les caractéristiques des patients entre les deux groupes.

La randomisation crée deux groupes de patients comparables en moyenne.

Cette allocation aléatoire des traitements est couramment appelée randomisation (« randomization *»). Un essai contrôlé randomisé est donc un essai dans lequel les patients sont répartis entre le groupe contrôle et le groupe expérimental de manière aléatoire. Comme la randomisation est un processus aléatoire et que les propriétés des processus aléatoires ne sont connues qu'en moyenne sur un grand nombre de répétitions, la randomisation ne garantit la comparabilité de deux groupes qu'en moyenne.*

La randomisation en pratique

L'allocation aléatoire des traitements ne s'effectue pas à l'aide de dés, ni en tirant un papier d'un chapeau au moment d'inclure un patient. Dans les essais en double aveugle, les moyens de randomisation donnent le numéro de la boite de traitement à délivrer au patient. Cette boite contient soit le traitement étudié soit le traitement contrôle. La correspondance entre numéro de traitement et nature réelle du traitement est consignée dans une liste qui ne sera utilisée qu'au moment de l'analyse statistique. Dans les essais en ouvert, le moyen de randomisation donne la nature du traitement en clair.

Le moyen de randomisation le plus simple est la liste constituée à l'avance à l'aide d'un programme informatique ou de tables de nombres au hasard. Cette liste donne pour chaque patient inclus successivement dans l'essai le numéro de la boite ou la nature du traitement. Dans les essais multicentriques, une liste est disponible dans chaque centre. La randomisation peut aussi s'effectuer à l'aide d'enveloppes scellées qui renferment le numéro ou la nature du traitement à donner au patient que l'on randomise. Ces deux moyens de randomisation présentent l'inconvénient d'être délocalisés dans chaque centre et de permettre la prise de connaissance anticipée de la nature du traitement que devrait recevoir le ou les prochains patients. Il devient alors possible d'arranger le moment ou un patient sera inclus afin qu'il reçoive le traitement souhaité.

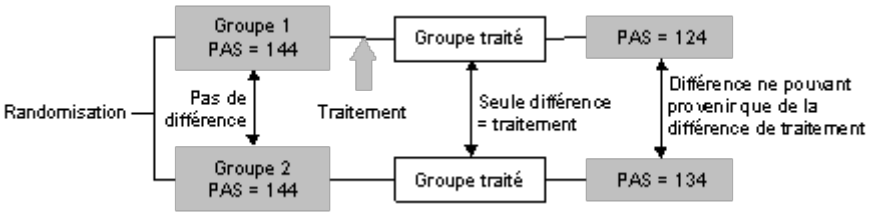


Figure 8 – La randomisation crée deux groupes initialement comparables. L'un de ces groupes recevra le traitement étudié, l'autre non. Ainsi durant l'essai, ces deux groupes ne différeront que par un seul facteur : le traitement. Si à la fin de l'essai une différence est notée entre les deux groupes, elle ne pourra pas provenir d'une différence initiale. Cette différence sera due en fait à la seule différence qui existe entre les groupes : la nature du traitement qu'ils reçoivent.

Seule une procédure centralisée de randomisation garantit suffisamment son imprévisibilité.

Les randomisations centralisées évitent ce problème. Le numéro ou la nature de traitement est obtenu par appel d'un centre de randomisation et après enregistrement du patient en train d'être inclus dans l'essai. Différents moyens de communication sont utilisables : téléphone, fax, Minitel en France, Internet, etc.

Imprévisibilité de la randomisation

Pour être la plus efficace possible, la randomisation doit être imprévisible [1, 2]. Autrement, il existe un risque de biais de sélection. Par exemple, si la nature du traitement que recevra le prochain patient est prévisible, l'investigateur a la possibilité de retarder le moment de l'inclusion d'un patient jusqu'à ce qu'il reçoive le traitement que consciemment ou inconsciemment il souhaite pour lui. Les listes ou les enveloppes donnant en clair la nature du traitement exposent à ce genre de problème.

Un investigateur participe à un essai d'un nouvel antihypertenseur. L'inclusion se fait en fonction d'une liste donnant en clair la nature du traitement à administrer. Ce médecin voit en consultation Monsieur Dupont qui présente une HTA sévère. Sachant que s'il incluait maintenant M Dupont, celui-ci recevrait le traitement standard qu'il considère comme moins efficace, ce médecin va retarder l'inclusion de M Dupont. En faisant cela, cet investigateur n'a pas la volonté de tricher, mais il privilégie "l'intérêt" de son patient au détriment de l'essai. Ainsi biaisé, cet essai ne permettra pas de savoir si l'intuition de ce médecin était bonne ! et même donnera un résultat forcément en défaveur du nouveau traitement.

L'inclusion centralisée d'évite qu'un patient n'ayant pas reçu le traitement souhaité ou ayant présenté un événement précoce ne soit "sorti de l'étude" en ne laissant aucune trace. La sortie de l'étude d'un patient en raison du traitement alloué entraîne un biais de sélection. Les sorties de l'étude dépendant de toutes autres raisons (événements cliniques, effets indésirables, etc.) font courir le risque d'un biais de suivi. Avec l'inclusion centralisée, tout patient inclus, quel que soit son devenir dans l'étude est connu. Ce procédé permet de détecter la possibilité du biais même s'il ne permet pas de l'éviter. En effet, si le critère de jugement est indisponible pour les patients "sortis de l'étude", l'analyse en intention de traiter ne sera pas possible. Seule l'analyse en per-protocole qui est potentiellement biaisée est réalisable.

Les mauvaises « randomisations »

Différents moyens qui ne sont pas basés sur un tirage au sort sont parfois utilisés comme procédés de randomisation.

Parmi ceux-ci figure l'attribution du traitement en fonction de la date. Le traitement A est donné les jours pairs et le contrôle les jours impairs. Cette façon de procéder est inacceptable comme randomisation car le groupe est entièrement prévisible et autorise l'apparition d'un biais de sélection. L'attribution du traitement en fonction de la date de naissance, du numéro de dossier, etc. possède les mêmes inconvénients. La nature du traitement ne doit pas non plus être déterminée en fonction du médecin qui prend en charge le patient ou par le service où

le patient est hospitalisé. Il est fort probable que les recrutements des médecins et/ou des services soient différents et que certains praticiens soignent des patients plus sévèrement atteints que d'autres.

Bibliographie

1. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408-12.

2. Moher D. Assessing the quality of randomized controlled trials: an annoted bibliography of scales and cheklists. Controlled Clinical Trials 1995;16:62-73.

Le double aveugle

Suivi des patients durant l'essai

La randomisation construit deux groupes initialement comparables. Encore faut-il que cette comparabilité soit maintenue au cours de l'essai et que rien ne vienne la détruire durant le suivi (« follow-up ») ou lors la mesure du critère de jugement, pour que ces deux groupes ne différèrent tout au long de l'essai que par la nature du traitement expérimental qu'ils reçoivent. L'influence des différents facteurs de confusion doit donc être maintenu identique entre les 2 groupes.

Par exemple, si dans le suivi, les patients d'un groupe reçoivent plus de traitements concomitants que ceux de l'autre groupe, il sera impossible de savoir si la différence obtenue en fin d'essai est bien due au traitement étudié ou si elle provient de la différence entre les traitements concomitants. Un biais est aussi possible si la mesure du critère de jugement ne s'effectue pas de la même façon dans les groupes. On parle de biais de suivi (ou de réalisation) et de biais de mesure.

Exemple

Un nouveau traitement N est comparé au traitement standard S dans la prophylaxie des thromboses veineuses profondes (TVP) en chirurgie. L'étude est réalisée en ouvert. La survenue d'une TVP est détectée à partir de la clinique puis confirmée par phlébographie.

Le nouveau traitement est considéré comme plus efficace que le standard, ce que doit confirmer l'essai.

Subjectivement, les TVP seront plus facilement suspectées avec le traitement standard qu'avec le nouveau traitement, considéré comme plus efficace, Les mêmes signes cliniques seront plus facilement mis sur le compte d'une TVP si l'on considère que le patient reçoit un traitement moyennement efficace que s'il reçoit un traitement considéré comme très efficace. De ce fait, la sensibilité de la recherche des TVP sera plus élevée dans le groupe du traitement standard où la phlébographie sera réalisée plus fréquemment, conduisant à un plus fort taux de détection de TVP.

	Sensibilité	*Incidence réelle*	*Test positif (incidence observée)*
Nouveau traitement	*70%*	*10%*	*7%*

Traitement standard	*90%*	*10%*	*9%*

Ainsi, bien que la vraie fréquence des événements soit de 10% dans les deux groupes, l'asymétrie dans la performance de la recherche du critère de jugement crée une différence apparente faisant croire à une supériorité du nouveau traitement. La subjectivité des investigateurs entraîne un biais

Le double insu

Le double insu évite les biais de suivi et de mesure.

Le double insu (« double blinding *») évite toutes différences dans le suivi et l'évaluation des deux groupes. Le principe du double insu consiste à faire que tous les patients, quelle que soit leur appartenance à l'un des groupes de l'essai, apparaissent identiques. Ceci est obtenu en ne révélant pas la nature exacte du traitement reçu par les patients. Les patients du groupe traité reçoivent un traitement strictement identique en apparence à celui reçu par les patients du groupe contrôle. Les patients du groupe traité sont donc indiscernables des patients du groupe contrôle. Ainsi, tous les agissements pouvant interférer avec l'effet du traitement (traitement concomitant, mesure du critère de jugement, etc.) sont appliqués de façon symétrique aux deux groupes.*

Nous avons vu qu'une utilisation asymétrique des traitements concomitants systématiquement plus fréquemment dans un groupe que dans l'autre entraîne un biais (à ce niveau, il convient de ne pas confondre les traitements concomitants donnés à l'entrée de l'étude et les traitement de secours (« rescue treatment *») donné en cas d'échec thérapeutique.). Cependant, il est impossible d'interdire le recours à ces traitements, entre autres par ce qu'il serait non éthique de ne rien faire pour un patient qui s'aggrave. Si les deux groupes sont indiscernables, il sera impossible de recourir plus fréquemment aux traitements concomitants dans un groupe que dans un autre.*

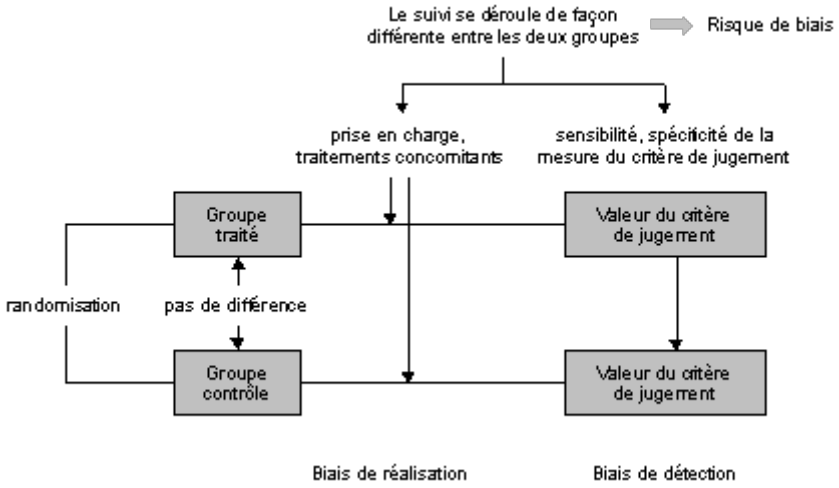


Figure 9 – La randomisation fournit deux groupes comparables, garantissant la valeur de la comparaison. Encore faut-il que cette comparabilité initiale se maintienne durant le suivi et qu'aucune autre différence systématique que le traitement étudié ne vienne détruire la comparabilité des groupes. De même le critère de jugement doit-être mesuré de la même façon dans les deux groupes. Si les deux groupes ne sont pas suivis de la même façon il y a risque de biais de suivi Un risque de biais de mesure apparaît quand le critère de jugement est mesuré de façon différente entre les deux groupes Certaines précautions doivent être prises durant le suivi afin d'éviter de créer des différences entre les deux groupes autres que les traitements étudiés

Le double insu évite les biais dus à l'observateur (interprétation subjective des résultats et hétéro suggestion) et au patient (autosuggestion).

Néanmoins si un traitement est moins efficace qu'un autre, il entraînera un plus grand nombre de patients en échec thérapeutique et donc un taux plus important de recours aux traitements concomitants. Grâce à ces traitements les deux groupes obtiendront finalement le même résultat au niveau du critère de jugement faisant croire à l'absence de différence entre les 2 traitements : le traitement le moins efficace ayant reçu le renfort de traitements concomitants. Dans un essai de supériorité où l'on cherche à montrer que le traitement étudié est supérieur au traitement de référence, ce résultat n'est pas gênant. Aucune différence n'étant notée, l'essai ne pourra pas servir d'argument à l'utilisation du nouveau traitement. Par contre dans un essai dont l'objectif est de montrer l'équivalence des deux traitements, ce mécanisme sera beaucoup plus gênant (ce point sera discuté plus avant dans le chapitre concernant les essais d'équivalence). Pour éviter ce biais, les échecs thérapeutiques ou les absences d'amélioration motivant le recours à un autre traitement doivent être pris en compte comme échec dans le critère de jugement (ce qu'ils sont réellement). Si le critère de jugement n'est pas un événement clinique, mais la valeur d'un paramètre, il devra être mesuré avant l'introduction du traitement concomitant.

Par exemple, dans l'essai d'un antihypertenseur contre placebo, le protocole doit prévoir un traitement de secours pour les patients chez lesquels une aggravation de l'hypertension apparaît. Il s'agit d'un traitement qui sera utilisée à la place du traitement de l'étude sans qu'il soit nécessaire de vérifier la nature du traitement que recevait auparavant le patient.

Le recours au double-aveugle est particulièrement important quand le critère de jugement est de nature subjective. Par exemple, un essai dans la sclérose en plaque a comparé au placebo deux traitements : l'un associant cyclophosphamide et prednisone et l'autre plasmaphérèse et prednisone. Le critère de jugement était l'évaluation par un médecin de l'évolution de la maladie [1]. Cette évaluation a été réalisée en double : une fois en aveugle par des médecins ne connaissant pas le traitement du patient (analyse primaire) et une autre fois par les médecins traitants qui connaissaient la nature du traitement reçu par les patients. Avec l'évaluation en aveugle, aucun effet statistiquement significatif n'a pu être mis en évidence, tandis que l'évaluation sans aveugle montre une efficacité du traitement le plus sophistiqué (plasmaphérèse) mais pas de l'autre traitement. Cet exemple illustre comment l'évaluation d'un critère de jugement subjectif non réalisée en aveugle peut favoriser le traitement dans lequel le plus d'espoirs est mis, et ainsi conduire à un biais.

Tableau 1 – Les différents types d'insu

Double insu (double aveugle « double blind ***»)***
• *ni le patient, ni le médecin investigateur ne connaissent la nature réelle du traitement*
Simple insu (simple aveugle)
• *le médecin connaît la nature du traitement mais pas le patient (les nécessités éthiques et réglementaires de l'information de patients font que le simple insu est en pratique impossible)*
Étude en ouvert (« open design ***»)***
• *le médecin et le patient connaissent la nature du traitement*

Le terme en ouvert est aussi parfois utilisé pour désigner une étude non comparative, sans groupe contrôle.

Le terme « triple aveugle » est parfois utilisé. Il désigne un essai dans lequel l'analyse statistique s'effectue sans avoir connaissance de la nature des traitements. Les traitements sont identifiés par un code du type « traitement A » et « traitement B » et ce n'est qu'une fois l'analyse terminée que la nature exacte des traitements A et B est révélée. Cette procédure n'a plus cours actuellement. Pour éviter une certaine subjectivité dans le choix des analyses statistiques [2], elle est remplacée par l'établissement, avant la fin de l'essai, d'un plan d'analyse détaillant la façon dont sera conduite l'analyse statistique correspondant aux principaux objectifs de l'essai.

Obtention du double insu

Le but du double insu est de faire que les patients des deux groupes soient indiscernables

Pour obtenir un double-insu il est nécessaire que tous les patients reçoivent apparemment le même traitement. Les patients du groupe contrôle doivent donc recevoir un traitement sans effet mais qui est indiscernable du traitement testé. Il s'agit d'un placebo qui a la même forme galénique que le traitement étudié (appelé parfois verum), qui est donné avec la même fréquence, pour la même durée et qui a le même goût, la même couleur, etc. Ainsi, le traitement administré ne permet pas de distinguer les deux groupes étant donné qu'il a la même apparence quel que soit le groupe auquel appartient le sujet.

Le placebo dans un essai à deux rôles : assurer l'indistinctibilité des deux groupes pour réaliser le double insu, et créer les conditions nécessaires à l'apparition de l'effet placebo chez les patients du groupe contrôle dans le but de prendre en compte ce facteur de confusion.

Traitements de natures différentes

Quand le traitement étudié est comparé à un traitement actif de référence il est plus difficile de rendre les deux traitements apparemment identiques. En particulier, si le nouveau traitement étudié s'administre par voie orale et si le traitement contrôle s'administre par voie intra veineuse.

Dans ce cas, le double insu est obtenu par la technique dite du "double placebo". Tous les patients reçoivent un traitement oral et un traitement intraveineux, mais l'un d'entre eux est un placebo. Cependant, les patients du groupe expérimental reçoivent le nouveau traitement par voie orale et le placebo du traitement de référence par voie intraveineuse. Ils bénéficient donc de l'effet du nouveau traitement. Les patients du groupe contrôle bénéficient de l'effet du traitement de référence (administré sous forme intraveineuse alors qu'ils reçoivent sous forme orale le placebo du nouveau traitement).

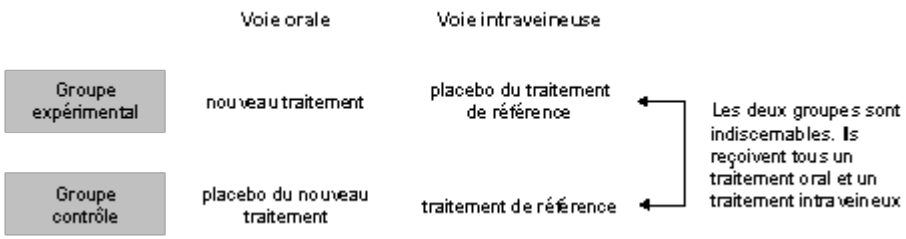


Figure 10 – Réalisation du double insu à l'aide d'un double placebo dans un essai comparant deux traitements actifs de voies d'administration différentes.

Nécessité d'ajustement

Lorsque le traitement évalué nécessite un ajustement de dose (comme par exemple avec un anticoagulant ajusté en fonction du temps de l'INR) la réalisation du double insu s'en trouve compliquée mais n'est pas impossible. Plusieurs solutions sont envisageables. Une tierce personne, différente du médecin ayant en charge le patient, reçoit les résultats des dosages nécessaires à l'ajustement puis communique à l'investigateur la conduite à avoir : augmenter ou diminuer les doses. Le résultat est communiqué à l'investigateur si celui-ci révèle une situation dangereuse pour le patient. Dans les essais contre placebo, des ajustements fictifs sont générés.

Exemple

Dans un essai de statines contre placebo [3], la dose initiale de 20 mg pouvait être doublée si la cholestérolémie restait supérieure à un seuil au bout des premiers mois de traitement. Pour maintenir l'aveugle, les dosages du taux de cholestérol étaient centralisés et seules les instructions relatives au maintien ou au changement de posologie étaient communiquées aux médecins traitants. En cas de cholestérolémie restant très élevée, la valeur des paramètres lipidiques était révélée mais pas la nature du traitement reçu.

Acceptabilité des essais en ouvert

La réalisation d'un double insu est parfois difficilement envisageable, principalement pour des raisons éthiques, par exemple, lors d'une comparaison d'intervention chirurgicale à un traitement médical. La réalisation d'un double insu impliquerait un simulacre d'intervention pour les patients alloués dans le groupe traitement médical ! Des raisons éthiques évidentes rendent irréalisables ce double insu qui a pourtant été pratiqué dans quelques cas.

Lorsque le double aveugle n'est pas possible, il convient que le critère de jugement soit le moins possible sujet à interprétation afin d'éviter l'apparition d'un biais d'évaluation. L'idéal est représenté par les événements cliniques ou le décès. De toute façon un comité d'adjudication travaillant en aveugle du traitement est nécessaire. Dans ces essais en ouvert il convient d'éviter les critères plus ou moins subjectifs (score, échelle, etc.), dont la mesure pourrait être influencée par la connaissance du traitement. Si le recours a de tels critères est néanmoins indispensable, leur évaluation devra être réalisée en aveugle du traitement, par un autre médecin que celui qui assure le traitement et le suivi médical du patient.

Une procédure d'inclusion et de randomisation centralisée est indispensable dans un essai en ouvert. En son absence, il est très facile de favoriser l'attribution d'un des traitements aux patients les plus sévèrement atteints et d'introduire ainsi un biais de sélection. Par exemple, l'utilisation d'enveloppe dans un essai en ouvert est inacceptable.

Exemple

L'étude CAPP [4] est un essai de morbi-mortalité dans l'hypertension qui a comparé le captopril au traitement standard par diurétique ou bêta-bloquants. L'essai a été réalisé en ouvert et a recruté 10 985 patients suivis en moyenne 6,1 ans. La randomisation s'est effectuée à l'aide d'enveloppe. L'analyse des caractéristiques de base révèle un déséquilibre important entre les groupes. La pression artérielle était plus élevée dans le groupe captopril que dans le groupe contrôle (166,6/103,6 vs 163,3/101,2 mm Hg, p<0·0001) ainsi que la proportion de diabétique. Ce déséquilibre introduit un biais de sélection et rend les résultats inexploitables.

Levée d'insu

Dans les essais en double insu, certaines situations d'urgence peuvent nécessiter la levée de l'insu (« code break *»), c'est-à-dire de connaître la nature exacte du traitement que reçoit un patient. C'est par exemple le cas en cas de tentatives de suicide avec le traitement de l'étude, ou d'anesthésie générale. Les essais prévoient donc des procédures permettant de le faire 24h sur 24h. Pour éviter des levées intempestives et sans justification, la procédure de levée d'insu doit être centralisée. La mise à disposition de l'investigateur d'une série d'enveloppes scellées à ouvrir en cas de besoin fait courir un risque de levée plus ou moins systématique de l'insu. Même si elles sont justifiées, des levées d'insu en nombre important risquent de biaiser les résultats. Elles peuvent aussi faire courir des rumeurs qui peuvent gêner la poursuite de l'essai.*

Bibliographie

1. Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. Neurology 1994;44(1):16-20.

2. Gotzsche PC. Blinding during data analysis and writing of manuscripts. Controlled Clinical Trials 1996;17(4):285-90; discussion 90-3.

3. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet 1994;344:1383-89.

4. Hansson L, Lindholm LH, Niskanen L, for the Captopril Prevention Project (CAPPP) study group. Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial. Lancet 1999;353:611-16.

Les différents principes méthodologiques ont été édictés pour éviter l'apparition de biais. L'essai contrôlé, randomisé, en double insu produit des résultats non biaisés. L'existence d'un groupe contrôle permet de prendre en compte l'effet des facteurs de confusion (évolution naturelle du traitement, effet placebo, régression à la moyenne, etc.). La randomisation permet d'obtenir deux groupes identiques comparables. Le suivi en double insu maintient cette comparabilité au cours de l'essai et lors de la mesure du critère de jugement. À coté de ces principes de bases, d'autres sont encore nécessaires : l'absence de perdu de vue et l'analyse en intention de traiter. Ces principes ne sont cependant pas suffisants : faut-il encore que l'essai soit correctement réalisé.

Représentation algébrique

Les différents effets agissant sur la valeur du critère de jugement peuvent être représentés de façon algébrique (

Figure 1). La randomisation assure que X0, l'état initial des patients, est identique en moyenne entre les deux groupes. Les facteurs de confusion E1…E5 agissent de la même manière entre les deux groupes. En raison du double aveugle, les deux groupes subissent le même effet des traitements concomitants et des erreurs de mesure.
Au total, les valeurs observées du critère de jugement se décomposent de la manière suivante :

$$\text{Groupe traité : } X^T = X_0 + E_1 + E_2 + E_3 + \ldots + T$$

$$\text{Groupe contrôle : } X^C = X_0 + E_1 + E_2 + E_3 + \ldots + 0$$

où X_0 désigne l'état initial des patients, E_1,E_2,… les effets des différents facteurs de confusion et T l'effet du traitement étudié (qui est nul dans le groupe contrôle). La comparaison du résultat du groupe expérimental avec celui du groupe contrôle revient à faire la différence :

$$X^T - X^C = X_0 - X_0 + E_1 - E_1 + \ldots + T - 0 = T$$

qui estime sans biais l'effet du traitement T.

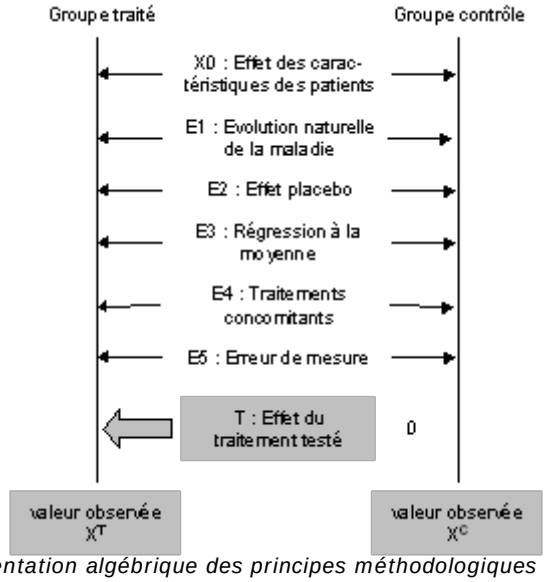


Figure 1 – Représentation algébrique des principes méthodologiques

Liste des biais

Les biais sont très nombreux. Sacket en dénombre plus de 35 [1]. La structure de l'essai, les méthodes de mesure, les observateurs ou les sujets peuvent être responsables des biais.
Les principaux biais qui peuvent affecter un essai thérapeutique sont les suivants.

1. Le **biais de confusion** correspond au biais basique introduit par les facteurs de confusion dans les études sans groupe contrôle. Le risque de biais de confusion est supprimé par l'utilisation d'un groupe contrôle.
2. Le **biais de sélection** survient quand les deux groupes ne sont pas comparables initialement. La randomisation empêche la survenue d'un biais de sélection.
3. Le **biais de suivi** (appelé aussi biais de réalisation) provient d'une destruction de la comparabilité des groupes au cours du suivi (ou de la réalisation de l'essai). Ce biais est évité par l'utilisation du double insu.
4. Le **biais d'attrition** apparaît à la faveur du retrait de certains patients de l'analyse. Le biais d'attrition est contrôlé par l'analyse en intention de traiter.
5. Le **biais d'évaluation** survient lorsque le critère de jugement n'est pas recherché de la même manière entre les groupes de traitement. le double-insu supprime le risque de ce biais.

Fondamentalement la liste des biais peut se résumer à 2 catégories : les biais de confusion et les biais de mesure.

1. Les **biais de confusion** sont produits par une prise en compte inappropriée des facteurs de confusion (conduisant par exemple à une action différente des facteurs de confusion sur les 2 groupes comparés). On retrouve donc dans cette catégorie le biais de confusion proprement dit, le biais de sélection et le biais de suivi.
2. Les **biais de mesure** surviennent lors de l'évaluation du critère de jugement, par exemple lorsque celui-ci n'est pas mesuré de la même manière entre les 2 groupes de l'essai. Le biais d'évaluation appartient à cette catégorie.

Bibliographie

1. Sackett DL. Bias in analytic research. J Chronic Dis 1979;32:51-63. *PMID:*

Plan factoriel

Introduction

Le plan factoriel (« factorial design *») est un plan d'expérience qui répond à deux questions différentes avec le même essai. Il permet un gain de temps et une « économie en patients ». Pour atteindre correctement son but, il est nécessaire que les traitements n'interagissent pas entre eux. Autrement, le plan factoriel perd sa puissance et se retrouve dans l'impossibilité de répondre à aucune des deux questions (1).*

Principe général

Le plan factoriel est l'utilisation des mêmes patients pour effectuer simultanément deux comparaisons. La première comparaison est celle du traitement A à son placebo, et la seconde celle du traitement B à son placebo. Les patients de l'essai seront randomisés une première fois entre A et son placebo. Puis une seconde fois, sans tenir compte de la nature du premier traitement reçu, entre B et son placebo. Ces deux randomisations simultanées créent en fait quatre groupes de patients (

Figure 1) :

- ¼ des patients recevront le traitement A et le traitement B
- ¼ recevront le traitement A et le placebo du traitement B
- ¼ recevront le placebo du traitement A et le traitement B
- ¼ recevront le placebo de A et le placebo de B

En pratique, la randomisation des patients se fait directement entre ces quatre groupes afin d'assurer l'équilibre des effectifs. Au total, la moitié des patients reçoit le traitement A et l'autre moitié le placebo de A. De même, la moitié des patients reçoit le traitement B et l'autre moitié le placebo de B.

La comparaison du traitement A à son placebo s'effectue en comparant tous les patients recevant A à tous ceux recevant son placebo. Ainsi A est comparé à son placebo chez des patients qui reçoivent B ou le placebo de B. La moitié des patients du groupe traité avec A reçoivent B et l'autre moitié le placebo de B. la même répartition est obtenue dans le groupe traité avec le placebo de A. La condition nécessaire pour que cette comparaison soit licite est que A et B agissent de façon indépendante. Ce mode d'analyse est appelée analyse factorielle car elle ne s'intéresse qu'aux facteurs principaux du plan factoriel : les deux traitements A et B.

Le même principe est appliqué pour la comparaison du traitement B à son placebo.

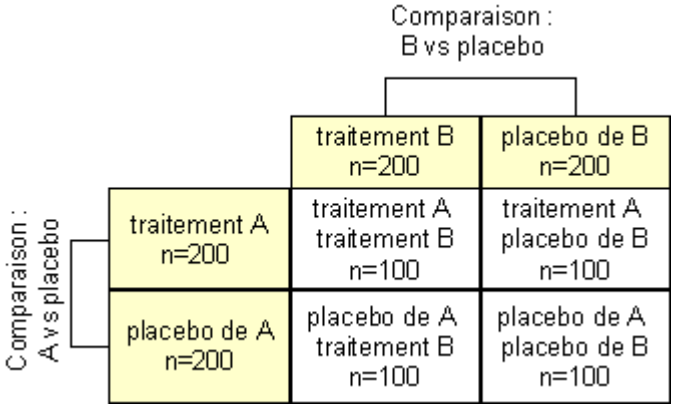


Figure 1 – Répartition des patients dans un essai en plan factoriel

L'analyse statistique d'un plan factoriel se fait en deux étapes. La première étape est la recherche d'une interaction (cf. infra). En l'absence d'interaction, l'effet des traitements est recherché en ajustant sur l'autre traitement. Pour cela une analyse stratifiée est réalisée où l'effet du traitement est déterminé dans chacune des deux strates définies par la nature de l'autre traitement, puis une sorte de méta-analyse de ces deux résultats est réalisée (2). La signification statistique du résultat est recherchée par un test statistique stratifié comme, par exemple, le test de Mantel Haenszel ou un test du Logrank stratifié pour les données de survie. Une autre possibilité est de faire une analyse multivariée à l'aide d'une régression logistique ou d'un modèle de Cox, en incluant dans le modèle les deux traitements et leur interaction.

Lorsque le critère de jugement est continu, une analyse de variance à 2 facteurs peut être réalisée, mais nécessite l'égalité des effectifs des 4 groupes (ou leur proportionnalité).

Dans la recherche de l'effet traitement, les deux strates qui sont regroupées peuvent avoir des niveaux de risque différents. Cette éventuelle différence est induite par le deuxième traitement. La stratification de l'analyse augmente la puissance des tests et la précision des estimations en prenant en compte la variabilité introduite par le deuxième facteur. La variabilité au sein de chaque strate est inférieure à la variabilité observée globalement lorsque l'on ne tient pas compte du deuxième facteur.

Un autre intérêt de l'ajustement est de corriger l'estimation d'un éventuel biais de confusion (similaire au paradoxe de Simpson, cf. chapitre Méta-analyse*). Cependant l'équilibre des effectifs induit par la randomisation limite la survenue de ce type de biais. Le plus souvent, l'analyse simple non ajustée donne un résultat très proche de celui de l'analyse ajustée.*

Exemple

Deux traitements A et B sont évalués par un plan factoriel. Les résultats obtenus sur le critère de jugement principal sont rapportés dans le tableau ci-dessus.

Tableau 1 – Résultat d'un essai en plan factoriel (N désigne l'effectif du groupe, n le nombre d'événements)

n / N	**Traitement B N=1000**	**Placebo B N=1000**
Traitement A (N=1000)	98/500	109/500
Placebo A (N=1000)	112/500	123/500

L'estimation de l'effet du traitement A s'effectue de la manière suivante :

Tableau 2 – Estimation de l'effet du traitement A

	A	**placebo de A**	**RR (IC95%)**	
strate B	98/500	112/500	0,87 [0,69 ;1,11]	
strate placebo de B	109/500	123/500	0,89 [0,71 ;1,11]	
résultat ajusté			0,88 [0,75;1,04]	p=0,13
résultat non ajusté	*207/1000*	*235/1000*	*0,88 [0,75 ;1,04]*	*p=0,15*

La dernière ligne de ce tableau rapporte le résultat de l'analyse non ajustée calculé directement en regroupant les deux strates B et placebo de B.

Pour le traitement B, les analyses brutes et ajustées obtiennent des résultats non significatifs.

Tableau 3 – Estimation de l'effet du traitement B

	B	**placebo de B**	**RR (IC95%)**	
strate A	98/500	109/500	0,90 [0,70 ;1,15]	
strate placebo de A	112/500	123/500	0,91 [0,73 ;1,14]	
résultat ajusté			0,91 [0,77;1,07]	p=0,24
résultat non ajusté	*219/1000*	*232/1000*	0,91 [0,77 ;1,07]	*p=0,26*

Interaction

Définition

Il y a interaction quand l'effet de l'association des traitements n'est pas la « somme »[1][1] des effets des traitements. En cas d'interaction, l'effet d'un traitement varie suivant qu'il est ou pas associé à l'autre.

Deux traitements sont aussi évalués dans un plan factoriel. Le tableau ci-dessus présente les résultats obtenus avec le critère de jugement principal.

Tableau 4 – Résultat d'un essai dans lequel existe une interaction (N désigne l'effectif du groupe, n le nombre d'événements)

	Traitement B N=1000	Placebo B N=1000
Traitement A N=1000	98/500	130/500
Placebo A N=1000	112/500	123/500

Le calcul de l'efficacité de A en présence ou en absence de B révèle l'existence d'une interaction car l'effet de A varie en fonction de l'association ou non du traitement A au traitement B.

Tableau 5 – Mise en évidence de l'interaction dans la recherche de l'effet du traitement A

	r_1	r_0	RR
Effet de A en présence de B	0,20	0,22	0,88
Effet de A en l'absence de B	0,26	0,25	1,06

L'existence de cette interaction est aussi révélée en calculant l'effet de B en présence ou en absence de A.

Tableau 6 – Mise en évidence de l'interaction dans la recherche de l'effet du traitement B

	r_1	r_0	RR
Effet de B en présence de A	0,20	0,26	0,75
Effet de B en l'absence de A	0,22	0,25	0,91

Interprétation

L'existence d'une interaction empêche l'estimation simultanée de l'effet des deux traitements sur l'ensemble des patients du plan factoriel. Comme l'efficacité d'un traitement est variable en fonction de l'association ou non à l'autre traitement, il convient de maintenir les quatre groupes séparés. De plus, l'analyse stratifiée donne des estimations biaisées de l'effet de chaque traitement. Trois estimations sont alors nécessaires pour caractériser l'ensemble des effets de A et B :

1. *A* vs *PBO A + PBO B,*
2. *B* vs *PBO A + PBO B,*
3. *A+B* vs *PBO A + PBO B.*

Cependant, les effectifs n'ont pas été calculé pour réaliser ces comparaisons. La recherche des effets s'effectue alors avec la moitié de l'effectif nécessaire et manque de puissance. Certains essais, pour éviter d'être dans cette situation en cas d'interaction, adoptent des tailles d'échantillons suffisantes pour garantir la puissance de l'analyse séparée des groupes.

Exemple

GISSI-prevenzione est un essai comparant dans un plan factoriel l'huile de poisson (n-3 PUFA) et la vitamine E dans la prévention des maladies cardiovasculaires (3). Les effectifs ont été déterminés de façon à garantir la puissance de chaque traitement au groupe contrôle : n-3 PUFA seul 2836 patients, vitamine E seule 2830, association n-3 PUFA et vitamine E 2830 et contrôle 2828. De cette façon la recherche de l'effet de la vitamine E peut être effectuée en comparant les 2830 patients en ayant reçus aux 2828 patients du groupe contrôle (analyse en 4 groupes) ; sans être obligé de recourir à la comparaison des 5666 patients ayant reçu de la vitamine E seule ou associée aux n-3 PUFA aux 5668 patients n'ayant pas reçu de vitamine E (analyse en deux groupes). Bien que les résultats de l'essai fassent suspecter une interaction, la recherche des effets des 2 traitements est encore possible en raison de ce surdimensionnement des effectifs.

Tests statistiques

L'existence d'une interaction se recherche de différente manière : test d'interaction de l'effet traitement entre les strates d'ajustement (équivalent à un test d'hétérogénéité en méta-analyse), test de Zelen d'homogénéité des odds ratio, terme d'interaction dans un modèle multivarié. Pour les critères continus l'interaction est recherchée à l'aide de l'analyse de variance.

Une difficulté importante surgit à ce niveau. L'obtention d'un test d'interaction non significatif ne permet pas d'affirmer l'absence d'interaction. Le résultat non significatif peut être dû à un manque de puissance (l'effectif d'un plan factoriel n'est pas calculé pour garantir la puissance de la recherche de l'interaction). Cependant, un certain degré d'interaction, même s'il n'est pas significatif, peut fausser l'estimation des effets des traitements.

Il convient donc, d'analyser soigneusement les tendances obtenues dans les 3 comparaisons simples d'un plan factoriel (A vs PBO, B vs PBO, A+B vs PBO) avant d'accepter le résultat de l'analyse stratifiée (4).

Par exemple, le tableau suivant rapporte des résultats d'un plan factoriel dans lequel les deux traitements A et B réduisent chacun la mortalité de 20% (RR=0.80). Mais l'efficacité de A disparaît lorsqu'il est associé à B et vis versa. Le test d'interaction n'a pas suffisamment de puissance pour détecter l'interaction et se révèle non significatif (p=0.12). L'analyse en deux groupes du plan factoriel conduit à des estimations biaisés des effets de A et de B en donnant un risque relatif de 1 pour chacun (effet de A : 90/1000 versus 90/1000, RR=1,00 ; effet de B : 90/1000 versus 90/1000, RR=1,00)

Tableau 7 – Résultat d'un essai dans lequel l'association de deux traitements efficaces ne l'est pas.

	A	**B**	**A+B**	**PBO**
risque relatif	0,80	0,80	1,00	-
effectif	500	500	500	500
décès	40	40	50	50

Exemple

Le Tableau 8 rapporte les résultats d'un essai (5) évaluant simultanément la vitamine D et le calcium pour la prévention secondaire des fracture chez le sujet agé (5). « In a factorial-design trial, 5292 people aged 70 years or older (4481 [85%] of whom were women) who were mobile before developing a low-trauma fracture were randomly assigned 800 IU daily oral vitamin D3, 1000 mg calcium, oral vitamin D3 (800 IU per day) combined with calcium (1000 mg per day), or placebo."

"The sample size was based on a factorial design to test calcium versus no calcium and vitamin D3 versus no vitamin D3. The anticipated incidence of new fractures in the control group was 15%, based on similar trials. The aim was to enrol 4200 participants to give 80% power (2P<0·05) to detect a decrease in incidence to 12%. Furthermore, the sample size was anticipated to have over 80% power to identify a 2% absolute difference in rates of hip fracture."

Bibliographie

1. *McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. Jama 2003;289(19):2545-53.*

2. *Stampfer MJ, Buring JE, Willet W, Rosner B, Eberlein K, Hennekens CH. The 2x2 factorial design: its application to a randomized trial of aspirin and carotene in US physician. Stat Med 1985;4:111-116.*

3. *GISSI-prevenzione investigators. Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: results of the GISSI-prevenzione trial. Lancet 1999;354:447-55.*

4. *Lubsen J, Pocock SJ. Factorial design in cardiology: pros and cons. Eur Heart Journal 1994;15:585-588.*

5. *Grant AM, Avenell A, Campbell MK, McDonald AM, MacLennan GS, McPherson GC, et al. Oral vitamin D3 and calcium for secondary prevention of low-trauma fractures in elderly people (Randomised Evaluation of Calcium Or vitamin D, RECORD): a randomised placebo-controlled trial. Lancet 2005;365(9471):1621-8.*

	Vitamin D3 and calcium (n=1306)	Vitamin D3 (n=1343)	Calcium (n=1311)	Placebo (n=1332)	With calcium (n=2617)	Without calcium (n=2675)	HR (95% CI)	With vitamin D3 (n=2649)	Without vitamin D3 (n=2643)	HR (95% CI)
New fractures	184 (14·1%)	212 (15·8%)	189 (14·4%)	196 (14·7%)	373 (14·3%)	408 (15·3%)	0·99 (0·86–1·15)	396 (14·9%)	385 (14·6%)	1·01 (0·88–1·17)
Confirmed fractures	179 (13·7%)	208 (15·5%)	185 (14·1%)	192 (14·4%)	364 (13·9%)	400 (15·0%)	0·99 (0·86–1·15)	387 (14·6%)	377 (14·3%)	1·01 (0·87–1·16)
Low-trauma fractures only	165 (12·6%)	188 (14·0%)	166 (12·7%)	179 (13·4%)	331 (12·6%)	367 (13·7%)	0·94 (0·81–1·09)	353 (13·3%)	345 (13·1%)	1·02 (0·88–1·19)
Proximal femur	46 (3·5%)	47 (3·5%)	49 (3·7%)	41 (3·1%)	95 (3·6%)	88 (3·3%)	..	93 (3·5%)	90 (3·4%)	..
Other leg and pelvic	37 (2·8%)	48 (3·6%)	41 (3·1%)	54 (4·1%)	78 (3·0%)	102 (3·8%)	..	85 (3·2%)	95 (3·6%)	..
Distal forearm	33 (2·5%)	33 (2·5%)	33 (2·5%)	28 (2·1%)	66 (2·5%)	61 (2·3%)	..	66 (2·5%)	61 (2·3%)	..
Other arm	46 (3·5%)	48 (3·6%)	33 (2·5%)	49 (3·7%)	79 (3·0%)	97 (3·6%)	..	94 (3·5%)	82 (3·1%)	..
Clinical vertebral	0	4 (0·3%)	3 (0·2%)	1 (0·1%)	3 (0·1%)	5 (0·2%)	..	4 (0·2%)	4 (0·2%)	..
Other	3 (0·2%)	8 (0·6%)	7 (0·5%)	6 (0·5%)	10 (0·4%)	14 (0·5%)	..	11 (0·4%)	13 (0·5%)	..
Deaths	221 (16·9%)	217 (16·2%)	243 (18·5%)	217 (16·3%)	464 (17·7%)	434 (16·2%)	..	438 (16·5%)	460 (17·4%)	..
Time to death(months)	24 (13·5–38·0)	22 (10·5–35·0)	22 (10·0–33·0)	24 (13·0–34·0)	24 (11·0–34·0)	23 (12·0–35·0)	1·13 (0·98–1·3)	23 (12–36)	23·5 (11–33)	0·92 (0·80–1·05)
Reported falls during window weeks*	161 (12·3%)	219 (16·3%)	185 (14·1%)	196 (14·7%)	346 (13·2%)	414 (15·5%)	0·89 (0·77–1·02)	380 (14·3%)	381 (14·4%)	0·97 (0·84–1·12)

Data are number of patients (%) or median (IQR). *Window weeks=the single weeks before completion of 4-monthly questionnaires.

Table 3: Main outcomes by randomised group and by supplement groups

Tableau 8 – Exemple de présentation des résultats d'un plan factoriel

L'essai croisé

PowerPoint

Principe

Dans un essai croisé, chaque patient reçoit tous les traitements de l'essai, administrés lors de périodes successives.

L'essai croisé (« cross-over ») utilise le patient comme son propre témoin. Tous les patients reçoivent le traitement étudié et le traitement contrôle dans un ordre aléatoire. L'avantage de ce plan d'expérience est d'assurer une forte comparabilité des groupes contrôle et traité étant donné que ce sont les mêmes patients que l'on retrouve dans ces deux groupes. La variabilité inter-patients est supprimée et remplacée par une variabilité intra-patients, qui est souvent plus petite.

Le temps de participation d'un patient à l'essai est divisé en deux périodes de temps. Durant chacune de ces périodes, le patient recevra un traitement différent. Par exemple, le traitement étudié E durant la première période et le traitement contrôle C durant la seconde. Deux séquences de traitement sont donc possibles : le traitement étudié en premier puis le traitement contrôle (séquence E-C) ou bien le traitement contrôle d'abord et ensuite le traitement étudié (séquence C-E).

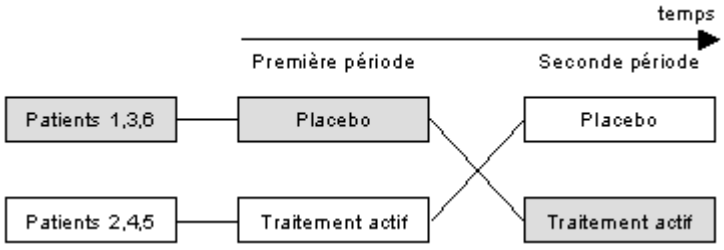


Figure 1 – Représentation schématique d'un essai croisé. A mi-parcours les patients d'un groupe de traitement croisent pour rejoindre l'autre groupe de traitement. À la fin tous les patients seront passés par chaque groupe de traitement. L'ordre d'administration des traitements crée deux groupes dont l'effectif est égal à la moitié du nombre de patients inclus dans l'essai.

La valeur du critère de jugement est mesurée à la fin de chaque période. Chaque patient produit donc une mesure du critère de jugement sous le traitement étudié et une avec le traitement contrôle. Par exemple, dans un essai d'un nouveau broncho-dilatateur contre placebo, chaque patient donne une valeur de peak-flow sous broncho-dilatateur et sous placebo. L'ensemble des patients permet donc de disposer d'une série de mesures sous broncho-dilatateur et d'une série sous placebo, qui permettent de calculer un effet du traitement par comparaison.

Valeur sous traitement actif	
patient 1	(2e période)
patient 2	(1er période)
patient 3	(2e période)
patient 4	(1er période)
patient 5	(1er période)
patient 6	(2e période)

Valeur sous placebo	
patient 1	(1er période)
patient 2	(2e période)
patient 3	(1er période)
patient 4	(2e période)
patient 5	(2e période)
patient 6	(1er période)

Figure 2 – Répartition des valeurs du critère de jugement par traitement. Chaque patient apparaît dans les deux groupes. Chaque groupe contient le même nombre de valeurs que de patients inclus dans l'essai.

Pour chaque patient, l'ordre d'application des traitements (séquence E-C ou C-E) est déterminé de façon aléatoire. Cette randomisation permet de prendre en compte les facteurs de confusion, en particulier ceux qui font que la première période est systématiquement différente de la seconde. Par exemple, dans un essai croisé dans l'hypertension artérielle, la régression à la moyenne fait que les valeurs de pression artérielle de la seconde période sont plus faibles en moyenne que celles de la première. Si le placebo était administré systématiquement durant la première période et le nouvel antihypertenseur durant la seconde, une baisse de pression artérielle moyenne entre ces deux périodes apparaîtrait en dehors de tout effet du traitement.

La compréhension du plan d'expérience croisé est rendue difficile par le nombre de groupes différents qu'il engendre. Les deux séquences d'administration du traitement définissent deux groupes : celui des patients recevant la séquence E-C et celui de ceux qui reçoivent C-E (

Figure 1). *Ces deux groupes sont issus de la randomisation et contiennent chacun la moitié des sujets de l'essai. Deux autres groupes peuvent être créés en fonction de la nature du traitement ; le groupe des mesures du critère de jugement obtenues avec le traitement étudié et celui de celles mesurées sous traitement contrôle (Figure 2). Ces groupes regroupent des valeurs du critère de jugement et non plus des patients. Leur effectif est égal au nombre de patients inclus dans l'essai (puisque tous les patients reçoivent les deux traitements). C'est à partir de ces groupes qu'est estimé l'effet traitement.*

Réduction de la variance dans l'essai croisé

Le plan d'expérience croisé a pour avantage de réduire la variance de la mesure de l'effet traitement par rapport au plan en groupes parallèles.

La variable *X* représente les mesures du critère de jugement observées avec le premier traitement et *Y* celles obtenues avec le second traitement. Considérons aussi que les variabilités des mesures sont les mêmes dans les deux groupes (les variances sont égales $\sigma_X^2 = \sigma_Y^2 = \sigma^2$).

Dans le plan d'expérience en groupe parallèle, la variance de l'effet traitement mesure la différence *X*-*Y* est $\sigma_x^2 + \sigma_y^2 = 2\sigma^2$ car les variables *X* et *Y* sont issues de deux groupes indépendants (mesurées chez des patients différents).

Dans un essai croisé, les valeurs *X* et *Y* sont obtenues chez les mêmes patients et sont donc corrélées entre elles. *X* et *Y* n'étant plus indépendantes, leur covariance $\mathrm{cov}(X,Y)$ n'est pas nulle. La variance de l'effet traitement mesuré par la différence *X*-*Y* est donc $\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$. En prenant un coefficient de corrélation de 0,5 qui est une valeur raisonnable, la variance de l'effet traitement devient $2\sigma^2 - 2\frac{1}{2}\sigma^2 = \sigma^2$. Dans ce cas, l'essai croisé nécessite 4 fois moins de sujets que l'essai en bras parallèles. Bien entendu s'il s'avère qu'il n'y a pas de corrélation entre les deux mesures effectuées chez le même patient, le plan d'expérience croisé n'apporte pas de gain [3].

L'essai croisé permet de comparer les traitements au sein du même malade (comparaison inter-individuelle) au lieu d'effectuer une comparaison entres groupes de sujets (comparaison inter-individuelle).

Étant donné que chaque patient apparaît à la fois dans le groupe traité et dans le groupe contrôle, des techniques statistiques employées pour les essais en bras parallèles ne peuvent pas être utilisées car les deux échantillons ne sont pas indépendants. L'analyse est réalisée avec les méthodes pour séries appariées. Pour chaque patient, la différence entre la valeur du critère de jugement sous traitement étudié et celle sous traitement contrôle est calculée. En l'absence d'effet traitement, ces différences vont fluctuer autour de zéro et dans le cas contraire autour de la vraie valeur de l'effet traitement. Ces différences sont indépendantes entre-

elles puisqu'elles ont été obtenues chez des patients distincts. La recherche d'un effet traitement est donc effectuée en comparant la moyenne de ces différences à zéro[2[1]] *et en utilisant la variance des différences.*

Assez souvent, le critère de jugement n'est pas la valeur prise par un paramètre en fin de période de traitement mais son changement entre le début et la fin de la période de traitement. Par exemple, l'effet d'un traitement antihypertenseur peut être évalué par le changement de pression artérielle entre avant et après la période de traitement. Pour chaque patient, deux changements sont mesurés : celui observé durant la première période et celui de la seconde. Ce sont ces changements qui sont comparés pour estimer l'effet traitement (cf. Tableau 1).

Exemples

Un nouveau traitement antihypertenseur est évalué contre placebo dans un essai croisé incluant 13 patients. Le critère de jugement est le changement de pression artérielle systolique (PAS) observée durant la période de traitement.

Le Tableau 1 présente les données recueillies durant les 2 périodes chez les 13 patients ainsi que le changement de PAS.

Tableau 1 – Données individuelles des patients par période

	Première période				**Seconde période**			
Sujet	**Traitement**	**PAS avant**	**PAS après**	**changement**	**Traitement**	**PAS avant**	**PAS après**	**changement**
1	Placebo	158	160	2	T. testé	158	142	-16
2	T. testé	140	117	-23	Placebo	149	153	4
3	Placebo	152	153	1	T. testé	149	128	-21
4	T. testé	159	138	-21	Placebo	162	163	1
5	Placebo	166	167	1	T. testé	165	147	-18
6	Placebo	162	161	-1	T. testé	152	127	-25
7	T. testé	168	149	-19	Placebo	151	151	0
8	T. testé	153	137	-16	Placebo	160	164	4
9	Placebo	148	150	2	T. testé	164	146	-18
10	T. testé	152	134	-18	Placebo	143	138	-5
11	T. testé	163	141	-22	Placebo	163	162	-1
12	Placebo	168	163	-5	T. testé	156	131	-25
13	Placebo	151	149	-2	T. testé	158	137	-21

Le Tableau 2 présente les changements de PAS par traitement. L'effet du traitement est calculé par différence entre le changement observé sous placebo et celui observé avec le nouvel antihypertenseur. Cet effet est donc la baisse de PAS induite par le traitement corrigé de la variation observée sous placebo (et représentant l'effet des facteur de confusion).

[2[1]] avec deux bras parallèles on compare directement la moyenne d'un groupe avec celle de l'autre étant donné que les deux groupes sont indépendants.

L'essai montre une baisse de 20 mmHg avec le traitement étudié par rapport au placebo, avec un écart type de 3,28 mmHg. Cette baisse est significativement différente de zéro.

Avantages et limites

Avantages

Le principal avantage de l'essai croisé est d'être, à effectif identique, plus puissant que l'essai en bras parallèles. En effet, la prise du sujet comme son propre témoin diminue la variabilité de la mesure de l'effet traitement. Avec des mesures moins variables il faut donc moins de sujets pour mettre en évidence un effet. Ceci n'est vrai que si, effectivement, il existe une corrélation forte entre les mesures faites chez le même sujet. Pour gagner, il faut que les mesures effectuées chez le même sujet soient moins variables que celles effectuées entre sujets. Si cela n'est pas vérifié, l'essai croisé n'est pas plus puissant.

Tableau 2 – Résultat par traitement

Sujet	***Nouveau traitement***	***Placebo***	***Effet traitement***
1	*-16*	*2*	*-18*
2	*-23*	*4*	*-27*
3	*-21*	*1*	*-22*
4	*-21*	*1*	*-22*
5	*-18*	*1*	*-19*
6	*-25*	*-1*	*-24*
7	*-19*	*0*	*-19*
8	*-16*	*4*	*-20*
9	*-18*	*2*	*-20*
10	*-18*	*-5*	*-13*
11	*-22*	*-1*	*-21*
12	*-25*	*-5*	*-20*
13	*-21*	*-2*	*-19*
moyenne	*-20*	*0*	*-20*
écart-type	*3,03*	*2,87*	*3,28*

La corrélation entre les mesures réalisées chez un même patient est la condition sine qua non *pour obtenir une réduction de variance avec le plan d'expérience croisé. Le calcul du nombre de sujets prend donc en compte ce paramètre et plus la corrélation est forte moins l'effectif nécessaire est important. Cependant, il n'est pas*

obligatoire que deux mesures successives faites chez les mêmes sujets soient fortement corrélées. Entre autres, si les deux mesures sont effectuées à distance l'une de l'autre leur corrélation risque d'être faible, ceci d'autant plus que le paramètre mesuré est variable d'un instant à l'autre. Par exemple, le taux sérique de créatinine varie chez le même individu dans de fortes proportions d'un moment à l'autre (en fonction de l'alimentation ou de degré de déshydratation). Deux mesures du taux de créatinine à 6 heures d'intervalles seront corrélées entre elles, mais pas deux mesures séparées par un mois. Apparaît ici un des dangers que fait courir l'essai croisé. Si la corrélation des mesures est surestimée, l'effectif sera sous-estimé et l'essai se révélera probablement négatif. Il est donc primordial de disposer d'une estimation correcte de la covariance lors de la planification d'un essai. Pire, dans le cas d'une corrélation négative des mesures, l'essai croisé perd de la puissance par rapport à l'essai en bras parallèle pour un même nombre de patients. Cleophas cite une dizaine d'essais croisés qui se sont certainement retrouvés dans cette situation [3].

Limites

L'essai croisé ne peut pas être utilisé dans toutes les situations. Pour être valable il nécessite que plusieurs conditions soient remplies.

1. *le critère de jugement doit pouvoir être mesuré à plusieurs reprises chez le même sujet. Ce qui exclut la plupart des événements cliniques comme la mortalité : un patient décédé en première période ne peut évidemment plus présenter le critère de jugement en seconde période ! L'essai croisé est donc inutilisable avec un critère de jugement basé sur la survenue d'un événement clinique. Par contre un critère binaire qui peut se répéter est utilisable dans un essai croisé comme la définition d'un critère succès/échec à partir de ce qui a été observé sur la période. Par exemple, avec un antihypertenseur, le succès peut être défini par le maintient de la pression artérielle en dessous d'un seuil durant le traitement. Ce critère est observable de manière indépendante sur les deux périodes.*
2. *L'effet des traitements ne doit pas être irréversible pour que les sujets se retrouvent en début de seconde période dans un état identique à celui qu'ils avaient en début de première période. Cela exclut par exemple les traitements qui guérissent.*
3. *Une période de lavage (« washing out ») doit être aménagée entre les deux périodes pour permettre au traitement administré en premier de disparaître ainsi que ses effets (lavage pharmacocinétique et pharmacodynamique).*
4. *La maladie de doit pas évoluer de façon notable entre les deux périodes. Par exemple, elle ne doit pas guérir spontanément avant la fin de la deuxième période.*
5. *Il ne doit pas y avoir d'interférence entre l'ordre d'administration des traitements et leur effet. L'effet d'un traitement doit être le même que celui-ci soit administré en premier ou en second.*
6. *Le nombre de perdus de vue doit être limité en première période.*
7. *La mesure répétée du critère de jugement doit être dépourvue d'effet de conditionnement, par exemple par apprentissage lorsque le patient doit remplir un questionnaire ou par accoutumance du sujet aux effets secondaires. Ces phénomènes introduisent un effet ordre qui rend la seconde période systématiquement non comparable à la première. L'égale répartition des traitements sur les deux périodes prévient la survenue d'un biais, mais cet effet ordre entraîne une augmentation de la variabilité des mesures.*

Concrètement l'essai croisé n'est utilisable que dans des situations particulières : maladie chronique évolution stable, critère de jugement intermédiaire, traitements dont les effets disparaissent rapidement à son arrêt, efficacité d'apparition rapide.

Formes particulières de l'essai croisé

Certaines situations se prêtent à des plans d'expériences similaires à l'essai croisé. En dermatologie, et encore plus en cosmétologie, les traitements topiques peuvent être appliquées simultanément chez le même patient sur deux champs cutanés différents. Encore faut-il que les traitements n'aient pas une diffusion systémique. Une généralisation du plan d'expérience croisé à plus de deux traitements, le carré latin, est couramment utilisé dans les essais de phase 1.

Le principe de l'essai croisé permet aussi la réalisation d'essai de taille 1 [1,2,5,6]. Ces essais qui ne portent que sur un patient et correspondent à une succession d'administration dans un ordre aléatoire de deux ou plusieurs traitements. Ces essais sont utilisables pour évaluer des traitements dans des maladies très rares ou pour rechercher le traitement le plus adapté à un patient.

Bibliographie

1. *Anonymous. Randomised controlled trials in single patients.* Drug Ther Bull *1998;**36:**40.*

2. *Chatellier G. Randomized study of n-of-1 trials versus standard practice.* Rev Epidemiol Sante Publique *1996;**44:**382-3.*

3. *Cleophas TJM. Crossover trials are only useful when there is a positive correlation between the response to different treatment modalities.* Br J Clin Pharmacol *1996;**41:**235-39.*

4. *Grizzle JE. The two-period change-over design and its use in clinical trials.* Biometrics *1965;**22:**467-80.*

5. *Porta M, Bolumar F, Hernandez I, Vioque J. N of 1 trials. Research is needed into why such trials are not more widely used.* BMJ *1996;**313:**427.*

6. *Saunders KB. n of 1 trials.* BMJ *1994;**309:**1584.*

7. *Senn SJ.* Cross-over trials in clinical research. *Chichester: John Wiley, 1993.*

Les autres plans d'expériences

Groupes parallèles

Introduction

L'essai en deux groupes parallèles, appelé aussi en bras parallèles (« parallel groups *» ou «* parallel arms *»), est l'archétype de l'essai thérapeutique. Le traitement étudié est comparé à un traitement contrôle (placebo ou traitement actif) à l'aide de deux groupes de patients constitués par randomisation de façon contemporaine et suivis* ***en parallèle****. Nous ne reviendrons pas sur ce plan d'expérience qui a été longuement décrit avec l'exposé des principes méthodologiques.*

Ce schéma peut être étendu à plusieurs groupes, autorisant la comparaison de plusieurs traitements ou modalités de traitements entre eux. Différents cas de figure sont possibles :

- deux ou plusieurs traitements concurrents sont comparés à un même contrôle (placebo ou traitement actif) et/ou entre eux,
- différentes doses du même traitement sont comparées à un placebo. Il s'agit d'une étude de relation dose-efficacité,
- un traitement étudié est comparé à plusieurs contrôles, par exemple son placebo et un traitement actif.

Dans tous les cas, les essais multigroupes posent des problèmes particuliers d'ordre statistique liés aux comparaisons multiples.

Relation dose-efficacité, essai de doses

Les essais de doses étudient différentes doses d'un même traitement à la recherche de la dose optimale : celle qui est à la fois la plus efficace et la mieux tolérée. Ce sont des essais qui comprennent un groupe contrôle (recevant le plus souvent un placebo) et plusieurs autres groupes recevant des doses différentes du même traitement. Ces essais sont réalisées dans deux situations différentes.

Tout d'abord à la phase précoce du développement d'un traitement où le but est d'établir la courbe de dose-réponse en utilisant le plus souvent un critère de jugement intermédiaire. L'estimation de cette courbe repose sur l'étude d'un nombre conséquent de doses.

À un stade plus avancé du développement, la dose optimale peut aussi être recherchée en utilisant un critère clinique. Cette recherche est nécessaire quand il est difficile d'extrapoler les informations obtenues au niveau du critère intermédiaire. C'est par exemple le cas avec les médicaments de la coagulation où les effets hémorragiques qui sont très dose-dépendant ne peuvent être étudiés valablement qu'avec les critères cliniques (accidents hémorragiques). Comme ces études de dose nécessitent plus de patients pour effectuer les comparaisons, elles sont conduites en utilisant seulement un petit nombre de doses (en général deux).

L'analyse de ce type d'étude repose sur la recherche d'une relation entre les doses et l'effet par un test de tendance. Ensuite, chaque dose est comparée au contrôle. Il est rare que les doses soient comparées par un test statistique, car ces comparaisons, pour être puissantes, nécessiteraient d'inclure un grand nombre de patients dans chaque bras. Au final, la détermination de la dose optimale nécessite que le test de tendance soit significatif et que la dose retenue soit effectivement supérieure au contrôle. La dose est déterminée en évaluant la balance bénéfice-risque de chaque dose.

> **Note** : Tests statistiques de tendance - Les tests de tendance (« *trend test* ») recherchent si la variable à expliquer varie en même temps que la variable explicative. La régression linéaire est un test de tendance pour les variables continues. Il existe des tests de tendance adaptés aux proportions, aux odds ratio, etc.

Plusieurs comparaisons statistiques sont réalisées dans une étude de doses. Pour limiter les conséquences de l'inflation du risque alpha, l'utilisation d'une méthode d'ajustement comme la méthode de Bonferroni est recommandée. Dans un essai où k *doses seront comparées au placebo, un seuil de signification statistique de* α/k *est utilisé pour chaque comparaison. Si, en plus, chaque dose est comparée aux autres le seuil ajusté est de* α/2k.

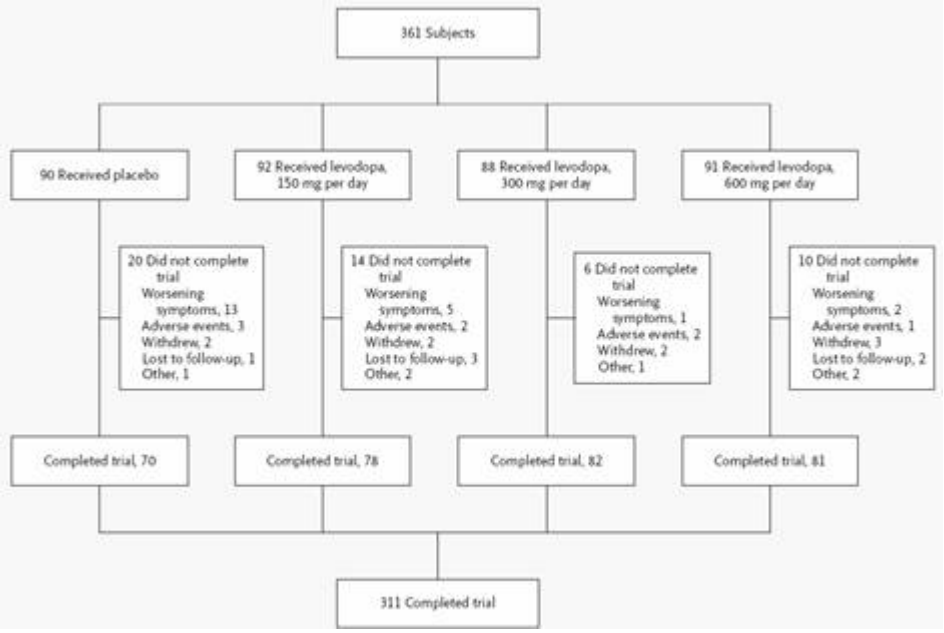
D'autres plans d'expérience sont utilisables pour les études de doses. Si ses conditions de validité sont remplies, le plan croisé permet aussi la comparaison de plusieurs doses. Il a l'avantage de mesurer la relation dose-effet au niveau individuel.

Des plans spécifiques sont aussi employés comme « la titration forcée » ou l'« optionnal titration ». Ils dépassent le cadre de cet ouvrage.

Exemple

Un essai a comparé 3 dose de levodopa dans la maladie de Parkinson (1).

« METHODS: In this randomized, double-blind, placebo-controlled trial, we evaluated 361 patients with early Parkinson's disease who were assigned to receive carbidopa-levodopa at a daily dose of 37.5 and 150 mg, 75 and 300 mg, or 150 and 600 mg, respectively, or a matching placebo for a period of 40 weeks … The primary outcome was a change in scores on the Unified Parkinson's Disease Rating Scale (UPDRS) between baseline and 42 weeks."

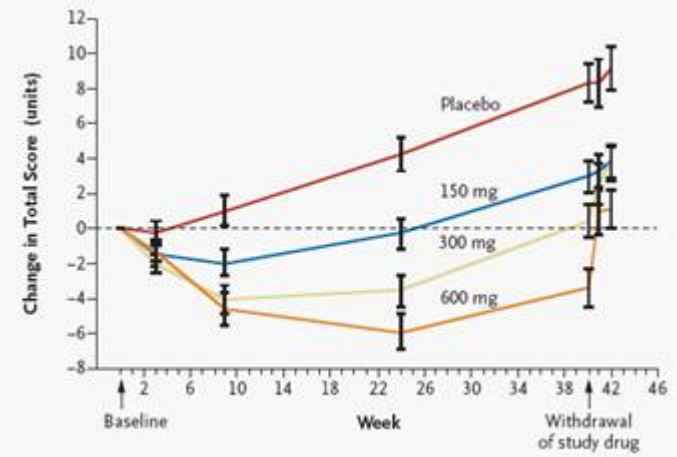


La relation dose effet a été recherché à l'aide du test de tendance. « Levodopa, in a dose–response pattern, significantly (P<0.001) reduced the worsening of symptoms of Parkinson's disease as reflected in the change between the total score on the UPDRS at baseline and that at week 42 (i.e., two weeks after washout of the study medication) »

Table 2. Changes in the Scores on the UPDRS between Baseline and Week 42.*

Characteristic	Placebo	Levodopa			P Value for Trend
		150 mg/day	300 mg/day	600 mg/day	
Evaluation by primary rater					
No. of subjects	70	78	82	81	
UPDRS score					
Total score	7.8±9.0	1.9±6.0	1.9±6.9	−1.4±7.7	<0.001
Mental component	0.3±1.5	0.0±1.5	0.1±1.2	0.1±1.4	0.18
ADL component	2.3±3.4	0.5±2.3	0.4±2.9	−0.3±3.0	<0.001
Motor component	5.2±6.4	1.4±5.5	1.4±5.3	−1.4±5.9	<0.001
Evaluation by treating investigator					
No. of subjects	70	78	82	81	
UPDRS score					
Total score	9.0±10.4	4.0±8.2	4.0±8.4	1.0±9.9	<0.001
Mental component	0.5±1.3	−0.1±1.4	0.1±1.4	0.1±1.6	0.31
ADL component	2.5±4.0	0.8±3.1	1.0±2.8	0.3±3.5	<0.001
Motor component	6.0±7.6	3.2±6.4	3.0±6.4	0.6±7.7	<0.001

* Plus–minus values are means ±SD. On the UPDRS, higher scores indicate greater severity of impairment. Negative numbers indicate improvement as compared with the baseline value. The total score on the UPDRS showed a significant trend toward the reduction of symptoms with higher doses of levodopa in the evaluations by both the primary raters and the treating investigators. The post hoc analysis showed that the effects of all three doses of levodopa differed significantly from the effect of the placebo. Scores on the UPDRS showed that treatment effects were significant for activities of daily living (ADL) and the motor component but not for the mental component.



Comparaison de plusieurs traitements

Le principe général de ces essais est le suivant. Deux nouveaux traitements A et B sont comparés au placebo pour montrer leur efficacité. À la question double de l'efficacité de A et B par rapport au placebo se pose aussi la question du traitement le plus efficace entre A et B.

Cet essai conduit donc à la réalisation de 3 comparaisons statistiques : A vs placebo, B vs placebo et A vs B. Le seuil de signification statistique de chacune de ces comparaisons devra donc être corrigé pour tenir compte de l'inflation du risque alpha lors de comparaisons multiples.

Le calcul du nombre de sujets nécessaires doit aussi être effectué de façon à garantir une puissance statistique correcte à toutes ces comparaisons. En particulier, la comparaison A vs B concerne deux traitements actifs et demande des effectifs plus importants qu'une comparaison par rapport au placebo. Toutes ces comparaisons doivent être clairement définies a priori comme objectif de l'essai. Autrement, les comparaisons non prévues sont purement exploratoires et sans valeur de confirmation.

Exemple

L'essai EPIC comparait deux modalités d'administration d'un anti agrégant plaquettaire anti-Gp2b3a, l'abciximab comme traitement adjuvant à l'angioplastie coronaire. L'objectif était d'éviter la survenue de l'occlusion précoce de l'artère dilatée. Deux modes d'administration de l'abciximab étaient comparés au placebo : un bolus de 0.25mg/kg et un bolus de 0.25mg/kg suivi par une perfusion de 10µg/min. La taille des trois groupes était d'environ 70 patients.

Fréquence du critère de jugement principal dans les 3 groupes

Placebo	Bolus	Bolus + perfusion
n=696	n=695	n=708
89 (12,8%)	79 (11,4%)	59 (8,3)%

Le test de tendance est significatif (p=0,009), mais seulement la comparaison individuelle du bolus associé à la perfusion vs placebo est significative p=0,008 ; la comparaison du bolus au placebo donne p=0,43. Le plan d'analyse ne prévoyait de faire les comparaisons vs placebo que si le test de tendance était significatif. Ces résultats conduisent à recommander l'association bolus+perfusion.

Stratification

La stratification consiste à séparer dans l'essai différents types de patients à l'aide de strates. Une randomisation indépendante est réalisée dans chaque strate, ce qui conduit à un équilibre des effectifs entre les 2 groupes de chaque strate.

Tableau 1 – Exemple d'un essai stratifié sur les hommes et les femmes.

	Groupe traité	Groupe contrôle
Strate des hommes	256	260
Strate des femmes	123	120
Ensemble de l'essai	379	380

La stratification est utilisée :

- avant tout pour ajuster sur un facteur pronostique important, ce qui réduit la variabilité de la mesure de l'effet et augmente la puissance.
- pour tester deux hypothèses simultanément (voir le chapitre consacré aux sous-groupes),

Stratification pour diminuer la variance résiduelle

La stratification sur un facteur pronostique augmente la puissance de la recherche de l'effet traitement en diminuant la variabilité résiduelle rattachée à la comparaison.

Imaginons que dans le groupe de patients présentant le facteur E, la moyenne du critère de jugement est différente de celle du groupe des patients ne présentant pas le facteur E. Dans ces groupes la variabilité des valeurs est la même. Si ces deux groupes de patients sont rassemblés, la variabilité des mesures devient plus forte que la variabilité observée au sein de chaque groupe : la variabilité inter-groupe va s'ajouter à la variabilité intra-groupe. (cf. chapitre Statistiques avancées*). L'ajustement consiste à rechercher l'effet du traitement dans chaque groupe puis à fusionner ces effets traitements. La recherche de l'effet traitement est plus puissante à l'intérieur de chaque groupe car la variabilité y est plus faible par construction. L'ajustement revient à faire une sorte de méta-analyse des résultats de chaque strate d'ajustement.*

Contrairement aux études épidémiologiques, le but de la stratification n'est pas d'éviter un biais de confusion car dans les essais, les effectifs des groupes au sein des strates sont équilibrés par la randomisation.

Valider le traitement dans deux populations de patients

L'autre utilisation de la stratification est d'effectuer la validation du traitement dans deux populations de patients différents. Ce plan d'expérience est utilisé quand il y a des arguments pour penser que l'effet du traitement ne sera pas identique dans ces deux populations. Cette utilisation de la stratification répond de façon satisfaisante à la problématique des analyses en sous-groupes : l'hypothèse est formulée a priori et le calcul du nombre de sujets nécessaires est effectué dans chaque strate.

Autres plans d'expérience

Différents plans d'expériences ont été proposés pour répondre à des questions spécifiques ou pour contourner les rares limites (inconvénients) du plan d'expérience en bras parallèle. Le Tableau 2 présente un récapitulatif des diverses possibilités.

Essai de remplacement

L'essai de remplacement s'adresse à des patients souffrant d'une maladie chronique et qui prennent déjà un traitement. Cet essai consiste à substituer après randomisation et en double aveugle, le traitement habituel des patients par le traitement testé. L'intérêt de cette substitution est d'éviter l'arrêt chez tous les patients du traitement habituel, évitant ainsi les désagréments d'un sevrage (par exemple avec les corticoïdes). Si besoin, la substitution peut avoir lieu avec un chevauchement des deux thérapeutiques pour effectuer un relais par exemple. De plus, ce plan d'expérience peut augmenter l'acceptabilité de l'essai.

Ce type d'essai peut aussi être utilisé pour évaluer chez ces patients un traitement habituel mais qui n'a pas fait la preuve de son efficacité. Après randomisation, le traitement habituel de la moitié des patients est remplacé par un placebo.

Exemple

La prophylaxie des pneumopathies à pneumocystis carinii dans l'infection par VIH est recommandée chez les patients présentant une immunodéficience importante (moins de 200 CD4/mm3). Cependant, se pose la

question de l'arrêt de la chimio-prophylaxie après la remontée du taux de CD4 en réponse à un traitement antiviral intensif. Cette question a été abordée par un essai de remplacement comparant par randomisation arrêt et poursuite de la prophylaxie chez 474 patients (2).

Essai avec sortie rapide (traitement de sauvetage)

Dans les essais avec sortie rapide, le traitement de l'essai (éventuellement le placebo) est promptement arrêté en cas d'aggravation du patient ou d'échec du traitement de l'étude dans l'obtention d'un certain but comme, par exemple : une pression artérielle non contrôlée au bout d'une durée pré déterminée ou s'élevant au dessus d'une valeur maximale ; une fréquence de crises d'épilepsie ou d'angine de poitrine supérieure à un seuil prédéfini ; l'absence de normalisation des enzymes au bout d'un certain temps dans une hépatite, etc. Le traitement peut aussi être changé à la première survenue d'un événement qu'il est censé prévenir : première récidive d'angor instable, de grand mal épileptique, crise de tachycardie supraventriculaire, etc…

Tableau 2 - Récapitulatif des principaux plans d'expérience

Deux bras parallèles, contre placebo
Deux bras parallèles, contre traitement actif
Trois bras parallèles contre placebo et contre traitement actif
Dose effet (plusieurs doses du traitement sont testées contre placebo ou traitement de référence actif)
Évaluation par-dessus (« *on top* », « *add-on* »)
Plan factoriel
Essai croisé
Essai croisé multi traitement ou multi période (carré latin)
Essai de taille 1
Essai de remplacement
Essai de sortie rapide
Essai de retrait

Le critère de jugement est la nécessité de changer de traitement. Les critères de décision de changement de traitement doivent être parfaitement définis ainsi que la périodicité des réévaluations afin d'éviter que des patients ne restent pas sous un traitement qui ne contrôle pas suffisamment leur maladie.

L'inconvénient de ce type d'essai est qu'il donne uniquement des informations sur l'efficacité à court terme

Essai de retrait

Dans les essais de retrait (« withdrawal trial *»), tous les patients reçoivent le traitement testé durant une certaine période puis celui-ci est arrêté et remplacé (retrait) par le placebo chez un certain nombre de patients déterminé par randomisation. Après retrait randomisé la période d'observation peut être de durée fixe ou s'étendre jusqu'à la survenue du critère de jugement (événement clinique)*

L'essai de retrait randomisé permet, par exemple, de déterminer l'effet préventif sur la rechute de la prolongation d'un traitement prescrit pour traiter un épisode aigu de maladie récidivante. Y a-t-il lieu de poursuivre un traitement après le traitement d'un épisode aigu ? Par exemple, la prolongation d'un traitement antiviral permet-elle de réduire la fréquence des récidives dans l'herpès.

La recherche de dose peut utiliser un plan d'expérience similaire dénommé essai d'escalade. Un effet rebond survenant à l'arrêt du traitement peut faire croire à la persistance de l'efficacité.

Exemple

L'essai RADIANCE (3) est un essai de retrait de la digoxine réalisé chez 178 patients porteurs d'une insuffisance cardiaque chronique de stade II ou III, traitée par diurétique, digoxine et IEC. Après randomisation et en double insu, la digoxine a été arrêtée chez 93 patients et remplacée par un placebo. Une aggravation de l'insuffisance cardiaque a été observée chez 23 patients après arrêt de la digoxine et chez seulement 4 patients pour ceux qui ont continué à recevoir la digoxine (p<0.001).

La taille de l'effet observé avec ce type d'essai est en général supérieure à celle vue dans une population non sélectionnée car la randomisation porte uniquement sur des patients qui tolèrent le traitement testé et qui ne s'aggravent pas. En effet, en fin de période initiale, les patients qui sont effectivement randomisée et sur lesquels portera la comparaison sont des patients sélectionnés, non représentatifs des patients tout venant.

Lorsque la question posée est la durée du traitement, un essai de retrait répond en fait à la question suivante : après une certaine durée de traitement y-a-t-il un intérêt à poursuivre un traitement chez les sujets qui sont toujours traités et qui n'ont donc pas présenté jusque là le critère de jugement et qui ont bien toléré le traitement. La réponse apportée ne s'applique qu'à des patients qui sont arrivés au terme d'une certaine durée de traitement sans présenter le critère de jugement (une récidive par exemple) et qui ont bien toléré le traitement durant cette période.

La même question initiale peut aussi être abordée avec un essai classique comparant une durée courte de traitement (correspondant à la durée de la période initiale de l'essai de retrait) à une durée longue (correspondant à la phase de prolongation de l'essai de retrait). Cet essai répond alors à la question : y-a-t-il un intérêt à envisager d'emblée un traitement prolongé par rapport à un traitement de courte durée.

Les extensions d'essais

Les extensions d'essais consistent à continuer l'observation des patients après la fin programmée de l'essai proprement dit. Cette extension est toujours réalisée de façon ouverte (autrement il s'agit d'un essai de retrait ou un essai croisé), sans groupe contrôle (tous les patients de l'essai reçoivent le traitement étudié). Les apports de ce type d'observation sont très limités. En particulier, ces extensions ne permettent pas d'obtenir des données fiables sur les effets indésirables ou sur l'efficacité en raison de l'absence de groupe contrôle et du fait que n'entrent dans la phase de prolongation que les patients satisfaits de l'efficacité et de la tolérance du traitement.

Randomisation non équilibrée

Le plus souvent la randomisation répartie les patients en nombre égal dans les groupes traité et contrôle d'un essai. On parle alors de randomisation 1:1, pour un patient alloué dans le groupe traité, un autre patient sera alloué dans le groupe contrôle[3[1]]*. Dans le cas d'une comparaison entre deux groupes, cette répartition assure une puissance statistique optimale. Cependant dans certains essais, la randomisation est conçue pour obtenir un effectif 2 fois ou 3 fois supérieur, rarement plus, dans un groupe par rapport à l'autre. Différentes raisons conduisent à ce choix.*

Dans un essai multibras, le groupe contrôle sert à plusieurs comparaisons. Il a été proposé que son effectif soit plus important que celui des autres groupes.

Une autre justification est de réduire le nombre de patients alloué au groupe du traitement que les investigateurs pensent être inefficace par exemple le placebo, afin de diminuer le nombre de patients qui auront une perte de chance du fait de leur participation à l'essai. Ce raisonnement est fallacieux car il sous-entend une intime conviction que le traitement est efficace. Or, soit les preuves de cette efficacité existent déjà et dans ce cas un nouvel essai à la recherche de l'efficacité est inutile. Soit cette démonstration n'existe pas et dans ce cas il y a équité entre les deux traitements. Dans ce cas, rien ne garantit que ce seront les patients du groupe traité avec le traitement étudié qui seront les plus chanceux. Si ce traitement s'avère délétère, ce seront les patients du groupe contrôle qui auront reçu le meilleur traitement.

Malgré ces réserves, une randomisation déséquilibrée peut être envisagée dans le deuxième essai réalisé pour confirmer le résultat favorable du premier essai. Le premier essai apporte une information a priori fiable. Il

---

3[1] Ce rapport est assuré en moyenne. Dans les essais multicentriques des déséquilibres modérés

n'est pas encore formellement démontré que le traitement étudié soit supérieur au traitement contrôle, mais cela est fort probable.

Bibliographie

1. *Fahn S, Oakes D, Shoulson I, Kieburtz K, Rudolph A, Lang A, et al. Levodopa and the progression of Parkinson's disease. N Engl J Med 2004;351(24):2498-508.*

2. *Bernaldo de Quiros JC, and the Grupo de Estudio del SIDA 04/98. A randomized trial of the discontinuation of primary and secondary prophylaxis against pneumocystis carinii pneumonia after highly active antiretroviral therapy in patients with HIV infection. NEJM 2001;344:159-67.*

3. *Packer M, Gheorghiade M, Young JB, Costantini PJ, Adams KF, Cody RJ, et al. Withdrawal of digoxin from patients with chronic heart failure treated with angiotensin-converting-enzyme inhibitors. NEJM 1993;329:1-7.*

Le contrôle du biais d'attrition : Analyse en intention de traiter et remplacement des données manquantes

Introduction

Le biais d'attrition est le biais induit par l'exclusion de patients au cours de l'étude. Cette perte (attrition) est susceptible d'induire un biais, surtout quant ces exclusions ne se font pas strictement au hasard mais avec une probabilité dépendant du traitement reçu et/ou de l'évolution du patient.

Pour éviter ce biais, tous les patients inclus dans l'essai doivent être pris en compte dans l'analyse. Autrement, en cas d'attrition, le résultat final cours le risque de faire conclure à tort à une différence entre les traitements comparés. Pourtant, le maintient dans l'analyse de patients qui n'ont pas été traités conformément avec le protocole, qui ont reçu par erreur le traitement de l'autre groupe ou qui ont eu un suivi incomplet, peut apparaître surprenant et ne pas permettre une estimation correcte de l'effet du traitement évalué. En fait, nous allons montrer que seul le maintient de ces patients dans leur groupe d'origine évite les biais. Cette stratégie d'analyse s'appelle l'analyse en ITT et s'oppose à l'analyse en per protocole qui sélectionne les patients analysés sur le respect du protocole et la disponibilité des données.

Mais l'analyse en intention de traiter n'est pas suffisante pour éviter le biais d'attrition. EN cas de données manquantes sur le critère de jugement principal, l'ITT n'assure pas que tous les patients inclus seront analysés. Pour atteindre cet objectif il faut prendre en compte ces patients pour lesquels aucune valeur du critère de jugement n'est disponible en remplaçant les données manquantes.

Contrôle du biais d'attrition = analyse en intention de traiter + remplacement des données manquantes

Les écarts au protocole

Dans un essai, malgré la prise de beaucoup de précautions, des écarts aux protocoles peuvent survenir chez certains patients. Les différents écarts au protocole possibles sont nombreux et comprennent entre autres :

l'arrêt prématuré du traitement de l'étude (« *withdrawal* »), voir l'absence de prise de ce traitement,
une mauvaise observance (« *compliance* ») du traitement étudié,
la prise du traitement de l'autre groupe (« *cross-over* »),
la prise d'un traitement interdit,
l'inclusion de patients ne répondant pas aux critères d'inclusion (« *eligibility violation* »),
la non présentation aux visites de suivi (« *lost to follow-up* »),
l'absence de données concernant le critère de jugement (« *missing data* »).

Comment doivent être prises en compte ces déviations au protocole ?

La première idée venant à l'esprit est de ne pas prendre en compte ces patients dans l'analyse car ils ne permettent pas de mesurer de façon optimale l'effet du traitement. En effet, soit ils n'ont pas pu bénéficier de la totalité du bénéfice apporté par le traitement, soit il n'a pas été possible de mesurer correctement l'effet de ce traitement chez eux. Même si elle paraît logique à première vue, cette façon de procéder est susceptible de biaiser les résultats.

Les retraits de l'essai (appelés aussi sorties de l'essai, « drop out *») risquent d'introduire un biais car ils entraînent la destruction de la comparabilité initiale des groupes issus de la randomisation. Par exemple, il peut arriver qu'un patient présente le critère de jugement juste après avoir été randomisé et avant même d'avoir reçu le traitement de l'essai. Le retrait de ce patient de l'étude pourrait paraître logique puisqu'il n'a pas eu le temps de bénéficier du traitement, mais cela reviendrait à favoriser ce groupe en lui retirant un patient à haut*

risque. L'autre mécanisme par lequel les « sorties d'étude » introduisent un biais provient du fait que les écarts au protocole peuvent être liés à l'effet du traitement.

Pour éviter ces biais et maintenir la comparabilité initiale des groupes, un essai doit être analysé en intention de traiter (« intention to treat analysis *») [1-3], c'est-à-dire en analysant tous les patients randomisés dans leur groupe de randomisation, quel que soit le traitement qu'ils ont effectivement reçu ou quel que soit leur devenir dans l'essai. Pour réaliser cette analyse, il convient que le critère de jugement soit disponible pour tous les patients et donc qu'il y ait aucun perdu de vue.*

Tableau 1 – Terminologie.

Arrêt de traitement	Patient qui arrête prématurément de prendre le traitement de l'étude. Les arrêts de traitement ne sont pas forcément des écarts au protocole ; Celui-ci peut très bien (et doit) prévoir l'arrêt des traitements en cas de survenue d'événements indésirables.
Sortie de l'essai	Patient inclus dans l'essai mais qui en est retiré pour une raison quelconque : arrêt de traitement, écarts au protocole, inclusion à tort, etc.
Écart au protocole	Patient, qui à un moment ou un autre de l'essai, n'a pas été traité ou suivi en conformité avec le protocole.
Inclus à tort	Patient inclus dans l'essai alors qu'il ne présentait pas tous les critères d'inclusions ou qu'ils présentait au moins un critère d'exclusion.
Perdu de vue	Au sens premier, patient dont le devenir est inconnu et pour lequel le critère de jugement est indisponible. Par extension peut définir tout patient pour lequel le critère de jugement est indisponible quelle qu'en soit la raison (patient perdu de vue, critère non mesuré).
Donnée manquante	Patients pour lequel le critère de jugement est indisponible.

Le principe de l'analyse en intention de traiter stipule que tous les patients randomisés soient suivis jusqu'à leur décès ou à la fin de l'essai, ou jusqu'à l'observation de l'événement critère de jugement, quelle que soit leur observance au traitement de l'étude.

L'analyse en intention de traiter se distingue donc des autres formes d'analyses, comme l'analyse per protocol (« per protocol analysis *»), où des patients sont exclus de l'analyse sur la base d'information acquise après randomisation. Ces procédés s'apparentent à une sélection post-hoc des observations participant à l'analyse et sont susceptibles d'introduire un biais. En effet, étant donné que cette sélection post-hoc s'effectue sur des critères observés après randomisation, les résultats ne peuvent pas s'appliquer à la population des patients initialement traités. De plus, étant donné que les patients analysés ne sont pas identifiés avant la randomisation, les propriétés de la randomisation ne s'appliquent plus à la sous-population sélectionnée et empêchant ainsi l'évaluation non biaisée de l'effet du traitement. Le fait que les critères de sélection aient été définis a priori dans le protocole et qu'il n'y a pas de déséquilibre significatif des caractéristiques des groupes ne permet pas d'éliminer l'existence de ces biais.*

Tableau 2 – Différentes causes d'arrêt de traitement.

Résultat thérapeutique insuffisant • le patient ressent un manque d'efficacité thérapeutique ce qui le conduit à arrêter le traitement de l'essai et/ou à ne plus se présenter aux visites de suivi • le médecin ayant en charge le patient juge le résultat thérapeutique insuffisant et arrête le traitement de l'étude pour recourir à une alternative thérapeutique (patient qui s'aggrave)
Résultat thérapeutique satisfaisant. le traitement est arrêté car l'état du patient s'est nettement amélioré ou la guérison a été obtenue
Traitement mal toléré
Évolution de la maladie (aggravation ou amélioration)
Le patient ne supporte plus de prendre un traitement de prendre un traitement inconnu

Biais entraîné par les sorties d'étude

L'exemple suivant illustre le type de biais que peut introduire la non-prise en compte dans l'analyse de la totalité des patients randomisés.

Un essai compare un nouveau traitement N à un traitement standard S en utilisant comme critère de jugement le taux d'échecs thérapeutiques (nombre de patients n'ayant pas répondu au traitement). En réalité le nouveau traitement a la même efficacité que le traitement standard. Le vrai risque relatif de la comparaison N vs S est donc RR=1. Par contre, le nouveau traitement est moins bien toléré que le traitement standard. Il s'avère aussi que la survenue d'un effet indésirable entraîne plus fréquemment l'arrêt de la prise du traitement chez les patients qui sont en échec thérapeutique que chez ceux qui répondent au traitement.

Le Tableau 3 présente les résultats de cet essai. En analysant tous les patients randomisés (qu'ils aient arrêté ou non la prise du traitement), il s'avère que N n'est pas différent de S (taux d'échecs thérapeutiques identiques dans les deux groupes 10%). Par contre en retirant de l'analyse les patients qui ont arrêté leur traitement avant la fin prévue, le taux d'échec thérapeutique s'avère plus faible avec le nouveau traitement (8,6%) qu'avec le traitement standard (9,8%), ce qui conduit à conclure faussement que N est plus efficace que S (risque relatif de 0,88). Ce biais provient du fait que les arrêts de traitement ne sont pas indépendants de l'effet du traitement. La sortie de l'essai de ces arrêts de traitement fait que ces patients ne sont pas évalués vis-à-vis de leur réponse thérapeutique ce qui rend impossible la détection du biais.

Par contre si aucun patient n'est sorti de l'étude, les arrêts de traitement sont enregistrés puis le patient évalué au terme. Une situation de ce type peut être imaginée, par exemple, avec un nouvel antidépresseur dont l'efficacité sur les troubles de l'humeur est strictement identique au traitement de référence, mais qui va s'accompagner de plus d'effets indésirables : sécheresses buccales, nausées. Un patient aura d'avantage tendance à arrêter un traitement qu'il supporte mal, s'il ne note pas d'amélioration de son état thymique. L'arrêt du traitement ne sera donc pas indépendant de l'effet du traitement.

Tableau 3 – Exemple de biais introduit par les sorties d'essais. Les deux traitements ont la même efficacité et les patients non-répondeurs arrêtent deux fois plus fréquemment leur traitement que les répondeurs. Cependant, le nouveau traitement est moins bien supporté que le traitement standard.

	Nouveau traitement	**Traitement standard**
Patients randomisés		
Effectif randomisé	1000	1000
Fréquence échec	10,0%	10,0%
Échecs thérapeutiques (non répondeurs)	100	100
Patients analysés		
taux d'arrêts chez les répondeurs	13%	2%
taux d'arrêts chez les non répondeurs	26%	4%
sortie d'étude chez les répondeurs	117	18
sorties d'étude chez les non répondeurs	26	4
répondeurs	783	882
non répondeurs	74	96
effectif analysé	857	978
fréquence échec	8,6%	9,8%
risque relatif	**0,88**	

Cet exemple théorique illustre la première justification de l'intérêt de l'analyse en intention de traiter : éviter le biais introduit par l'exclusion de patients dont le taux n'est pas indépendant de la réponse thérapeutique en prenant en compte la totalité des patients randomisés.

L'analyse en intention de traiter peut paraître parfois « choquante » et conduire à des résultats peu pertinents. C'est par exemple le cas où une forte proportion des patients d'un groupe a reçu le traitement de l'autre groupe. L'analyse en intention de traiter conduit à comparer deux groupes où les patients ont quasiment reçu le même traitement. Bien entendu cette situation n'est pas optimum pour la mise en évidence de l'effet du traitement. L'analyse en intention de traiter entraîne une perte de puissance, mais c'est la seule façon de gérer les écarts aux protocoles sans prendre un risque important de biais. Pour éviter de se retrouver dans cette situation les écarts aux protocoles doivent être exceptionnels et tout doit être fait dans le déroulement de l'essai pour les éviter le plus possible, sans, bien entendu, que cela aille à l'encontre de l'intérêt du patient. Il n'est pas acceptable de poursuivre l'administration d'un traitement mal supporté sous prétexte que l'arrêt du traitement va perturber la recherche de l'effet du traitement.

De plus il faut remarquer que certains écarts au protocole sont des événements se produisant dans la vie de tous les jours et qu'ils reflètent ce qui se passe en pratique.

Exemples

Exemple 1 - *L'objectif de l'essai Coronary Drug Projet était de savoir si la prise d'un hypocholestérolémiant pouvait réduire, par rapport à la prise d'un placebo, la mortalité à 5 ans de patients atteints de coronaropathies [4]. Les résultats furent décevants : la mortalités était de 20% dans le groupe hypolipémiant contre 21% dans le groupe placebo (différence non significative). Il était tentant de ne considérer que les patients qui avaient pris régulièrement le traitement hypocholestérolémiant et de les comparer aux patients du groupe placebo. Cette mortalité est alors de 15% et la différence devient statistiquement significative avec le placebo. Cependant, dans le groupe placebo, le taux de mortalité des patients observants était aussi de 15% !*

Exemple 2 - *Sacket relate l'exemple d'un essai comparant la chirurgie au traitement médical pour la prévention des accidents vasculaires cérébraux en cas de sténoses carotidiennes bilatérales [5] [6]. Le critère de jugement utilisé combinait AVC, accident ischémique transitoire et décès.*

L'analyse des données montrait une diminution de fréquence du critère de jugement de 27%, statistiquement significative (p=0,02). Mais cette analyse excluait 16 patients qui avaient présenté un AVC ou qui étaient décédés avant de sortir de l'hôpital. Tous les patients exclus sauf 1 appartenaient au groupe chirurgie. L'analyse de tous les patients randomisés ne montre pas de différence significative p=0,09 (Tableau 4).

Cet exemple montre comment l'exclusion des patients peut transformer la nature des conclusions d'un essai.

Tableau 4 – Différence de résultat entre l'analyse initial et l'analyse en intention de traiter dans un essai de chirurgie carotidienne.

Traitement	**Ev / n**	**Fréquence**	
	Analyse initiale		
médical	53/72	74%	p=0,02
chirurgie	43/79	54%	
	Analyse en intention de traiter		
médical	54/73	74%	p=0,09
chirurgie	58/94	62%	

Exemple 3 - *Dans un essai dans l'angor comparant sur le pronostic, traitement médical contre chirurgie, les patients du groupe traitement médical qui nécessitent un recours à la chirurgie (par exemple pour une dégradation subite de leur état) doivent être analysés dans leur groupe d'origine pour lequel ils représentent un échec. Les passer dans le groupe chirurgie favorise le traitement médical en faisant disparaître certains de ses échecs et défavorise le traitement chirurgical si ces patients sont à plus mauvais pronostic.*

Type d'analyse

En fonction de la façon dont sont pris en compte les écarts au protocole, plusieurs types d'analyse sont possibles : analyse en intention de traiter, en per protocol, en traitement reçu. Chacune de ces analyses s'appuie sur une population (« data set *») différente qui sont respectivement pour les deux premières : la population en intention de traiter («* full data set *»), la population per protocol («* per protocol data set *»).*

L'analyse per protocole donne une meilleure estimation de l'effet potentiel du traitement (celui obtenu dans une situation idéale). Mais cette estimation est potentiellement biaisée. L'analyse en intention de traiter garantit l'absence de biais et permet l'estimation de l'effet du traitement dans les conditions proches de celle de la vie réelle. L'analyse principale d'un essai doit donc être faite en intention de traiter, complétée, éventuellement par une analyse en per protocol ou en traitement reçu.

L'application du principe de l'intention de traiter nécessite un suivi complet de tous les patients randomisés pour tous les critères de l'étude

Tableau 5 – Les différentes types d'analyses possibles : intention de traiter, per protocole et en traitement reçu

Intention de traiter
Analyse de tous les patients randomisés dans le groupe où ils furent randomisés ♦ quelle que soit leur observance au traitement alloué ♦ quel que soit le traitement réellement reçu ♦ quel que soit l'éventuel retrait du patient de l'étude ou d'éventuelle déviation au protocole
Per protocole
Analyse uniquement des patients qui ont été traités en pleine conformité avec le protocole. Les inclusions à tort, les patients non observants, les patients traités avec le traitement de l'autre groupe sont exclus de l'analyse.
Analyse en traitement reçu
Les patients sont analysés en fonction de la nature du traitement qu'ils ont reçu, même s'il ne s'agit pas du traitement qui leur fut alloué par la randomisation

Aspect pragmatique de l'intention de traiter

L'autre justification du principe de l'intention de traiter est qu'il permet d'analyser l'essai conformément à ce qui se passe dans la pratique médicale courante. Par exemple certains patients stoppent prématurément leur traitement car ils le supportent mal. Cette mauvaise tolérance peut limiter l'intérêt du traitement et il est impératif que ces patients soient pris en compte dans l'analyse.

Premier exemple

Pour pouvoir être efficace, l'administration de fibrinolytique à la phase aiguë de l'infarctus du myocarde doit être effectuée le plus précocement après l'obstruction coronarienne. Une reperfusion trop tardive, survenant au delà du stade d'ischémie réversible, ne permet pas de limiter la taille de la nécrose et ne peut donc pas avoir un effet sur la mortalité.

Ainsi, ce traitement doit être administré au stade de suspicion d'infarctus, avant que le diagnostique n'ait pu être confirmé par l'observation de l'élévation enzymatique qui se produit plus tardivement. Cette nécessité conduit donc à inclure dans les essais des patients qui s'avèreront ultérieurement ne pas présenter des infarctus. Comment analyser ces patients ?

Certains pourraient les exclure en argumentant que ces patients n'apportent aucune information sur l'effet du fibrinolytique étant donné qu'ils ne souffraient pas de la pathologie visée.

Cependant, parmi les tableaux cliniques pouvant se présenter comme une suspicion d'infarctus figurent les péricardites et les dissections aortiques. Ces pathologies sont à haut risque de complications graves lors de l'administration d'un fibrinolytique. Il est donc indispensable de conserver ces patients dans l'analyse pour vérifier que les effets délétères qu'ils pourraient présenter ne contrebalancent pas la totalité du bénéfice dégagé chez les infarctus confirmés. En effet, dans la « vraie vie » tous ces patients seront traités.

L'essai ne cherche pas à montrer que le traitement est efficace dans l'absolu (en cas d'infarctus confirmé et uniquement dans ce cas là) , mais que l'utilisation de ce traitement s'accompagne d'une réduction de la mortalité chez les patients traités, dans les conditions où il sera utilisé en pratique, ce que seule l'analyse en intention de traiter permet de le vérifier.

Tableau 6 – Comparaison du bénéfice clinique et de l'efficacité « pharmacologique »

Objectif	Type d'analyse	Approche	Intéresse le
Bénéfice clinique ("*clinical effectiveness*")	Analyse en intention de traiter	Essai pragmatique	Clinicien, santé publique
Efficacité « pharmacologique »	Analyse per-protocole	Essai explicatif	Pharmacologue, chercheur

En conclusion, l'analyse en intention de traiter produit la réponse la plus proche de la réalité et la moins biaisée à la question la plus pertinente celle de l'efficacité clinique.

Deuxième exemple

Dans les années 1980 plusieurs anticorps monoclonaux dirigés contre l'endotoxine des bactéries gram négatif ont été développés comme traitement adjuvant des septicémies à gram négatif (G-). L'évaluation clinique de l'un d'entre eux, le HA-1A (Centoxin), illustre les problèmes posés par les analyses per-protocole. Pour pouvoir être efficace, ces anticorps monoclonaux doivent être administrés très précocement en cas de suspicion de septicémie à G-, bien avant la confirmation bactériologique par les hémocultures.

Dans un premier essai [7], publié en février 1991 dans le New England Journal of Medicine, la mortalité à 28 jours passait de 49% sous placebo à 30% chez les patients traités par HA-1A, réduction significative (p=0,014). Dans l'essai, 543 patients ont été randomisés (262 dans le groupe HA-1A et 281 dans le groupe placebo) mais seulement 200 d'entre eux se sont avérés porteurs d'une bactériémie à gram négatif. L'analyse de l'efficacité n'a été réalisée que chez ces 200 patients (95 placebo et 105 HA-1A) ce qui revient à exclure 343 sujets (63%) de l'analyse. Malgré ses limites méthodologiques ce résultat fut accepté et justifia la prescription de ce traitement coûtant au alentours de 20 000 FF ($3 500) dans plusieurs pays (Angleterre et France entre autres). En fait, l'analyse en intention de traiter ne montre pas de bénéfice pour l'ensemble des patients inclus.

La sélection après randomisation des patients participant à l'analyse détruit la comparabilité issue de la randomisation comme en témoigne l'analyse des caractéristiques des bases des patients. Le groupe traité se trouve favorisé avec, dans le groupe placebo, plus de CIVD, de détresse respiratoire, d'insuffisance hépatique aiguë et de défaillance rénale aiguë (Tableau 7). Le pronostic des patients du groupe traité était donc, initialement, meilleur que celui des patients du groupe contrôle.

Tableau 7 – Sévérité de l'état septique à l'entrée chez les patients avec bactériémie Gram négatif.

Variable	Placebo (n=95)	HA-1A (n=105)
Score APACHE II	25,7+/-8,1	23,6+/-9,0
	pourcentage	
Hypotension	51	51
Intubation endotrachéale	55	54
Coagulation intravasculaire disséminée	21	18
Détresse respiratoire de l'adulte	13	9
Défaillance hépatique aiguë	26	19
Insuffisance rénale aiguë	46	35

En conclusion, l'interprétation qui fût faite en 1991 des résultats de cet essai était abusive. Ce résultat n'apporte pas la preuve que ce traitement réduit la mortalité des patients traités. Au mieux, et sans garantie méthodologique, le traitement n'apporte un bénéfice qu'à une faible proportion des patients.

Un second essai (CHESS) a été réalisé et publié en 1994 dans les Annals of Internal Medicine [8]. 2199 patients furent inclus dans cet essai dont 621 (28,2%) s'avérèrent a posteriori être porteurs d'une bactériémie à gram négatif. Les taux de mortalité à 14 jours furent de 32% dans le groupe placebo et de 33% avec l'anticorps monoclonal (p=0,864), tandis qu'ils furent respectivement de 37% et 41% chez les sujets sans bactériémie à gram négatif (p=0,073). Cet essai ne confirme donc pas le premier et ne donne pas d'argument pour l'utilisation de ce traitement. Au contraire, il suggère même une tendance à une surmortalité chez les sujets sans infection à gram -.

L'exemple du Centoxin illustre aussi les limites de l'approche physiopathologique pour la validation ultime d'un traitement. L'essai de 1991 était un essai qui se voulait explicatif : l'HA-1A réduit-il la mortalité de ses patients cibles ? Malheureusement la réponse à cette question nécessite une détermination a posteriori des patients cibles, ce qui expose à plusieurs sources de biais (destruction de la randomisation, absence analyse en intention de traiter, multiplicité des comparaisons). Or, la véritable question qui se pose en pratique médicale est : l'utilisation de l'HA-1A chez les patients suspects de développer une septicémie à G- permet-elle d'en réduire la mortalité

« Efficacité pure »

Des traitements peuvent être plus ou moins bien supportés par les patients mais cependant apporter un bénéfice aux patients qui le tolèrent. C'est par exemple le cas avec certaines chimiothérapies anticancéreuses ou les traitements antiviraux.

La question est donc de montrer que ces traitements sont efficaces chez les patients les supportant. Nous avons vu qu'une l'analyse post-hoc limitée aux patients supportant le traitement ne répond pas à cette question. Un plan expérimental non biaisé mais non entièrement satisfaisant consiste à d'abord tester la tolérance au traitement dans une phase pré randomisation puis de randomiser les patients qui tolèrent le traitement entre un groupe expérimental et un groupe contrôle.

Ce plan permet de répondre à la question de l'efficacité chez les sujets tolérant le traitement, mais il ne prend pas en compte le devenir des patients non tolérants et ne répond pas à la question pratique. En effet, que deviennent les patients ne tolérant pas le traitement : le retard de mise en place d'un traitement induit par le test de tolérance leur est peut-être dommageable ? L'intolérance au traitement est peut-être le fait d'effets indésirables sérieux ?

La réponse satisfaisante à cette question peut être obtenue dans un essai de stratégie. Une stratégie où les patients reçoivent le nouveau traitement tant qu'ils le supportent puis le traitement de référence en cas d'apparition d'intolérance est comparée à une stratégie où tous les patients reçoivent le traitement de référence. Cette approche intègre les éventuels effets délétères qu'auraient pu subir les patients n'ayant pas toléré le nouveau traitement et répond à la question très pragmatique : y a-t-il un intérêt à débuter le traitement par la nouvelle thérapeutique. En effet, en pratique, on propose le traitement à un sujet sans savoir s'il va le tolérer ou non. Ce dont on veut s'assurer c'est, qu'a priori, il ne pourra que bénéficier de cette stratégie ; c'est-à-dire, s'il tolère le traitement augmenter ses chances de succès thérapeutique et s'il ne tolère pas, avoir les mêmes chances que s'il avait été traité d'emblée avec le traitement de référence. Utiliser un traitement qui a montré qu'il apportait un bénéfice supplémentaire chez les patients qui le supportent est une approche centrée sur la fascination de trouver des traitements intéressants, mais n'est pas une approche centrée sur l'intérêt des patients. On obtient un gain chez certains patients au détriment d'autres. Il faut montrer que le traitement apporte un bénéfice supplémentaire chez ceux qui le tolèrent et que, chez les patients ne tolérant pas le traitement, la période nécessaire pour le constater n'a pas entraîné un retard de mise en place d'un traitement pouvant être dommageable pour le patient et que les effets liés à l'intolérance n'ont pas été délétères.

Ainsi, en chirurgie, il serait aberrant de ne mesurer l'efficacité d'une intervention chirurgicale que chez les patients chez lesquelles elle a réussi et d'écarter ainsi les décès péri-opératoires, les complications ou les échecs du geste chirurgical.

Conclusion

L'analyse en intention de traiter peut parfois paraître « insolite ». C'est pourtant la seule approche qui évite l'apparition de biais liés aux écarts au protocole ou aux arrêts de traitements. L'analyse en intention de traiter peut être complétée à titre documentaire d'une analyse per-protocole permettant d'approcher l'efficacité « théorique » d'un traitement utilisé dans des conditions idéales. Mais cette efficacité « théorique » surestime souvent l'efficacité qui sera obtenue en pratique, car elle ne prend pas en compte les errements de la « vraie vie ». Par contre, l'analyse en intention de traiter mesure le bénéfice qu'apporte un traitement dans des conditions proches de celles de son utilisation en pratique courante.

Lecture critique

En lecture critique, pour l'évaluation de la validité interne, il convient de déterminer si tous les patients randomisés sont évalués et entrent dans le calcul des résultats finaux. Il est donc nécessaire de chercher s'il existe des patients randomisés qui ont été exclus de l'analyse.

Les différents points à analyser sont les suivants :

- *le chapitre méthode statistique à la recherche de la mention : analyse en intention de traiter*
- *la première partie de la section « résultats » dévolue classiquement à la description du nombre de patients, les retraits, les perdus de vue*
- *le « flow chart » (ou «* trial profile *») qui représente sous forme graphique les différents flux de patients entre la randomisation, et l'analyse [9, 10] (cf. chapitre suivant).*
- *la concordance entre les effectifs randomisés annoncés dans le texte et les effectifs rapportés dans le tableau de description des caractéristiques de bases et celui des critères de jugement.*

Assez souvent et surtout dans les publications anciennes, le nombre de patients exclus après randomisation n'apparaît pas. Dans ce cas, il est impossible d'écarter un certain degré de suspicion. Seule la mention explicite de l'absence de perdus de vue ou d'exclusion après randomisation évite les suspicions. Cette information positive est donc présentée dans les publications « modernes ».

Parfois l'analyse est conduite suivant le principe de l'intention de traiter (aucun patient n'est sorti de l'essai par l'investigateur) mais l'existence de nombreux perdus de vue entraîne l'exclusion de ces patients de l'analyse. Même si cette analyse est en intention de traiter les perdus de vue font courir le risque de biais, ce qui implique de prendre en compte les données manquantes (voir chapitre : Données manquantes*).*

Bibliographie

1. Lachin JL. Statistical considerations in the intent-to-treat principle. Controlled Clinical Trials 2000;21(5):526. *PMID:*
2. Lewis JA, Machin D. Intention to treat--who should use ITT? Br J Cancer 1993;68(4):647-50. *PMID:*
3. Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB. Analysis of clinical trials by treatment actually received: is it really an option? Stat Med 1991;10(10):1595-605. *PMID:*
4. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. JAMA 1975;231:360-381. *PMID:*
5. Fields WS, Maslenikov V, Meyer JS, Hass WK, Remington RD, Macdonald M. Joint study of extracranial arterial occlusion. V. Progress report of prognosis following surgery or nonsurgical treatment for transient cerebral ischemic attacks and cervical carotid artery lesions. JAMA 1970;211(12):1993-2003. *PMID:*
6. Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. NEJM 1979;301:1410-1412. *PMID:*
7. Ziegler EJ, Fisher CJ, Sprong CL. Treatment of gram-negative bacteraemia and septic shock with HA-1A human monoclonal antibody against endotoxin. NEJM 1991;324:429-36. *PMID:*
8. McCloskey RV, Straube RC, Sanders C, Smith SM, Smith CR. Treatment of septic shock with human monoclonal antibody HA-1A. A randomized, double-blind, placebo-controlled trial. CHESS Trial Study Group. Ann Intern Med 1994;121(1):1-5. *PMID: 8198341*.
9. Egger M, Jüni P, Bartlett C, for the CONSORT Group. Value of the flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996-1999. *PMID:*
10. Moher D, Schulz KF, Altman DG, for the CONSORT group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med 2001;134:657-662. *PMID:*

Les flux de patients

Les différentes catégories de patients

Durant le déroulement d'un essai apparaissent différentes catégories de patients [1].

*Les **patients présélectionnés** (« screened patients »)* *correspondent aux patients qui ont été évalués en vue d'une inclusion dans l'essai. Après exclusion des patients non éligibles, les patients restants sont effectivement inclus dans l'essai. La différence entre le nombre de patients envisagés et ceux effectivement inclus permet d'évaluer grossièrement si la population des patients inclus est représentative de la population ciblée. Les raisons de non-éligibilité doivent être précisées : critères d'exclusions, refus de consentement, autres raisons. Un fort taux de non-inclusion sans raison particulière peut faire suspecter des exclusions arbitraires.*

Ces non inclusion n'introduisent pas de biais mais peuvent fausser la représentativité des patients inclus dans l'essai. La généralisabilité d'un résultat se juge donc non seulement sur les critères d'inclusion et d'exclusion de l'essai mais aussi sur ce qui c'est passé durant cette phase de sélection.

> Par exemple, dans de nombreux essais de statines [2] les patients ont eu, avant inclusion définitive, un traitement d'essai de quelques semaines par la statine (« run-in ») afin de vérifier leur tolérance à ce médicament. Durant cette période, les patients qui présentaient des signes musculaires cliniques ou biologiques ont été exclus. Lors de l'analyse de la balance bénéfice –risque, il convient de ne pas oublier que les bons résultats obtenus l'ont été chez des patients qui ne présentaient pas de signe précoce d'intolérances musculaires. En pratique, il convient d'instaurer un traitement au long court qu'après avoir évaluer la tolérance, de la même manière que dans l'essai.

*Le nombre de **patients randomisés** dans chaque groupe représente la valeur de référence à partir de laquelle se positionnent tous les autres effectifs. Le nombre de patients randomisés doit être similaire entre les groupes de l'essai (en dehors d'une randomisation volontairement déséquilibrée). Un déséquilibre important fait suspecter une défaillance de la randomisation et/ou des exclusions secondaires cachées. Dans tous les cas, la comparabilité initiale des groupes est remise en cause.*

*Les **patients évalués** (« analyzable patients »)* *sont ceux qui sont effectivement pris en compte dans l'analyse statistique et qui contribuent à l'estimation de l'effet traitement. Nous avons vu que ce nombre doit être identique au nombre de patients randomisés pour éviter les biais (biais d'attrition, cf. analyse en intention de traiter). Une différence entre nombre randomisé et nombre évalué peut être due :*

- aux patients **perdus de vue** ou pour lesquels le critère de jugement est indisponible,
- aux **sorties d'essai**, c'est-à-dire à des patients qui, après randomisation, sont retirés de l'essai et exclus de l'analyse. De telles exclusions ne sont pas acceptables au niveau de l'analyse primaire car elles font courir un risque important de biais (cf. analyse en intention de traiter). L'expression « sortie d'essai » devrait être bannie du vocabulaire de l'essai thérapeutique !

D'autres catégories de patients documentent le devenir de la population incluse au cours du déroulement de l'essai.

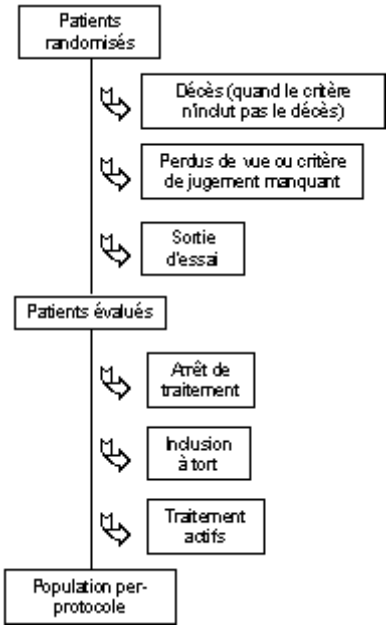


Figure ***Erreur ! Signet non défini.*** *– Les principaux flux de patients dans un essai thérapeutique*

Le nombre ***d'arrêts de traitement prématurés*** *représente le nombre de patients qui ne sont pas allé au bout du traitement prévu par le protocole. Les raisons d'arrêt prématuré permettront d'identifier s'il s'agit d'un problème de faisabilité (traitement trop lourd à mettre en œuvre, difficile à suivre par le patient, etc.), de tolérance ou d'efficacité (arrêt des traitements pour inefficacité ou aggravation de la maladie, voir décès).*

Le nombre de ***patients traités uniquement avec le traitement alloué,*** *c'est-à-dire chez lesquels il n'y a pas eu d'interruption du traitement ou de recours à une autre thérapeutique, permet d'apprécier le contraste existant entre les deux groupes en terme de traitement.*

Dans les essais contre placebo, c'est le nombre de patients ***ayant reçu un traitement actif*** *qui témoigne d'une éventuelle atténuation du contraste entre les groupes. Dans le groupe placebo, ce nombre est celui des patients qui ont reçu un traitement concomitant actif, tandis que dans le groupe traité il s'agit du nombre de patients recevant soit le traitement étudié soit un traitement concomitant actif. Les patients ayant arrêté prématurément le traitement étudié sans recevoir d'autre traitement actif ne sont pas comptés.*

Le nombre de patients ***traités à la fin du suivi*** *est informatif dans les essais le longue durée, comme les essais de prévention (par exemple un essai de morbi-mortalité avec un anti-hypertenseur) où il peut y avoir une forte proportion de patients prenant un traitement actif en fin d'essai.*

Ces nombres d'arrêts de traitement prématurés, de patients ayant reçu un traitement actif permettent de discuter les raisons d'un résultat négatif : les deux groupes ont reçu en même proportion un traitement actif ou de nombreux patients du groupe expérimental n'ont pas été traités de façon optimale.

Le nombre de ***patients ayant terminé le suivi prévu*** *permet d'évaluer le recul moyen d'observation et donc d'apprécier si les résultats sont le reflet de l'effet du traitement au bout du suivi moyen ou s'ils représentent majoritairement un effet à plus court terme (cf. analyse des courbes de survie).*

Lecture critique

En lecture critique, la grille proposée ci-dessous apporte une aide à l'extraction ou à la reconstitution de ces effectifs.

Tableau 1 – Tableau récapitulatifs des différents effectifs et flux de patients identifiable dans un essai thérapeutique

RECRUTEMENT	
Nombre de patients dont l'éligibilité a été évaluée (patients présélectionnés)	(1)
Nombre total de non inclusion	a+b+c
critère d'exclusion	a
refus de consentement	b
autres raisons	c
Nombre de patients inclus	(2) = (1) - (a+b+c)

RANDOMISATION		
	Groupe expérimental	**Groupe contrôle**
Patients randomisés	(1)	(1)
Décès	a	a
perdu de vue	b	b
sortie d'essai (exclusion secondaire)	c	c
donnée manquante	d	d
Patients évalués	(2)=(1)-a-b-c-d	(2)=(1)-a-b-c-d
arrêt de traitement	e	e
prise d'un traitement « actif » hors comparaison	f	f
inclusion à tort	g	g
autres déviations au protocole	h	h
Population per-protocole	(3)=(2)-e-f-g-h	(3)=(2)-e-f-g-h
patients ayant terminé le suivi prévu	-	-
patients ayant reçu la totalité du traitement prévu	-	-

La recommandation « CONSORT »[3] (cf. chapitre Lecture critique*) propose qu'un graphique de flux («* flow chart *») soit systématiquement présent dans les comptes rendus d'essais cliniques [1].*

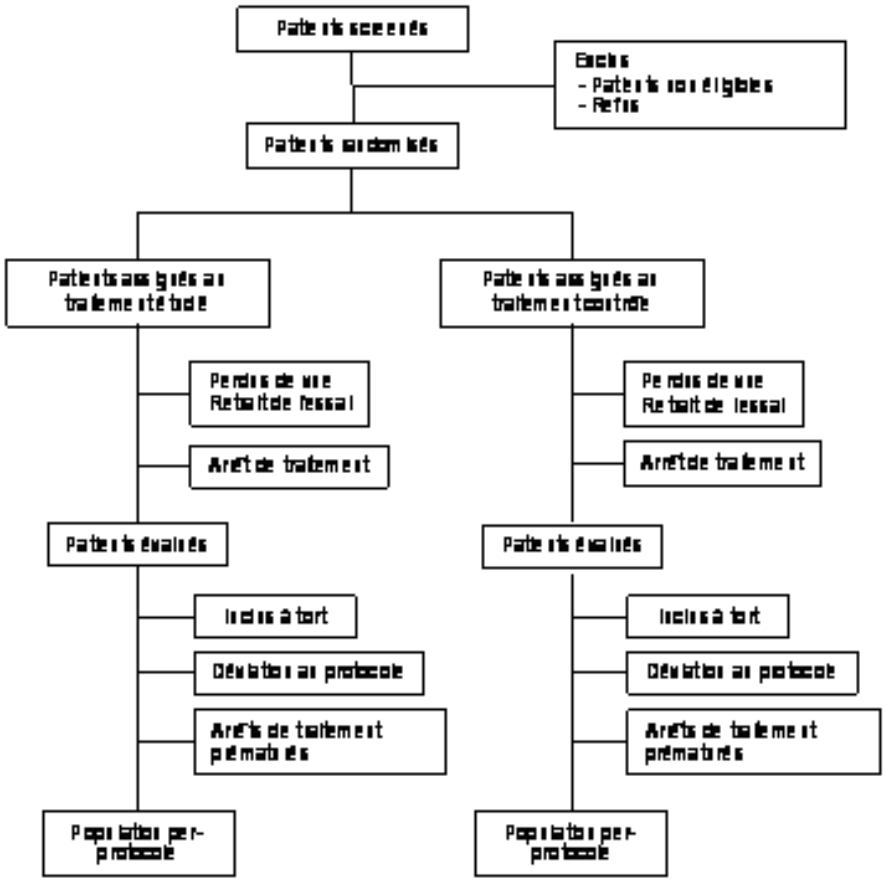


Figure 2 – Modèle de graphique de flux.

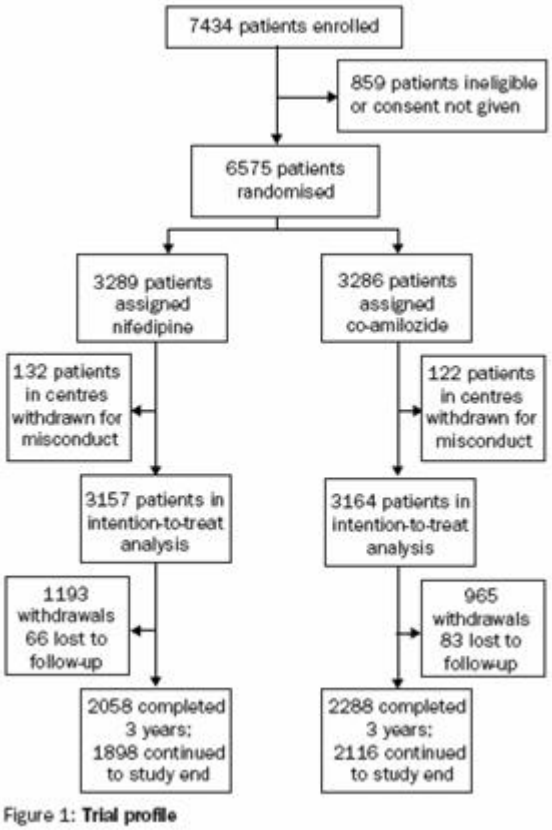


Figure 1: **Trial profile**

Figure 3 – Exemple de flowchart (INSIGHT)

Bibliographie

1. *Egger M, Jüni P, Bartlett C, for the CONSORT Group. Value of the flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996-1999.* PMID:

2. *MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360(9326):7-22.* PMID: 12114036.

3. *Moher D, Schulz KF, Altman DG, for the CONSORT group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med 2001;134:657-662.* PMID:

CONCEPTS STATISTIQUES

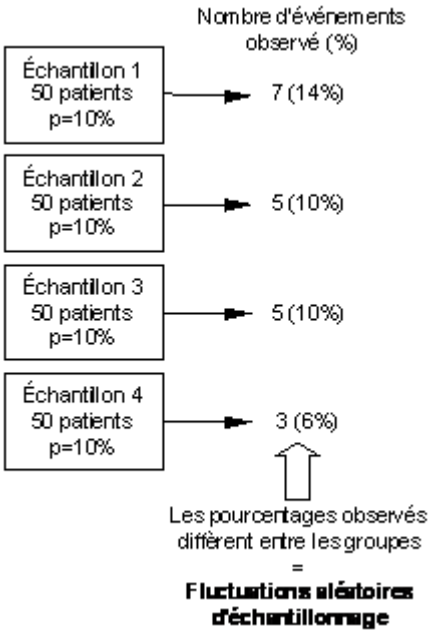
Principe général des tests statistiques

Powerpoint

Les fluctuations aléatoires

La survenue d'un événement clinique chez un patient est en partie imprévisible et s'apparente donc à un phénomène aléatoire. Pour un patient donné, il est impossible de prévoir avec certitude la survenue ou non de l'événement. Par exemple, la survenue sur une période de 5 ans d'un accident cardio-vasculaire chez un sujet hypertendu est imprévisible.
Si l'on surveille plusieurs groupes regroupant des sujets ayant tous la même probabilité de faire l'événement, disons 10%, les différents pourcentages observés vont fluctuer autour de cette valeur. Comme dans ces groupes tous les sujets ont le même risque, appelé **vraie valeur** dans la terminologie statistique, ces différences observées sont à mettre uniquement sur le compte du hasard. Ces fluctuations du paramètre d'intérêt (ici la fréquence de survenue de l'événement clinique) observées entre différents échantillons et dues entièrement au hasard sont appelées fluctuations aléatoires d'échantillonnage.

Figure 1 – Parmi 4 groupes de patients ayant la même probabilité p (appelée aussi risque) de faire l'événement, les pourcentages d'événements observés varient d'un groupe à l'autre Ces différences sont dues au hasard et sont appelées fluctuations aléatoires d'échantillonnage.



Les erreurs statistiques

Les fluctuations aléatoires d'échantillonnage ont des conséquences sur la comparaison de deux groupes (à la recherche d'une différence numérique dans le paramètre considéré). Elles peuvent, entre autres, faire apparaître entre les groupes une différence qui en réalité n'existe pas. Dans une situation où le risque est identique dans les 2 groupes, par hasard, le pourcentage observé dans un groupe pourra être inférieur à ce qu'il aurait du être tandis que dans l'autre groupe, le hasard conduit à une valeur observée surestimant la vraie valeur. Par cette double action du hasard en sens contraire, apparaît une différence entre les deux pourcentages **observés** alors qu'en réalité ils auraient du être identiques puisque les patients des deux groupes ont tous le même risque.
Le but pratique de la comparaison est de conclure, à partir de l'observation, sur l'existence (ou non) d'une vraie différence entre les deux groupes. Comme la réalité est inconnue, l'observation d'une différence apparente va faire conclure, à tort, à l'existence d'une différence vraie entre ces deux groupes. Dans l'essai thérapeutique, la constatation d'une différence suggère l'existence d'un effet non nul du traitement.

Ainsi les fluctuations aléatoires sont susceptibles de conduire à des conclusions erronées à partir de l'observation. L'observation fait conclure à l'existence d'une différence qui, en réalité, n'existe pas. Il s'agit d'une erreur statistique car elle est induite par les fluctuations aléatoires. Elle est appelée erreur statistique de première espèce, ou erreur alpha.

Dans un essai thérapeutique, l'erreur alpha est de conclure à l'efficacité d'un traitement qui, en fait, est inefficace.

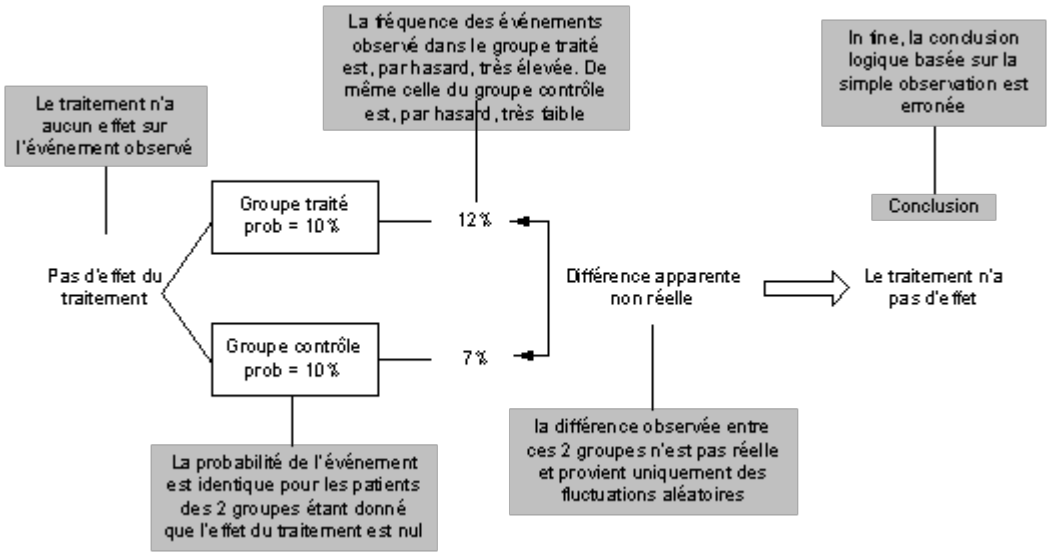


Figure 2 – Illustration du mécanisme conduisant à l'erreur statistique alpha

À l'opposé, les fluctuations aléatoires peuvent aussi faire disparaître une différence qui existe pourtant. Lors de la comparaison d'un paramètre d'intérêt entre deux groupes pour lesquels il existe une réelle différence, le hasard peut conduire à ce que les observations se rapprochent les unes des autres, annulant ainsi la différence. L'observation conduit à conclure, à tort, à l'absence de différence, conclusion là aussi erronée du fait des fluctuations aléatoires. Cette erreur est appelée erreur statistique de deuxième espèce ou erreur bêta.

Dans un essai thérapeutique, l'erreur statistique bêta fait courir le risque de ne pas mettre en évidence l'efficacité d'un traitement.

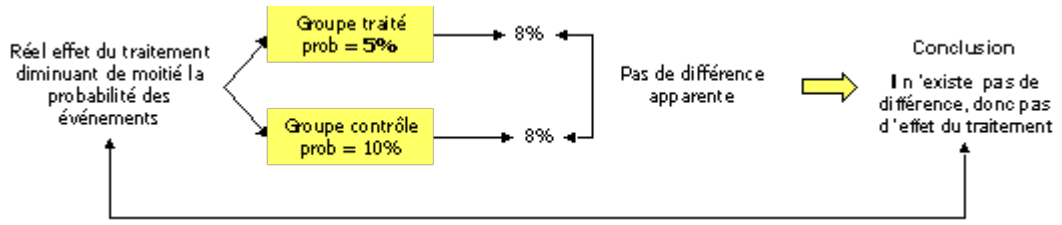


Figure 3 – Illustration du mécanisme conduisant à l'erreur statistique beta

Le test statistique

Il découle de ce que nous venons de voir concernant l'erreur statistique alpha que devant une différence observée il existe deux possibilités : 1) cette différence est uniquement due au hasard et en réalité elle n'existe pas ; 2) cette différence observée est la conséquence directe d'une réelle différence entre les deux groupes.

Les comparaisons sont effectuées pour chercher à faire des conclusions à partir des observations. Dans l'essai thérapeutique, on cherche à conclure ou non à l'efficacité du traitement utilisé en comparant les résultats obtenus dans chaque groupe. De plus ces conclusions vont être à la base de décision, dont les conséquences sont parfois très larges. A partir des conclusions d'un essai thérapeutique, on prendra ou non la décision de recommander l'utilisation d'un traitement.
S'il n'existait aucun moyen de faire la part des choses entre ces deux possibilités, aucune conclusion et décision ne seraient possibles en pratique. Un risque d'erreur inconnu serait constamment présent, laissant planer un doute sur toute conclusion. La solution à ce dilemme est apportée par le test d'hypothèse.

> Le test statistique est un moyen qui permet de rechercher s'il existe une réelle différence entre 2 groupes

Devant une différence observée, le test statistique permet de calculer la probabilité que l'on aurait d'observer ce résultat si en réalité il n'y avait pas de différence entre les deux groupes. Cette probabilité est appelée *p*.
Avec un peu moins de rigueur, il est possible de dire qu'elle correspond à la probabilité que la différence observée soit due au hasard en l'absence d'effet du traitement. Elle permet ainsi une quantification du risque de faire une erreur de première espèce si l'on décidait de conclure à l'existence d'une différence entre les deux groupes.
En pratique, on avancera effectivement cette conclusion que si le risque que l'on a de se tromper est suffisamment petit. Classiquement, il a été convenu que le risque acceptable d'erreur alpha est de 5%. Ainsi, devant une différence observée, on conclura à l'existence d'une réelle différence seulement si le risque de se tromper pris en faisant cette conclusion est inférieur à 5%, c'est-à-dire, si la valeur de *p* donnée par le test est inférieure au seuil de 5%.
Le test statistique est donc un moyen de contrôler le risque d'erreur alpha. Il ne prend pas directement en compte le risque d'erreur bêta.

> Le risque alpha est le risque numérique (probabilité) de commettre une erreur statistique alpha. Le risque bêta est celui de commettre une erreur bêta.

La signification statistique

Lorsque $p \leq 5\%$, la différence est dite « statistiquement significative ». C'est-à-dire qu'elle est suffisamment importante par rapport aux fluctuations aléatoires pour que sa probabilité d'être observée en l'absence de réelle différence soit inférieure au seuil préalablement choisi de 5% (seuil de la signification statistique).
Quand $p > 5\%$, la différence n'est pas « statistiquement significative ». En simplifiant, « elle n'est pas suffisamment importante par rapport aux fluctuations aléatoires pour pouvoir raisonnablement exclure qu'elle soit un artefact dû au hasard ». Une différence non significative n'est pas synonyme d'absence d'effet. La comparaison est peut-être insuffisamment puissante pour mettre en évidence la différence qui existe. L'absence de preuve n'est pas la preuve de l'absence. Le problème du risque bêta et de la puissance statistique sera envisagé dans une autre rubrique.

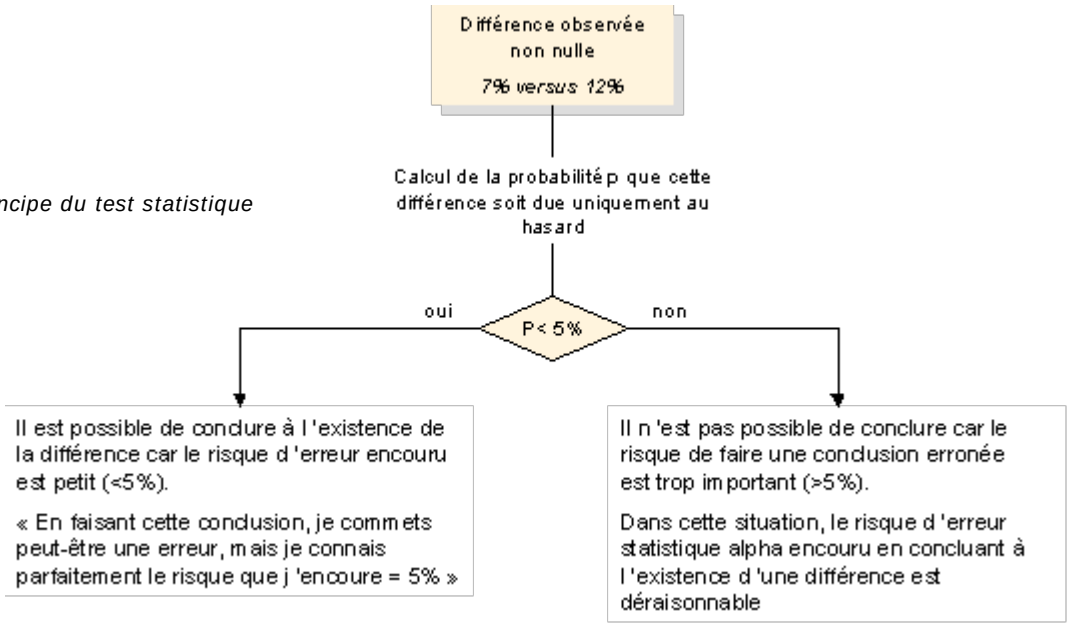


Figure 4 – Principe du test statistique

Un résultat statistiquement significatif signifie seulement que le risque d'erreur alpha est faible, il ne signifie pas qu'il n'y a aucun risque d'erreur et que la conclusion que l'on fait est une certitude. Avec un seuil de 5%, avec un résultat significatif il reste encore 5% de risque de se tromper.

Un seuil de risque α de 5% est-il acceptable ?

Classiquement le seuil de la signification statistique est fixé à 5%. Une autre valeur peut être utilisée, en particulier plus contraignante, comme 1%. En effet, un risque de 5% n'est pas totalement négligeable. Par exemple, supposons qu'il existe environ 400 spécialités différentes dans la pharmacopée et que chacune n'a été évaluée que par un seul essai thérapeutique. Avec un risque alpha de 5%, 20 de ces produits seraient présents à tort dans notre arsenal thérapeutique.
Avec un traitement qui sera très largement diffusé, comme un vaccin par exemple, prendre un risque de conclure à tort à son efficacité de 5% est trop important. Un risque de 1% serait le bienvenu. Par contre, avec une maladie très rare pour laquelle aucun traitement efficace n'est encore disponible, consentir un risque alpha de 10% est peut être envisageable.
Il est difficile de définir des normes pour le choix du seuil de la signification statistique. Il s'agit d'un choix de valeur. L'important est de se souvenir de la signification de ce choix et du fait que la valeur habituelle de 5% est arbitraire et qu'elle n'est pas immuable. Le choix d'une autre valeur plus restrictive est tout à fait possible.

Faut-il un ou deux essais ?

Un seuil de signification inférieur à 5% est de plus en plus utilisé dans les essais thérapeutiques comme par exemple dans l'essai HPS [1] qui comparaient la simvastatine au placebo dans la prévention des maladies cardiovasculaires chez des patients à haut risque. Cet essai de morbi-mortalité de grande taille a choisi un seuil de signification statistique de 1% en partie car il avait de forte chance d'être unique.
En effet, deux essais sont en général demandés pour apporter la preuve de l'efficacité. Cette redondance diminue le risque de conclusion globale erronée. Avec deux essais significatifs à 5%, le risque de conclure à tort à l'efficacité est de 5%*5%=0.25%. Cette règle des deux essais représente donc, entre autre, un moyen de réduire le risque d'erreur de première espèce, sans exiger un seuil de signification pour chaque essai plus strict que la valeur « habituelle » de 5%.
Cependant dans le cas où la recherche de l'effet nécessite de très nombreux patients (plusieurs milliers), il est difficile de réaliser deux essais. Dans ce cas, il est fortement souhaitable que l'essai unique qui est réalisé adopte un seuil de signification plus petit que 5% ; 2.5‰ dans l'idéal ce qui serait équivalent à la réalisation de 2 essais ; 1% au minimum (comme HPS).

Approche formelle du test d'hypothèse

Le test statistique cherche à départager deux hypothèses, l'une appelée hypothèse nulle (H0) et l'autre hypothèse alternative (H1). Dans un essai thérapeutique, l'hypothèse nulle correspond à l'absence d'effet du traitement étudié. L'hypothèse alternative est l'hypothèse que l'on cherche à « prouver » : l'effet du traitement n'est pas nul.
Ainsi dans un essai, on recherche l'effet d'un traitement en comparant deux proportions de survenue d'événements P_1 et P_0 :

H0 : $P_1 = P_0$
H1 : $P_1 \neq P_0$

Il existe deux risques d'erreur attachés au choix de H1 ou de H0. Il est ainsi possible d'accepter H1 alors que H0 est vraie (résultat faux positif) ou d'accepter H0 alors que H1 est vraie (résultat faux négatif).

α = Pr(accepter H1 si H0 est vraie) *faux positif*
β = Pr(accepter H0 si H1 est vraie) *faux négatif*

Alors que l'hypothèse nulle est unique, l'hypothèse alternative correspond à une infinité de situations $P_1-P_0 = \Delta$ où Δ peut prendre n'importe quelle valeur. Le risque β ne peut donc être déterminé que pour une certaine valeur de Δ, correspondant à une hypothèse H1 particulière.

Le départage des hypothèses se fait à l'aide d'une valeur, noté *p,* déterminée à partir des données observées. La valeur *p* est la probabilité d'observer des résultats au moins aussi en désaccord avec l'hypothèse nulle que ceux qui ont été effectivement notés. Ainsi *p* chiffre le degré de désaccord existant entre l'observation et l'hypothèse nulle.

À partir de la valeur de *p* calculée, le choix final de l'hypothèse se base sur la règle suivante :

Si $p \leq \alpha$, H0 est rejetée et H1 est acceptée.

Si $p > \alpha$, aucune conclusion n'est faite (en particulier H0 n'est pas accepté car il n'est pas possible de contrôler le risque d'erreur bêta).

Interprétation erronée du *p* ou d'un test significatif

Les tests statistiques et le degré de signification *p* font souvent l'objet d'interprétations erronées [2].

Ainsi, on dit fréquemment à l'issue d'un test de comparaison des moyennes statistiquement significatif qu'il y a 95% de chance pour que les moyennes des deux groupes diffèrent. En réalité, une telle affirmation n'a aucun sens puisque les moyennes des populations sont des constantes et non des variables aléatoires. La probabilité *p* n'est pas relative à la différence entre les moyennes considérées mais bien au jugement que l'on émet au sujet de l'égalité de ces moyennes. Tout ce que l'on peut dire, en concluant à l'existence d'une différence avec un test statistiquement significatif, c'est que l'on a 5 chances sur 100 seulement d'aboutir à une telle conclusion par le simple fait du hasard.

En toute rigueur, il n'est pas possible non plus de dire que la valeur de *p* représente la probabilité que les résultats de l'essai soient dus à la chance. En fait, la valeur de *p* est la probabilité d'observer un résultat sous l'hypothèse que seule la chance (les fluctuations aléatoires d'échantillonnage) explique ce résultat. Il ne s'agit pas de la probabilité que la chance donne ce résultat, puisque la chance donne ce résultat avec une probabilité de 1 (par définition du risque α et du *p* on se place dans la situation où en réalité il n'y a pas de différence).

Ce n'est pas non plus la probabilité de l'absence de différence. La valeur de *p* est la probabilité d'observer un résultat en l'absence de différence, ce n'est pas la probabilité qu'il n'y ait pas de différence compte tenu du résultat observé. Il est donc inexact de dire que le degré de signification *p* mesure la probabilité d'absence de différence.

Tableau 1 – Interprétations erronées du p

le *p* n'est pas	**le *p* est**
p n'est pas la probabilité de l'hypothèse nulle	*p* est la probabilité d'obtenir le résultat observé si l'hypothèse nulle est vraie
p n'est pas la probabilité d'absence de différence	*p* est la probabilité d'observer une différence au moins aussi importante si en réalité il n'y a pas de différence
p n'est pas la probabilité que le traitement n'ait pas d'effet	*p* est la probabilité d'obtenir le résultat qui a été observé si le traitement est en réalité inefficace
$p<0.05\%$ ne signifie pas qu'il y a moins de 5% de chance que le traitement soit sans effet	il y a moins de 5% d'observer le résultat obtenu si le traitement est sans effet
p n'est pas $Pr(H_0)$ ou $1-Pr(H_1)$	$p = Pr(\text{résultat}/H_0)$
p n'est pas la probabilité de l'hypothèse nulle	*p* est la probabilité conditionnelle du résultat sous l'hypothèse nulle

Alpha et bêta vus comme des taux de filtration

Le test statistique peut être vu comme un filtre que l'on utilise pour extraire de l'ensemble des résultats produits par les essais cliniques ceux que l'on retiendra comme argument de l'efficacité des traitements évalués.

Ce filtre laisse passer α% des résultats produits avec un traitement sans effet (ce qui peut être vu comme un taux de filtration de α% des faux positifs) et 1-β% des résultats produits avec un traitement efficace (soit un taux de filtration de 1-β% des vrais positifs).
Ainsi un risque alpha de 5% signifie que 5% des essais réalisés avec un traitement sans effet sera finalement retenu comme argument de l'efficacité du traitement testé. Une puissance de 80% signifie que 80% des essais réalisés avec un traitement ayant l'efficacité attendue sera retenu comme preuve de l'efficacité du traitement.
Le nombre de faux positifs retenus à l'issu de cette procédure dépend donc du taux de filtration alpha mais aussi de la quantité de résultats issus de traitement sans effet que l'on a soumis à la filtration. A taux de filtration constant, plus la quantité de résultats issus de traitement sans effet est importante, plus il y aura de faux positifs de l'autre coté du filtre. Le même phénomène se produit pour les vrais positifs.

Si dans l'ensemble de résultats que l'on passe par le filtre du test statistique, il y a p% de résultats issus d'un traitement ayant l'efficacité attendue et 1-p% de résultats obtenus avec un traitement sans effet, a l'issu de la filtration nous aurons $\alpha \times (1-p)\%$ de faux positifs et $p \times (1-\beta)\%$ de vrais positifs.
En termes de probabilité, après avoir obtenu un résultat qui passe le filtre (c'est à dire statistiquement significatif), la probabilité que le traitement ai l'efficacité attendue est égale à $\frac{(1-\beta)p}{(1-\beta)p+\alpha(1-p)}$, p étant dans ce cas la probabilité a priori que le traitement soit efficace.
Ce raisonnement est identique à celui que l'on peut faire en faisant le parallèle entre tests statistiques et tests diagnostiques (cf. infra).

Bibliographie

1. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002;**360**(9326)**:**7-22.
2. Sterne JAC, Davey Smith G. Sifting the evidence—what's wrong with significance tests? *BMJ* 2001;**322:**226-31.

Ressources WEB

Study design and choosing a statistical test (http://bmj.bmjjournals.com/statsbk/13.shtml)

Elementary Concepts in **Statistics** (http://www.statsoft.com/textbook/esc.html)

Sampling distribution (http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html)

La problématique des comparaisons multiples

Powerpoint

Introduction

La principale implication du risque alpha dans l'essai thérapeutique est de garantir une relative solidité à la conclusion sur l'effet du traitement en écartant raisonnablement le risque d'une conclusion erronée du fait d'une erreur de 1er espèce. Le test permet de limiter le risque alpha à un niveau choisi (en général 5%)

Lorsque plusieurs tests statistiques sont réalisés simultanément pour chercher à répondre à la question de l'efficacité du traitement, le risque global d'erreur de première espèce s'accroît. La répétition à chaque test du risque d'obtenir un résultat significatif par hasard augmente le risque global de conclure à tort à l'efficacité du traitement. C'est par exemple le cas si l'on a la possibilité de conclure que le traitement est efficace a l'issu d'un premier test portant sur un premier critère de jugement mais aussi lors d'un deuxième sur un autre critère de jugement ou bien lors d'un troisième, etc. In fine, le risque alpha global de conclure à tort à l'efficacité à l'issu de cet essai n'est plus de 5% (même si c'est le seuil retenu pour chaque test) mais il est bien supérieure (Tableau 1).

Tableau 1 – Augmentation du risque de conclure à tort à l'efficacité du traitement (risque de première espèce) en fonction du nombre de tests réalisés.

Nombre de tests (seuil α=5%)	Risque global d'erreur
1	5%
2	10%
10	40%
50	92%
k	$1-0.95^k$

Il y a inflation du risque alpha. Aux dés, la probabilité d'obtenir un six est plus importante avec le lancer de 3 dés qu'avec un seul de (figure 1). Avec un risque alpha à 5%, la réalisation de 100 tests indépendants produits en moyenne 5 résultats statistiquement significatifs.

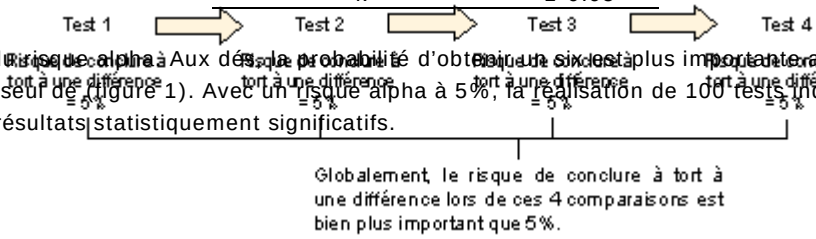


Figure 1 – L'inflation du risque alpha en cas de répétition des tests statistiques

Mathématiquement on montre qu'à l'issu de *k* tests réalisés avec un seuil de signification α, le risque global d'erreur $\alpha_{global} = 1 - (1 - \alpha)^k$.

Le problème de l'inflation du risque alpha survient lorsque l'on s'autorise à conclure à partir du moment où au moins un test est significatif. Par contre, si on exige que les *k* tests soient significatifs simultanément pour conclure, il n'y a pas inflation du risque alpha, mais au contraire une diminution du risque alpha global. En effet avec 2 tests significatifs à 5%, le risque alpha global descend à 5% * 5% = 2,5 ‰.

Conséquence en lecture critique

En lecture il convient d'être particulièrement attentif au problème de l'inflation du risque alpha. En effet, une situation de multiplicité des comparaisons enlève presque toute valeur à un résultat statistiquement significatif

[1, 2] puisqu'il est possible, en répétant les tests, d'obtenir un $p<0.05$ avec n'importe quel traitement, même sans effet.
Un résultat significatif obtenu dans un contexte où il est impossible de savoir le nombre de tests réalisés au total n'apporte aucune preuve statistique : « les données ont été torturées jusqu'à ce qu'elles avouent ! ». Les anglo-saxons parlent de « data dredging ».

Les situations de comparaisons multiples dans l'essai thérapeutique

Cette problématique des comparaisons multiples est présente dans l'essai thérapeutique à plusieurs niveaux :

- la multiplicité des critères de jugement
- les analyses en sous-groupes
- les analyses intermédiaires
- les doses multiples (ou d'une manière générale les essais avec plus de 2 bras)
- et dans une certaine mesure au niveau de la comparaison des caractéristiques de base

Contrôle de l'inflation du risque alpha

Comme cela est abordé dans les chapitres suivants plusieurs méthodes ont été proposées pour éviter l'inflation du risque alpha lorsque plusieurs comparaisons statistiques sont nécessaires.

La méthode la plus simple est celle de Bonferoni [3] appelée aussi méthode de Bonferroni-Holm. Elle consiste à réaliser les tests avec un seuil de signification plus petit que 5% et de choisir cette valeur de telle sorte qu'après inflation due aux comparaisons multiples, le niveau global atteint soit de 5%.
Les tests sont donc réalisés avec un seuil de $0.05/k$ où k désigne le nombre de comparaisons effectuées.

Justification de la méthode de Bonferroni

Après *k* tests, le risque alpha global est $1-(1-\alpha)^k$.
Quand α est petit, $(1-\alpha)^k \approx 1-\alpha k$
donc $1-(1-\alpha)^k \approx \alpha k$.
En prenant comme seuil de chaque test $\alpha' = \alpha / k$, le risque global est maintenu approximativement à α étant donné que l'inflation liée à *k* tests revient approximativement à multiplier par *k* le risque consenti au niveau de chaque test. La consultation du Tableau 1 montre que cette approximation est assez grossière mais fonctionne parfaitement bien par exemple pour *k=2*. La méthode de Bonferroni est donc à réserver aux situations où le nombre de tests réalisés est petit.

D'autres méthode de contrôle de l'inflation du risque existent comme la méthode de Dunn-Sidak qui utilise comme seuil de signification pour les tests $\alpha' = 1-(1-\alpha)^{1/k}$.
Le Table 1 compare les valeurs obtenues par la méthode de Dunn-Sidak avec celles données par la méthode de Bonferroni. Ces 2 méthodes donnent des valeurs très proches.

Table 1 – Comparaison des valeurs obtenues par la méthode de Dunn-Sidak avec celles données par la méthode de Bonferonni pour un risque alpha global de 5%.

k	Dunn-Sidak	Bonferroni
1	5.00%	5.00%
2	2.53%	2.50%
3	1.70%	1.67%
4	1.27%	1.25%
5	1.02%	1.00%
10	0.51%	0.50%
20	0.26%	0.25%

Une autre façon de générer la multiplicité des comparaisons au niveau d'un ensemble de critère de jugement est de suivre une procédure de tests hiérarchisés (« *hierarchical test procedure* ») [4]. Cette *test procedure* », « *closed* hiérarchiser, a priori, dans le procédure consiste à comparaisons multiples qui protocole de l'étude, les groupes ou des critères de seront réalisées (sous

jugement par exemple). Ensuite il est possible de conclure pour toutes les comparaisons pour lesquels la signification statistique est obtenue jusqu'à la première non significative (en descendant dans l'ordre préétabli par la hiérarchie). Cette procédure permet ainsi de conclure simultanément sur plusieurs comparaisons.

Bibliographie

1. Lord SJ, Gebski VJ, Keech AC. Multiple analyses in clinical trials: sound science or data dredging? Med J Aust 2004;181(8):452-4. *PMID: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15487966 15487966*.
2. Huque MF, Sankoh AJ. A reviewer's perspective on multiple endpoint issues in clinical trials. J Biopharm Stat 1997;7(4):545-64. *PMID: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9358328 9358328*.
3. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. Bmj 1995;310(6973):170. *PMID: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7833759 7833759*.
4. James J. Chen S-JW. Testing for Treatment Effects on Subsets of Endpoints. Biometrical Journal 2002;44(5):541-557. *PMID: http://dx.doi.org/10.1002/1521-4036(200207)44:5<541:AID-BIMJ541>3.0.CO;2-0*

Critère de jugement principal et critères de jugements secondaires

Généralités

Plusieurs critères de jugement sont souvent envisageables pour mettre en évidence l'effet d'un traitement d'un point de vue clinique. Un effet obtenu au niveau de n'importe lequel de ces critères justifierait l'utilisation de ce traitement en pratique. Cependant si aucune précaution n'est prise, cette multiplicité des critères va conduire à des comparaisons statistiques multiples. En effet, on conclura à l'existence d'un effet du traitement dès lors qu'un des tests statistiques rattachés à ces critères est significatif. Le risque de conclure à tort à l'effet du traitement augmente.

Le moyen le plus simple pour maintenir le risque d'erreur alpha au niveau choisi est de ne faire qu'une seule comparaison. Pour cela, un critère de jugement est privilégié a priori, c'est le ***critère de jugement principal****. La conclusion sur l'efficacité du traitement sera prise uniquement sur cette comparaison. Le risque d'erreur alpha de la conclusion est donc alors parfaitement contrôlé et égal à 5%.*

En l'absence de différence significative sur le critère principal, une différence pourtant significative sur l'un ou plusieurs des critères secondaires ne permet pas de conclure. Le risque d'erreur est alors trop grand (

Figure 1).

Pour la même raison, la définition du critère de jugement doit préciser le moment de sa mesure. La répétition au cours du temps des comparaisons implique une inflation du risque alpha. L'interprétation des critères de jugement secondaires présente des aspects assez similaires à celle des analyses en sous-groupes (1, 2).

	Candesartan (n=1514)	Placebo (n=1509)	Adjusted hazard ratio (95% CI)*	p
Cardiovascular death or hospital admission for CHF	333 (22·0%)	366 (24·3%)	0·86 (0·74–1·00)	0·051
Cardiovascular death	170 (11·2%)	170 (11·3%)	0·95 (0·76–1·18)	0·635
Hospital admission for CHF	241 (15·9%)	276 (18·3%)	0·84 (0·70–1·00)	0·047
Cardiovascular death, hospital admission for CHF, MI	365 (24·1%)	399 (26·4%)	0·87 (0·75–1·00)	0·051
Cardiovascular death, hospital admission for CHF, MI, stroke	388 (25·6%)	429 (28·4%)	0·86 (0·75–0·99)	0·037
Cardiovascular death, hospital admission for CHF, MI, stroke, coronary revascularisation procedure	460 (30·4%)	497 (32·9%)	0·91 (0·80–1·03)	0·130

MI=Myocardial infarction. *Covariate-adjusted model for variables shown in table 1.

Table 2: **Primary and secondary outcomes**

Figure 1 – Exemple de situation où aucune différence significative n'est obtenue au niveau du critère de jugement principal (Cardiovascular death or hospital admission for CHF). Les différences significatives observées au niveau de certains critères de jugement secondaires(Hospital admission for CHD ou Cardiovascular death or hospital admission for CHF, MI, stroke) ne permettent pas d'affirmer l'efficacité du traitement en raison de la multiplication des comparaisons (6 tests au total).

Définition

Le critère de jugement principal (« primary endpoint *» ou «* main endpoint *») est le critère qui va servir à la mise en évidence de l'efficacité du traitement étudié. Il est unique afin de permettre le contrôle du risque de conclure à tort à l'efficacité (erreur statistique alpha). En effet pour éviter les effets pervers de la multiplication des tests statistiques, il convient de ne baser la conclusion de l'essai que sur un et un seul test statistique – celui qui sera fait sur le critère de jugement principal.*

Du fait de son unicité, le critère de jugement doit être soigneusement choisi et doit correspondre au critère le plus cliniquement pertinent vis-à-vis de l'objectif thérapeutique de la maladie. En effet, nous verrons par la suite que seul le critère principal permet de conclure. L'utilisation d'un critère non cliniquement pertinent comme critère principal enlève à l'essai le pouvoir décisionnel et le rend par là sans grand intérêt.

Parfois, d'autres contingences, moins valides scientifiquement, conduisent à considérer d'autres éléments dans le choix du critère de jugement, comme une fréquence de base plus élevée afin de réduire la taille (et le coût) de l'essai ou une plus grande simplicité d'acquisition.

> La réalisation de plusieurs tests statistiques avant de conclure à l'effet du traitement augmente le risque de faire cette conclusion à tort. En effet, cette conclusion sera faite dès qu'un des tests sera significatif. On prend donc un risque d'erreur de 5% au premier test, puis encore 5% au second, etc. À l'issue de tous les tests, le risque d'erreur alpha est bien supérieur à 5%. Avec 5 critères indépendants, la probabilité de trouver au moins une différence significative à tort est de 23%.

À côté du critère de jugement principal, d'autres critères peuvent aussi être analysés. Ils sont dénommés critères de jugement secondaires (« secondary endpoints *»). Ces critères secondaires peuvent être :*

- des critères utilisés pour documenter les bénéfices secondaires du traitement (par exemple, si le critère principal est la mortalité totale, le traitement peut aussi réduire la fréquence des événements non mortels, ou augmenter la qualité de vie),
- des critères complémentaires utilisés pour documenter le mécanisme d'action de l'effet obtenu (par exemple, les causes spécifiques de mortalité afin d'expliquer comment est obtenue une réduction de mortalité totale),

- des critères intermédiaires (par exemple le taux de reperfusion coronaire dans les essais de fibrinolyse à la phase aiguë de l'infarctus du myocarde),
- des critères correspondant à des effets délétères du traitement,
- ou les composantes d'un critère de jugement composite utilisé comme critère principal.

Nous verrons cependant dans la section consacrée à l'interprétation, qu'il n'est pas possible de conclure de façon formelle sur un critère de jugement secondaire (même si son analyse été prévue d'emblée dans le protocole). Les critères secondaires sont présents à titre documentaire.

Tableau 1 – Deux exemples de choix de critères principaux et secondaires.

	Exemple 1	**Exemple 2**
	Éradication de *Helicobacter pylori* dans la dyspepsie	Antiagrégant dans la prévention cardiovasculaire
Critère de jugement principal	Disparition des symptômes de dyspepsie	Décès + infarctus + AVC
Critères de jugement secondaires	• Score de sévérité de la dyspepsie (Glasgow) • Score de qualité de vie • Éradication de *H. pylori* • Recours à un traitement antisecrétoire	• Mortalité totale • Mortalité coronarienne • Mortalité cardiovasculaire • Infarctus mortel et non mortel • Infarctus non mortels • AVC mortel et non mortel • AVC non mortels

Corrélation entre les critères

Les résultats obtenus au niveau des différents critères d'un essai ne sont pas en général indépendants car il existe une corrélation plus ou moins forte entre les critères de jugement. Par exemple, la mortalité cardiovasculaire est incluse dans la mortalité totale qui sera modifiée si la mortalité cardiovasculaire l'est. Les effets observés sur ces deux mortalités sont donc corrélés. Il en est de même par exemple entre les événements mortels et les événements non mortels. Un traitement qui diminue la fréquence des événements mortels le fait souvent par une diminution de la fréquence des événements mortels et non mortels.

Critères de jugement principaux multiples

Il est parfois nécessaire de prendre plusieurs critères de jugement principaux, en général deux. Dans ce cas l'utilisation d'une méthode statistique, comme la méthode de Bonferroni, est indispensable. Le coût à payer est un plus grand nombre de sujets nécessaires (car le calcul s'effectue avec un alpha plus petit qu'avec un seul test).

Les justifications possibles pour l'utilisation de deux critères de jugement sont :

- *la possibilité de démontrer simultanément l'efficacité et la sécurité du traitement, en concluant que le nouveau traitement est supérieur au traitement standard à la fois en efficacité et en sécurité*
- *la recherche d'une porte de secours en aménageant la possibilité de pouvoir conclure sur un critère moins cliniquement pertinent en cas d'absence de résultats sur le critère clinique dur. C'est par exemple le cas d'un essai où le critère le plus intéressant cliniquement est la mortalité mais où on lui adjoint un critère morbidité comme critère de secours. En effet, un bénéfice est en général plus facile à obtenir sur un critère de morbidité que sur la mortalité. De plus, ces événements sont bien plus*

fréquents que les décès, et avec l'effectif nécessaire pour la recherche d'un effet sur la mortalité, l'étude sera surpuissante au niveau du critère de morbidité, maximisant ainsi les chances d'obtenir un résultat significatif.

Exemple

L'essai ValHefT est un essai comparant le valsartan au placebo dans l'insuffisance cardiaque (3). "The primary outcomes were mortality and the combined end point of mortality and morbidity, defined as the incidence of cardiac arrest with resuscitation, hospitalization for heart failure, or receipt of intravenous inotropic or vasodilator therapy for at least four hours"

Une méthode statistique appropropriée a été utilisée pour prendre en compte ce double critère de jugement principal : "Statistical analyses were performed at an overall significance level of 0.05, adjusted for the two primary end points. Each primary end point was tested at a two-sided significance level of 0.02532, on the basis of the Dunn–Sidak inequality: $\alpha' = 1-(1-\alpha)^{1/2}$".

Aucune différence significative n'a été observée au niveau de la mortalité, contrairement à ce qui est observé au niveau du 2ème critère principal, le critère composite.

TABLE 2. INCIDENCE AND RELATIVE RISK OF THE PRIMARY END POINTS.

EVENT	VALSARTAN GROUP (N=2511)	PLACEBO GROUP (N=2499)	RELATIVE RISK (CI)*	P VALUE†
	no. with event (%)			
Death from any cause (during entire trial)	495 (19.7)	484 (19.4)	1.02 (0.88–1.18)	0.80
Combined end point	723 (28.8)	801 (32.1)	0.87 (0.77–0.97)	0.009
Death from any cause (as first event)	356 (14.2)	315 (12.6)		
Hospitalization for heart failure	346 (13.8)	455 (18.2)		
Cardiac arrest with resuscitation	16 (0.6)	26 (1.0)		
Intravenous therapy	5 (0.2)	5 (0.2)		

*The 98 percent confidence interval (CI) was calculated for the mortality end point (death from any cause), and the 97.5 percent confidence interval was calculated for the combined mortality–morbidity end point.

†P values were calculated by the log-rank test from time to first event.

L'existence de ce critère de secours permet finalement de conclure que «Valsartan significantly reduces the combined end point of mortality and morbidity ».

Puissance au niveau des critères de jugements secondaires

La taille d'un essai est déterminée pour garantir la puissance statistique de la comparaison au niveau du critère de jugement principal. Par contre, la puissance de l'essai sur les critères de jugements secondaires n'est pas contrôlée et peut être faible dans certains cas : lorsque la fréquence de base du critère secondaire est inférieure à celle du critère principal ou quand l'effet attendu du traitement est plus petit pour le critère secondaire que pour le critère principal. Ces situations sont courantes. Les événements fréquents sont choisis préférentiellement comme critère principal car ils permettent de réduire la taille de l'essai. Par exemple, un critère composite de morbi-mortalité est plus facilement choisi que la mortalité totale car sa fréquence de base est plus élevée. Ainsi les événements moins fréquents (mais peut être plus cliniquement pertinents) sont envisagés comme critères secondaires mais à leur niveau la puissance et la précision de l'estimation de l'effet seront faibles. Ce point permet de re-insister sur le fait que le critère de jugement principal doit être le critère le plus pertinent et qu'un essai doit être centré sur l'intérêt des patients (choix du critère le plus pertinent cliniquement) et non pas sur celui du traitement (choix du critère permettant d'obtenir le plus facilement un résultat significatif pour le traitement).

Le manque de puissance potentiel sur les critères de jugements secondaires est à prendre en compte dans l'interprétation des résultats.

Ainsi, il n'est pas paradoxal de ne pas mettre en évidence l'effet d'un traitement sur un critère secondaire moins fréquent que le critère principal, alors qu'un effet statistiquement significatif a été observé au niveau de ce dernier. Cette discordance provient simplement du fait que la recherche de l'effet au niveau du critère secondaire manque de puissance statistique et non pas d'un manque d'efficacité du traitement sur ce critère.

De même, les intervalles de confiance obtenus au niveau des critères de jugement secondaires peuvent être plus larges que ceux du critère principal pour la même raison.

Exemple

Dans un essai, 4000 patients ont été nécessaires pour assurer une puissance de 90% à la recherche de l'effet (RR=0,8) sur le critère de jugement principal, dont la fréquence de base était de 20%. Le tableau suivant donne la puissance de la recherche du même effet (RR=0,8) sur des critères secondaires moins fréquents.

Critère	Risque relatif	Fréquence de base	Puissance
Critère principal	0,8	20%	90%
Critère secondaire 1	0,8	16,5%	83%
Critère secondaire 2	0,8	10%	60%
Critère secondaire 3	0,8	7,1%	45%
Critère secondaire 4	0,8	3,2%	27%

Prise en compte de la multiplicité des critères de jugements secondaires

D'une manière générale, aucune démonstration n'est à attendre au niveau des critères secondaires. Celle-ci est impossible en toute rigueur en raison de l'absence de contrôle strict du risque d'erreur de première espèce et d'hypothèse formulée a priori.

Il est cependant possible de chercher à conclure au niveau des critères secondaires en prenant en compte la multiplicité des tests statistiques résultant de l'analyse des critères de jugements secondaires par une méthode d'ajustement du seuil de signification statistique (4, 5).

Utilisation de la méthode de Bonferroni

La méthode de Bonferroni peut être utilisée pour contrôler l'inflation du risque alpha au niveau des critères de jugement secondaires. Un seuil ajusté de $\alpha/(k+1)$ *est utilisé pour chaque critère secondaire (k comparaisons pour chaque critère secondaire + une comparaison pour le critère principal), ce qui maintient un risque global d'erreur égal à* α *au niveau de l'ensemble des critères secondaires. L'analyse du critère principal est effectuée avec un seuil de* α*. Cette approche revient à corriger les valeurs de p obtenues au niveau des critères secondaires en les multipliant par* (k+1)*. Si un critère secondaire est jugé comme statistiquement significatif après l'application de ces règles, il est alors possible de considérer que l'essai démontre l'efficacité du traitement sur ce critère (1).*

Exemple

L'essai SOLVD prevention (6) avait pour objectif d'évaluer si un inhibiteur de l'enzyme de conversion (l'enalapril) pouvait réduire la mortalité des patients porteurs d'une insuffisances cardiaque ventriculaire gauche (IVG) asymptomatique. Six critères secondaires étaient prévus par le protocole : mortalité cardiovasculaire, mort subite, infarctus, AVC, hospitalisation pour insuffisance cardiaque et la qualité de vie. L'essai a inclus 4228 patients qui ont été suivis en moyenne 37 mois.

Le résultat non significatif (p=0,30) obtenu sur la mortalité globale a débouché sur la conclusion qu'il n'y avait pas de preuve que l'enalapril puisse augmenter la survie des patients avec IVG asymptomatique.

Par contre, au niveau des critères secondaires, une réduction significative des hospitalisations pour insuffisance cardiaque (risque relatif de 0,64, p<0,001) ainsi que des infarctus (RR=0,76, p<0,01) était observée.

L'application de la règle d'ajustement requiert que les critères secondaires soient significatifs au seuil de 0,05/7=0,0071. Ainsi il est possible de conclure pour les hospitalisations, mais pas pour l'infarctus.

Tests hiérarchisés

Une autre façon de générer la multiplicité des comparaisons au niveau d'un ensemble de critère de jugement est de suivre une procédure de tests hiérarchisés (« hierarchical test procedure *», «* closed test procedure *») (7). Cette procédure consiste à hiérarchiser, a priori, dans le protocole de l'étude les critères de jugement. Ensuite il est possible de conclure pour tous les critères pour lesquels la signification statistique est obtenue jusqu'au premier non significatif (en descendant dans l'ordre préétabli par la hiérarchie).*

Cette procédure permet ainsi de conclure simultanément sur plusieurs critères et autorise ainsi, en toute rigueur, des critères de jugements primaires multiples.

Bibliographie

1. *Davis CE. Secondary endpoints can be validly analyzed, even if the primary endpoint does not provide clear statistical significance. Controlled Clinical Trials 1997;18:557-560.*

2. *O'Neill RT. Secondary endpoints cannot be valid analyzed if the primary endpoint does not demonstrate clear statistical significance. Controlled Clinical Trials 1997;18:550-556.*

3. *Cohn JN, Tognoni G. A randomized trial of the angiotensin-receptor blocker valsartan in chronic heart failure. N Engl J Med 2001;345(23):1667-75.*

4. *Moye LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium. Stat Med 2000;19(6):767-79.*

5. *Moye LA. Response to commentaries on 'Alpha calculus in clinical trials: considerations for the new millennium'. Stat Med 2000;19(6):795-9.*

6. *The SOLVD investigators. Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduces left ventricular ejection fraction. NEJM 1992;327:685-691.*

7. *James J. Chen S-JW. Testing for Treatment Effects on Subsets of Endpoints. Biometrical Journal 2002;44(5):541-557.*

Analyse en sous groupe

Introduction

Des analyses en sous-groupes (« by sub-groups analysis ») sont fréquemment réalisées en complément de l'analyse principale d'un essai thérapeutique. Bien que ces analyses présentent un réel intérêt dans la recherche de facteurs modifiant l'effet du traitement, elles ne permettent pas de conclure [1, 2]. En effet, leurs résultats sont de nature exploratoire et sont exposés aux risques des comparaisons statistiques multiples. Ainsi, les analyses en sous-groupes ne génèrent que des nouvelles hypothèses qui devront être confirmées par de nouveaux essais.

Définition et but

L'analyse en sous groupes consiste à rechercher l'effet du traitement dans une sous-population des patients d'un essai.

Une analyse en sous-groupes consiste à subdiviser la population d'un essai thérapeutique en deux ou plusieurs sous-groupes et à étudier l'efficacité du traitement dans chacun de ces sous-groupes. Le but est de rechercher une interaction entre l'effet du traitement et une ou plusieurs variables (cf. infra). Par exemple, une analyse en sous-groupes suivant le sexe revient à mesurer séparément l'effet du traitement chez les hommes et chez les femmes. Une analyse suivant l'âge entraîne souvent la division de la population en plusieurs sous-groupes et en autant d'estimations de l'effet du traitement.

En général, plusieurs analyses en sous-groupes sont réalisables en fonction de différentes variables (Tableau 1 et

Figure 1) : le sexe, l'âge, le poids, le stade de la maladie, les antécédents, les traitements concomitants, etc.

L'objectif recherché par ces analyses en sous-groupes est différent suivant le résultat de l'essai : résultat concluant (basé sur une différence statistiquement significative) ou résultat non concluant (absence de différence statistiquement significative).

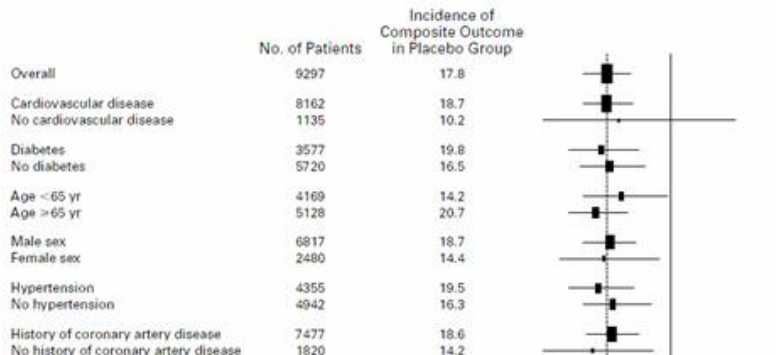


Figure 1 – Exemple de représentation graphique des résultats d'analyses en sous groupe.

Analyse en sous-groupes dans un essai non concluant

Dans un essai qui n'a pas montré de différence statistiquement significative, le but des analyses en sous-groupes est de rechercher le ou les sous-groupes dans lesquels existerait un effet du traitement statistiquement significatif. L'idée est de dire que l'effet du traitement n'existe pas chez tous les types de patients mais seulement chez certains d'entre eux. Le mélange de patients qui bénéficient du traitement avec d'autres n'en bénéficiant pas, conduit, au niveau de l'essai, à une dilution de l'effet et à l'absence de différence significative (Tableau 1). L'analyse en sous-groupes pourrait permettre de trouver les patients chez lesquels le traitement « marche ».

Tableau 1 – Dans cet essai non concluant les analyses en sous-groupes semblent suggérer que bien qu'il ne soit pas possible de mettre en évidence un effet tout patient confondu, le traitement serait efficace chez les sujets diabétiques. En fait cette observation ne peut pas être considérée comme une preuve mais seulement comme une nouvelle hypothèse à tester dans un nouvel essai.

Sous-groupe	**Effet du traitement (risque relatif)**	**p**
Essai en entier	0,92	NS
Age<75	0,92	NS
Age>75	0,95	NS
Hommes	0,92	NS
Femmes	0,99	NS
Antécédents d'infarctus	0,87	NS
Pas d'antécédents d'infarctus	1,03	NS
Diabétiques	0,78	p<0,01
Non diabétiques	1,09	NS

Le résultat d'une analyse en sous-groupes est de nature exploratoire.

En fait, les analyses en sous-groupes se heurtent à plusieurs difficultés méthodologiques qui ne permettent pas de déboucher sur des conclusions sûres : répétitions des tests statistiques, démarche exploratoire, destruction de la randomisation.

Les analyses en sous-groupes font courir le risque d'une inflation non contrôlée du risque d'erreur statistique a.

Rappel statistique

La réalisation de plusieurs tests statistiques avant de conclure à l'effet du traitement augmente le risque de faire cette conclusion à tort. En effet, cette conclusion sera faite dès qu'un des tests sera significatif. On prend donc un risque d'erreur de 5% au premier test, puis encore 5% au second, etc. À l'issue de tous les tests, le risque d'erreur alpha est bien supérieur à 5%. Avec 5 critères indépendants, la probabilité de trouver au moins une différence significative à tort est de 23% (cf. chapitre *Les risques d'erreur statistiques*).

La multiplication des tests statistiques (un par sous-groupe) augmente la probabilité d'obtenir un test significatif uniquement par hasard. Un résultat de sous-groupe significatif est toujours suspect car il est impossible de savoir si ce test révèle l'effet réel ou s'il s'agit simplement d'un artefact lié à la répétition des tests. Un résultat significatif obtenu dans ces conditions fait courir un risque d'erreur dans la conclusion bien supérieur aux 5% habituellement consentis.

De plus, rien ne permet d'être sûr que les patients recevant le traitement ou le contrôle sont comparables à l'intérieur des sous groupes.

En fait, dans cette situation, les analyses en sous-groupes ne sont pas totalement dénuées d'intérêt. Elles génèrent de nouvelles hypothèses, qui seront vérifiées dans de nouveaux essais de confirmation. Cependant il est fréquent que ces hypothèses ne se vérifient pas quand elles sont testées de façon indépendante dans une étude ad-hoc [3].

Exemples

Exemple 1 *– L'essai Suvimax [4] a évalué l'impact d'un apport supplémentaire en vitamines et minéraux antioxydants (bêta-carotène, vitamines E et C, zinc et sélénium), à doses nutritionnelles, dans la prévention des cancers et des maladies cardiovasculaires. Le résultat mis en avant est une réduction de l'incidence des cancers uniquement chez les hommes. Ce résultat est issu d'une analyse en sous groupe. En effet, le protocole [5] ne prévoit aucune stratification de l'essai sur le sexe (le nombre de sujets est calculé pour l'ensemble de l'essai). Aucun sous groupe n'est pré spécifié. Ce sous groupe est donc issu d'une analyse post hoc. De plus, aucun critère de jugement principal unique n'a été défini. Ainsi ce résultat a été obtenu dans un contexte de forte inflation du risque alpha (en ne considérant pas l'analyse post hoc, le résultat mis en avant est issu de 9 tests statistiques).*

Exemple 2 *-Greffe neurone Parkinson*

Un essai randomisé en double aveugle contre placebo a évalué la greffe de cellules embryonnaires dans la maladie de Parkinson sévère [6].

" The primary outcome was a subjective global rating of the change in the severity of disease, scored on a scale of -3.0 to 3.0 at one year, with negative scores indicating a worsening of symptoms and positive scores an improvement."

Au niveau des résultats "The mean (+/-SD) scores on the global rating scale for improvement or deterioration at one year were 0.0+/-2.1 in the transplantation group and -0.4+/-1.7 in the sham-surgery group (P=0.62). Among younger patients (60 years old or younger), standardized tests of Parkinson's disease revealed significant improvement in the transplantation group as compared with the sham-surgery group(P=0.01 for scores on the Unified Parkinson's Disease Rating Scale; P=0.006 for the Schwab and England score). There was no significant improvement in older patients in the transplantation group."

Et c'est le résultat de cette analyse en sous groupe qui est mise en avant dans la conclusion générale de l'étude : " Human embryonic dopamine-neuron transplants survive in patients with severe Parkinson's disease and result in some clinical benefit in younger but not in older patients". *Cette conclusion est donc abusivement forte pour un résultat reposant uniquement sur une analyse en sous groupe.*

Analyse en sous-groupes dans un essai concluant

Dans un essai concluant (où une différence statistiquement significative a été obtenue) le but des analyses en sous-groupes serait de rechercher ceux dans lesquels le traitement serait le plus efficace et surtout ceux dans lesquels il serait inefficace. L'objectif est de mieux définir la population cible en restreignant éventuellement la diffusion du traitement par rapport à la population qui a été incluse dans l'essai (Tableau 2).

Comme dans le cas d'un essai non significatif, les analyses en sous-groupes se heurtent à des difficultés méthodologiques qui les empêchent d'aboutir aux conclusions qu'elles recherchent.

Il est impossible de conclure qu'un traitement est sans efficacité chez certains patients sous prétexte qu'aucune différence significative n'existe dans ce sous-groupe. L'absence de différence significative ne signifie pas qu'il y a absence d'effet car la puissance de la comparaison au niveau d'un sous-groupe n'est pas assurée. En effet, la taille des sous-groupes est inférieure à la taille nécessaire pour mettre en évidence un effet qui est la taille de l'essai tout entier. La probabilité de ne pas conclure à une différence qui existe pourtant est forte (risque d'erreur statistique de deuxième espèce, erreur bêta).

Par exemple, dans l'essai ISIS-2, l'aspirine administrée à la phase aiguë de l'infarctus du myocarde produit une réduction significative très importante de la mortalité à 1 mois. Mais, lors de l'analyse en sous-groupes en fonction des signes astrologiques, l'aspirine apparaît inefficace pour les sujets du signe de la balance ou des gémeaux et plus efficace que la moyenne pour le signe du capricorne [7]. Dans les paradigmes scientifiques actuels, aucune théorie ne permet de penser que ces différences sont réelles !

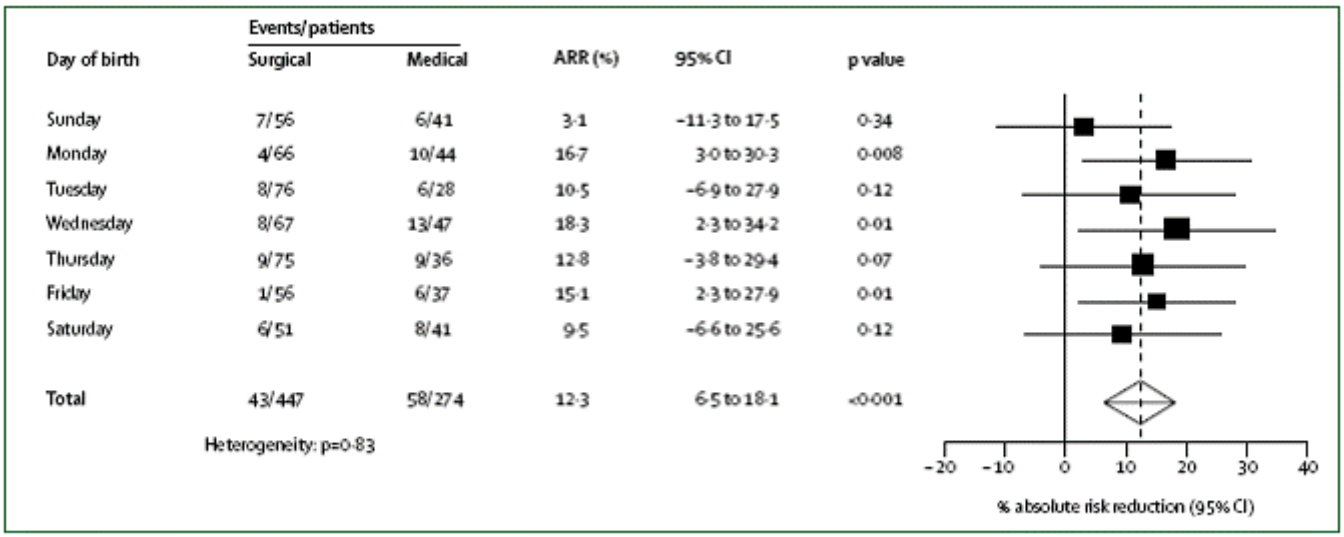
Tableau 2 – A partir de ces analyses en sous groupes, il semblerait que les sujets âgés ainsi que les sujets aux antécédents d'infarctus ne bénéficient pas du traitement et que les diabétiques bénéficient bien plus que les non diabétiques.

Sous groupe	**Effet du traitement (risque relatif)**	**p**
Essai en entier	0,78	p<0,05
Age<75	0,65	p<0,01
Age>75	0,90	NS
Hommes	0,76	p<0,05
Femmes	0,78	p<0,05
Antécédent d'infarctus	0,97	NS
Pas d'antécédent d'infarctus	0,70	p<0,01
Diabétique	0,50	p<0,001
Non diabétique	0,91	p<0,05

De plus, pour conclure qu'il ne convient pas d'utiliser le traitement chez les patients de certains sous-groupes, il faut pouvoir démontrer que l'efficacité dans ce sous-groupe est insuffisante. Une telle démonstration appartient au domaine de la recherche de l'équivalence (ou de la non-infériorité). Les analyses en sous-groupes sont réalisées en dehors de ce contexte, qui exige que l'objectif de l'essai soit la mise en évidence d'une efficacité insuffisante dans ces sous groupes, et la définition a priori de ce qu'est une efficacité insuffisante (définition d'une borne d'efficacité minimale intéressante ou seuil d'équivalence).

En dehors de ces conditions, le résultat d'une analyse en sous-groupes n'est pas opposable au résultat de l'essai tout entier. Il ne peut pas être objecté à un médecin décidant de traiter un patient en se basant sur l'estimation « tous patients confondus » que l'analyse en sous-groupes suggère une efficacité insuffisante pour

ce patient. Le niveau de la preuve du résultat du sous-groupe est inférieur à celui du résultat « tous patients confondus ». Lorsqu'un traitement a montré son efficacité dans un essai, le résultat d'un sous-groupe en dehors des conditions énoncées précédemment ne justifie pas l'abstention pour ces patients.



Day of birth	Events/patients: Surgical	Medical	ARR (%)	95% CI	p value
Sunday	7/56	6/41	3·1	−11·3 to 17·5	0·34
Monday	4/66	10/44	16·7	3·0 to 30·3	0·008
Tuesday	8/76	6/28	10·5	−6·9 to 27·9	0·12
Wednesday	8/67	13/47	18·3	2·3 to 34·2	0·01
Thursday	9/75	9/36	12·8	−3·8 to 29·4	0·07
Friday	1/56	6/37	15·1	2·3 to 27·9	0·01
Saturday	6/51	8/41	9·5	−6·6 to 25·6	0·12
Total	43/447	58/274	12·3	6·5 to 18·1	<0·001

Heterogeneity: p=0·83

Figure 2: Effect of carotid endarterectomy in patients with ≥70% symptomatic stenosis in ECST[126] according to day of week on which patients were born

Test d'interaction

On parle d'interaction quand une covariable influence la taille de l'effet du traitement

Il y a interaction quand l'effet d'un traitement varie entre les sous-groupes. Quand le traitement est bénéfique ou délétère dans tous les sous-groupes mais avec une variation de la taille de l'effet, l'interaction est dite quantitative. Par exemple : le traitement A entraîne une réduction relative de risque de 10 % chez les sujets de moins de 65 ans et de 20% chez ceux de plus de 65 ans. On parle alors d'interaction entre l'âge et l'effet traitement.

Une interaction qualitative est une interaction où le traitement est bénéfique dans un sous-groupe et délétère dans un autre. Par exemple : le traitement B augmente de 1 an la survie chez les femmes mais la réduit de 6 mois chez les hommes.

Seul un test d'interaction significatif permet de conclure que l'effet du traitement est différent entre des sous-groupes : la différence existant entre les sous-groupes est trop large pour pouvoir être expliquer raisonnablement par le seul fait du hasard.

Il existe des tests statistiques recherchant les interactions quantitatives [8, 9]. Ces tests recherchent si les différences observées entre les tailles des effets dans les différents sous-groupes peuvent être expliquées par le seul hasard ou non. Lorsqu'ils sont significatifs il n'est plus raisonnable de conclure que les différences observées sont dues au seul fait du hasard. Il est alors possible de conclure que l'effet du traitement varie effectivement entre les sous-groupes. Par contre, la constatation d'un résultat non significatif dans un sous-groupe et d'un résultat significatif dans l'autre ne permet pas de conclure que l'effet du traitement varie. Il se peut que les intervalles de confiances se chevauchent largement.

Le raisonnement est le même que celui qui est mis en œuvre pour juger de la discordance de deux résultats (cf. section : validité externe *du chapitre :* Lecture critique*). Le test d'interaction est similaire au test d'hétérogénéité en méta-analyse.*

La Figure 2 présente la représentation graphique, maintenant standard, des analyses en sous groupes. Pour chaque modalité de sous groupe, l'effet traitement et son intervalle de confiance est représenté. En regard, la

valeur de p du test d'interaction est indiqué. Le test d'interaction est un test d'hétérogénéité ou un test de tendance. Le test de tendance est un test d'hétérogénéité qui recherche, non seulement s'il y a une différence d'effet traitement entre les modalités, mais aussi si cette variabilité suit une tendance linéaire en fonction de la valeur de la modalité.

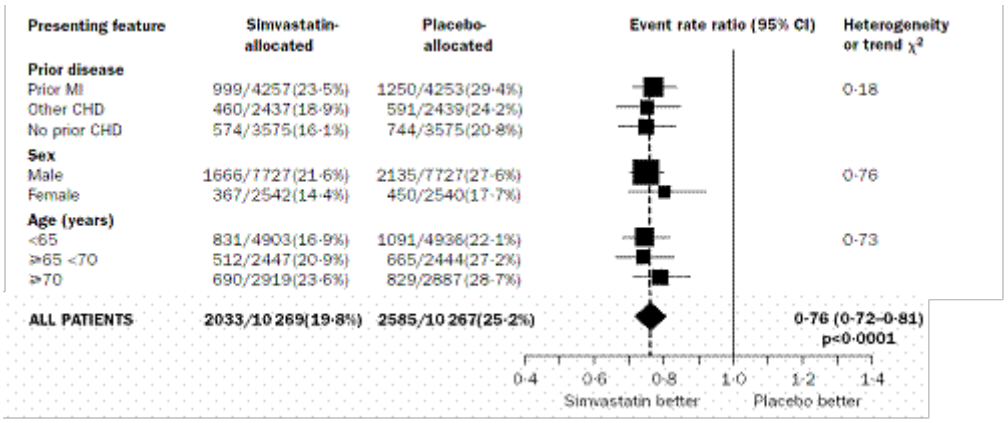


Figure 2 – Représentation graphique complète des analyses en sous groupes présentant les tests d'interaction (hétérogénéité entre les modalités des sous groupes) (Extrait de l'essai HPS [10]).

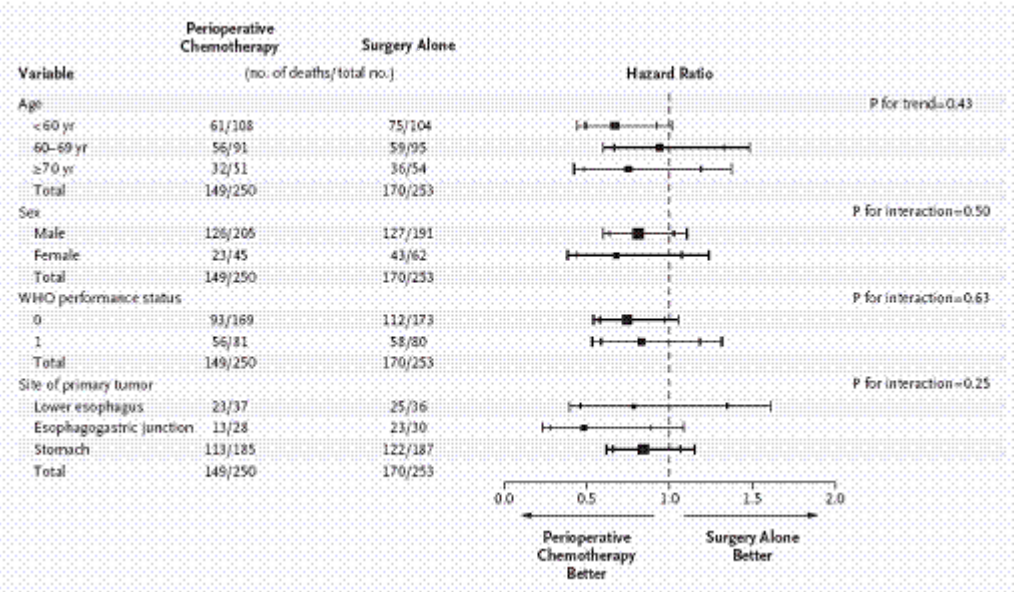


Figure 3 – Présentation complète des analyses en sous groupe. Les hazard ratio sont représentés entourés de leur intervalle de confiance à 95% (marques internes) et à 99% (marques externes). L'intervalle de confiance à 99% réalise en quelque sorte un ajustement statistique contre l'inflation du risque alpha (pour 5 comparaisons simultanée suivant la correction de Bonferroni)

Les recherches d'interaction sont elles aussi soumises au problème d'inflation du risque alpha. La réponse à la question « y-a-t'il une situation clinique où l'effet du traitement varie » conduit à une analyse exploratoire et à la réalisation de nombreux tests favorisant l'identification d'une telle situation a tort.

Dans un exemple papier pédagogique du Lancet, Peter M Rothwell liste au moins 11 domaines dans lesquels les résultats d'une analyse en sous groupe n'a pas été confirmé par un nouvel essai [3].

Observation	Refutation
Aspirin is ineffective in secondary prevention of stroke in women[29,30]	31
Antihypertensive treatment for primary prevention is ineffective in women[32,33]	34
Antihypertensive treatment is ineffective or harmful in elderly people[35]	36
Angiotensin-converting enzyme inhibitors do not reduce mortality and hospital admission in patients with heart failure who are also taking aspirin[37]	38
β blockers are ineffective after acute myocardial infarction in elderly people,[39] and in patients with inferior myocardial infarction[41]	40
Thrombolysis is ineffective >6 hours after acute myocardial infarction[42]	43
Thrombolysis for acute myocardial infarction is ineffective or harmful in patients with a previous myocardial infarction[42]	44
Tamoxifen citrate is ineffective in women with breast cancer aged <50 years[45]	46
Benefit from carotid endarterectomy for symptomatic stenosis is reduced in patients taking only low-dose aspirin due to an increased operative risk[47]	48
Amlodipine reduces mortality in patients with chronic heart failure due to non-ischaemic cardiomyopathy but not in patients with ischaemic cardiomyopathy[49]	50

Table 1: **Examples of subgroup analyses that have shown apparently clinically important heterogeneity of treatment effect which has subsequently been shown to be false**

Approche exploratoire

Les analyses en sous-groupes décidées après le recueil des données sont des analyses "post-hoc" dont la valeur est purement exploratoire : l'expérience servant au recueil des données a eu un autre objectif que celui de l'analyse en sous groupe ; aucune hypothèse préalable n'a été formulée; aucun calcul de nombre de sujets nécessaires n'a été réalisé pour garantir une puissance suffisante. Le problème méthodologique provient du fait que les analyses « post-hoc » génèrent à la fois l'hypothèse et la vérification de l'hypothèse. Cette situation tautologique ne peut pas être utilisée comme preuve mais seulement pour générer de nouvelles hypothèses. Les analyses exploratoires exposent ainsi au risque d'ériger un artefact en loi. Il convient ainsi de faire la distinction entre les hypothèses formulées a priori (« prior hypothesis ») et les hypothèses dérivées des données (« data-derived hypothesis »). Il est souhaitable de ne pas calculer de valeur de p pour les hypothèses non définies a priori car en général ces valeurs de p ne sont pas retrouvées quand l'hypothèse est testée de façon indépendante dans une autre étude [3].

La définition a priori, dans le protocole de l'essai, des analyses en sous-groupes pressenties ne suppriment que partiellement ces réserves méthodologiques. La multiplication des tests persiste.

> La situation la plus problématique est représentée par les sous-groupes définis par une variable mesurée après la randomisation et qui est donc potentiellement influencée par le traitement. C'est par exemple le cas du sous-groupe des patients ayant une artère coronaire perméable après randomisation dans un essai de fibrinolytique à la phase aiguë de l'infarctus du myocarde. Chez ces patients, un faible taux de mortalité ne reflète pas seulement l'efficacité du traitement mais aussi le bon pronostic spontané des sujets qui ont une reperfusion spontanée ou un infarctus de petite taille.

Essais stratifiés

Les essais stratifiés répondent de façon fiable à une question de type analyse en sous-groupes.

Pour répondre de façon satisfaisante aux questions que l'on se pose dans les analyses en sous-groupes, il est nécessaire de recourir aux essais stratifiés. Ces essais étudient simultanément deux ou plusieurs strates qui

sont l'équivalent des sous-groupes. Les réserves méthodologiques des analyses en sous-groupes sont levées grâce aux mesures suivantes :

- formulation explicite des hypothèses concernant les différentes strates dans l'objectif de l'essai,
- calcul du nombre de sujets nécessaires pour chacune des strates,
- randomisation indépendante par strate,
- prise en compte de la multiplicité des tests statistiques.

Ces essais "stratifiés" permettent, par exemple, de tester de façon fiable des hypothèses du type : le traitement est efficace dans le sous groupe 1 et dans le sous groupe 2 ou du type : le traitement est efficace tout groupe confondu et l'effet du traitement est différent entre le sous-groupe1 et le sous-groupe 2 (interaction).

Recherche de conclusion dans les sous groupes

La conclusion pour chaque sous groupe est possible si une méthode statistique de contrôle de l'inflation du risque alpha a été mis en œuvre. Il est aussi possible de faire un véritable essai stratifié.

Une méthode possible pour le contrôle de l'inflation du risque alpha est la méthode des tests hiérarchisés.

Exemple

Une procédure de test hiérarchisés a été mis en œuvre dans l'essai TARGET (Lancet 2004; 364: 665–74) pour montrer une meilleure tolérance en terme d'événements cardiovasculaires du lumiracoxib par rapport aux AINS chez des patients souffrant d'une arthrose. La prise simultanée de faible dose d'aspirine pour un problème cardiovasculaire est un élément important de la question posée. L'essai a donc cherché à répondre à cette question pour les 2 populations de patients prenant et ne prenant pas d'aspirine.

"The primary endpoint was analysed with a closed test procedure applying a hierarchical testing process. In the first step, this endpoint was tested in the population of patients not taking low-dose aspirin. If this test was positive the second step was to analyse the endpoint in the overall population. If this test was positive the third and final step was to do the analysis in the population of patients taking low-dose aspirin."

	Number of patients with events/number at risk (%)	Hazard ratio (95% CI)	p*
Overall population			
Both substudies†			
Lumiracoxib	29/9117 (0·32%)	0·34 (0·22–0·52)	<0·0001
Non-steroidal anti-inflammatory drugs	83/9127 (0·91%)		
Lumiracoxib vs ibuprofen substudy‡			
Lumiracoxib	10/4376 (0·23%)	0·29 (0·14–0·59)	0·0006
Ibuprofen	33/4397 (0·75%)		
Lumiracoxib vs naproxen substudy‡			
Lumiracoxib	19/4741 (0·40%)	0·37 (0·22–0·63)	0·0002
Naproxen	50/4730 (1·06%)		
Non-aspirin population			
Both substudies§			
Lumiracoxib	14/6950 (0·20%)	0·21 (0·12–0·37)	<0·0001
Non-steroidal anti-inflammatory drugs	64/6968 (0·92%)		
Lumiracoxib vs ibuprofen substudy¶			
Lumiracoxib	5/3401 (0·15%)	0·17 (0·07–0·45)	0·0003
Ibuprofen	28/3431 (0·82%)		
Lumiracoxib vs naproxen substudy¶			
Lumiracoxib	9/3549 (0·25%)	0·24 (0·12–0·50)	0·0001
Naproxen	36/3537 (1·02%)		
Aspirin population			
Both substudies§			
Lumiracoxib	15/2167 (0·69%)	0·79 (0·40–1·55)	0·4876
Non-steroidal anti-inflammatory drugs	19/2159 (0·88%)		
Lumiracoxib vs ibuprofen substudy¶			
Lumiracoxib	5/975 (0·51%)	0·92 (0·27–3·20)	0·9008
Ibuprofen	5/966 (0·52%)		
Lumiracoxib vs naproxen substudy ¶			
Lumiracoxib	10/1192 (0·84%)	0·73 (0·32–1·65)	0·4502
Naproxen	14/1193 (1·17%)		

*Based on Wald χ^2 statistic for treatment group comparison. Cox proportional-hazards models include, in addition to treatment group, the factors: †substudy, low-dose aspirin, and age; ‡low-dose aspirin and age; §substudy and age; and ¶age.

Table 3: Incidence of upper gastrointestinal ulcer complications (definite or probable), by substudy and aspirin use (modified intention-to-treat analysis)

Étude de la généralisabilité

En pratique, *les analyses en sous-groupes sont utilisées pour vérifier s'il y a lieu de suspecter une modification de l'efficacité de l'effet du traitement en fonction des caractéristiques des patients. Cette vérification est utile pour se faire une idée de* ***l'extrapolabilité*** *du résultat. En effet, s'il n'y a pas d'argument pour suspecter une telle variabilité de l'efficacité, le résultat de l'essai est certainement représentatif de l'efficacité du traitement sur une population de patients plus large. Cette analyse ne tient pas compte du degré de signification statistique obtenu pour les modalités des sous groupes mais seulement du test d'interaction.*

Par exemple dans l'essai dont une partie de l'analyse en sous groupes est représentée Figure 2, il est raisonnable de conclure à une bonne généralisabilité du résultat de l'essai aux hommes et aux femmes, pour tous les âges et quels que soient les antécédents.

Par contre pour l'essai dont l'analyse en sous groupe est représentée sur la

Figure 4 la situation est tout autre. ValVeFT est un essai comparant le valsartan au placebo dans l'insuffisance cardiaque par dessus la stratégie thérapeutique habituelle comprenant IEC ou bétabloquants [11]. Sur l'ensemble de l'essai " Valsartan significantly reduces the combined end point of mortality and morbidity

and improves clinical signs and symptoms in patients with heart failure, when added to prescribed therapy *". Cependant, au niveau de l'analyse en sous groupe* "the post hoc observation of an adverse effect on mortality and morbidity in the subgroup receiving valsartan, an ACE inhibitor, and a beta-blocker raises concern about the potential safety of this specific combination*". Bien qu'il s'agisse d'un résultat obtenu sur un sous groupe, cette conclusion est raisonnable par application du principe de précaution car l'effet suspecté est sérieux (augmentation de mortalité) et pourrait concerner un nombre important de patients (les bétabloquants sont un traitement standard de l'insuffisance cardiaque). Au niveau réglementaire en Europe ce principe est repris dans le point to consider sur la multiplicité (CPMP/EWP/908/99 http://www.emea.eu.int/pdfs/human/ewp/090899en.pdf).*

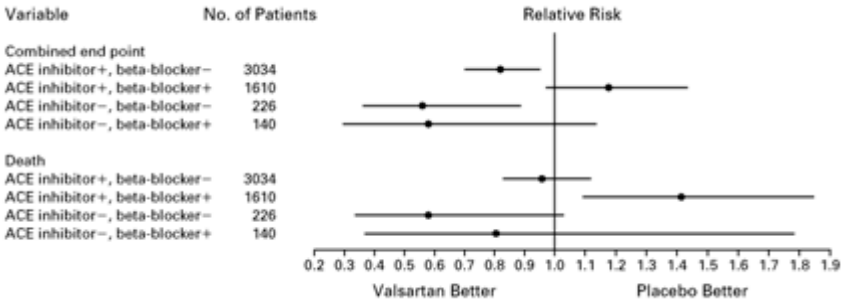


Figure 4 – Analyse en sous groupe de l'essai ValHeFT [11] montrant une augmentation de la mortalité induite par le valsartan chez les patients sous beta-bloquant et IEC (2ème ligne,n=1610).

Sous études issues des analyses en sous groupes

Une analyse en sous groupes peut parfois faire l'objet d'une publication à elle toute seule, par exemple l'analyse chez les diabétiques, chez les femmes, etc. Ces papiers s'intitulent souvent « sub-study » car ces analyses étaient prévues au protocole de l'étude. Les réserves à émettre dans leur interprétation sont identiques à celles listées ci-dessus.

Exemple

***The effect of perindopril on cardiovascular morbidity and mortality in patients with diabetes in the EUROPA study: results from the PERSUADE* substudy.**

*AIMS: The aim of this study was to assess the effect of the angiotensin converting enzyme inhibitor perindopril on cardiovascular events **in diabetic patients** with coronary artery disease. METHODS AND RESULTS: A total of 1502 diabetic patients with known coronary artery disease and without heart failure **of 12 218 overall** in the EUropean trial on Reduction Of cardiac events with Perindopril in stable coronary Artery (EUROPA) disease were randomized in a double-blinded manner to perindopril 8 mg once daily or placebo. Follow-up was for a median of 4.3 years. The primary end point was cardiovascular death, non-fatal myocardial infarction, and resuscitated cardiac arrest. Perindopril treatment was associated with a non-significant reduction in the primary endpoint in the diabetic population, 12.6 vs. 15.5%, relative risk reduction 19% [(95% CI, -7 to 38%), P=0.13]. This was of similar relative magnitude to the 20% risk reduction observed in the main EUROPA population. CONCLUSION: Perindopril tends to reduce major cardiovascular events in diabetic patients with coronary disease in addition to other preventive treatments and the trend towards reduction was of a similar relative magnitude to that observed the general population with coronary artery disease. D'après l'abstract PubMed de la réf. [12]*

Bénéfice absolu / bénéfice relatif dans les sous groupes

L'interprétation des résultats des sous-groupes peut encore être compliquée dans le cas où le sous-groupe pour lequel l'efficacité semble la plus faible est celui où le risque est le plus grand. C'est fréquemment le cas avec l'âge. Dans l'évaluation de l'efficacité de l'alteplase à la phase aigu de l'infarctus du myocarde, l'étude GUSTO [13] débouchent sur les estimations suivantes :

Tableau 3 – Résultat de l'analyse en sous groupe en fonction de l'age de l'essai GUSTO

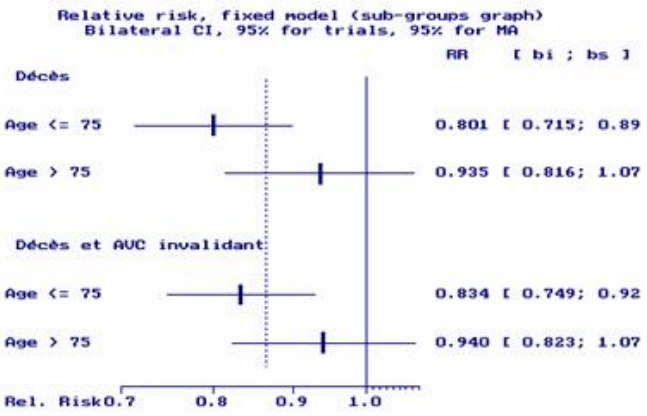
	Bénéfice relatif sur la mortalité à 30j RR (IC95%)	Bénéfice relatif sur la mortalité ou AVC invalidant à 30j RR (IC95%)
Age ≤ 75	0,80 (0,71 ;0,90)	0,83 (0,75 ;0,93)
Age > 75	0,94 (0,82 ;1,07)	0,94 (0,82 ;1,07)

Bien que le test d'interaction ne soit pas significatif (p=0.098), il est couramment dit dans les textes cardiologiques que le bénéfice de la perfusion accélérée d'alteplase est plus marquée chez le sujets de moins de 75 ans [111]. Ces résultats sont cependant ceux obtenus en termes de bénéfice relatif (risque relatif). Comme il s'avère que le risque spontané n'est pas le même en fonction de l'âge, il est important de présenter aussi l'effet du traitement sous forme de bénéfice absolu.

Tableau 4 – Bénéfice absolu en fonction de l'age dans l'essai GUSTO

	mortalité à 30j		mortalité ou AVC invalidant à 30j	
	Risque spontané (sous SK)	Bénéfice absolu (DR IC95%)	Risque spontané (sous SK)	Bénéfice absolu (RR IC95%)
Age ≤ 75	5,5%	-1,10% (-1,64% ;-0,64%)	6,0%	-1,00% (-1,57% ;-0,43%)
Age > 75	20,6%	-1,34% (-4,03% ;+1,36%)	21,5%	-1,30% (-4,04% ;+1,44%)

Figure 5 – Représentation graphique des résultats obtenus avec les risques relatifs.



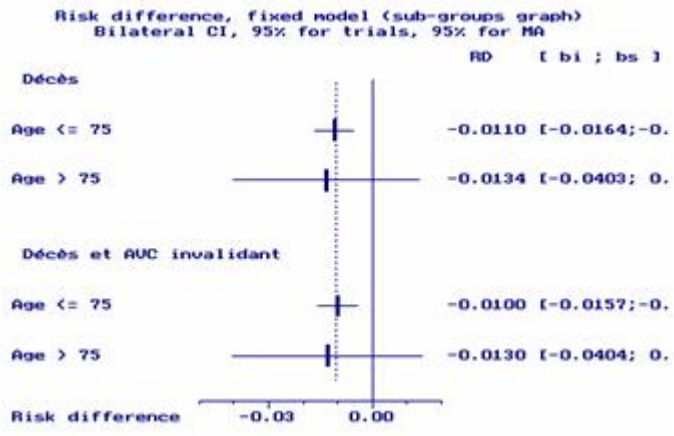


Figure 6 – Représentation graphique des résultats obtenus avec les différences de risques.

Les différences d'efficacité observées avec le risque relatif entre les groupes d'âges disparaissent en terme de bénéfice absolu. Chez les personnes de plus de 75 ans, l'efficacité du t-PA est moins importante, mais en raison d'un risque de base plus grand, le bénéfice absolu est de même ampleur que celui obtenu chez les sujets de moins de 75 ans. L'alteplase administrée en perfusion accélérée à la phase aiguë de l'infarctus pourrait sauver autant de vies en traitant 1000 patients de moins de 75 ans qu'en traitant 1000 patients de plus de 75 ans. À partir de ce constat est-il raisonnable de limiter l'utilisation de ce nouveau produit aux seuls patients de moins de 75 ans ?

Bibliographie

1. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266(1):93-8. *PMID:*
2. Sleight P. Subgroup analyses in clinical trials - fun to look at, but don't believe them! Curr Control Trials Cardiovasc Med 2000;1:25-27. *PMID:*
3. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365(9454):176-86. *PMID: 15639301*. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15639301
4. Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D, et al. The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. Arch Intern Med 2004;164(21):2335-42. *PMID: 15557412*. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15557412
5. Hercberg S, Preziosi P, Briancon S, Galan P, Triol I, Malvy D, et al. A primary prevention trial using nutritional doses of antioxidant vitamins and minerals in cardiovascular diseases and cancers in a general population: the SU.VI.MAX study--design, methods, and participant characteristics. SUpplementation en VItamines et Mineraux AntioXydants. Control Clin Trials 1998;19(4):336-51. *PMID: 9683310*. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9683310
6. Freed CR, Greene PE, Breeze RE, Tsai WY, DuMouchel W, Kao R, et al. Transplantation of embryonic dopamine neurons for severe Parkinson's disease. N Engl J Med 2001;344(10):710-9. *PMID: 11236774*. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11236774
7. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither, among 17187 cases of suspected acute myocardial infarction. Lancet 1988;2:349-360. *PMID:*
8. Matthews JN, Altman DG. Interaction 3: How to examine heterogeneity. BMJ 1996;313(7061):862. *PMID:*
9. Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not P values. BMJ 1996;313(7060):808. *PMID:*
10. Collins R, Peto R, Armitage J. The MRC/BHF Heart Protection Study: preliminary results. Int J Clin Pract 2002;56(1):53-6. *PMID: 11831837*. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11831837

11. Cohn JN, Tognoni G. A randomized trial of the angiotensin-receptor blocker valsartan in chronic heart failure. N Engl J Med 2001;345(23):1667-75. *PMID: 11759645*. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11759645
12. Daly CA, Fox KM, Remme WJ, Bertrand ME, Ferrari R, Simoons ML. The effect of perindopril on cardiovascular morbidity and mortality in patients with diabetes in the EUROPA study: results from the PERSUADE substudy. Eur Heart J 2005;26(14):1369-78. *PMID: 15860521*. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15860521
13. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. NEJM 1993;329:673-682. *PMID:*

Intervalle de confiance

Introduction

Dans l'interprétation d'un essai thérapeutique, la signification statistique est un élément important qui assure que le résultat obtenu a de forte chance d'être réel et non pas d'être le fruit du hasard. Cependant la signification statistique ne préjuge en rien de l'intérêt clinique du résultat [1].

Signification statistique n'est pas synonyme de signification clinique ou de pertinence clinique

Un test statistique ne se prononce que sur l'existence, probable ou non, d'un effet du traitement, et ne donne aucune information directe sur l'importance de celui-ci. La valeur de p ne représente pas l'intensité de l'efficacité. Un traitement n'est pas d'autant plus efficace que la valeur de p est petite. En effet, toute différence, aussi petite soit-elle, peut-être rendue aussi significative que souhaitée en augmentant le nombre de sujets. Un p significatif peut être obtenu avec un effet dont la taille est cliniquement pertinente, mais aussi avec un effet de petite taille, sans intérêt en pratique, si un très grand nombre de patients a été inclus dans l'essai. Une différence statistiquement significative *n'est pas forcément une différence* cliniquement significative.

Estimation

La pertinence clinique d'un résultat dépend de la taille de l'effet qui est estimé par l'essai thérapeutique. Cette estimation est fournie par la valeur observée dans l'essai (estimation ponctuelle) entourée de son intervalle de confiance (« confidence interval *»). L'intervalle de confiance traduit la précision statistique du résultat.*

Le but de l'estimation est de déterminer la vraie valeur d'un paramètre, par exemple, la vraie réduction relative de mortalité. Cependant, la valeur estimée dans un échantillon peut être assez loin de la vraie valeur du fait des fluctuations aléatoires d'échantillonnage, c'est-à-dire du fait du hasard. L'intervalle de confiance permet de prendre en compte cette incertitude aléatoire dans la présentation des estimations.

*Tout résultat d'essai thérapeutique est rapporté en mentionnant les valeurs du critère de jugement observées dans chaque groupe de traitement, l'estimation de la taille de l'effet entourée de son intervalle de confiance et la valeur du p du test statistique de l'existence d'un effet non nul : "*A total of 46 patients in the rofecoxib group had a confirmed thrombotic event during 3059 patient-years of follow-up (1.50 events per 100 patient-years), as compared with 26 patients in the placebo group during 3327 patient-years of follow-up (0.78 event per100 patient-years); the corresponding relative risk was 1.92 (95 percent confidence interval,1.19 to 3.11; P=0.008)*".*

Définition de l'intervalle de confiance

L'intervalle de confiance (IC) à 95% est un intervalle de valeurs qui a 95% de chance de contenir la vraie valeur du paramètre estimé. Avec moins de rigueur, il est possible de dire que l'IC représente la fourchette de valeurs à l'intérieur de laquelle nous sommes certains à 95% de trouver la vraie valeur recherchée. L'intervalle de confiance est donc l'ensemble des valeurs raisonnablement compatibles avec le résultat observé (l'estimation ponctuelle). Il donne une visualisation de l'incertitude de l'estimation.

Des intervalles de confiance à 99% ou à 90% sont parfois utilisés. La probabilité (degré de confiance) de ces intervalles de contenir la vraie valeur est respectivement de 99% et 90%.

L'intervalle de confiance est constitué des valeurs qui ne sont pas statistiquement significativement différentes du résultat observé. Les bornes supérieures et inférieures sont donc les valeurs les plus éloignées du résultat qui ne lui sont pas statistiquement différentes. Par contre les valeurs situées à l'extérieur de l'intervalle sont statistiquement différentes du résultat observé. Ainsi, la borne supérieure est la plus grande valeur non significativement différente de la valeur observée.

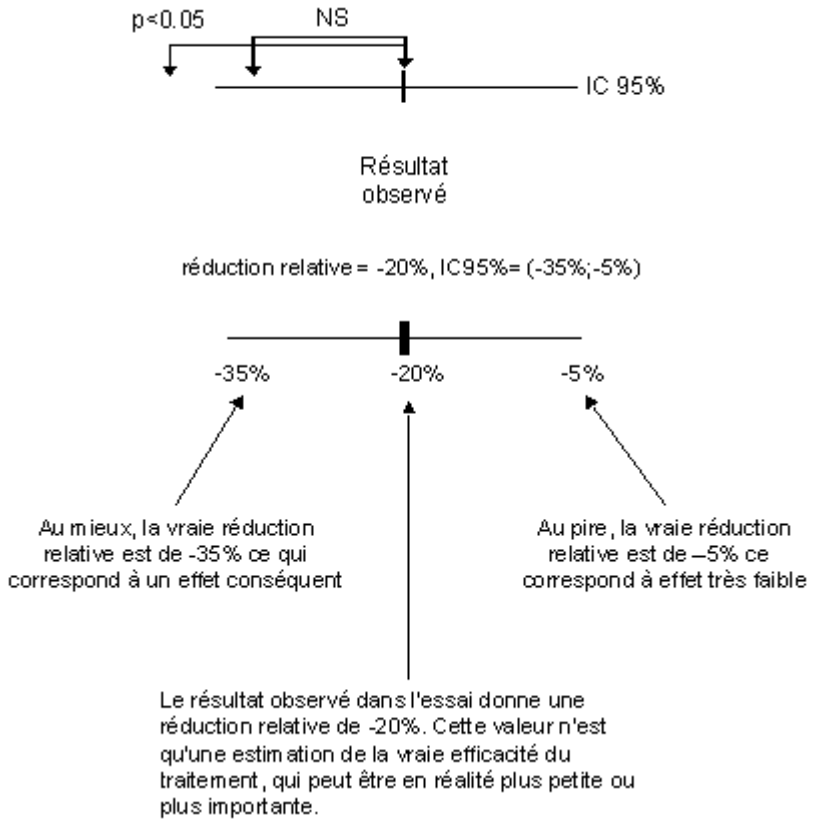


Figure 1 – Interprétation des bornes d'un intervalle de confiance

Exemple

Par exemple (

Figure 1), une réduction de mortalité de 20% avec un IC 95% de [35% ; 5%] signifie que bien qu'une baisse de 20% ait été observée ponctuellement dans l'essai, il n'est pas possible d'exclure que l'efficacité du traitement soit en réalité plus petite (au pire elle peut être de 5%) ou plus grande (au mieux de 35%).

En d'autre terme, dans cet essai une réduction de 5% n'est pas statistiquement différent de 20%.

Relation entre intervalle de confiance et test statistique

Dans un essai, l'intervalle de confiance visualise la précision avec laquelle l'effet du traitement est connu. La valeur de p est d'interprétation difficile car elle combine à la fois une information sur la taille de l'effet et une sur la précision de l'estimation de la taille de l'effet. Par contre, l'intervalle de confiance présente ces deux informations de manière distincte.

Lorsque l'intervalle de confiance contient la valeur caractéristique de l'effet nul (risque relatif de 1 ou différence de 0), il n'est pas possible d'exclure le fait que la vraie valeur soit cet effet nul. Ainsi la différence observée ne peut pas être considérée comme statistiquement significative.

À l'inverse, un test significatif au seuil de 5% conduit à dire qu'il y a 95% de chance que la vraie valeur de l'effet soit différente de l'effet nul. C'est-à-dire que l'intervalle de confiance à 95% ne contient pas la valeur de l'effet nul.

Ainsi, lorsqu'un test est significatif au seuil a (par exemple 5%), l'intervalle de confiance à 100-α% (c'est-à-dire dans notre exemple 95%) ne contient pas la valeur correspondant à l'absence d'effet (1 pour un risque relatif ou un odds ratio, 0 pour une différence de risque ou de moyenne). À l'opposé, lorsqu'un test n'est pas significatif, l'intervalle de confiance contient cette valeur (

Figure 2).

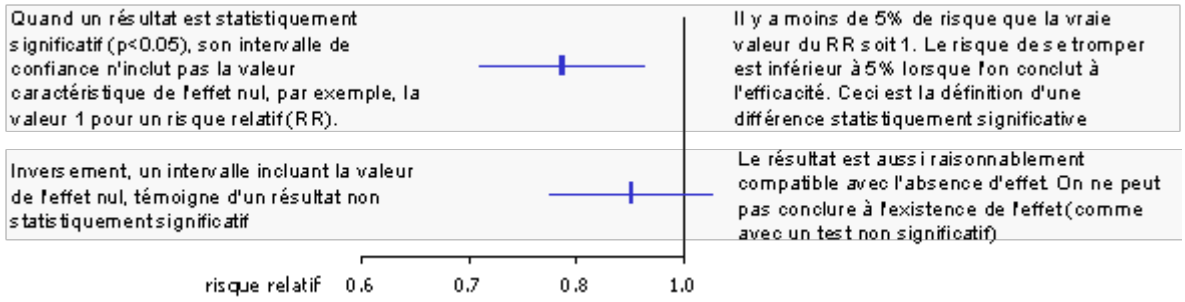


Figure 2 – Correspondance entre intervalle de confiance et test statistique

Interprétation

La borne péjorative de l'intervalle de confiance (le plus souvent la borne supérieure) représente le plus petit effet du traitement que l'on ne peut pas raisonnablement exclure.

Les intervalles de confiance permettent de visualiser le plus petit effet du traitement que l'on ne peut pas raisonnablement exclure [2-4]. Ce plus petit effet est la borne péjorative de l'intervalle de confiance (borne supérieure le plus souvent quand l'effet est bénéfique).

Cette logique qui cherche à exclure le pire est celle du test statistique. Pour accepter une conclusion d'efficacité du traitement, les données doivent permettre d'exclure avec une « quasi-certitude » (c'est-à-dire avec un risque d'erreur statistique minimal) la survenue du pire (le traitement n'a pas d'effet ou il a un effet délétère). Cette formulation visualise une fois de plus que le seuil classique de 5% est peut être trop élevé vis à vis des interprétations auxquelles il sert de substratum. En effet, peut-on parler de quasi certitude avec un risque d'erreur de 5% ?

Interprétation des intervalles de confiance dans le cas d'un résultat significatif

Premier cas de figure

Dans l'essai A (cf.

Tableau 1), le traitement entraîne une « réduction » relative du risque (RRR) de -23% (IC95% [-30%,-16%]). Pour cet exemple, une valeur de RRR négative signe une réduction du risque, à l'inverse une valeur positive une augmentation. Cette convention a été adoptée pour mettre sur la partie gauche du graphique les effets correspondant à un effet bénéfique. De ce fait, le graphique des RRR s'interprète de manière similaire à celui des risques relatifs. L'interprétation de ce résultat est qu'il existe un effet statistiquement significatif, de taille importante et connu avec précision. Ce traitement est intéressant en pratique car quelle que soit la valeur réelle de l'effet, celle-ci reste intéressante. Dans le pire des cas, cet effet est encore de -16% ce qui correspond à une réduction relative du risque satisfaisante.

Tableau 1 – Exemple de 5 situations différentes (ces données sont représentées graphiquement sur la Figure 3).

Essai	*RRR*	*IC 95%*	*p*
A	*-23%*	*[-30%;-16%]*	*0,000*
B	*-6%*	*[-10% ;-1%]*	*0,024*
C	*-23%*	*[-41% ;-1%]*	*0,043*
D	*0%*	*[-4% ;4%]*	*1,000*
E	*-19%*	*[-48% ;27%]*	*0,362*

RRR : « réduction » relative de risque. Par convention dans cet exemple, une RRR négative signe une réduction de risque. A l'opposé, une valeur positive témoigne d'une augmentation.

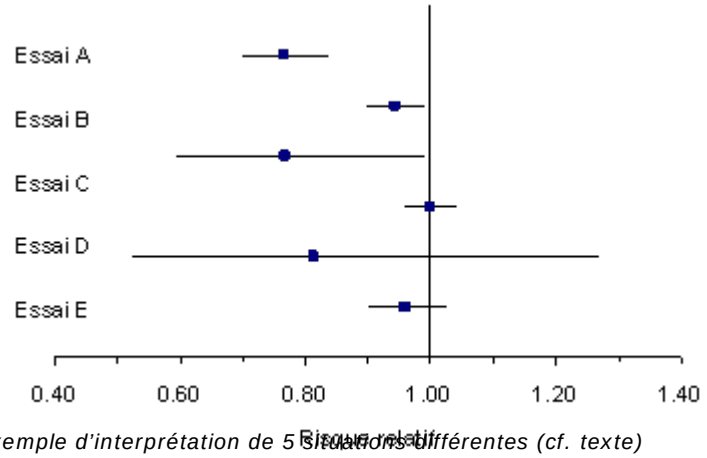


Figure 3 – Exemple d'interprétation de 5 situations différentes (cf. texte)

Deuxième cas de figure

Le traitement dans l'essai B entraîne une « réduction » relative du risque de -6% (IC95% [-10% ; -1%]). L'interprétation de ce résultat est qu'il existe un effet statistiquement significatif, que l'effet du traitement est connu avec précision (l'intervalle de confiance est étroit) mais qu'il n'est pas formellement prouvé que le traitement soit intéressant en pratique. En effet, même dans la meilleure des situations, c'est-à-dire celle où l'effet réel serait proche de la borne inférieure (-10%), la taille de l'effet reste faible et peu intéressante en pratique.

Troisième cas de figure

Le traitement dans l'essai C entraîne une « réduction » relative du risque de -23% (IC95% [-41% ;-1%]). L'interprétation de ce résultat est qu'il existe un effet statistiquement significatif , la taille de l'effet n'est pas connue avec précision mais il se pourrait que cet effet soit de taille intéressante. En effet, l'estimation ponctuelle (-23%) témoigne d'un effet substantiel de même que la borne inférieure de l'intervalle (-41%). Cependant l'incertitude sur ce résultat est grande, et il est aussi possible que l'effet réel soit quasiment nul (proche de la borne supérieure, -1%). En pratique, il est difficile de recommander l'utilisation de ce traitement car il existe une possibilité qu'il soit peu efficace. Un essai supplémentaire qui permettra d'améliorer la précision de l'estimation de l'effet par une méta-analyse pourrait être souhaitable. Tableau ***Erreur ! Signet non défini.*** *– Exemple de 5 situations différentes (ces données sont représentées graphiquement sur la*

Figure 3).

Interprétation des intervalles de confiance dans le cas d'un résultat non-significatif

Premier cas de figure

Dans l'essai D, le traitement n'entraîne pas de modification relative du risque (RRR=0%, IC95% de [-4% ;+4%]). Ce résultat n'est pas significatif (p=1,00). Au mieux, il pourrait exister une réduction très faible de 4% qui ne présente pas beaucoup d'intérêt en pratique. Bien qu'en toute rigueur, il ne soit pas possible de conclure à l'absence d'efficacité, l'interprétation de l'intervalle de confiance autorise à conclure que très probablement ce traitement ne serait d'aucune utilité en pratique. Cet exemple montre la supériorité de l'approche par les intervalles de confiance sur celle utilisant uniquement des tests statistiques. En utilisant

l'approche des tests statistiques il est impossible de formuler une conclusion (une différence non significative ne permet pas de conclure). Par contre, avec l'approche basée sur les intervalles de confiance et étant donné la précision du résultat, il est licite de conclure à l'absence d'intérêt de ce traitement : même si celui-ci a une efficacité non nulle, la taille de l'effet est trop petite pour être intéressante en pratique.

Deuxième cas de figure

Le traitement dans l'essai E entraîne une « réduction » relative non significative de -19% (IC à 95% de [-48%,+27%]). Il apparaît clairement que ce résultat non significatif n'autorise pas à conclure à l'absence d'effet. En effet, ce résultat est compatible avec une « réduction » relative de -48%, effet de taille conséquente. De plus l'intervalle est en très grande partie du côté favorable ce qui renforce la possibilité de l'existence de l'effet. En conclusion, il est possible que le traitement soit efficace et que cette efficacité soit suffisamment importante pour être intéressante en pratique. Ce résultat encourage à réaliser un nouvel essai de plus grande puissance.

Exemple

La méta-analyse des essais évaluant la vitamine E en prévention des événements cardiovasculaires regroupe 81 788 patients et donne le résultat suivant : « Vitamin E did not significantly lower cardiovascular mortality compared with control treatment (6·0 vs 6·0%, relative risk 1·0 [0·94–1·06], p=0·94 ». Bien que non significatif, il est possible que la vitamine E n'a aucun intérêt en prévention des maladies cardiovasculaires. Au mieux se serait éventuellement une réduction de 6% sans grand intérêts.

Remarques diverses

Nécessité d'essais surpuissants

Pouvoir écarter qu'un traitement possède une efficacité trop petite pour être cliniquement pertinente nécessite d'avoir des intervalles de confiance excluant largement l'absence d'effet. Cette configuration nécessite davantage de puissance que pour simplement exclure l'absence d'effet. En fait le raisonnement devrait être similaire à celui de l'essai de non-infériorité (cf. chapitre Les essais d'équivalence clinique*). Il n'est pas suffisant de montrer qu'un traitement a un effet non nul, il conviendrait plutôt de montrer que l'efficacité du traitement est suffisamment importante pour être cliniquement pertinente, c'est-à-dire qu'elle n'est pas inférieure au plus petit bénéfice cliniquement intéressant. Ce point suggère que les essais doivent être surpuissants par rapport à la puissance nécessaire pour rejeter l'absence d'effet. Plusieurs essais ont suivi cette approche.* [5-7]

Analyses intermédiaires

La pratique des analyses intermédiaires se généralise. Elles permettent d'arrêter les essais au plus tôt, dès que le nombre de patients inclus est suffisant pour mettre en évidence l'effet. Ce type d'analyse complique le calcul de l'intervalle de confiance (cf. chapitre Analyses intermédiaires*).*

Par son principe, ce mode d'analyse conduit, dans les essais arrêtés précocement, à des intervalles de confiance dont la borne supérieure est à la limite de l'absence d'effet. Il devient alors difficile d'analyser si l'effet obtenu est cliniquement pertinent.

Multiplicité des intervalles de confiance et inflation du risque alpha

Un problème d'inflation du risque alpha survient lorsque l'on considère simultanément plusieurs intervalles de confiance, de manière similaire au phénomène survenant lors des comparaisons multiples (cf. chapitre sur les tests statistiques).

Un intervalle de confiance nous assure que si l'on fait le parie que la vraie valeur est comprise entre la borne inférieure et la borne supérieure d'avoir raison dans 95% des cas.

Certaines situations conduisent à s'intéresser simultanément à plusieurs intervalles de confiance. C'est par exemple le cas lorsque l'on décrit l'effet d'un traitement sur tous les critères de jugement mesurés dans l'essai. Ces situations conduisent à une inflation du risque alpha et si l'on fait le pari que toutes les vraies valeurs sont comprises dans leur intervalle de confiance respectif, le risque d'avoir raison n'est plus de 95% mais il est plus faible. D'autant plus faible que le nombre d'intervalle est important. Avec 5 critères, la probabilité que les 5 intervalles de confiances inclus simultanément les 5 vraies valeurs n'est que de 77%.

Pour contrait ce phénomène, il est possible d'ajuster la largeur des intervalles de confiance à l'aide de la méthode de Bonferroni, en prenant pour IC à 95% des intervalles en fait à $(1-\alpha_{aj})\%$ (où α_{aj} représente le seuil ajusté par la méthode de Bonferroni).

Bibliographie

1. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ Clin Res 1996;292:746-750.
2. Borenstein M. The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials* 1994;15:411-428.
3. Rothman KJ, Yankauer A. Confidence intervals vs significance tests: quantitative interpretation. *Am J Public Health* 1986;75:587-588.
4. Bulpitt CJ. Confidence intervals. *Lancet* 1987;1:494-497.
5. Collins R, Peto R, Armitage J. The MRC/BHF Heart Protection Study: preliminary results. *Int J Clin Pract* 2002;56(1):53-6.
6. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002;360(9326):7-22.
7. Pfeffer MA, McMurray J, Leizorovicz A, et al. Valsartan in acute myocardial infarction trial (VALIANT): rationale and design. *Am Heart J* 2000;140(5):727-50.

Analyses intermédiaires

Introduction

Les analyses intermédiaires sont des analyses des données réalisées en cours d'essai, avant que tous les patients prévus aient été recrutés et/ou avant la fin de la période de suivi initialement prévue.

Au cours d'un essai, l'information s'accumule progressivement au fur et à mesure des inclusions et du suivi des patients. Mais c'est seulement au terme de l'essai, après avoir recruté l'effectif prévu et suivi les patients avec le recul prévu, que la quantité d'information est suffisante et que les données peuvent être analysées. Cependant, des analyses intermédiaires (« interim analysis ») à la recherche de l'effet du traitement en cours d'essai sont envisageables pour diverses raisons avant que tous les patients prévus aient été recrutés et/ou avant la fin de la période de suivi initialement prévue.

Par exemple, dans un essai devant recruter 300 patients avec un suivi d'un mois, deux analyses intermédiaires sont réalisées : la première après l'inclusion des 100 premiers patients, la seconde avec 200 patients. L'analyse finale porte, comme prévu, sur 300 patients. Dans un essai de 1000 patients avec un suivi de 5 ans, une analyse intermédiaire est réalisée après recrutement de l'ensemble des patients, mais à mi-parcours, c'est-à-dire avec un recul de 2,5 ans.

Le but de ces analyses intermédiaires est triple.

1. *Le premier est de pouvoir détecter au plus tôt le bénéfice du traitement afin d'éviter de traiter des patients par un placebo alors que les données amassées sont suffisantes pour conclure à l'efficacité du traitement étudié (**arrêt pour efficacité**). De plus, la confirmation au plus tôt du bénéfice apporté par un traitement permet de faire bénéficier du traitement tous les patients hors essai le plus rapidement possible.*
2. *Le deuxième objectif est de se donner les moyens de détecter au plus tôt un éventuel effet délétère du traitement afin de limiter le nombre de patients exposés au risque (**arrêt pour toxicité**). Dans ces deux circonstances, le but des analyses intermédiaires est d'éviter de continuer à inclure des patients alors que l'on dispose d'une réponse suffisamment fiable à la question posée.*
3. *Le troisième objectif est d'arrêter une étude dont on peut prédire avec une certitude raisonnable qu'elle ne pourra pas aboutir (**arrêt pour futilité**). L'arrêt précoce permettra de diriger les ressources vers le test de nouvelles hypothèses.*

La réalisation de ces analyses posent cependant un certain nombre de problèmes méthodologiques et nécessitent une méthodologie adaptée [1, 2]. Mais avant d'aborder l'exposé de ces problèmes et de leur solution, voyons les circonstances qui peuvent conduire à un arrêt prématuré d'un essai lors d'une analyse intermédiaire.

Situations conduisant à un arrêt prématuré

Dans un chapitre précédent, nous avons vu que l'effectif d'un essai est le nombre de sujets minimal nécessaire pour garantir une probabilité élevée de mettre en évidence l'effet du traitement. Dans ce cas comment peut-on espérer, lors d'une analyse intermédiaire, pouvoir faire la même chose avec moins de patients ?

En fait, si l'effet réel du traitement est bien supérieur à l'effet initialement suspecté ou que le risque de base des patients inclus est bien supérieur à celui attendu, il sera possible de mettre en évidence l'effet du traitement avec moins de sujets que l'effectif prévu. Dans les deux cas, il y a eu sous-estimation de l'un ou de ces deux paramètres dans le calcul du nombre de sujets nécessaires et l'effectif initialement calculé est surdimensionné par rapport à la réalité.

Cette situation est imaginable car, assez souvent, des données fiables manquent pour faire les hypothèses du calcul du nombre de sujets. Il est alors envisageable que les valeurs retenues soient éloignées de la réalité. Les mêmes raisons expliquent aussi l'échec de certains essais qui avaient surestimé la taille de l'effet et donc sous estimé l'effectif nécessaire pour le mettre en évidence.

Inflation du risque alpha

Exposé de la problématique

La comparaison répétée de l'efficacité de deux traitements par des tests statistiques successifs accroît le risque de conclure à tort à la supériorité de l'un par rapport à l'autre.

La réalisation de plusieurs analyses statistiques dans la même expérience, pour tester la même hypothèse, conduit à des comparaisons statistiques multiples. À chaque analyse intermédiaire un test statistique est réalisé pour rechercher un effet du traitement. La répétition à chaque test du risque d'obtenir un résultat significatif par hasard augmente le risque global de conclure à tort à l'efficacité du traitement lors de cet essai. In fine le risque alpha n'est plus de 5% (même si c'est le seuil retenu pour chaque test) mais il est bien supérieure.

L'utilisation de techniques statistiques adaptées est nécessaire pour empêcher cette augmentation du risque alpha, appelée en jargon statistique « inflation du risque alpha ». Le but de ces méthodes est de garantir un risque global, sur l'ensemble des comparaisons effectuées, de conclure à tort à l'efficacité du traitement de 5%. Sur l'ensemble des comparaisons effectuées le risque d'obtenir au moins un résultat significatif par le fait du hasard est contrôlé et garde sa valeur prédéfinie de 5%.

Principe de la solution

Plusieurs solutions sont possibles qui sont à la base de différentes méthodes. L'une d'entre elles consiste à diminuer le seuil de signification statistique de chacune des comparaisons intermédiaires, par exemple en divisant le risque alpha global α par le nombre de comparaisons effectuées n. C'est la méthode de Bonferroni [3]. Ainsi malgré l'inflation du risque alpha, le risque final de conclure à tort à l'efficacité restera compris dans les valeurs habituelles.

Avec 3 analyses intermédiaires prévues, le nombre total de comparaisons qui seront effectuées est de 4 : les 3 intermédiaires plus la comparaison finale. Le seuil à utiliser pour chacune de ces analyses est de 5%/4=1,25%. Si un p inférieur à 1,25% est obtenu à l'une des analyses intermédiaires, il est alors possible de conclure et d'arrêter l'essai sans attendre la fin du recrutement prévu.

Études de cas

Cas 1. *Dans la situation dépeinte par le tableau ci dessous, l'essai peut être arrêté à la 2ème analyse intermédiaire. Le p obtenu lors de cette analyse est inférieur au seuil de signification corrigé et l'essai peut donc être arrêté prématurément. Cette situation met en avant tout l'intérêt des analyses intermédiaires.*

Analyses intermédiaires			Analyse finale
1	2	3	
p = 0,10	p = 0,011		

Cas 2. *Dans ce deuxième exemple, le p<5% de la troisième analyse intermédiaire ne permet pas de conclure à une différence significative car la valeur obtenue reste supérieure au seuil corrigé pour 4 tests (1,25%). L'essai va donc à son terme et lors de l'analyse finale le p devient inférieur au seuil corrigé ce qui donne donc finalement un résultat statistiquement significatif.*

Analyses intermédiaires			Analyse finale
1	2	3	
p = 0.25	p = 0.08	p = 0.04	P = 0.012

Cas 3. *Le cas suivant peut paraître déroutant. Aucune analyse intermédiaire ne conduit à interrompre prématurément l'essai. Lors de l'analyse finale un p de 4% est obtenu. Cette valeur, bien qu'elle soit inférieure à 5% n'autorise pas à conclure à un résultat statistiquement significatif car elle reste supérieure au seuil corrigé. Il ne peut pas être considéré comme significatif car du risque alpha a été consommé au cours des analyses précédentes, effritant le contrôle du risque d'erreur de première espèce apporté par un p<5% au niveau d'une comparaison donnée (le coté gênant de ce résultat a conduit au développement d'une méthode qui évite de se retrouver dans cette situation, la méthode de Peto, cf. infra)*

Analyses intermédiaires			Analyse finale
1	2	3	
p = 0,42	p = 0,28	p = 0,12	p = 0,04

Cas 4. *Dans le dernier cas de figure, aucune analyse n'atteint le seuil corrigé de signification statistique. L'essai n'obtient donc pas de résultat statistiquement significatif.*

Analyses intermédiaires			Analyse finale
1	2	3	
P = 0,89	p = 0,48	p = 0,25	p = 0,10

Le coté paradoxale du cas n°3

Le cas n° 3 est un perturbant. A la dernière analyse, le p de 4% de la dernière analyse ne permet pas de conclure alors que s'il avait été obtenu sans aucune analyse intermédiaire l'essai serait concluant !

En fait, 4% à l'issu d'un seul test à la fin d'un essai ce n'est pas la même chose qu'un 4% à la 4ème analyse car dans ce dernier cas il y a eux les 3 autres tests.

Pour mieux comprendre cela, nous allons faire du dénombrement.

En ne faisant qu'une et une seule analyse en fin d'étude, sur 100 essais réalisés avec un traitement sans effet, quel est le nombre d'essais permettant de conclure à tort à l'efficacité ? C'est 5 (pour un seuil de signification de 5%) par définition du risque alpha. Le risque effectif de conclure à tort est donc bien celui que l'on attend (5/100).

Toujours avec un traitement sans effet, mais en faisant 3 analyses intermédiaires et une analyse finale (soit 4 analyses au total), combien avons-nous de possibilités d'obtenir un résultat concluant (c'est à dire d'obtenir un résultat significatif au 1er test OU au 2ème OU au 3ème OU au 4ème et dernier test.) :? Le calcul est un peu plus fastidieux. C'est 5 au premier test (5% des 100 essais initiaux) puis c'est, au deuxième test, 5% des 95 essais qui n'ont pas été significatifs au premier (soit 4.75), puis 5% des 100-5-4.75=90.25 essais qui n'ont pas été significatifs à la première et à la seconde analyse intermédiaire soit 5.5125, etc. (cf tableau ci dessous)

Au total, en procédant de cette façon, sur 100 essais il y a en aura 18,5 qui donneront une possibilité de conclure à l'efficacité du traitement (soit lors de la 1er analyse intermédiaire, soit lors de la seconde, soit etc.). Le risque globale d'erreur de 1er espèce est donc de 18,5/100 soit 18,5%, bien supérieur au 5% que l'on est prêt à consentir.

Analyse intermédiaire (AI) n°	*Nombre d'essais arrivant à l'AI*	*Nb de résultats significatifs à l'AI (tests avec un seuil de 5%)*	*Cumul des conclusions à l'efficacité*
1	*100*	*5*	*5*
2	*95*	*4.75*	*9.75*
3	*90.25*	*4.5125*	*14.2625*
4 (Analyse finale)	*85.7375*	*4.286875*	*18.549375*

Une méthode de Bonferroni permet de solutionner ce problème. En prenant comme seuil pour les tests 5%/4=1.25% ont obtient le décompte suivant :

Analyse intermédiaire (AI) n°	Nombre d'essais arrivant à l'AI	Nb de résultats significatifs à l'AI (tests avec un seuil de 1.25%)	Cumul des conclusions à l'efficacité
1	100	1.25	1.25
2	98.75	1.234375	2.484375
3	97.515625	1.218945313	3.703320313
4 (Analyse finale)	96.29667969	1.203708496	4.907028809

Ainsi, en exigeant pour conclure à l'efficacité d'avoir un p inférieur à 1.25% à une de ces 4 analyses, on obtiendra sur 100 essais que 4,9 essais concluants. Ce qui donne un risque d'erreur global de 1er espèce de 4,9/100=4,9% proche des 5% recherché. En fait, la méthode de Bonferroni est conservatrice car elle autorise légèrement moins de conclusion à tort que ce que l'on est prêt à accepter.

p global

Dans les cas 1 et 2, les valeurs numériques de p obtenues lors des tests ne peuvent pas être retranscrites comme telle. Par exemple, dans le cas 1 il n'est pas possible de dire qu'un résultat significatif avec p=0,011 a été obtenu. La procédure utilisée ne garantit que le risque alpha de la conclusion globale, et non pas celle de chaque comparaison. la conclusion correcte est de dire que dans le cas 1 un résultat statistiquement significatif a été obtenu avec p<5%. Des techniques statistiques spécialisées peuvent être utilisées pour estimer le p global qui dans l'exemple 1 sera proche de 5%.

Ce point complique aussi le calcul d'un intervalle de confiance. L'intervalle de confiance à 95% calculé directement est trop étroit. De la même façon que l'on utilise un seuil de risque alpha plus petit que 5% à chaque comparaison, il est nécessaire de prendre un niveau de confiance plus élevé qui pourrait être en première approximation (100%-1,25% = 98,75%). En effet, d'après le lien existant entre test statistique et intervalle de confiance, l'intervalle de confiance de 98,75% correspond à un test réalisé avec un seuil de signification de 1,25%. Là aussi des techniques statistiques adaptées existent pour calculer les intervalles de confiances.

Les différentes méthodes

La première méthode proposée (Pocock) utilisait un seuil constant à chaque analyse [4] et était une application de la méthode de Bonferroni aux analyses intermédiaires. Cette stratégie n'est plus conseillée actuellement car elle conduit à interrompre trop facilement un essai peu de temps après son démarrage. Elle expose aussi à la situation du cas n°3 de l'étude de cas précédente (cf. supra). Pour éviter ces écueils, les autres méthodes utilisent des seuils de signification croissants au cours des analyses intermédiaires. Par exemple, dans la méthode de O'Brien et Flemming les seuils sont très faibles au moment des premières analyses [5].

Le précédent cas n°3 peut apparaître paradoxal. Lors de l'analyse finale le p de 4% ne permet pas de conclure à un résultat statistiquement significatif. Pourtant si aucune analyse intermédiaire n'avait été réalisée, ces données auraient conduit à un résultat statistiquement significatif. Même si la conclusion adaptée s'explique par l'inflation du risque alpha, la publication d'un résultat final d'essai où l'on conclut à un résultat non significatif avec un p de 4% risque d'être mal comprise et interprétée. Pour éviter cette situation, Peto et Haybittle ont proposés une méthode où les comparaisons intermédiaires s'effectuent avec un seuil très bas (de l'ordre de 0,001) [6], ce qui consomme peu de risque alpha et permet de prendre un seuil très proche de 5% pour l'analyse finale. Avec cette méthode, un essai ne peut être interrompu prématurément que si on obtient un résultat très hautement significatif lors d'une analyse intermédiaire, c'est-à-dire si l'effet réel s'avère très supérieur à celui attendu.

D'autres méthodes, en particulier celles proposées par Lan et Demets [7], sont intermédiaires entre celles de Peto- Haybittle et de Pocock (Tableau 1). La méthode de Peto- Haybittle est actuellement très prisée. D'autres approches statistiques ont été proposées. L'approche du stochastique curtailment a pour principe d'extrapoler, à partir des résultats observés lors d'une analyse intermédiaire, ce que pourrait être le résultat final de l'essai et de calculer les probabilités d'obtenir une différence significative sous différentes hypothèses d'effet du traitement. Des méthodes bayesiennes ont également été proposées 19. Armitage P. Interim analysis in clinical trials. Stat Med 1991;10:925-37.

Tableau 1 – Comparaison des différentes méthodes

	Analyses intermédiaires				**Analyse**
	1	**2**	**3**	**4**	**finale**
Pocock	0,017	0,017	0,017	0,017	0,017
O'Brien et Flemming	0,00005	0,004	0,012	0,025	0,04
Lan et Demets 1	0,015	0,016	0,017	0,018	0,019
Lan et Demets 2	0,00001	0,002	0,011	0,025	0,041
Peto - Haybittle	0,001	0,001	0,001	0,001	0,05

Exemple

"The data and safety monitoring board monitored the incidence of the primary outcome to determine the benefit of clopidogrel, using a modified Haybittle–Peto boundary of 4 SD in the first half of the study and 3 SD in the second half of the study. The boundary had to be exceeded at two or more consecutive time points, at least three months apart, for the board to consider terminating the study early. There were two formal interim assessments performed at the times when approximately one third and two thirds of the expected events had occurred."

Les deux monitorages

L'objectif des analyses intermédiaires étant double, deux surveillances sont réalisées simultanément : celle de l'efficacité (« efficacy *») et celle de la sécurité («* safety *»). Un essai pouvant être arrêté prématurément soit par la surveillance de l'efficacité soit par celle de la sécurité. Il est devenu habituel d'utiliser des règles d'arrêt et des méthodes différentes pour les deux surveillances. La surveillance des essais d'équivalence présente de nombreuses particularités et fait appel à des adaptations des méthodes classiques.*

Exemple

« The protocol specified that the independent data and safety monitoring board would undertake an interim analysis when 25%, 50%, and 75% of the total anticipated primary endpoints had accrued. The interim analyses used an asymmetric (Peto-Haybittle) type rule and we prespecified that the board might advise termination if a significant difference emerged in favour of atorvastatin (at $p<0{\cdot}0005$ one-sided, $p<0{\cdot}001$ twosided at any analysis) or in favour of placebo (at $p<0{\cdot}005$, 0·1, and 0·2 one-sided, for the three interim analyses, respectively). »

Autres objectifs des analyses intermédiaires

Les analyses intermédiaires s'intègrent dans un processus global de surveillance des essais. À coté de la recherche anticipée d'un effet du traitement et de la protection des personnes incluses dans l'essai, cette surveillance a pour objectif de vérifier le bon déroulement de l'essai. Il s'agit d'éviter des dérives dans la réalisation de l'essai, qui, si elles n'étaient détectées qu'à la fin, rendraient l'essai inutilisable en raison de défauts de qualité rédhibitoires.

Les éléments à surveiller sont les suivants :

- le taux d'écart au protocole : l'essai est-il de qualité ?
- le taux d'inclusion : est-ce que l'essai pourra être réalisé dans un délai acceptable ?

- les caractéristiques des patients inclus : le risque de base des patients effectivement inclus correspond t-il à celui initialement prévu et utilisé dans le calcul du nombre de sujets nécessaire ? Les patients recrutés correspondent-ils à la population cible de l'essai ?

Cette surveillance permet de prendre au plus tôt des mesures correctrices. Les centres investigateurs ayant des difficultés à suivre le protocole pourront rectifier le tir. En cas de taux de recrutement insuffisant, d'autres centres investigateurs pourront être recrutés afin d'éviter qu'un essai dure trop longtemps. En effet, une durée excessive limite l'intérêt d'un essai.

Cette surveillance par analyse des données amassées se superpose à la surveillance « de terrain » de l'essai (appelé parfois « monitorage ») qui est focalisée sur le contrôle de qualité des données (visite de centres, contrôles des données, audit).

Les analyses intermédiaires en pratique

La réalisation d'une analyse statistique implique la levée de l'insu. Une organisation particulière est donc nécessaire pour éviter que la réalisation d'analyses intermédiaire perturbe la réalisation de l'essai en double insu et conduise à l'introduction de biais. En particulier le résultat de ces analyses doit rester inconnu de toutes les personnes impliquées dans la réalisation de l'essai : investigateur, personnels de coordinations, promoteurs. En effet, la divulgation des résultats des analyses intermédiaires pourrait avoir de nombreuses conséquences délétères pour l'essai : arrêt des inclusions en cas de tendance favorable et utilisation en pratique d'un traitement sans que la démonstration de l'efficacité n'ait pu être obtenue.

En pratique, les analyses intermédiaires sont réalisées par une structure indépendante de la coordination de l'essai. Les résultats de l'analyse sont communiqués à un comité de surveillance (« safety committee , Independent Data Monitoring Committee (IDMC), Data and Safety Monitoring Board (DSMB), Monitoring Committee, Data Monitoring Committee *») composé de personnes indépendantes. Ce comité émettra au vu des résultats des analyses statistiques une recommandation destinée au comité directeur de l'essai. Cette recommandation peut être de poursuivre le recrutement, d'interrompre l'essai à ce stade, de modifier le protocole.*

Les analyses séquentielles

Les méthodes séquentielles permettent de répéter de nombreuses fois en cours d'essais la comparaison statistique et d'arrêter l'essai dès qu'il est possible de rejeter l'hypothèse nulle. L'analyse séquentielle est équivalente à la réalisation de nombreuses analyses intermédiaires. À chaque nouvelle paire de patients (un patient traité avec le traitement étudié et un avec le traitement contrôle) on calcule une nouvelle valeur du test que l'on compare à une valeur maximum et une valeur minimale. Si la valeur maximale est dépassée, on rejette l'hypothèse nulle. Si la valeur est inférieure à la valeur minimum, on rejette l'hypothèse alternative ce qui permet de conclure à l'équivalence à un delta après. Dans ces deux cas, l'essai s'arrête. Tant que la valeur du test reste entre ces deux limites, on continue à recueillir des observations.

La réalisation de cette approche repose en pratique sur un graphique représentant les limites de décision sous forme de triangle (figure 1). Pour cela la méthode est appelé test triangulaire.

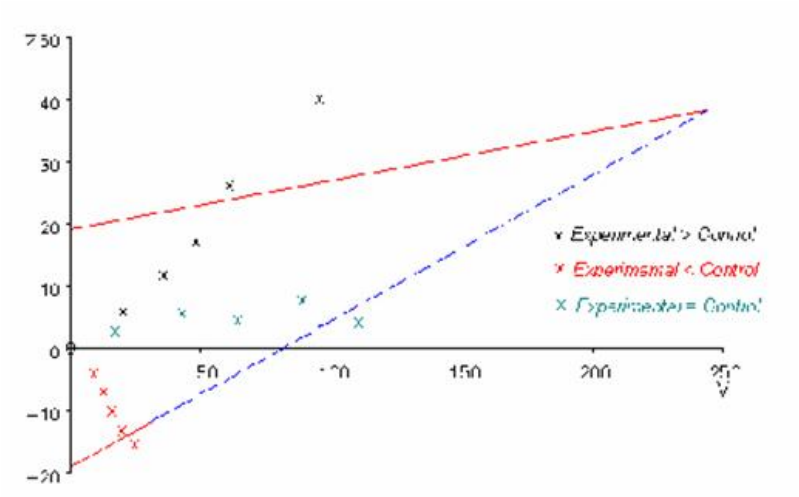


Figure 1- Représentation graphique du test triangulaire qui est un type d'analyse séquentielle.

Un inconvénient est que la méthode suppose que l'observation de la paire précédente soit terminée quand, est admise dans l'essai, la paire de patients suivante. Cela implique en pratique que la durée d'observation soit courte par comparaison à l'intervalle entre les admissions des sujets dans l'essai.

Un des avantages des analyses séquentielles est de permettre de conclure en moyenne avec moins de sujets qu'une approche classique. L'autre intérêt est l'assurance d'arrêter l'essai au plus tôt dès que la preuve de la supériorité d'un des traitements comparés est atteinte.

Dans de nombreux cas, l'analyse séquentielle est utilisée sous forme d'analyses séquentielles groupées : l'analyse n'est pas réalisée à chaque paire de patients mais après l'inclusion d'un petit nombre de patients.

Lecture critique et l'interprétation

Les questions à se poser lors de la lecture critique sont les suivantes en ce qui concerne les analyses intermédiaires :

- Si plusieurs analyses successives du même essai ont été réalisées, est-ce que ces analyses étaient de véritables analyses intermédiaires prévues a priori et utilisant une méthode de protection contre les risques des comparaisons multiples ? Pour le lecteur non statisticien il est difficile de juger de la méthode utilisée. D'une manière générale, toutes les méthodes couramment utilisées sont satisfaisantes. Il suffit donc de vérifier de le chapitre analyses statistiques précise bien l'utilisation d'une méthode sans qu'il soit utile de rentre dans les détails.
- Le p rapporté est-il correct ? Est-ce le p trouvé directement au niveau d'une des la comparaison ou le seuil de signification statistique ajusté pour tenir compte de l'ensemble des comparaisons ?
- Si des intervalles de confiance sont rapportés, sont-ils corrigés pour prendre en compte l'inflation du risque alpha ?

Pour en savoir plus

[8]

Bibliographie

1. *Buyse M. Interim analyses, stopping rules and data monitoring in clinical trials in Europe. Stat Med 1993;12(5-6):509-20.* PMID:

2. *Pocock SJ. When to stop a clinical trial. BMJ 1992;305(6847):235-40.* PMID:

3. *Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.* PMID:

4. *Pocock SJ. Group sequential methods for clinical trials. Biometrics 1977;35:549-56.* PMID:

5. *O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549-56.* PMID:

6. *Peto R, Pike MC, Artmitage P. Design and analysis of randomized clinical trials requiring prolonged observation of each patients. Br J cancer 1976;34:585-612.* PMID:

7. *Lan KKG, Demets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70:659-63.* PMID:

8. *Mueller PS, Montori VM, Bassler D, Koenig BA, Guyatt GH. Ethical issues in stopping randomized trials early because of apparent benefit. Ann Intern Med 2007;146(12):878-81.* PMID: 17577007.

Analyse ajustée

But de l'ajustement

L'ajustement (« adjustement *») consiste à corriger la mesure de l'effet du traitement des effets parasites induits par des covariables qui influencent aussi le critère de jugement. Cela diminue le bruit de fond induit par ces covariables et améliore ainsi la précision de l'estimation. L'ajustement sépare aussi l'effet de ces covariables de celui du traitement pour tenter de supprimer l'effet des facteurs de confusion. En théorie l'ajustement a donc ces 2 intérêts potentiels mais dans l'essai thérapeutique l'ajustement est seulement utilisé pour augmenter la précision.*

Augmenter la précision

> *L'ajustement permet de prendre en compte dans les calculs statistiques un facteur qui augmente la variabilité du critère de jugement.*

Critères de jugement binaires

Le

Tableau 1 illustre le gain en précision et en puissance apportée par une analyse ajustée. L'essai analysé repose sur une allocation aléatoire des traitements stratifiée en fonction du pronostic des patients.

> **Note** :Une allocation des traitements stratifiée consiste à faire des sous allocations distinctes dans chaque strate. Le terme strate désigne en fait des catégories différentes de patients. L'allocation stratifiée a pour objectif d'assurer l'équilibre entre les 2 groupes de ces différentes catégories de patients. Les strates sont constituées en fonction du pronostic des patients afin d'éviter au maximum un déséquilibre entre les groupes des facteurs de confusion.

La mortalité des patients à mauvais pronostic est de 60% tandis qu'elle est de 5% pour les patients à bon pronostic. L'effet du traitement est le même quel que soit le pronostic des patients (réduction relative de 50%). Lorsque l'analyse est réalisée sans tenir compte du pronostic, le résultat s'avère non statistiquement significatif. Par contre un résultat statistiquement significatif est obtenu après ajustement sur le pronostic. Le gain de l'ajustement est d'autant plus important qu'il existe un contraste prononcé entre les strates.

Tableau 1 – Comparaison des résultats des analyses non ajustées et ajustées en fonction d'une variable pronostique illustrant le gain en précision et puissance apportée par un ajustement sur une variable pronostique.

	Décès / n		RR	*p*
	G. traité	G. contrôle	[IC 95%]	
Bon pronostic	5 / 200 2,5%	10 / 200 5,0%	0,50 [0,17 ; 1,44]	-
Mauvais pronostic	6 / 20 30%	12 / 20 60%	0,50 [0,23 ; 1,07]	-
Analyse non ajustée	11 / 220 5%	22 / 220 10%	0,50 [0,25 ; 1,01]	*p* = 0,052
Analyse ajustée	**-**	**-**	**0,50 [0,27 ; 0,93]**	***p* = 0,027**

Exemple

La Figure 1 montre les résultats des analyses ajustée et non ajustée d'une étude de forte puissance et où il n'existe pas de déséquilibre entre les groupes, l'ajustement apporte peu de gain en puissance. Les estimations des 2 approches sont similaires.

	Studied (n=3803)	***Placebo (n=3796)***	***Unadiusted ratio (95% CI)***	***P***	***Adiusted ratio (95% CI)****	***P***
Cardiovascular death or hospital admissions for CHF	*1150 (30-2%)*	*1310 (34-5%)*	*0-84 (0-77-0-91)*	*<0-0001*	*0-82 (0-75-0-88)*	*<0-0001*
Cardiovascular death	*691 (18-2%)*	*769 (20-3%)*	*0-88 (0-79-0-97)*	*0-012*	*0-87 (0-78-0-96)*	*0-006*
Hospital admission for CHF	*757 (19-9%)*	*918 (24-2%)*	*0-79 (0-72-0-87)*	*<0-0001*	*0-77 (0-70-0-84)*	*<0-0001*
Cardiovascular death, hospital admission for CHF, MI	*1213 (31-9%)*	*1369 (36-1%)*	*0-84 (0-78-0-91)*	*<0-0001*	*0-82 (0-76-0-89)*	*<0-0001*
Cardiovascular death, hospital admission for CHF, MI, stroke	*1269 (33-4%)*	*1420 (37-4%)*	*0-85 (0-79-0-92)*	*<0-0001*	*0-83 (0-77-0-90)*	*<0-0001*

Cardiovascular death, hospital admission for CHF, MI, stroke, coronary revascularisation procedure	*1404 (36-9%)*	*1549 (40-8%)*	*0-86 (0-80-0-93)*	*<0-0001*	*0-85 (0-79-0-92)*	*<0-0001*

Figure 1 – Tableau de résultat de l'analyse ajustée et non ajustée d'un essai thérapeutique

Critère de jugement quantitatif

Dans l'exemple représenté dans la

Figure 2, en l'absence d'ajustement, les mesures des deux groupes se recouvrent largement et il est difficile de mettre en évidence une différence. Par contre, lorsque la valeur de x *est prise en compte, les deux groupes apparaissent séparés (les points se retrouvent alignés suivant 2 droites distinctes). La variable* x *explique en totalité la variabilité des mesures* y *et la différence entre les deux groupes devient évidente.*

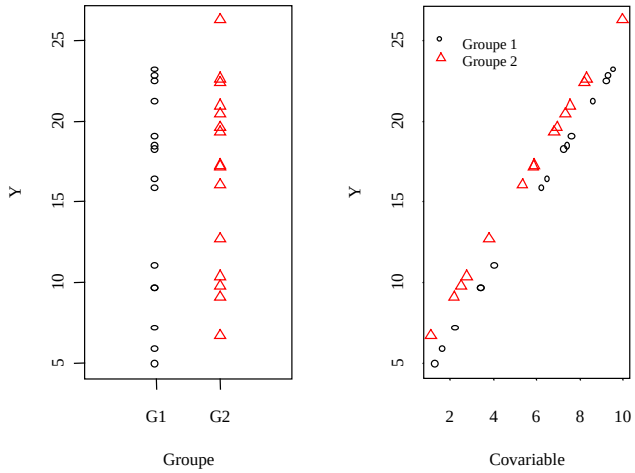


Figure 2 – Sans ajustement (figure de gauche) les mesures y des deux groupes se chevauchent largement. Après ajustement sur la covariable x, la différence entre les deux groupes devient évidente car le traitement et la covariable x expliquent la totalité de la variabilité des mesures y. La différence existant entre les groupes est visualisée le fait que les points des 2 groupes s'alignent sur 2 droites distinctes. A valeur de X identique, la valeur de Y dans le groupe 2 est toujours plus importante que dans le groupe 1.

Dans la réalité, la ou les covariables n'expliquent pas la totalité de la variabilité des mesures mais leur prise en compte peut réduire la variance résiduelle et augmenter ainsi la précision de l'estimation de la différence entre les deux groupes (

Figure 3). Dans cette figure, les doubles flèches verticales matérialisent la variabilité des mesures. Sans ajustement sur la covariable x, *la variabilité des mesures comparées est importante donnant une mauvaise précision dans l'estimation de l'effet traitement. Après ajustement sur la covariable* x *par une analyse de covariance, l'effet traitement se traduit par la distance verticale entre les deux droites de régression. La comparaison des deux droites utilise une erreur dépendant de l'erreur d'estimation de la pente des droites et, surtout, de la variabilité résiduelle des points autour de chaque droite de régression qui est bien plus faible que la variabilité des mesures avant ajustement. La précision dans l'estimation de l'effet traitement est ainsi bien plus élevée. Le lecteur intéressé par le développement mathématique se référera à un manuel de statistique (par exemple page 302 de la réf (1))*

Ainsi, l'ajustement sur une ou des covariables liées aux mesures, en expliquant une partie de la variabilité totale des mesures, réduit la variabilité qui bruite la recherche de l'effet du traitement.

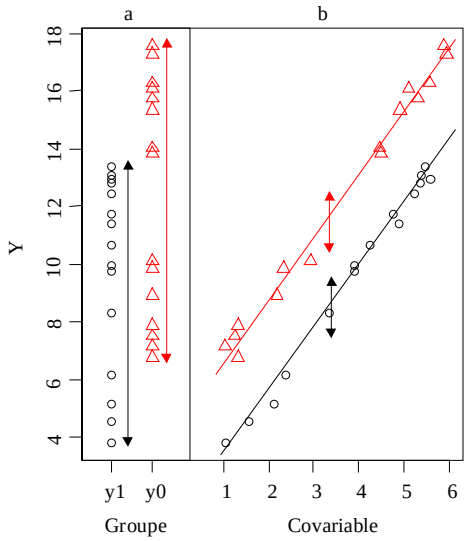


Figure 3 – Ajustement sur la covariable x par une analyse de covariance. Les doubles flèches visualisent la variabilité des mesures. La sous-figure a de gauche représente les mesures brutes sans ajustement. La sous-figure b de droite représente l'ajustement sur x par analyse de covariance.

Supprimer l'effet des facteurs de confusion

La différence observée au niveau du critère de jugement entre les deux groupes d'un essai peut être influencée par un déséquilibre initial dans les variables pronostiques (ce qui conduit à un biais de sélection). Ces covariables ont une valeur de facteur de confusion. Le but de l'ajustement est d'essayer de corriger la différence totale observée de ce qui est due à un déséquilibre initial au niveau des variables pronostiques. Une fois corrigé, il reste la différence due au traitement, celle qui ne peut pas être expliquée par les covariables.

Critère de jugement quantitatif

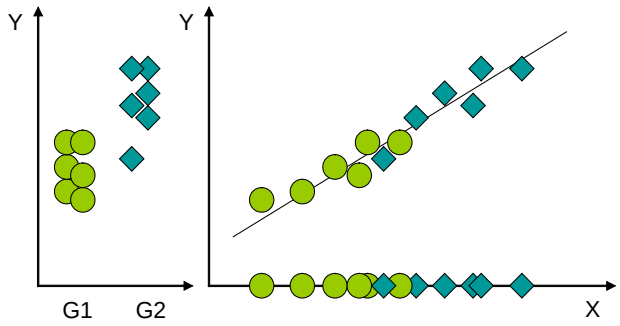


Figure 4 – Illustration d'un cas où la covariable X est « confondante » pour la recherche d'une différence entre les 2 groupes G1 et G2 et la valeur du critère de jugement Y. Explication dans le texte.

La

Figure 4 illustre une situation où la covariable X entraine une confusion dans la recherche d'une différence entre les 2 groupes de traitement G1 et G2. Sans ajustement (partie gauche de l'illustration) il semble qu'il existe une différence de Y entre G1 et G2. Cependant (partie droite de l'illustration) cette apparente différence

est entièrement due au fait que les valeur de X dans G1 sont plus petite que les valeur de X dans G2 ***et*** *qu'il existe une relation très forte entre la valeur de X et la valeur de Y.*

En fait la différence que l'on observe entre G1 et G2 au niveau de Y est entièrement due à la différence qu'il existe entre G1 et G2 au niveau de X. Le groupe ne conditionne en rien la valeur de Y. Deux points, l'un de G1 et l'autre de G2, ayant la même valeur de X (ce qui est presque le cas au niveau du nuage de point) ont la même de Y (ce qui montre que le groupe n'influence pas la valeur de Y).

Critère de jugement binaire

Dans le Tableau 2, un déséquilibre de gravité de l'état des patients existe entre les deux groupes traité et non traité : il y a bien plus de patients à mauvais pronostic dans le groupe traité (141/200 = 70%) que dans le groupe contrôle (61/199 = 31%). Cette différence de pronostic en défaveur du groupe traité va contrer l'effet du traitement. Dans une analyse non ajustée, aucun effet du traitement n'est mis en évidence avec un RR de 0,86 non statistiquement significatif. Par contre, en corrigeant de l'effet du déséquilibre de pronostic par une analyse ajustée, un effet statistiquement significatif apparaît avec un risque relatif de 0,50. Le résultat de l'analyse non ajustée apparaît d'ailleurs incohérent avec les résultats trouvés dans chaque strate (RR=0,50), incohérence qui disparaît avec le résultat ajusté.

Une situation de ce type est rare dans un essai randomisé. La randomisation assure, en moyenne, la répartition harmonieuse entre les groupes des variables pronostiques. En l'absence de déséquilibre, l'analyse non ajustée donne la même estimation que l'analyse ajustée.

La randomisation, éventuellement stratifiée, contrôle les facteurs de confusion a priori. L'ajustement tente de contrôler les facteurs de confusion a posteriori.

Tableau 2 – Apport de l'ajustement pour corriger un effet de confusion.

	Décès / n		RR	*p*
	G. traité	G. contrôle	[IC 95%]	
Mauvais pronostic	35 / 141 25%	30 / 61 49%	0,50 [0,34 ; 0,74]	-
Bon pronostic	3 / 59 5%	14 / 138 10%	0,50 [0,15 ; 1,68]	-
Analyse non ajustée	38 / 200 19%	44 / 199 22%	0,86 [0,58 ; 1,27]	NS
Analyse ajustée	**-**	**-**	**0,50 [0,35 ; 0,73]**	***p* < 0,001**

Le principe de l'ajustement est de rechercher l'effet du traitement pour chaque niveau de la variable d'ajustement (ici, les deux classes de pronostic), puis de regrouper ces estimations. Le principe est identique à celui d'une méta-analyse. D'autres techniques statistiques peuvent être utilisées comme les méthodes de régression multivariée (par exemple la régression logistique).

Outils d'ajustement

Différentes techniques réalisent des analyses ajustées. Les méthodes les plus simples sont représentées par les tests ajustés : test de Mantel-Haenszel (2, 3) pour les critères binaires, test du logrank stratifié pour les données de survie, test t stratifié pour les variables continues. Ces tests stratifiés ne permettent de prendre en compte qu'un petit nombre de covariables (1 à 3). Pour réaliser des ajustements sur un nombre plus important de covariables, il est plus pratique de recourir à l'analyse multivariée comme la régression logistique, le modèle de Cox ou la régression linéaire multiple.

Règles d'ajustement

L'ajustement dans un essai randomisé n'est pas aussi indispensable que dans les études d'observation. Dans les essais stratifiés, un ajustement sur les variables de stratification est à réaliser systématiquement afin d'obtenir le gain en précision qu'engendre la stratification.

Il n'est pas légitime d'ajuster sur les variables qui s'avèrent a posteriori déséquilibrées entre les groupes ou fortement liées au critère de jugement.

Si un ajustement est souhaité, les covariables d'ajustement doivent être décidées a priori et non pas choisies en fonction de leur déséquilibre entre les groupes ou de leur liaison avec le critère de jugement observée dans l'essai. La détermination des covariables d'ajustement à partir des données de l'essai conduit à un risque de biais et perturbe l'inférence statistique. Ainsi les variables d'ajustement doivent être choisies a priori, uniquement à partir de la connaissance de leur valeur pronostique (4). En effet, dans un essai, la randomisation répartit les facteurs de risque de façon équilibré en moyenne sur l'ensemble des facteurs de risque. Tous les facteurs de risque ne sont pas connus ou mesurés. En cas d'ajustement sur les facteurs mesurés dont le déséquilibre est compensé par des facteurs de risque non mesurés ou non connus, l'ajustement peut conduire à une fausse estimation de l'effet du traitement.

Avec le modèle de Cox, il est souhaitable d'ajuster systématiquement sur les variables pronostiques connues car il a été montré que l'analyse brute non censurée conduit à un biais dans l'estimation de la taille de l'effet du traitement (5).

Conclusion

Dans un essai thérapeutique randomisé, l'ajustement pour tenir compte d'un déséquilibre des variables pronostiques entre les groupes n'est en général pas nécessaire lorsque les effectifs sont importants. Si malgré la randomisation un déséquilibre existe, l'ajustement ne corrige pas totalement le problème. Il n'est possible que sur les facteurs de confusion connus. Or il est rare que les facteurs connus expliquent une forte proportion de la variabilité totale. Ainsi même après ajustement, il reste la possibilité de biais dû aux facteurs de confusion inconnus. La nécessité de recourir à un ajustement doit donc toujours faire suspecter un biais dans le résultat et faire émettre des réserves.

L'utilisation, non prévue a priori, de l'ajustement pose un problème dans l'analyse d'un essai thérapeutique, en particulier lorsque le recours à l'ajustement est indispensable pour atteindre la signification statistique (6).

L'ajustement est donc seulement à utiliser dans l'essai thérapeutique pour augmenter puissance et précision, en ajustant sur des covariables définies a priori dans le protocole.

L'ajustement à 2 intérêts donc deux utilisations possibles :

1. *corriger un résultat du biais induit par un déséquilibre des facteurs de confusion entre les groupes,*
2. *augmenter la précision (donc la puissance) de l'estimation de l'effet du traitement.*

Dans l'essai thérapeutique, la première utilisation est problématique surtout si les variables d'ajustement sont choisies a posteriori, en fonction des déséquilibres observés.

La deuxième utilisation ne pose pas de problème dans un essai thérapeutique mais doit être un ajustement décidé ***a priori*** *reposant sur les variables de stratification de la randomisation ainsi que sur d'autres variables connues comme étant fortement pronostiques.*

Bibliographie

1. *Armitage P, Berry G. Statistical methods in medical research. Oxford: Blackwell Scientific Publications; 1994.*

2. *Bouyer J, Hémon D, Cordier S, Derriennic F, Strücker I, Stengel B, et al. Epidémiologie. Principes et méthodes quantitatives. Paris: Les Editions INSERM; 1995.*

3. *Cucherat M, Boissel JP, Leizorovicz A. La méta-analyse des essais thérapeutiques. Paris: Masson; 1997.*

4. *Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Controlled Clinical Trials 2000;21:330-342.*

5. *Chastang C, Byar DP, Piantadosi S. A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. Stat Med 1988;7:1243-1255.*

6. *Buyse M. Analysis of clinical trial outcomes: some comments on subgroup analyses. Controlled Clinical Trials 1989;10:187S-194S.*

Puissance et nombre de sujets nécessaires

Powerpoint

Définition

La puissance statistique d'un essai thérapeutique mesure son aptitude à mettre en évidence l'effet d'un traitement si celui-ci existe.

La puissance statistique d'un essai clinique est son aptitude (en termes de probabilité) d'obtenir un résultat statistiquement significatif si le traitement est réellement efficace. La puissance est égale à 1-b, où b est le risque de deuxième espèce, celui de ne pas mettre en évidence un effet qui existe pourtant. Le risque b est la probabilité d'obtenir un faux résultat négatif (ne pas mettre en évidence l'efficacité d'un traitement qui existe pourtant). La puissance est donc la probabilité d'obtenir un vrai résultat positif (mettre en évidence l'efficacité d'un traitement).

Un essai suffisamment puissant a une forte probabilité d'obtenir un résultat significatif si le traitement a l'efficacité escomptée. Un essai insuffisamment puissant a une faible probabilité de mettre en évidence l'effet du traitement qui existe pourtant.

La puissance est similaire au pouvoir grossissant d'un microscope (

Figure 1). Un grossissement suffisant est nécessaire pour montrer que deux points très proches l'un de l'autre, mais cependant séparés, sont distincts. Avec un grossissement insuffisant, ces deux points paraissent ne faire qu'un. Plus la distance entre les 2 points est petite, plus le pouvoir grossissant devra être élevé pour visualiser 2 points distincts. Il en est de même avec la recherche d'une différence entre deux groupes. Une puissance statistique suffisante est nécessaire pour montrer qu'il existe effectivement une différence entre les 2 groupes. Plus la différence entre les 2 groupes est petite, plus il faudra de puissance statistique pour montrer que les 2 groupes sont différents.

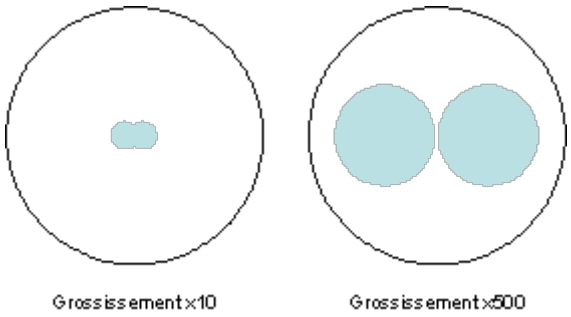


Figure 1 – Illustration de l'analogie entre puissance et pouvoir grossissant du microscope

La puissance dépend de plusieurs paramètres

La puissance statistique d'un essai utilisant un critère de jugement binaire dépend de plusieurs paramètres : la taille de l'effet à mettre en évidence, le nombre de sujets, le risque de base (risque sans traitement) et le risque d'erreur statistique alpha consenti.

La taille de l'effet à mettre en évidence est le paramètre qui conditionne en premier la puissance d'un essai. Plus l'effet du traitemer *our le mettre en évidence. Un même essai sera d'auta* *(figure 1). Ce paramètre n'est pas contrôlable par l'investi* *, en quelque sorte sa "puissance pharmacologique" (ou "*

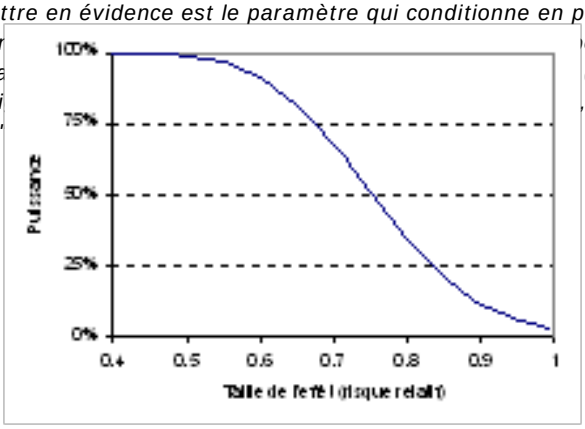


Figure 2 – Relation entre la puissance et la taille de l'effet (pour un effectif par groupe de 1000 patients et un risque de base de 10%). Plus le risque relatif est proche de 1, moins le traitement est efficace.

La puissance dépend aussi du nombre de sujets inclus dans l'essai. Plus le nombre de patients est important plus l'essai est puissant (figure 2). L'effectif de l'essai est le paramètre sur lequel l'investigateur peut le plus directement agir pour contrôler la puissance de son essai. En particulier lorsque l'effet recherché est petit, il est nécessaire d'inclure un grand nombre de patients. Par contre un effectif plus faible est suffisant pour mettre en évidence des effets conséquents.

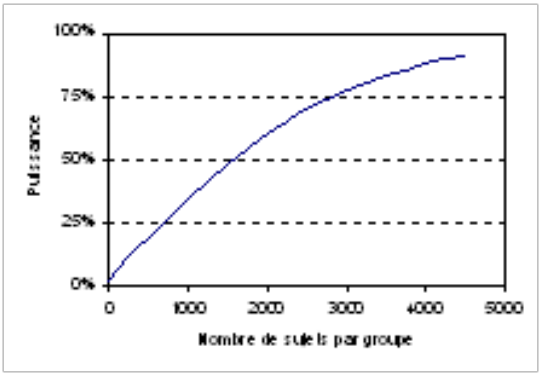


Figure 3 – Relation entre la puissance et le nombre de sujets par groupe (pour un risque de base de 10% et un risque relatif de 0,8).

La fréquence de base des événements (le risque de base) est un autre paramètre qui conditionne la puissance d'un essai. Il faut plus de puissance pour mettre en évidence un même effet sur un événement rare que sur un événement fréquent. Il faut donc plus de patients à faible risque que de patients à haut risque pour mettre en évidence un effet . Le risque de base est un paramètre sur lequel l'investigateur peut partiellement agir. En recrutant des patients à haut risque il se met dans une situation où il sera plus facile de mettre en évidence un effet.

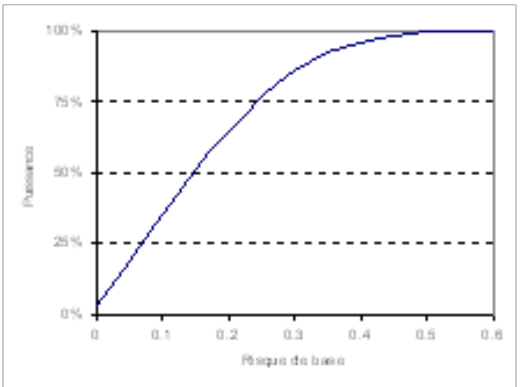


Figure 4 – Relation entre la puissance et le risque de base (pour un effectif par groupe de 1000 patients et un risque relatif de 0,8).

En dernier, la puissance dépend du risque alpha choisi. Risque alpha et puissance varient en sens inverse. Ainsi adopter un risque alpha inférieur à 5% nécessite plus de patients, ce qui explique pourquoi cela est rarement fait.

Lorsque le critère de jugement est continu, la variance du critère remplace la fréquence de base. Plus la variabilité entre sujet du critère de jugement est faible, plus la puissance est importante. Ainsi un même essai sera d'autant plus puissant que la variabilité du critère de jugement est faible. Pour maximiser la puissance de l'essai, il convient donc de réduire au maximum la variabilité des valeurs, en utilisant, par exemple, des groupes très homogènes de patients et en réduisant les erreurs de mesure, par exemple, en ayant recours à un laboratoire centralisé. Un essai utilisant le patient comme son propre témoin nécessite en général moins de patients qu'un essai en bras parallèles car la variabilité intra-sujet est inférieure (ou égale) à la variabilité inter-sujet.

La puissance d'un essai augmente avec :

- *le nombre de patients inclus*
- *l'importance de l'effet recherché*
- *la fréquence sans traitement de l'événement*

Puissance et intervalle de confiance

La largeur de l'intervalle de confiance reflète la puissance statistique de l'essai : plus la puissance statistique est élevée, plus l'intervalle de confiance est étroit. Ainsi, à vrai risque relatif et à risque de base constant, la largeur de l'intervalle de confiance dépend du nombre de sujets : plus l'effectif est important, plus l'intervalle de confiance est étroit. La précision de l'estimation de l'effet traitement est inversement proportionnelle à la largeur de l'intervalle de confiance. Donc plus la taille d'un essai est importante, plus il estime avec précision l'effet traitement.

Avec le risque relatif, un résultat est statistiquement significatif à partir du moment où l'intervalle de confiance ne contient pas la valeur 1 (marquant l'absence d'effet). Ainsi pour qu'un résultat soit statistiquement significatif, la largeur de l'intervalle de confiance doit donc être d'autant plus petite que le risque relatif est proche de 1 (

Figure 5). Comme la largeur de l'intervalle de confiance est directement liée au nombre de patients, il devient clair qu'un plus grand effectif est nécessaire pour mettre en évidence un petit effet qu'un effet plus important.

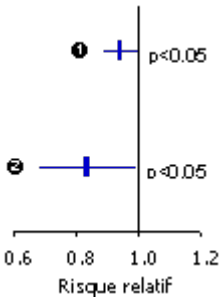


Figure 5 – Avec un traitement peu efficace (1), un résultat significatif est obtenu avec une largeur d'intervalle de confiance plus petite qu'avec un traitement très efficace (2).

Détermination du nombre de sujets nécessaires

Principe

Un nombre de sujets adapté à la taille de l'effet à mettre en évidence garantit à un essai une puissance suffisante.

Un essai peu puissant n'a généralement pas d'intérêt car il a peu de chance de mettre en évidence l'effet du traitement. Il représente donc un investissement non rentable. Afin de ne pas réaliser des essais sans intérêt, il convient de leur assurer une puissance statistique suffisante. Cela est fait en calculant a priori l'effectif nécessaire.

En effet, dans un contexte donné (taille de l'effet recherché et fréquence de base de l'événement), la puissance ne dépend plus que du nombre de sujets. Ce nombre de sujets est déterminé a priori afin de garantir la puissance statistique de l'essai. Des formules et des logiciels existent pour faire ce calcul. Ces formules nécessitent de connaître ou de faire des hypothèses sur les paramètres conditionnant la puissance : risque de base, taille de l'effet à mettre en évidence, risque alpha (en général 5%) et puissance souhaitée (en général 90%).

En fait le calcul d'un nombre de sujets ne garantit pas à 100% que l'essai aura la puissance nécessaire. Tout dépend de l'exactitude des hypothèses faites pour son calcul. Un effet traitement qui s'avère en réalité plus petit que celui initialement prévu fait que l'essai devient insuffisamment puissant. De même si le risque des patients effectivement inclus dans l'essai est inférieur à l'hypothèse utilisée pour le calcul, l'essai n'a plus la puissance nécessaire. La difficulté du calcul du nombre de sujets est dans l'estimation a priori de ces paramètres. En particulier, la taille de l'effet du traitement est souvent difficile à déterminer. Le traitement entraîne-t-il une réduction de fréquence du critère de jugement de 10%, 15% ou bien 20% ? Une solution consiste à prendre un effet relativement faible, en disant que si en réalité le véritable effet est encore plus petit, il sera sans intérêt en pratique, et donc, dans ces conditions, il n'est pas dramatique de ne pas pouvoir le mettre en évidence.

Plus l'effet recherché est petit, plus le nombre de sujets nécessaires est important. Il faut bien plus de patients pour comparer deux traitements actifs que pour comparer un traitement actif contre placebo. Lorsque les patients du groupe contrôle reçoivent déjà un traitement actif, la taille de l'effet est le risque de base sont plus petits. La mise en évidence du bénéfice apporté par la thrombolyse à la phase aiguë de l'infarctus du myocarde a nécessité entre 11806 patients (GISSI (1)) et 17187 patients (ISIS-2 (2)) lorsque le comparateur était le placebo. Par contre, lorsqu'il a été nécessaire de comparer les fibrinolytiques entre eux, un nombre plus considérable de patients a été nécessaire : GUSTO-1 41 021 patients (3), ISIS-3 41 299 patients (4). Contre placebo, la streptokinase entraîne une réduction relative de mortalité à 30 j de -23%. Mais lorsque la streptokinase est utilisée comme comparateur, la réduction supplémentaire de mortalité apportée par d'autres fibrinolytiques est bien plus faible, -10% pour l'alteplase par exemple. La mortalité sous placebo est de 12% (ISIS 2) mais s'abaisse à 7,2% sous streptokinase (GUSTO-1).

Plus l'événement critère de jugement est rare, plus le nombre de sujets nécessaires est important. D'un point de vue purement statistique, il semblerait donc avantageux de sélectionner soigneusement les patients recrutés afin qu'ils soient le plus à risque possible. Cette pratique a cependant pour principal inconvénient "d'hyper sélectionner" les patients et de rendre la population de l'essai non représentative de la population des patients

tout venant. De plus, des critères sélectifs rendent les patients recherchés rares et augmentent la durée de recrutement.

Le nombre de sujets nécessaires augmente quand :
• *la taille de l'effet à mettre en évidence diminue* • *la fréquence de base de l'événement diminue*

Méthode de calcul du nombre de sujets nécessaires

Le raisonnement détaillé du calcul du nombre de sujets nécessaires est le suivant.

D'une manière générale, la largeur de l'intervalle de confiance dépend du nombre de sujets. Plus l'effectif est important plus l'intervalle est étroit (

Figure 6).

Par exemple avec la moyenne la borne supérieure bs *d'un intervalle de confiance (de la moyenne) est*

$$bs = \overline{x} + 1.96\frac{s}{\sqrt{n}}$$

où s *désigne l'écart type, et* $\overline{x}$ *la moyenne. Ces 2 paramètres étant constant, la borne supérieure est d'autant plus éloignée de* $\overline{x}$ *que* n *est petit.*

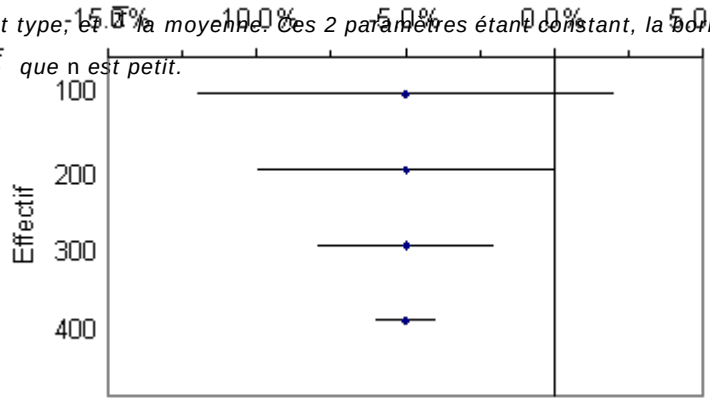


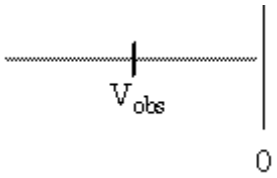
Figure 6 – Influence du nombre de sujets sur la largeur de l'intervalle de confiance illustrée ici avec la différence des risques (mais cette relation est universelle et se retrouve avec tous les indices d'efficacités)

Il y a une relation entre signification statistique et intervalle de confiance : quand l'intervalle de confiance inclus la valeur de l'absence d'effet (0 avec la différence des risques ou la différence des moyennes, 1 avec le risque relatif) le test n'est pas statistiquement significatif. Par contre quand l'intervalle exclu cette valeur de l'absence d'effet, le test est significatif (à un seuil de signification égal à 1-le degré de confiance de l'intervalle, 5% pour un intervalle de confiance à 95%).

Dans la suite nous allons illustrer le propos en utilisant comme indice d'efficacité la différence des risques, mais tout le raisonnement s'applique aussi au risque relatif, à la différence de moyennes, etc. Ainsi pour atteindre la

signification statistique, il convient que la borne supérieure de la différence de risque soit juste inférieure à zéro. Le but du calcul de l'effectif est de déterminer le nombre de sujets n *qui donne un intervalle de confiance de largeur telle que la borne supérieure soit juste inférieure à zéro (*bs<*0).*

Sans rentrer dans le détail calculatoire, il est possible de dériver une formule qui donne n *en fonction de la valeur de la borne supérieure (le lecteur avide de formules mathématiques trouvera celles-ci en annexe).*



En plus de l'effectif n*, la valeur de la borne supérieure dépend de la valeur* V_{obs} *de la différence de risque qui sera observée dans l'essai. Cette valeur reflète la vraie efficacité du traitement mais elle est soumise aux fluctuations aléatoires d'échantillonnage. Par hasard, la valeur observée dans l'essai peut surestimer ou sous-estimer la vraie valeur. En fait,* V_{obs} *fluctue autour de la vraie valeur* V *avec une certaine distribution comme cela est illustré*

Figure 7.

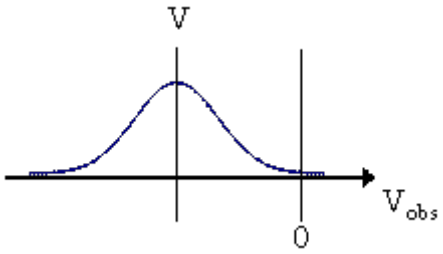


Figure 7 – Distribution des valeurs observée autour de la vraie valeur V*. Les valeurs observées dans l'essai fluctuent autour de la vraie valeur de l'efficacité du traitement* V*. Par hasard, certains essais peuvent surestimer l'efficacité (*V_{obs}>V*) d'autre la sous estimer (*V_{obs}<V*). Les valeurs très éloignées de* V *sont moins probables que les valeurs proches de* V*.*

Ces fluctuations influencent le calcul du nombre de sujets nécessaires. En effet, plus la valeur observée est importante (proche de zéro), plus il faudra un intervalle de confiance étroit pour atteindre la signification statistique. Comme a priori on ne sait pas qu'elle valeur sera observée dans l'essai (cela dépend du hasard), l'idée est de calculer l'effectif pour une valeur particulièrement défavorable de V_{obs}*, que l'on estime peu probable. Cette valeur de référence du calcul est notée* V_{ref}*. Ainsi dans la réalité, on aura de forte chance que la valeur effectivement observée soit inférieure à celle choisie pour le calcul, et le test sera donc forcément statistiquement significatif.*

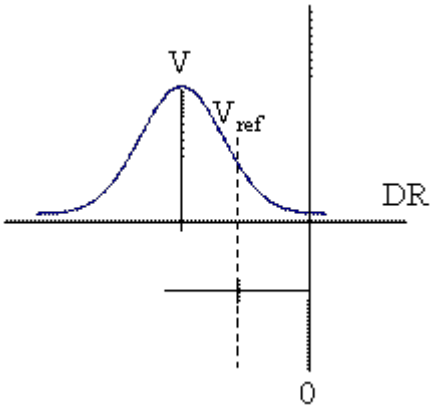


Figure 8 – Le calcul du nombre de sujet est fait pour une valeur particulière de V_{obs}*, notée* V_{ref}*, que l'on estime peu probable. On calcul donc n de telle façon que la largeur de l'intervalle de confiance pour cette valeur de référence donne une borne supérieure juste inférieure à 0, donc un test juste significatif (p<0.05 pour un intervalle de confiance à 95%). Ainsi le résultat sera statistiquement significatif lorsque l'essai par hasard sous estimera la vraie efficacité du traitement jusqu'à une sous estimation de* V_{ref}*.*

Si par hasard la valeur observée tombe au dessus de cette valeur de référence V_{ref}*, le résultat obtenu ne sera pas significatif (*

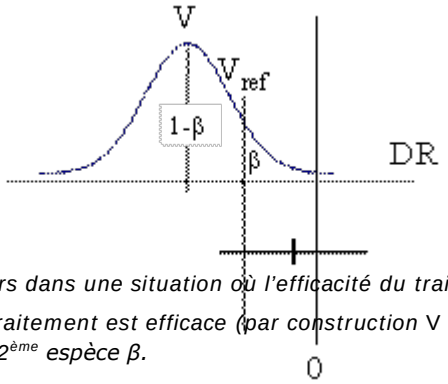


Figure 9). On se retrouve alors dans une situation où l'efficacité du traitement n'est pas mise en évidence (test non significatif) alors que le traitement est efficace (par construction V *est non nul). Cette situation est donc celle du risque statistique de* $2^{ème}$ *espèce β.*

Figure 9 – Situation où la valeur observée tombe au-delà de la valeur de référence (V_{ref}*) ayant servie au calcul du nombre de sujets nécessaires (cf. figure précédente), conduisant ainsi à un résultat non statistiquement significatif.*

V_{ref} *sera donc déterminé à partir de la vraie valeur de telle façon qu'il y ai qu'une probabilité β que la valeur observée soit supérieure à* V_{ref} *et donc que le résultat ne soit pas significatif. La puissance est alors égale à*

$1-\beta$. V_{ref} *est donc égal au (1-β)ème percentile de la distribution considérée et sa valeur est déterminée à partir d'une table ou d'un logiciel.*

> Par exemple, pour une distribution normale centrée sur zéro et d'écart type égale à 1, le 80ème percentile est égal à 0,84. Le 97,5ème est égal à 1,96.

Au total pour faire ce calcul d'effectif il faut disposer d'une valeur pour la vraie efficacité V, *de la distribution des valeurs observées autour de* V *(c'est-à-dire connaître la variabilité des valeurs observées, donc l'écart type ou la variance), et fixer un risque bêta consenti (et un risque alpha).* V, *l'écart type de* V *et bêta permettent de calculer* V_{ref}. *Puis on calcul l'effectif* n *nécessaire pour que la borne supérieure de l'intervalle de confiance autour de* V_{ref} *soit juste inférieure à zéro.*

Conséquences

Il apparait ainsi que la puissance est un paramètre de ***protection contre les sous estimations*** *du vrai effet traitement liées aux fluctuations aléatoires. Une puissance élevée permet de conclure même en cas de sous estimation importante de l'efficacité. L'essai montre alors une faible efficacité du traitement débouchant sur un résultat peu cliniquement pertinent. Pour cette raison il est inutile de rechercher une puissance démesurée car ce surcroît de protection coûte cher en nombre de patients et, s'il s'avère nécessaire, débouche sur un résultat peu favorable au traitement. En général une puissance de 80% est raisonnable.*

Lorsqu'un essai à une puissance inférieure à 80% il peut bien évidement obtenir un résultat significatif. Il sera seulement « moins à l'abri » des sous-estimations induites par le hasard. Avec une puissance de 50%, l'essai ne sera significatif que si l'essai estime correctement ou surestime l'effet du traitement. Apparaît ici un point important de l'interprétation des essais significatifs de faible puissance : il risque fort de donner une image trop optimiste de l'efficacité du traitement.

Annexe : formules mathématiques

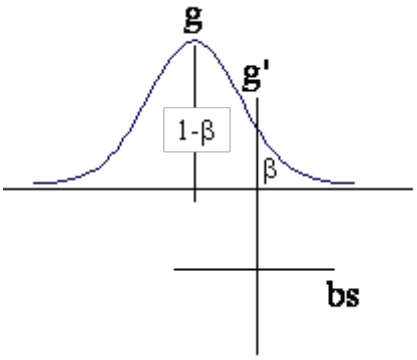
Le développement mathématique permettant d'obtenir les formules de calcul de l'effectif est donné à titre documentaire. En pratique, les calculs s'effectuent avec des logiciels que l'on trouve maintenant sur Internet (comme MfCalc à l'adresse www.spc.univ-lyon1.fr/mfcalc *).*

D'une manière générale, les bornes bi *et* bs *d'un intervalle de confiance d'un indice d'efficacité quelconque* g *s'obtiennent par*

$$bi, bs = g \pm z_{\alpha/2} SE(g)$$

où g *est la valeur observée du paramètre d'intérêt et* SE(g) *son erreur standard,* $Z_{\alpha/2}$ *étant la valeur du* $1-\frac{\alpha}{2}$ *ème percentile de la distribution normale. Par exemple pour α=5%,* $Z_{\alpha/2}=1,96$.

Le paramètre g peut être une différence de moyenne, une différence de risque, un risque relatif, un odds ratio, etc.



En adoptant les notations de la figure ci-dessus, la valeur de référence du calcul g' *s'obtient par*

$$g' = g + z_{\beta} SE(g)$$

La borne supérieure de l'intervalle de confiance autour de g' est

$$bs = g' + z_{\alpha/2} SE(g')$$

qui peut s'écrire

$$bs = g + z_{\beta} s + z_{\alpha/2} s$$

en posant

$$s = SE(g) = SE(g')$$

C'est-à-dire

$$bs = g + \left(z_{\beta} + z_{\alpha/2}\right) s$$

L'effectif $n = n_1 + n_0$ *que l'on cherche à calculer est contenu dans l'expression de* s. n_1 *désigne l'effectif du groupe traité et* n_0 *celui du groupe contrôle.*

- *Pour une différence de moyenne* SE(g) *est*

$$s = SE(g) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}$$

avec s_1 *et* s_0 *qui désignent respectivement les écarts types inter sujets des groupes traité et contrôle.*

- *Pour une différence de risque,* SE(g) *est*

$$s = SE(g) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

où p_1 *et* p_0 *représente les fréquences observées de l'événement dans des groupes traité et contrôle.*

- *pour un risque relatif on prend comme paramètre* g *le logarithme du risque relatif*

$$g = Log(RR)$$
$$s = SE(g) = \sqrt{\frac{1}{x_1} - \frac{1}{n_1} + \frac{1}{x_0} - \frac{1}{n_0}}$$

où x_1 *et* x_0 *représente le nombre d'événements observés dans les groupes traité et contrôle.*

En général, on souhaite que les effectifs des groupes soient identiques ce qui revient à écrire $n_1 = n_0 = n$. 2n *désignant alors l'effectif total de l'essai (des 2 groupes).*

Le calcul de l'effectif revient donc à solutionner l'expression donnant bs *pour* n

$$bs = 0$$

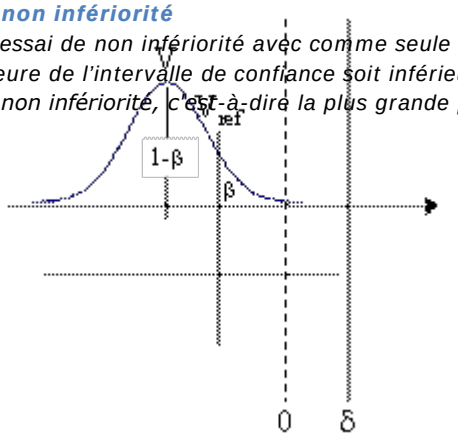
Par exemple pour le risque relatif cela revient à résoudre pour n *l'équation :*

$$\begin{aligned} bs &= g + \left(z_{\beta} + z_{\alpha/2}\right)\left[\frac{1}{x_1} - \frac{1}{n} + \frac{1}{x_0} - \frac{1}{n}\right]^{1/2} \\ &= g + \left(z_{\beta} + z_{\alpha/2}\right)\left[\frac{1}{p_1 n} + \frac{1}{p_0 n} - 2\frac{1}{n}\right]^{1/2} \\ &= 0 \end{aligned}$$

ce qu'un logiciel fait très bien !

Transposition au cas de l'essai de non infériorité

Le même raisonnement s'applique à l'essai de non infériorité avec comme seule différence le fait que l'on cherche plus à ce que la borne supérieure de l'intervalle de confiance soit inférieure à zéro (effet nul) mais qu'elle soit inférieure à δ (la limite de non infériorité, c'est-à-dire la plus grande perte d'efficacités acceptable).



Puissance a posteriori

Une puissance a posteriori est parfois calculée à partir des résultats de l'essai. L'idée est éventuellement de conclure à l'absence l'effet s'il s'avère que la puissance était malgré tout forte. Ce calcul s'avère délicat. En effet il est tentant de faire le calcul de puissance en utilisant comme estimation du risque de base et de l'effet du traitement les valeurs observées dans l'essai. Pour le risque de base cela est assez logique et permet d'affiner les hypothèses initiales du calcul du NSN. Par contre, prendre comme hypothèse pour l'effet du traitement la valeur observée n'a pas grand sens. En effet, faire cela conduit forcément à une puissance inférieure ou égale à 50% comme le montre la Figure 10. Il est donc inutile de faire le calcul étant donné que ce que l'on recherche (montrer que la puissance était forte) est forcément inatteignable.

En pratique, le seul calcul digne d'intérêt que l'on peut faire a posteriori est de déterminer l'effectif d'un nouvel essai que l'on pourrait éventuellement envisager en réactualisant les hypothèses avec les valeurs observées. L'hypothèse faite sur le vrai effet traitement pourra conserver la valeur utilisée pour le précédent essai ou la modifier en tenant compte du résultat de l'essai.

L'essai réalisé est non significatif (figure du haut). g est le vrai effet traitement utilisé pour le calcul a priori du NSN. g' est la valeur de l'effet traitement observé, qui est non significatif comme le témoigne l'intervalle de confiance qui inclut la valeur 0.

Pour le calcul de la puissance a posteriori (figure du bas), si l'on prend comme nouveau vrai effet traitement la valeur observée g', la distribution des valeurs observée se centre sur g' (flèche 1). Dans ce cas, la puissance de l'essai se visualise en repositionnant la borne supérieure de l'IC sur zéro (flèche 2). Le centre de l'IC permet de visualiser la puissance qui par construction est forcément inférieure à 50%.

Au mieux, si le résultat observé est juste non significatif (p=0.05), cad la borne supérieure de l'intervalle de confiance est égale à 0, la puissance est de 50%.

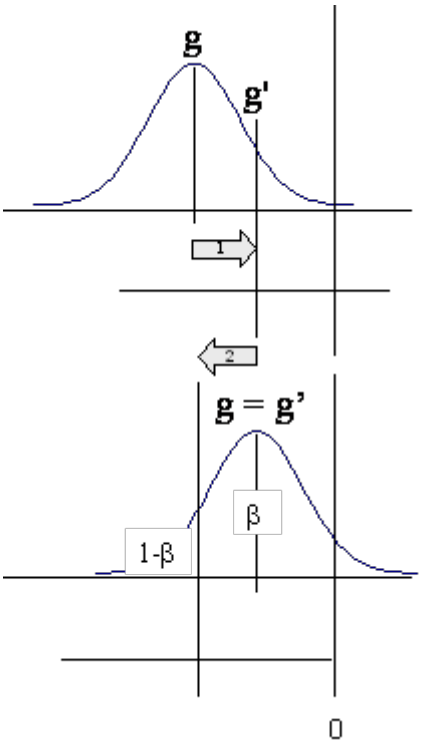


Figure 10 – Illustration du fait que la puissance a posteriori est inférieure ou égale à 50% si l'on prend comme hypothèse pour le vrai effet traitement la valeur observée dans l'essai.

Lecture critique

En lecture critique, les problèmes de puissance perdent un peu de leur importance. En effet, lorsque l'on est devant un résultat significatif, la puissance est un paramètre secondaire. Même si l'étude n'avait pas une puissance jugée comme satisfaisante, le résultat est ce qu'il est, et un manque de puissance ne peut suffire à le récuser. L'essai avait certes une probabilité modérée de mettre en évidence l'effet du traitement, mais il l'a mis en évidence et à partir de ce moment le résultat significatif a la même valeur que s'il avait été obtenu avec un essai de très forte puissance, avec cependant quelques réserves située sur un autre plan.

Si l'essai est de très petite taille dans un domaine où les patients ne sont pas rares, il est alors possible de suspecter que de nombreux essais de cette taille ont été réalisés et que l'essai que l'on est en train d'analyser est celui qui, par hasard (risque alpha), à donner une différence significative. Il ne peut donc pas constituer une preuve formelle de l'existence de l'efficacité. Il s'agit ici d'un problème de biais de publication.

Un essai de très faible puissance (moins de 50%) significatif surestime forcément l'efficacité du traitement et incite à un excès d'optimisme quant à l'efficacité du traitement. Il existe de nombreux exemples où les essais précoces de petite taille (comme des phases II ou IIb) ont donnés des résultats très encourageant non confirmés par les essais de grande taille ultérieur.

Une autre réserve est de nature Bayesienne. La probabilité d'existence de l'effet du traitement après un résultat significatif mais de faible puissance est inférieure à celle obtenue après un essai significatif de forte puissance (cf. section parallèle entre tests diagnostiques et tests statistiques et présentation des risques statistiques comme des taux de filtration).

Par contre, la présence des hypothèses du calcul du nombre de sujets nécessaires est indispensable pour vérifier que l'essai a été à son terme et qu'il n'a pas été arrêté prématurément lors de la réalisation d'analyses intermédiaires « sauvages », c'est-à-dire non protégées contre l'inflation du risque alpha (cf. section sur la répétition des comparaisons statistiques). En effet, si l'effectif initialement visé n'est pas indiqué, il est impossible d'exclure cette possibilité. L'essai a pu être analysé régulièrement au fur et à mesure du recrutement des sujets jusqu'à l'obtention d'une différence statistiquement significative. Le risque alpha n'est plus contrôlé dans ce cas et il est impossible de connaître la réalité statistique du résultat obtenu.

Par contre, si l'effectif calculé a priori est mentionnée dans la publication, il est facile de vérifier que l'essai a été à son terme et que l'analyse présentée correspond bien à celle initialement prévue.

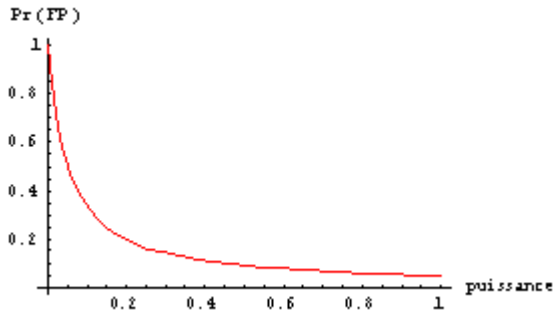
Encart : Relation entre puissance et probabilité d'obtenir un résultat faux positif
Un résultat faux positif est un résultat d'essai statistiquement significatif obtenu avec un traitement sans effet. La probabilité qu'un résultat significatif ne soit en fait qu'un faux positif est d'autant plus élevé que la puissance de l'étude est faible. De ce fait, un résultat pourtant significatif d'un essai de faible puissance doit être pris avec beaucoup de précaution. *La probabilité d'obtenir un faux positif dépende de la puissance de l'essai (cf. section parallèle entre tests diagnostiques et tests statistiques et la section présentation des risques statistiques comme des taux de*

filtration).

En effet il est possible de démontrer que la probabilité d'obtenir un faux positif est égale à :

$$\Pr(FP) = 1 - \frac{W\nu}{W\nu + \alpha(1-\nu)}$$ *où W est la puissance de l'étude, α est le risque alpha et ν la probabilité a priori que le traitement soit efficace.*

La figure suivante illustre l'évolution de la probabilité de faux positifs en fonction de la puissance pour une probabilité a priori de 50%.



Différence non significative

Un résultat non significatif ne permet pas de conclure car il peut correspondre à deux situations différentes qui sont impossibles à départager avec certitude (figure 4). Une différence non significative peut être le reflet d'une réelle absence d'effet du traitement mais peut aussi provenir d'un manque de puissance de l'essai qui n'a pas été en mesure de mettre en évidence une différence qui existe pourtant. Un résultat non statistiquement significatif ne signifie pas que le traitement est sans effet : "L'absence de preuve n'est pas la preuve de l'absence". Devant un résultat non significatif, il n'est pas possible de conclure à l'absence d'effet. La démonstration de l'absence d'effet demande bien plus qu'une simple différence non significative et se base sur un outil spécifique, l'essai d'équivalence.

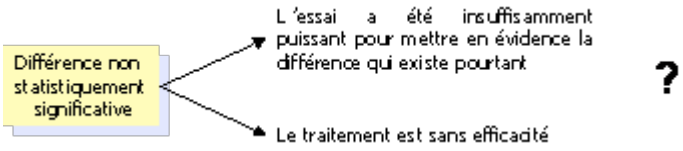


Figure 11 – Signification d'une différence non significative.

Ce point met en évidence l'importance de la puissance statistique d'un essai. Si celle-ci est insuffisante, l'essai ne pourra pas conclure et il n'aura servi à rien. Étant donné le coût, financier et en énergie, d'un essai, il convient de minimiser au maximum le risque de réaliser un essai insuffisamment puissant. En pratique, il est donc nécessaire d'assurer a priori une puissance suffisante à un essai.

Critères continus

Avec les critères continus la puissance dépend des paramètres suivants :

- La différence entre les moyennes, c'est-à-dire la taille de l'effet traitement. Plus l'effet traitement est important, plus la différence entre les moyennes des deux groupes sera large. À nombre de patients identiques, plus cet effet sera important plus l'essai sera puissant.

- La variance des mesures : plus les mesures sont variables, moins l'essai est puissant à nombre de sujets constant. En effet, une variabilité importante des mesures entraîne à son tour une variabilité importante de l'estimation des moyennes.
- Le nombre de patients : la puissance augmente avec le nombre de sujets car la précision d'estimation d'une moyenne augmente avec la taille des échantillons. Plus le nombre de sujets est important, plus les moyennes sont connues avec précision et donc plus il est facile de montrer qu'elles sont différentes (si c'est effectivement le cas).
- Et naturellement le risque alpha consenti.

Corollairement, le nombre de sujets nécessaire avec un critère continu dépend :

- Du risque alpha consenti et de la puissance recherchée.
- De la taille de l'effet traitement à mettre en évidence qui conditionne l'importance de la différence entre les moyennes. Plus l'effet est important moins il faudra de patients pour obtenir une différence significative.
- De la variabilité des mesures. Plus la variabilité du critère entre les unités statistiques est importante, plus il faut de patients pour estimer avec une bonne précision les moyennes, donc pour montrer qu'elles sont différentes. En fait, il s'avère que ce qui conditionne le nombre de sujets est la taille de la différence à mettre en évidence rapportée à la variabilité du critère. Quelle que soit la valeur des moyennes il faut le même nombre de patients pour mettre en évidence une différence de 1 écart type.

Références

1. GISSI (Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto miocardico). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. Lancet 1986;i:397-401.
2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither, among 17187 cases of suspected acute myocardial infarction. Lancet 1988;2:349-360.
3. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. NEJM 1993;329:673-682.
4. ISIS-3 (Third International Study of Infarct Survival) Collaborative Group. ISIS-3: A randomized trial of streptokinase vs tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs aspirin alone among 41299 cases of suspected acute myocardial infraction. Lancet 1992;339:753-770.

Test unilatéral – bilatéral

Test statistique bilatéral - unilatéral

Les tests statistiques peuvent être soit bilatéraux (« *two-tailed, two-sided* ») soit unilatéraux (« *one-tailed, one-sided* ») en fonction des types de conclusions que l'on cherche à faire.

L'objectif d'un test statistique est de montrer qu'il existe une différence entre 2 groupes par exemple en termes de comparaisons de 2 moyennes m_1 et m_2. Deux moyennes peuvent être différentes de 2 manières : soit m_1 est supérieure à m_2 soit m_1 est inférieure à m_2. Avec un test bilatéral on pourra montrer que $m_1 > m_2$ ou que $m_1 < m_2$. Avec un test unilatéral on ne pourra monter que $m_1 > m_2$. On utilise un test bilatéral quand les 2 sens de la différence nous intéresse et un test unilatéral quand seulement la supériorité nous intéresse (ou seulement l'infériorité).
Dans un essai thérapeutique (

Figure 1), un test bilatéral donne la possibilité de conclure de façon statistiquement significative, en fonction de ce qui est observé, soit à la supériorité du traitement étudié soit à son infériorité. Aucune conclusion n'est possible quand la différence observée n'est pas suffisamment importante pour être statistiquement significative dans un sens ou dans un autre.
Avec un test unilatéral, une seule conclusion est possible : celle de la supériorité du traitement étudiée à son contrôle. Le test conclura à une différence non statistiquement significative dans deux situations : 1) la différence observée est trop petite pour être statistiquement significative ou 2) les résultats obtenus serait en faveur de l'infériorité du traitement étudiée à son contrôle (si l'on avait fait un test bilatéral).

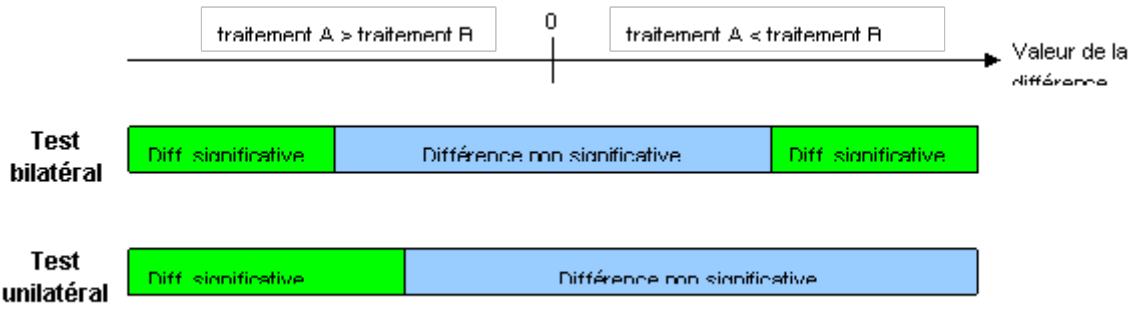


Figure 1- Zone de conclusion à une différence statistiquement significative avec un test bilatéral et avec un test unilatéral.
Le test bilatéral permet de conclure de manière significative aussi bien en cas d'observation d'une différence positive ou d'une différence négative. Il permet donc de conclure, en fonction de ce qu'il sera effectivement observé, soit à la supériorité de A sur B, soit à l'infériorité de A.. Le test unilatéral permet de conclure de manière significative uniquement en cas d'observation d'une différence négative. Il ne permet donc de conclure qu'à la supériorité de A sur B. Pout toutes les autres valeurs de différences observées, le résultat est non significatif.

Avec un test bilatéral on cherche à montrer que le traitement étudié est soit supérieur, soit inférieur. Avant le test, ces deux conclusions diamétralement opposées sont potentiellement possibles et ce sont les résultats

observés qui permettront de conclure à l'une ou à l'autre. Avec un test unilatéral on ne cherche à montrer que la supériorité. Si le traitement est inférieur le test ne sera pas significatif.

Ainsi dans un essai de supériorité utilisant un test bilatéral à 5%, le risque de conclure à tort à la supériorité est de 2.5% et celui de conclure à tort à l'infériorité est de 2.5%. Un essai de supériorité utilisant un test unilatéral à 5% a un risque de conclure à tort de 5% (ce qui est moins conservateur qu'un test bilatéral qui montre la supériorité).

Une différence non significative en bilatéral peut s'avérer significative en unilatéral (Tableau 1 et

Figure 2). Ainsi un test unilatéral possède une plus forte puissance qu'un test bilatéral. Le recours à un test unilatéral réduit le nombre de sujets nécessaires. Ce point peut être illustré grâce aux intervalles de confiance bilatéraux et unilatéraux (cf. section suivante).

Tableau 1 – Conclusions des tests bilatéraux et unilatéraux dans 4 situations différentes qui sont représentées graphiques dans la figure 2.

Résultat observé	***p* et conclusion avec un test bilatéral**	***p* et conclusion avec un test unilatéral**
RR = 0,70	$p<0,05$ Conclusion à la supériorité du traitement par rapport au placebo (effet bénéfique)	$p<0,025$ Conclusion à la supériorité du traitement par rapport au placebo (effet bénéfique)
RR = 0,79	$p=0.09$ Pas de conclusion possible	$p<0,05$ Conclusion à un effet bénéfique
RR = 0,98	$p=0,89$ Pas de conclusion possible	$p=0,86$ Pas de conclusion possible
RR = 1,45	$p<0,05$ Conclusion à l'infériorité du traitement par rapport au placebo (effet délétère)	$p=0,99$ Pas de conclusion

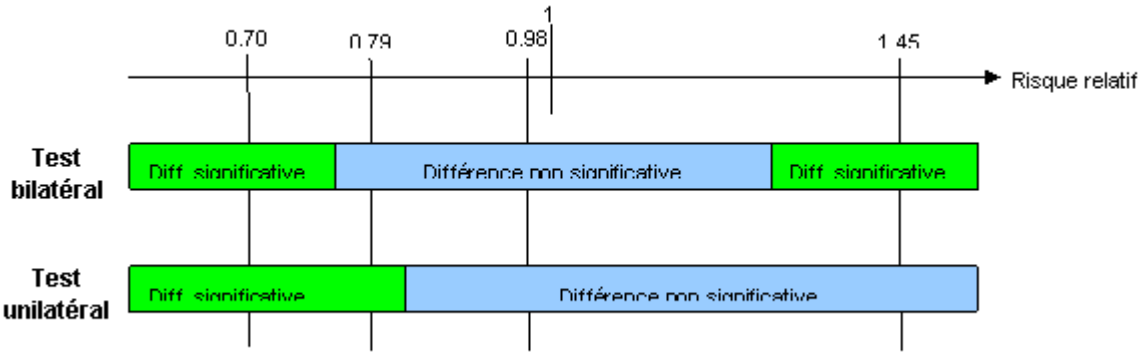


Figure 2 – Illustration des différents cas de figures décrits dans le tableau 1.

Le lecteur intéressé pourra trouver une présentation théorique des tests uni et bilatéraux à l'adresse Internet http://www.educ.necker.fr/cours/poly/biostatistique/Test_unilat_ral_ou_bilat_ral.htm.

Intervalles de confiance

Les intervalles de confiance (IC) peuvent être aussi bilatéraux ou unilatéraux. Un IC unilatéral est défini par une seule borne (supérieure ou inférieure suivant le sens de la différence que l'on cherche à mettre en évidence). Par exemple, un IC à 95% unilatéral de la réduction relative du risque (RRR) apporté par un nouveau traitement a pour borne supérieure une RRR de 20%. On peut raisonnablement exclure que l'efficacité du traitement soit inférieure à une réduction relative du risque de 20%. Il est raisonnable de conclure que la vraie efficacité du traitement est au moins une réduction relative de 20% du risque. Mais, on ne peut rien dire sur l'efficacité maximale.
Cette interprétation est à comparer avec celle d'un IC bilatéral, du type [20% ; 60%], avec lequel il est possible de conclure raisonnablement que la RRR apportée par le traitement est au moins de 20% et qu'elle n'excède pas 60%.
La borne d'un IC unilatéral est inférieure à la borne correspondante d'un IC bilatéral comme l'illustre le schéma de la Figure 3.

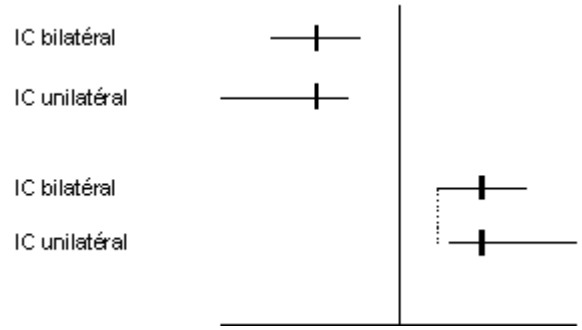


Figure 3 – Comparaison des IC uni et bilatéraux.

La borne supérieure d'un IC unilatéral à 95% a la même valeur que la borne supérieure d'IC bilatéral à 90%. En simplifiant, il est possible de dire que l'IC bilatéral à 95% exclut de chaque coté 2,5% des valeurs possibles sous l'effet des fluctuations aléatoire, tandis que l'IC bilatéral à 90% exclu 5% de ces valeurs de chaque coté. Un IC unilatéral à 95% exclut les valeurs « peu vraisemblables » que d'un seul coté. Contrairement à l'IC bilatéral, les 5% de valeurs exclues ne sont pas réparties également de part et d'autre mais le sont en bloc que d'un seul coté.

Figure 4 – Comparaisons des bornes supérieures d'un IC bilatéral à 90% et d'un IC unilatéral à 95%.

Choix du type de test

Au premier abord, les tests unilatéraux semblent correspondre le plus aux hypothèses thérapeutiques. Le but d'un essai thérapeutique est de montrer qu'un traitement est supérieur au placebo (ou supérieur à un traitement de référence) et ce n'est que si l'on montre cette supériorité de façon statistiquement significative que le nouveau traitement sera utilisé. Ainsi, mettre en évidence que le nouveau traitement est inférieur ou ne pas mettre en évidence sa supériorité (test non significatif) revient au même en pratique et s'accompagne des mêmes conséquences : le traitement n'est pas utilisé.

Cependant ce sont les tests bilatéraux qui sont utilisés de façon quasi exclusive en pratique et cela pour deux principales raisons. La première est qu'il peut être important de mettre en évidence qu'un traitement est délétère pour éviter d'entreprendre des essais supplémentaires.
La seconde est qu'un test bilatéral diminue le risque d'erreur alpha sur la conclusion de supériorité. Avec un test en 5%, le risque alpha de conclure à tort à la supériorité est en fait de 2,5%. Avec un test bilatéral, le même risque alpha serait de 5%. De ce fait utiliser un test bilatéral réduit le risque alpha de conclusion erronée à la supériorité du traitement étudié tout en conservant la valeur consacrée de 5%.
Si un test unilatéral est utilisé, il est important que ce choix ait été fait avant de connaître les résultats. En effet, choisir a posteriori un test unilatéral afin de rendre significative une différence qui ne l'est pas avec un test bilatéral pose le problème de toute analyse post-hoc.

Exemple 1

OBJECTIVE: The aims of this study were to substantiate the previously reported activity of ifosfamide in patients with advanced, persistent, or recurrent carcinosarcoma (mixed mesodermal sarcoma) of the uterus, and to determine whether the addition of cisplatin results in an improved response or survival. Secondarily, we sought to determine the toxicity of ifosfamide-cisplatin in this patient population. METHODS: Patients were randomized to receive ifosfamide (1.5 g/m(2)/day) times 5 days every 3 weeks for eight courses with mesna uroprotection, with or without cisplatin (20 mg/m(2)/day) times 5 days. No patient had received previous chemotherapy. RESULTS: Of 224 patients entered on this study, 30 were ineligible for a variety of reasons, leaving 194 evaluable patients. … Percentages of adverse effects reported in 191 patients receiving chemotherapy included (ifosfamide/cisplatin-ifosfamide) grade 3 or 4 granulocytopenia (36/60), grade 3 or 4 anemia (8/17), grade 3 or 4 central nervous system toxicity (19/14), and grade 3 or 4 peripheral neuropathy (1/12). Treatment may have contributed to the deaths of 6 patients treated with full doses of ifosfamide and cisplatin for 5 days. … The relative odds ratio of response adjusted for measurable sites of disease was 1.82 (P = 0.03, **one-tailed test**; 95% lower confidence limit, 1.06). Progression-free survival (PFS) and survival data suggest that the combination offers a slight prolongation of PFS (relative risk, 0.73; 95% upper confidence limit, 0.94; P = 0.02, **one-tailed test**), but no significant survival benefit (relative risk, 0.80, 95% upper confidence limit, 1.03; P = 0.071, one-tailed test). … (d'après réf (1)).

Exemple 2

OBJECTIVE: To compare the effectiveness of 100 mg of HA-1A and placebo in reducing the 14-day all-cause mortality rate in patients with septic shock and gram-negative bacteremia in the Centocor: HA-1A Efficacy in Septic Shock (CHESS) trial, and to assess the safety of 100 mg of HA-1A given to patients with septic shock who did not have gram-negative bacteremia. DESIGN: Large, simple, group-sequential, randomized, double-blind, multicenter, placebo-controlled trial. … PATIENTS: Within 6 hours before enrollment, the patients had been in shock with a systolic blood pressure of less than 90 mm Hg after adequate fluid challenge or had received vasopressors to maintain blood pressure. These episodes of shock began within 24 hours of enrollment. A presumptive clinical diagnosis of gram-negative infection as the cause of the shock episode and a commitment from the patients' physicians to provide full supportive care were required. ... RESULTS: 2199 patients were enrolled; 621 (28.2%) met all enrollment criteria, received HA-1A or placebo, and had confirmed gram-negative bacteremia. Mortality rates in this group were as follows: placebo, 32% (95 and HA-1A, 33% (109 of 328) (P = 0.864, Fisher exact test, **two-tailed**; 95% CI for the difference, -6.2% to 8.6%). Mortality rates in the patients without gram-negative bacteremia were as follows: placebo, 37% (292 of 793) and HA-1A, 41% (318 of 785) (P = 0.073, Fisher exact test, **one-tailed**; CI, -0.8% to 8.8%). CONCLUSIONS: In this trial, HA-1A was not effective in reducing the 14-day mortality rate in patients with gram-negative bacteremia and septic shock. These data do not support using septic shock as an indication for HA-1A treatment. If HA-1A is effective in reducing the mortality rate in patients dying from endotoxemia, these patients must be identified using other treatment criteria. (D'après ref (2))

Bibliographie

1. Sutton G, Brunetto VL, Kilgore L, Soper JT, McGehee R, Olt G, et al. A phase III trial of ifosfamide with or without cisplatin in carcinosarcoma of the uterus: A Gynecologic Oncology Group Study. Gynecol Oncol 2000;79(2):147-53.

2. McCloskey RV, Straube RC, Sanders C, Smith SM, Smith CR. Treatment of septic shock with human monoclonal antibody HA-1A. A randomized, double-blind, placebo-controlled trial. CHESS Trial Study Group. Ann Intern Med 1994;121(1):1-5.

Parallèle entre test statistique et test diagnostique

Introduction

Il est possible d'établir un parallèle entre les tests statistiques réalisés à la recherche de l'effet du traitement dans un essai thérapeutique et les tests diagnostiques utilisés pour poser le diagnostique d'une maladie (1). Ce parallèle permet d'appréhender ce qu'est la valeur prédictive d'un résultat d'essai significatif. Cette façon de voir les tests statistiques introduit aussi l'approche Bayesienne (cf. ci-dessous).

L'hypothèse nulle (H0) d'inexistence de l'effet du traitement peut être assimilée à l'absence de la maladie (noté M-), tandis que l'hypothèse alternative (H1) d'existence de l'effet est à mettre en parallèle avec la présence de la maladie (M+).
Le résultat de l'essai, significatif (R+) ou non significatif (R-), est assimilable à l'existence ou non du signe (recherché par le test diagnostic). Le signe recherché est présent (S+) quand l'essai obtient un résultat significatif, le signe est absent (S-) en cas de résultat non significatif.
Le risque α est la probabilité d'obtenir un résultat significatif sous l'hypothèse nulle, c'est-à-dire la probabilité de présence du signe en l'absence de la maladie : $\alpha = \Pr\left(S+\middle|M-\right)$. α est donc égal à $1-Sp$ où Sp désigne la spécificité du test (*Sp=Pr(S-/M-)*).
Le risque β est la probabilité de ne pas conclure, c'est-à-dire d'obtenir un résultat non significatif sous l'hypothèse alternative. Il correspond donc à la probabilité d'absence du signe en cas de maladie = *Pr(S-/M+)*, *β* est donc égal à *1-Se* où Se désigne la sensibilité (*Se=Pr(S+/M+)*). La puissance statistique *W=1-β* d'un essai est donc assimilable à la sensibilité d'un test diagnostique (détecter un effet lorsqu'il existe ou détecter un signe quand la maladie est présente).

Tableau 1 : Parallèle entre les table 2x2 du test diagnostic et du test d'hypothèse

	M+	M-
S+	Se	1-Sp
S-	1-Se	Sp

	H1	H0
R+	1-β	α
R-	β	1- α

Avec un test diagnostique, la valeur prédictive positive (VPP) est la probabilité de la maladie lorsque l'on est en présence du signe. C'est la probabilité de la maladie a posteriori, lorsque l'on a connaissance du résultat du test diagnostique. La VPP dépend de la probabilité a priori de la maladie *v*. Au niveau du test statistique d'un essai, la VPP correspond à la probabilité de l'existence de l'effet du traitement lorsque l'essai a obtenu un résultat significatif. Ainsi :

$$VPP = \frac{Se\ v}{Se\ v + (1-Sp)(1-v)} = \frac{(1-\beta)v}{(1-\beta)v + \alpha(1-v)} = \frac{Wv}{Wv + \alpha(1-v)}$$

Cette formule est identique à celle obtenue en raisonnant avec les risques statistiques vus comme des taux de filtration (cf. chapitre *principe général du test statistique*). La probabilité a priori *v* est une notion difficile à cerner avec précision. Il s'agit plutôt d'une probabilité subjective, d'une sorte de degré de croyance en l'existence de l'efficacité du traitement avant de l'évaluer. En moyenne, cette probabilité a priori peut être assimilée à la fréquence moyenne des résultats positifs obtenus avec les essais thérapeutiques.

> Ce développement est identique à celui que nous avons fait avec les risques statistiques vus comme des taux de filtration.

Le Tableau 3 donne la probabilité de l'existence de l'effet du traitement après avoir obtenu un résultat significatif dans un essai en fonction du seuil de signification statistique, de la puissance statistique de l'essai *W* et de la probabilité a priori d'existence de l'effet du traitement (*v*).

Ainsi un résultat significatif n'a pas la même valeur (prédictive de l'effet du traitement) quand la probabilité a priori est faible ou forte. Dans des situations très spéculatives où l'essai est réalisé sans qu'il y ait de justification, la probabilité a priori est très faible. Dans ce cas, un essai significatif n'a pas beaucoup de valeur prédictive, même en cas de résultats hautement significatifs.

Tableau 2 – Parallèle entre test statistique d'un essai et test diagnostique

Test statistique	**Test diagnostique**
Absence d'effet hypothèse nulle H0	Absence de la maladie M-
Existence de l'effet hypothèse alternative H1	Existence de la maladie M+
Résultat de l'essai significatif R+	Présence du signe recherché par le test diagnostique S+
Résultat non significatif R-	Absence du signe S-
Risque alpha (ou valeur de *p*) Pr(R+/H0)	Pr(S+/M-)=1-Pr(S-/M-) =1-Sp (Sp=spécificité)
Risque beta Pr(R-/H1)	Pr(S-/M+)=1-Pr(S+/M+) =1-Se (Se= sensibilité)
Puissance W=1-β Pr(R+/H1)	Se (sensiblité) Pr(S+/M+)
Probabilité que le traitement soit efficace si le résultat est significatif	Valeur prédictive positive *P*(M+/S+) = VPP
Probabilité a priori que l'hypothèse testée soit vraie	Probabilité a priori de la maladie V

Avec les traitements médicamenteux, les essais d'efficacité (phase 3) sont entrepris après les essais de pharmacologie animale et les essais de phase 1 et 2. La probabilité a priori d'existence d'un effet est alors forte. Environ 5 à 10% des essais de phase 3 n'obtiennent pas de résultats significatifs. Avant essai, il est possible de dire, qu'en moyenne, la probabilité qu'un traitement issu des phases précédentes soit efficace est d'au moins 90%. La valeur prédictive d'un résultat significatif est alors importante.
La valeur prédictive d'un résultat significatif dépend aussi de la puissance de l'essai : elle est moins importante avec un essai de faible puissance qu'avec une étude dont la puissance statistique est élevée. La probabilité d'avoir à faire à un résultat faux positif (probabilité qui est 1-VPP) est d'autant plus importante que la puissance de l'essai est faible. Un résultat faux positif étant un résultat statistiquement significatif mais obtenu avec un traitement inefficace.

Tableau 3 – Probabilité de l'existence de l'effet du traitement après avoir obtenu un résultat significatif dans un essai en fonction du seuil de risque alpha, de la puissance statistique de l'essai W et de la probabilité a priori d'existence de l'effet du traitement (v).

	Alpha		
	0.05	0.01	0.001
Probabilité a priori d'existence de l'effet v=90%			
20%	97.3%	99.4%	99.9%
50%	98.9%	99.8%	100.0%
80%	99.3%	99.9%	100.0%
Probabilité a priori d'existence de l'effet v=50%			
20%	80.0%	95.2%	99.5%
50%	90.9%	98.0%	99.8%
80%	94.1%	98.8%	99.9%
Probabilité a priori d'existence de l'effet v=10%			
20%	30.8%	69.0%	95.7%
50%	52.6%	84.7%	98.2%
80%	64.0%	89.9%	98.9%
Probabilité a priori d'existence de l'effet v=1%			
20%	3.9%	16.8%	66.9%
50%	9.2%	33.6%	83.5%
80%	13.9%	44.7%	89.0%

La fragilité possible de certaines preuves, issues pourtant d'un résultat statistiquement significatif, que laissent entr'apercevoir certaines valeurs de VPP du Tableau 3, provient simplement du fait que les preuves disponibles en médecine sont parfois insuffisamment puissantes et concluantes. Cette approche rejoint les constatations que l'on peut être amené à faire avec l'analyse de la pertinence clinique des intervalles de confiance.
Si les preuves apportées par les essais peuvent paraître fragiles dans certains cas (car obtenues avec un faible degré de signification statistique et dans une petite étude peu puissante), c'est parce que les études réalisées sont trop petites ou insuffisamment justifiées.

A priori neutre

Le choix de la probabilité a priori est le maillon faible de cette approche. Sa détermination est constamment entachée de subjectivité. Ainsi il peut être considéré comme gênant que la valeur d'un résultat recueilli de façon rigoureuse, au prix de nombreux efforts soit ensuite malaxée avec un « a priori » très subjectif. Pour contourner cette difficulté, il est possible de prendre un « a priori » totalement neutre (on dit non informatif), correspondant à v=50%. On ne favorise, a priori, aucune hypothèse et on laisse les données décider entièrement de la conclusion. Dans ce cas, cette approche Bayesienne (cf. ci-dessous) revient à l'approche

fréquentiste simple. La VPP d'un résultat significatif au seuil de 5% sera égal à 95% (avec une étude puissante où la puissance est de 95%).

Approche fréquentiste, approche Bayesienne

Deux courants de pensée coexistent dans les statistiques inférentielles : l'approche fréquentiste et l'approche Bayesienne.
L'approche Bayesienne (2) est celle que nous venons de voir en faisant ce parallèle entre test statistique dans un essai et test diagnostique.
L'approche fréquentiste est l'utilisation simple des tests statistiques, sans chercher à exprimer la probabilité de l'hypothèse en fonction du résultat (probabilité a posteriori). C'est l'approche classique des tests statistiques. Cependant le résultat qu'elle produit, la probabilité d'observer les données sous l'hypothèse nulle, peut apparaître insatisfaisant, car ne répondant pas directement à la question que l'on se pose : qu'elle la probabilité que le traitement soit efficace.

	Approche fréquentiste	Approche Bayesienne
Information apportée	Pr(résultat observé/H0) *P* value	Pr(H1/résultat) Valeur prédictive

L'approche Bayesienne cherche à estimer la probabilité de la conclusion, c'est-à-dire la probabilité de l'hypothèse alternative (probabilité a posteriori). En cherchant à estimer la probabilité a posteriori de l'hypothèse, l'approche Bayesienne nécessite l'introduction de la probabilité a priori. La probabilité a priori est difficile à obtenir. Elle ne provient pas, le plus souvent, de données mais d'une appréciation subjective. À ce niveau, se situe la principale difficulté et le point de faiblesse de cette approche. En effet, à partir du même essai, il va être possible de faire des conclusions parfois opposées en fonction de l' « a priori » choisi. Or, il s'avère que, dans la majorité des cas, ce le choix de la probabilité a priori est arbitraire et qu'aucune valeur ne s'impose par elle-même. Ainsi, il devient de plus en plus fréquent qu'un essai ait simultanément le double objectif de mettre en évidence la supériorité du traitement étudié ou son équivalence. Ce qui montre qu'a priori une hypothèse particulière ne s'impose pas naturellement. C'est, par exemple, le cas dans l'essai VALIANT qui compare un inhibiteur de l'angiotensine 2 à un inhibiteur de l'enzyme de conversion (3).

Références

1. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987;257:2459-63.
2. Press SJ. Bayesian statistics: principles, models and applications. New-York: John Wiley & Sons; 1989.
3. Pfeffer MA, McMurray J, Leizorovicz A, Maggioni AP, Rouleau JL, Van De Werf F, et al. Valsartan in acute myocardial infarction trial (VALIANT): rationale and design. Am Heart J 2000;140(5):727-50.

Les courbes de survie

Introduction

L'objet de ce document est de présenter l'intérêt et l'interprétation des courbes de survies (« survival curve »). Nous ne rentrerons pas dans le détail des calculs statistiques que le lecteur trouvera dans de nombreux ouvrages de statistiques (1, 2), mais nous insisterons sur l'analyse des courbes de survie et des tests qui leur sont rattachés, et sur les risques d'interprétations erronées.

Dans de nombreuses situations, l'un des objectifs thérapeutiques envisageables est de retarder la survenue d'un événement clinique. Dans le cas du décès, cet objectif revient à vouloir augmenter la durée de survie des patients. Dans la suite, nous n'envisagerons que le cas du décès et du temps de survie, mais la totalité de ce qui sera présenté s'applique aussi à tout autre type d'événements cliniques (infarctus, greffes, hospitalisation, etc.). On parle alors de survie sans événement.

> *En cancérologie, les survies sans progression de la maladie («*progression-free survival *») ou sans rechute («* recurrence-free survival *», «* relapse-free survival *») sont couramment utilisée.*

Cet objectif revient à montrer qu'un traitement augmente la durée moyenne de survie des patients traités par rapport à celle des patients contrôles. L'importance de l'effet du traitement s'apprécie alors par l'augmentation de la durée moyenne de survie qu'il engendre, mesurée de façon absolue ou relative.

En pratique, cependant, l'estimation des durées moyennes de survie se révèle impossible dans de nombreuses situations car il est rare de suivre tous les patients jusqu'à la survenue de l'événement considéré. Pour pallier cette difficulté les techniques d'analyse de survie ont été développées. Elles n'apportent cependant qu'une réponse indirecte et partielle à la question posée, celle de la détermination de l'effet d'un traitement sur la durée moyenne de survie. Les techniques d'analyse de survie décrivent aussi la dynamique de survenue des décès.

Exemple

*Dans un essai, la survie **moyenne** obtenue sans traitement dans le groupe contrôle a été de 24 mois. Dans le groupe recevant le traitement étudié, la durée **moyenne** de survie a été de 28 mois. Le traitement a donc apporté un gain en durée moyenne de survie de 28-24 = 4 mois. Ce gain absolu peut aussi être exprimé en gain relatif : (28-24)/24 = 16,7%. Le traitement augmente l'espérance de vie de 17%, c'est à dire que le traitement la multiplie par 1,17 (=28/24).*

Suivi partiel des patients

*La mesure de la durée de survie **moyenne** d'un groupe de patients nécessite de suivre tous les sujets jusqu'à la survenue de leur décès. Ceci n'est que rarement envisageable car le plus souvent cela nécessiterait une durée d'étude parfois très longue et reculerait de façon non acceptable le moment d'obtention de la réponse à la question posée. Pour des pathologies de gravité moyenne, il serait nécessaire de suivre les patients durant des dizaines d'années ou de n'inclure que des sujets très âgés.*

> *Une censure survient lorsque l'on arrête de suivre un patient avant la survenue de l'événement.*

En pratique, certains patients sont toujours vivants à la fin de l'étude. Leur suivi est dit « censuré » (« censured *») dans la terminologie statistique et par abus de langage, ces patients sont appelés « patients censurés ». Cependant, dans certains domaines médicaux où les taux de mortalité sont extrêmement élevés, comme avec certains cancers ou dans certaines situations de réanimation, il est possible de déterminer la durée de survie de presque tous les patients.*

L'existence de censure empêche le calcul de la durée de survie moyenne, car toutes les durées de survie ne sont pas connues.

Exemple

Cinq patients ont été inclus dans une étude, leur survie respective est de : 6 mois, 7 mois, 9 mois, 11 mois, 18 mois. La survie moyenne est donc de 10.2 mois. Si l'étude n'avait durée que 12 mois, la survie du dernier patient n'aurait pas été connue. La moyenne des survies disponibles à 12 mois (4 premiers patients), 8.25 mois, n'a aucun sens et sous estime fortement la survie moyenne réelle.

Origine des censures

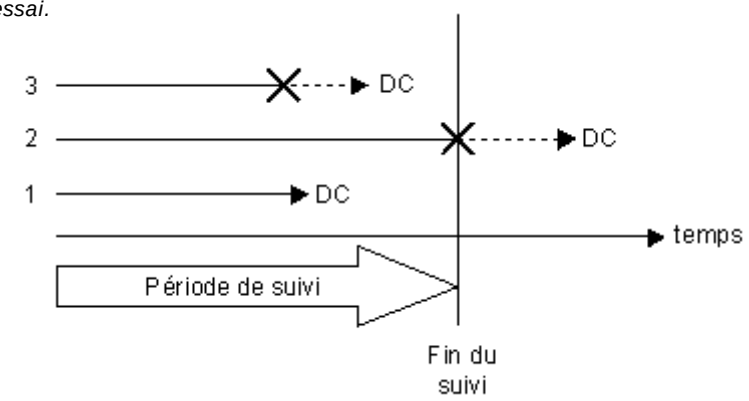
Dans un essai thérapeutique, les censures surviennent dans deux circonstances.

La première correspond aux patients qui sont toujours vivants au terme de la durée de l'essai (cf. section suivante). Cette censure est inévitable et provient uniquement du fait que, chez ces patients, la durée de suivi dans l'essai est inférieure à la durée de survenue du décès. Ces censures n'entraînent pas de biais.

L'autre circonstance conduisant à une censure concerne les patients perdus de vu en cours d'essais (ou retirés de l'essai avec un arrêt de leur suivi). Le suivi de ces patients est interrompu avant la survenue du décès et, à ce titre, il s'agit d'une censure. Cependant, ces censures sont susceptibles d'introduire des biais comme tous perdus de vue ou retraits de l'essai en général (cf. chapitre Analyse en intention de traiter*), car leur survenue n'est peut-être pas aléatoire. Ce type de censure correspond aisni à une* ***donnée manquante****.*

Les censures liées aux patients perdus de vue ou aux retraits de l'essai sont prises en considération par les méthodes statistiques de la même façon que celles correspondant aux patients n'ayant pas présenté l'événement durant la durée de l'essai, mais elles sont susceptibles d'introduire un biais. Ces deux types de censures ne doivent donc pas être confondus. Ce n'est pas parce que les méthodes d'analyses de survie permettent d'exploiter les censures correspondant aux perdus de vue, qu'elles évitent que ces perdus de vue biaisent les résultats (cf. infra). Avec les techniques d'analyse des données de survie, ***les patients perdus de vue doivent être considérés comme tels et non pas comme de simples censures****. Leur nombre doit être donné dans le rapport, et lors de la réalisation de l'essai tout doit être fait pour les éviter.*

Figure 1 – Illustration des différents types de censures rencontrées dans un essai. Dans le cas 1, le décès survient durant la période de suivi de l'essai. Dans le cas 2, le décès survient au-delà de la fin de l'essai. À la fin de la période de suivi, le patient est vivant. Il est pris en considération comme un censuré vivant. Il ne gênera pas l'estimation de la courbe de survie durant la période de l'essai. Le cas 3 est celui d'un patient qui est décédé durant la période de suivi mais ce décès n'est pas connu car le patient a été préalablement perdu de vue. Il s'agit aussi d'une censure, mais celle-ci soustrait de l'information et fausse l'estimation de la courbe de survie durant la période de l'essai.



Étalement des inclusions

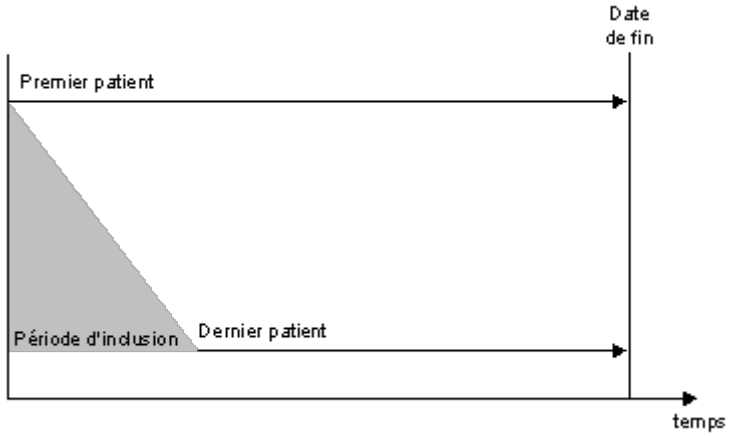
Les inclusions dans un essai thérapeutique s'étalent sur une période plus ou moins longue. Il est exceptionnel que tous les patients soient inclus le même jour. La durée de la période de recrutement est variable en fonction du nombre de centres investigateurs participants, de la rareté de maladie.

La fin du suivi d'un patient peut être définie de deux façons : à une date fixe ou après une durée de suivi déterminée.

Lorsque l'essai est arrêté à une date donnée (appelée date de point), la durée de suivi des patients est variable : le premier patient inclus dans l'essai a la durée de suivi la plus importante, le dernier patient la plus

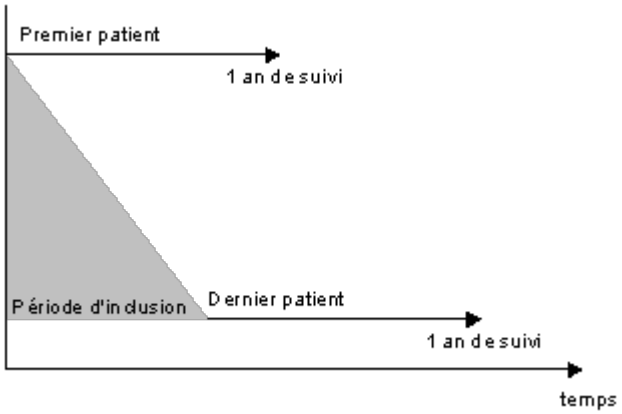
courte. Pour caractériser la durée pendant laquelle les patients ont été suivis, il convient alors de préciser la durée moyenne de suivi accompagnée des valeurs du suivi le plus court et de celui le plus long.

Figure 2 – Essai à date de point



Le suivi d'un patient peut aussi se terminer au bout d'une période de temps donnée, identique pour tous les patients. Par exemple, le suivi des patients prend fin lors de la visite réalisée à 1 an. Cette visite permet de mesurer le critère de jugement et met un terme à la participation du patient à l'essai. Dans ce type d'essai, tous les patients ont la même durée de suivi.

Figure 3 – Essai assurant la même durée de suivi pour tous les patients.



Assez souvent, l'attitude adoptée combine c'est deux possibilités. Une date de point est choisie de telle façon qu'elle assure pour tous les patients le suivi minimum recherché, par exemple 1 an. Les premiers patients sont alors pris en considération avec des durées de suivi plus importantes. Les résultats de cet essai pourront ainsi être évalués à deux moments différents : au bout du suivi recherché (résultat à 1 an par exemple) et à la fin de l'essai (avec des suivis variables pour les patients). Le résultat en fin d'étude se rapporte à la durée moyenne de suivi.

Exemple

*"*Enrollment in the trial began on February 17, 1983, and ended on June 30, 1986. … The patients underwent a final evaluation after the study's completion (cutoff date, June 30, 1987). All the patients were followed for at least 12 months, to a maximum of 52 months; the average duration of follow up was 25 months*".*

Note : Attention a ne pas confondre la durée moyenne de suivi et la durée moyenne de survie. Le suivi est la durée pendant laquelle un patient est suivi dans l'essai (de sa date d'inclusion à la date de fin de l'étude ou à sa date de décès). La survie est la durée qui s'écoule entre la date d'inclusion et la date de son décès (Évidement si le patient est toujours vivant à la fin de l'étude, sa survie est inconnue).

L'utilisation des courbes de survie permet de calculer l'effet du traitement pour des reculs où seulement une partie des patients a été suivie aussi longtemps. Par exemple, dans un essai où la durée de suivi moyenne est de 1,9 ans il est possible de calculer l'effet à 2 ans, à 3ans. Environ la moitié des patients sont « complètement » informatifs pour le calcul de l'effet à 2 ans, et seulement une faible proportion pour celui à 3ans. La précision de ces estimations diminue donc au court du temps. Pour cette raison, les représentations graphiques des courbes de survie doivent rapporter l'évolution du nombre de patients exposés au risque au cours du temps (cf. Figure 15).

Ainsi, l'effet du traitement à 3 ans donné par l'analyse des courbes de survie d'un essai où les durées de suivi s'étalent entre 2 et 4 ans, n'a pas la même précision, et donc, pas la même valeur, que l'estimation apportée par un essai où tous les patients ont été suivis 3 ans. En d'autres termes, il convient de toujours analyser la proportion des patients qui ont apporté de l'information pour le calcul d'un effet traitement à un moment donné.

Distribution asymétrique des temps de survie

La distribution des durées de survie est asymétrique dans presque tous les cas (cf. Figure 6). La durée de survie étant une variable continue, sa distribution peut être caractérisée par les paramètres de position et de dispersion habituels tels que : moyenne, médiane, écart-type ou étendue. Cependant le caractère asymétrique des distributions des durées de survie incite à préférer la médiane et la distance inter-quartile à la moyenne et à l'écart type.

Représentations graphiques des données de survie

Plusieurs types de courbes permettent de représenter les données de survie et de décrire la dynamique d'apparition des événements au cours du temps.

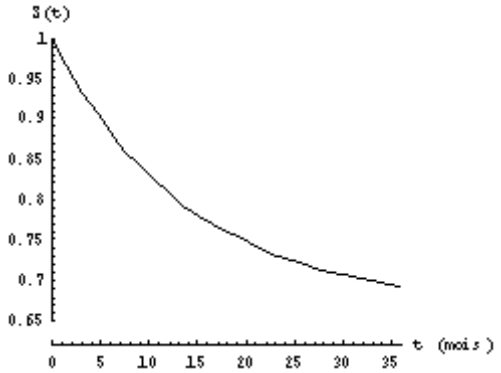
Définitions

Courbe de survie

Le taux de survie au temps t représente la proportion de patients toujours vivants après une durée de suivi t.

La courbe de survie est la représentation la plus employée pour décrire la dynamique de survenue au cours du temps des décès. Elle représente en fonction du temps le taux de survie (« survival rate »), c'est à dire la proportion des sujets initialement inclus dans l'essai toujours vivants au temps t. C'est la probabilité de survivre au moins jusqu'au temps t.

Figure 4 – Exemple de courbe de survie. Le taux de survie S(t) est la proportion des sujets initialement inclus dans l'essai toujours vivant au temps t.

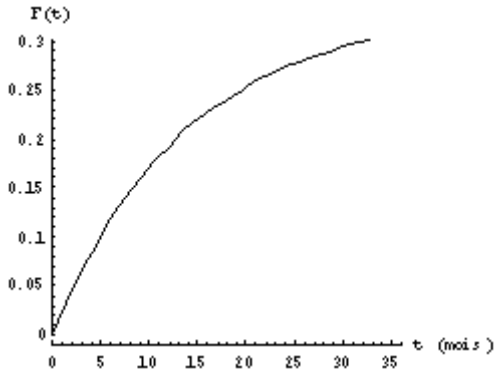


D'autres représentations de la dynamique d'apparition des événements sont également disponibles qui présentent d'une manière différente la même information.

Courbe des taux cumulés d'événements

La courbe de survie S(t) *est le complément à 1 du taux cumulé d'événements en fonction du temps* F(t) *(Figure 5). En effet, si, à un temps* t *le taux de survie est de 20%, le taux d'événement (décès) à ce temps est de 1-20% = 80%. Le taux cumulé d'événement n'est rien d'autre que le risque.*

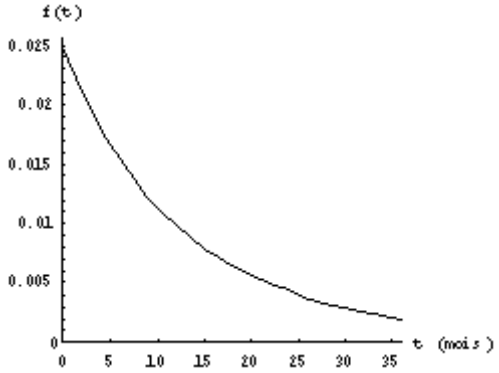
Figure 5 – Courbe du taux cumulé d'événements correspondant à la courbe de survie de la figure 4.



Distribution des temps de survie

Une autre représentation de la même information est la distribution f(t) des temps de survie. La distribution cumulative correspondante est la courbe F(t).

Figure 6 – Distribution des durées de survie correspondant à la courbe de survie de la figure 4.

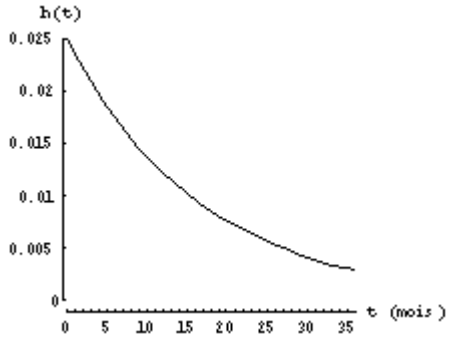


Courbe du risque instantané

Le risque instantané (« hazard *»)* h(t) *est le taux de décès observé sur une courte période du suivi entre* t *et* Δt*. Bien que mathématiquement il corresponde au risque de décéder durant une période de temps infinitésimale, en pratique, on peut l'associer, par exemple, au risque de décéder durant 1 jour ou 1 semaine ou 1 mois, en fonction de la gravité de la pathologie.*

Le nombre de patients qui décèdent entre t *et* t+dt *est donc égal au nombre de survivants en début de période multiplié par le risque instantané de cette période* $x_t = n_t * h(t)$.

Figure 7 – Représentation graphique du risque instantané.



Exemple numérique

Le Tableau 1 représente ces différents indicateurs dans un exemple numérique où le temps est découpé en années. Pour la clarté de l'illustration, le risque instantané est assimilé au risque annuel.

Tableau 1 – Exemple numérique illustrant les différents paramètres utilisés pour décrire la dynamique de survenue des événements. De plus amples informations sont données par le texte.

Année	Nombre de survivants	Risque instantané	Nombre de décès dans l'année	Nombre cumulé de décès	Taux cumulé de mortalité	Taux de survie
t	*n(t)*	*h(t)*			*F(t)*	*S(t)*
1	200	0.1	20	20	10.0%	90.0%
2	180	0.1	18	38	19.0%	81.0%
3	162	0.1	16	54	27.1%	72.9%
4	146	0.1	15	69	34.4%	65.6%
5	131	0.1	13	82	41.0%	59.0%
6	118	0.1	12	94	46.9%	53.1%
7	106	0.1	11	104	52.2%	47.8%
8	96	0.1	10	114	57.0%	43.0%
9	86	0.1	9	123	61.3%	38.7%
10	77	0.1	8	130	65.1%	34.9%
11	70	0.1	7	137	68.6%	31.4%
12	63	0.1	6	144	71.8%	28.2%
13	56	0.1	6	149	74.6%	25.4%
14	51	0.1	5	154	77.1%	22.9%
15	46	0.1	5	159	79.4%	20.6%

Le nombre de survivants n(t) *est le nombre de sujets toujours vivants en début d'année. Pour la première année ce nombre est le nombre de sujets inclus dans l'étude. À chaque début d'année, ce nombre diminue du nombre de décès survenus durant l'année écoulée.*

Le risque instantané h(t) *qui représente le taux de décès annuel est ici constant au cours du temps et égal à 10%. Cette situation est celle d'une maladie chronique où le risque de décéder varie peu au cours du temps. La simplicité de cette situation a été choisie pour simplifier la compréhension des mécanismes cachés derrière une courbe de survie. Dans la réalité des formes plus complexes de courbe de risque instantané sont observées.*

Le nombre de décès dans l'année diminue au fil du temps étant donné qu'il est proportionnel au nombre de sujets toujours vivants en début d'année et que ce dernier diminue au cours des années. Ainsi, ***sans que la gravité de la maladie ne diminue*** *(le risque instantané est toujours le même)* ***la pente de décroissance de la courbe de survie se réduit****. Cette situation, où le risque instantané est constant au cours du temps, est appelée modèle exponentiel car il conduit à une décroissance exponentielle du nombre de survivants. Par analogie avec la pharmacocinétique, elle peut être représentée par un modèle compartimental. A chaque instant, la « quantité » de sujets qui quittent la population observée (qui décèdent) est proportionnelle au*

nombre de sujets présents dans cette population. Le coefficient représentant la vitesse de « sortie » de la population est le risque instantané. La dynamique de ce système est représentée par l'équation différentielle :

$$dN/dt = -h(t)\,N(t)$$

Le nombre de décès cumulés est obtenu en cumulant le nombre de décès survenus durant les années précédentes.
Le taux cumulé d'événements est le rapport du nombre cumulé de décès divisé par l'effectif initial. Le taux de survie est le complément à 100% du taux d'événements cumulés.

Risque annuel

Le risque annuel ou taux annuel de mortalité ne s'obtient pas en divisant le taux cumulé observé à un moment donné par la durée de suivi. Dans notre exemple, le taux cumulé de décès à 10 ans est de 65,5%, ce qui pourrait laisser penser que le taux annuel de décès est de 65,5%/10=6,55%, valeur sousestimant la vraie valeur qui est 10%. L'écart est dû au fait qu'une interpolation linéaire est utilisée sur une exponentielle. En fait, l'interpolation linéaire n'est pas trop fausse que lorsque le taux de décès est faible. Par exemple à deux ans, le calcul donne 19%/2=9,5.

Modèle probabiliste de survie

Un modèle probabiliste représente bien la variabilité des durées de survie observée dans la réalité. Avec ce modèle, deux sujets de mêmes caractéristiques auront le même risque instantané. C'est le hasard qui fera que l'un décédera précocement et l'autre beaucoup plus tardivement. Ainsi dans le Tableau 1 où tous les patients ont exactement le même risque instantané, certains décèdent dans la première année et d'autres au bout de 15 ans. Certains favorisés par le hasard sont d'ailleurs survivants à cette date. Ce n'est plus les durées de survie, trop empreintes de variabilité structurelle, qu'il convient d'expliquer par des covariables, mais le risque instantané

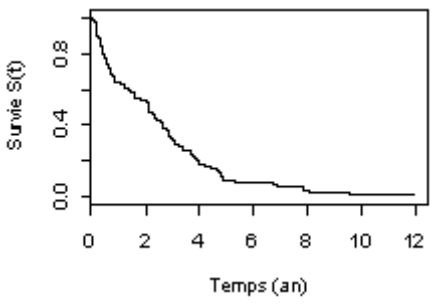
Calcul des courbes de survie en pratique

En pratique, il convient d'estimer les courbes de survie à partir de données censurées.

L'estimation des courbes de survie fait appel principalement à la technique de Kaplan Meier. Le taux de survie est réactualisé à chaque temps de survenue d'un décès, ce qui donne un aspect en marche d'escalier aux courbes (Figure 8). Comme tout estimateur, la courbe estimée de « Kaplan Meier » est connue avec une certaine imprécision qui doit être prise en compte dans son analyse.

Une autre technique, la méthode actuarielle, consiste à découper le temps en intervalles réguliers de largeur fixée arbitrairement. Avec cette technique, la courbe est composée d'une série de segment de droite et non plus de marche d'escalier. Cette technique rend les calculs plus simples, mais elle n'est plus guère d'actualité en raison de la généralisation des moyens de calculs modernes.

Figure 8 – Exemple d'une courbe de survie estimée avec la méthode de Kaplan-Meier.



Le calcul des courbes

Si au temps t *, survient un décès parmi 123 patients (qui représentent 68,33% des patients inclus (n=180), c'est-à-dire que le taux de survie juste avant* t *est de 68,33%), la survie après ce décès est de*

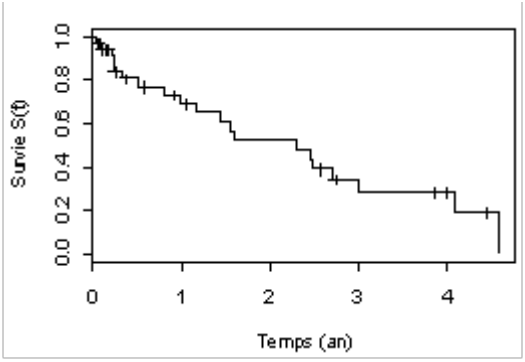
0,6833(122/123)=0,677745 soit 67,7745%. En effet après ce décès la proportion des patients vivants avant* t *qui reste vivant après* t *est de 122/123=99.187%. Après le temps* t *il reste vivants 99,187% des patients vivants juste avant* t*, comme juste avant* t *il restait vivants 68,33% des patients initialement inclus, la proportion des patients initialement inclus toujours vivants après* t *est 99,187% de 68,33% soit 67,7745%.*

Si maintenant, une censure s'est produite entre la date t de survenu du décès qui nous intéresse et le décès précédent, le nombre de patient à risque au moment du décès du temps t *(donc potentiellement informatif pour estimer la survie à ce temps) est de 123-1=122. A l'issu de ce décès, seul 121/122=99,18% des patients vivants avant t l'est toujours. Par rapport au nombre initialement inclus, le taux de survie est 99,18% de 68,33% soit 67,7699%. Soit une valeur légèrement inférieure à celle obtenu en l'absence de censure.*

Représentation des censures

Les moments où surviennent des censures sont souvent représentés par une croix sur les courbes de survie. Le nombre de censures et leur répartition le long des courbes permettent d'apprécier leurs conséquences sur la quantité d'information et un éventuel biais.

Figure 9 – Visualisation des censures par des croix sur une courbe de survie estimée.



Les conséquences des censures ne sont pas les mêmes suivants qu'il s'agit de censure liée à un suivi partiel ou de censure introduite par des perdus de vue (cf. Tableau 2).

Tableau 2 – Conséquences des deux types de censures sur l'interprétation des courbes de survie.

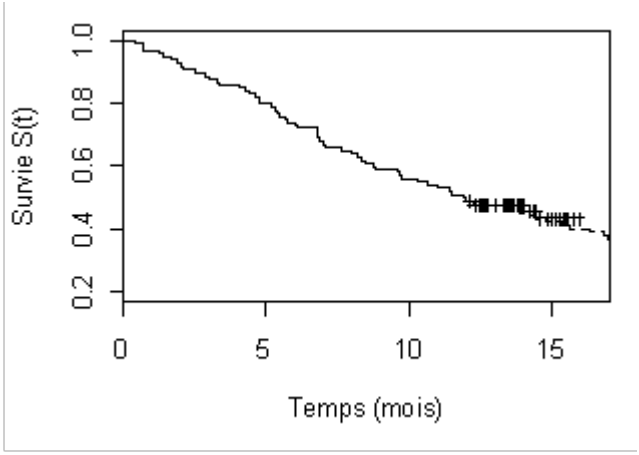
Type de censure	**Conséquences**
Censures liées à l'interruption du suivi avant le décès de tous les patients	◆ ces censures sont regroupées à la fin de la courbe ◆ elles ne faussent pas la partie de la courbe où il n'y a pas de censure ◆ elles ne soustraient pas d'information sur la partie de la courbe que l'on veut estimer (correspondant à la durée de l'essai)
Censures survenant durant la période de suivi, induites par des patients perdus de vue (censure non aléatoire))	◆ elles entraînent une perte d'information en soustrayant un certain nombre de décès concernant la partie de la courbe à laquelle on s'intéresse ◆ elles faussent potentiellement l'estimation de la courbe de survie ◆ elles faussent la comparaison de deux traitements si elles sont liées à l'effet du traitement ◆ elles limitent l'interprétation des courbes

Censures survenant à la fin de la période de suivi

Les censures qui correspondent aux patients toujours vivants à la fin de la durée de suivi prévu par l'essai apparaissent vers la fin de la courbe de survie. Ces censures s'étendent à la fin de la courbe sur une période

reflétant l'étalement des inclusions dans le temps. Elles ne perturbent pas la majeure partie de la courbe, en particulier celle que l'essai cherche à estimer.

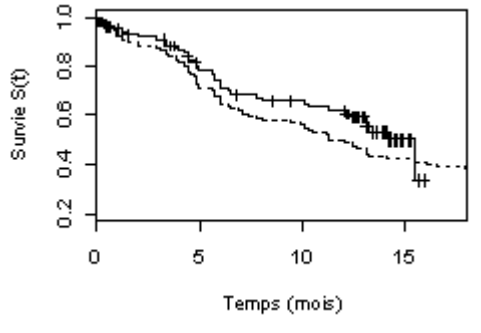
Figure 10 – Les censures survenant à la fin de la période de suivi correspondent à des patients vivants à la fin de l'essai. Elles ne perturbent pas la partie initiale de la courbe de survie. La durée minimale de suivi a été de 12 mois, la courbe obtenue estime donc correctement la survie durant les 12 premiers mois. La courbe obtenue en l'absence de toute censure est représentée en pointillé.



Censures correspondant à des perdus de vue

Les censures survenant durant la période de suivi de l'essai et correspondant à des perdus de vue sont en général en nombre limité. Lorsqu'elles surviennent au hasard, ces censures n'entraînent qu'une perte de puissance statistique, mais ne fausse pas l'estimation. Par contre si leur survenue n'a pas lieu au hasard, (on dit qu'elles sont informatives), elles faussent l'estimation de la courbe de survie (Figure 11).

Figure 11 – Censure informative. La probabilité de survenue d'une censure dépend de l'état du patient, donc de son risque de décédé durant la période d'observation. La courbe obtenue en l'absence de toute censure est représentée en pointillé.

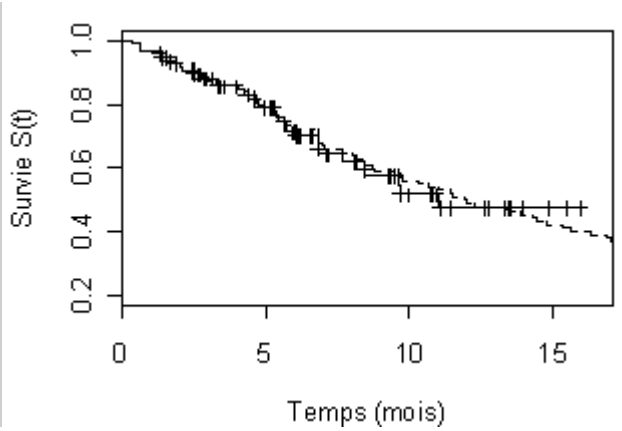


Essai à date de point

Dans un essai « à date de point », les censures s'étalent tout le long de la période d'observation, mais, contrairement au cas précédent, elles correspondent à l'étalement des inclusions et non pas à des perdus de vue. Les derniers patients inclus sont suivi de façon brève. Ces censures ne faussent pas l'estimation de la courbe de survie, mais réduisent la précision de l'estimation (cf. ci-dessous).

La description du suivi s'effectue à l'aide de la médiane des temps de suivi (jusqu'au décès ou à la censure) accompagnée des durées minimale et maximale. Le nombre de perdu de vue doit être précisé. Fréquemment, pour ne pas surcharger les graphiques, les censures ne sont pas représentées sur les courbes de survie. Le seul moyen alors disponible pour se faire une idée du suivi est cette description numérique.

Figure 12 – Répartition des censures dans un essai « à date de point ». La fin de l'essai survenant très rapidement après l'inclusion du dernier patient, les censures reflètent l'étalement des inclusions. La courbe obtenue en l'absence de toute censure est représentée en pointillé.



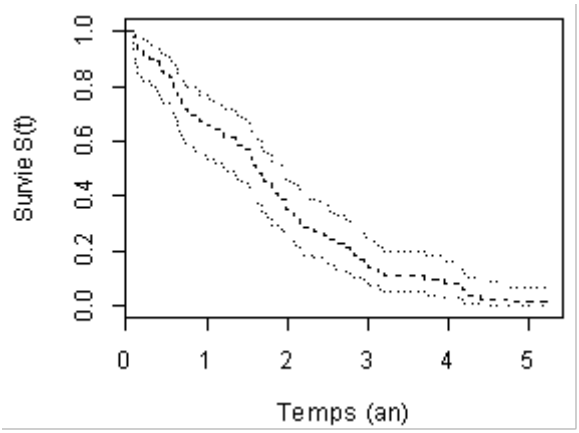
Durée de suivi (mois) médiane (min-max) : 5,9 (1,4-14,9)

Perdus de vue : 2

Précision de l'estimation

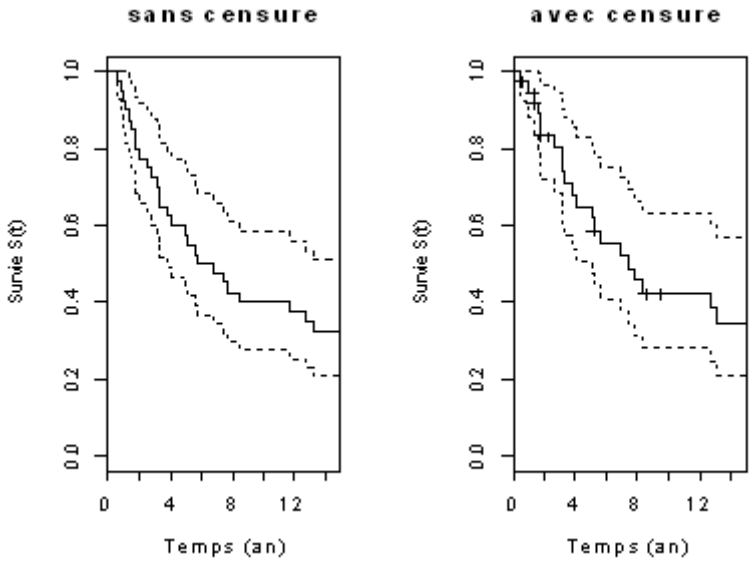
Comme avec toutes estimations statistiques, la précision de l'estimation d'une courbe de survie est représentée par un intervalle de confiance (le plus souvent à 95%). La Figure 13 représente une courbe de survie estimée, entourée de son intervalle de confiance à 95%. Cet intervalle prend la forme de deux courbes correspondant à la limite supérieure et à la limite inférieure. Cet intervalle a une probabilité de 95% d'inclure la vraie courbe de survie.

Figure 13 – Courbe de survie (en trait plein) estimée entourée de son intervalle de confiance (trait pointillé).



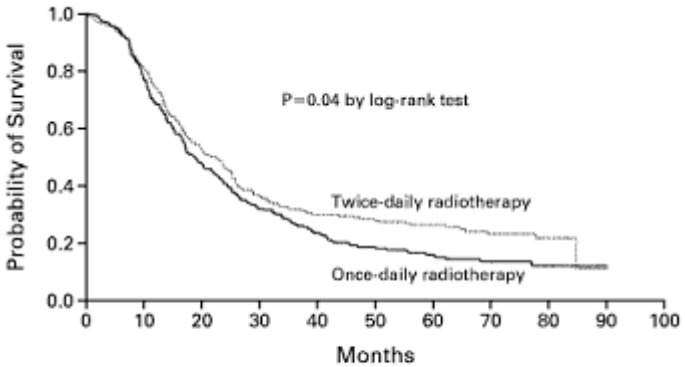
La survenue de censures réduit la précision de l'estimation comme en témoigne la Figure 14, où sont représentées les même données avec et sans censures. En présence de censures (sous-figure de droite), l'intervalle de confiance en fin de courbe est plus large que celui obtenu en l'absence de censure (sous-figure de gauche).

Figure 14 – Conséquence des censures sur la largeur de l'intervalle de confiance. Les deux courbes représentent les mêmes données. Celle de gauche est estimée à partir de l'ensemble des données, tandis que pour celle de droite certains suivis ont été censurés.



Il est utile de trouver sur un graphique de courbe de survie l'évolution du nombre de patients exposés au risque ainsi que le nombre d'événements. Ces données numériques permettent d'apprécier la quantité d'information disponible à un temps donné, et donc d'apprécier la précision et la fiabilité de l'estimation de la survie à ce temps.

Figure 15 – Graphique présentant l'évolution du nombre de patients exposés au risque (3). Cette information est aussi véhiculée par la représentation des censures ou par la représentation des intervalles de confiances.

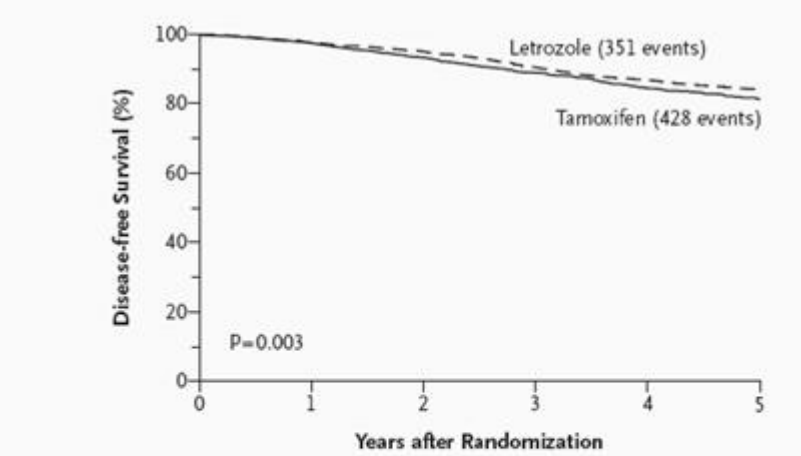


Treatment Group	0–20 Mo	20–40 Mo	40–60 Mo	60–80 Mo	80–100 Mo
	no. of deaths/no. at risk				
Once daily	108/206	48/96	15/47	4/21	0/5
Twice daily	100/211	47/109	7/62	5/42	1/14

Exemple

La figure suivante montre une estimation de la DFS (disease free survival) durant 5 ans. Cependant cette DFS n'est vraiment estimée avec précision que durant la 1er – 2ème année. En effet, par exemple dans le groupe

*letrozole, il ne reste plus que 2964 sujets à risque à la fin de la 2ème année alors que le taux de survie est de 95.1%. En l'absence de censure, le nombre à risque devrait donc être 4003*95.1%=3807. Il y a donc eu 3807-2964=843 censure durant les 2 premières années (21%). A partir de la fin de la 3ème année, le taux de censure dépasse les 60%, ce qui limite fortement l'informativité de la courbe au-delà de la 2ème – 3ème année.*



Letrozole						
No. at risk	4003	3892	2964	1261	892	567
Disease-free survival (%)		97.7	95.1	90.5	86.8	84.0
Tamoxifen						
No. at risk	4007	3896	2926	1238	866	544
Disease-free survival (%)		97.6	93.4	89.0	84.6	81.4

Risque instantané

Risque instantané et gravité de la maladie

Le risque instantané représente la gravité de la maladie à un moment donné. Le taux de survie mesuré après un temps de suivi donné est la conséquence du cumul des risques instantanés de chaque instant.

Dans les maladies chroniques le risque instantané peut être considéré comme constant au cours du temps. C'est par exemple le cas, du risque instantané d'AVC mortel chez les patients porteurs d'une sténose carotidienne. À chaque moment, un sujet a la même probabilité de décéder de cette pathologie.

Dans les maladies aiguës, le risque immédiat (instantané) est particulièrement élevé en début d'évolution de la maladie puis s'amenuise. Dans les maladies qui guérissent, il redevient nul après la guérison. Dans les maladies qui laissent des séquelles, le risque immédiat diminue après la phase aiguë mais ne disparaît pas. Par exemple, à la phase aiguë de l'infarctus du myocarde, le risque immédiat est le plus élevé au cours des 24 premières heures, puis il diminue régulièrement lors de la première semaine, puis encore au cours du premier mois. À distance, il demeure supérieur à celui d'un sujet comparable sans antécédent.

Le risque est la proportion d'événements survenus dans un groupe au bout d'un temps t *de suivi. Le risque est le reflet de l'exposition cumulative au risque instantané au cours de la période de suivi. Mathématiquement* $S(t) = \exp\left[-\int_0^t h(u)\,du\right]$*, la relation entre risque instantané et nombre de décès sur une période est donc complexe. Il est donc quasiment impossible de déduire l'évolution du risque instantané à partir de l'observation de la courbe de survie.*

Comparaison de deux traitements

Le bénéfice d'un traitement est assez souvent recherché à travers la comparaison de la courbe de survie du groupe traité à celle du groupe contrôle. Des tests statistiques existent, comme le test du logrank, pour montrer qu'il existe une différence statistiquement significative entre ces courbes, témoignant alors de l'effet du

traitement. En effet, l'effet du traitement ne peut pas être recherché en comparant directement les durées de survie. Ni le test t, ni les tests non-paramétriques ne prennent en compte les censures. De plus les distributions des temps de survie étant fortement asymétriques, l'hypothèse de normalité nécessaire au test t n'est pas vérifiée.

Test du logrank

Le test du Logrank est le test standard de comparaison de deux courbes de survie. Lorsqu'il est significatif, il permet de rejeter l'hypothèse que les deux courbes sont superposées. Il analyse si, globalement au niveau de chaque décès, la distance entre les deux courbes est plus grande que ce que pourrait expliquer le hasard. C'est-à-dire si le cumul des écarts entre les courbes mesurés chaque fois qu'un décès survient est plus grand que la valeur attendue uniquement du fait du hasard. Ainsi, le test du Logrank analyse les courbes dans leur globalité. Il peut être significatif même si les deux courbes se rejoignent en fin de suivi, aboutissant ainsi au même nombre de décès dans chaque groupe. Cependant le test perd de son efficacité lorsque les deux courbes n'évoluent pas de façon proportionnelle, en particulier lorsque les courbes se croisent. L'analyse visuelle des courbes doit donc toujours accompagner l'interprétation d'un test du Logrank.

D'autres tests existent, comme le test de Gehan (appelé aussi test de Wilcoxon) ou le test de Peto et Prentice plus sensibles aux décès précoces qu'aux décès tardifs. Le test du logrank est équivalent au test de Mantel Haenszel de combinaison de données stratifiées. C'est un test non paramétrique. Le test du Logrank est plus puissant que le test du risque relatif qui ne tient pas compte des censures.

Quantification de l'effet du traitement

Indices d'efficacité

Les données de survie conduisent à des indices d'efficacité spécifiques. La taille de l'effet traitement est mesuré le plus souvent à l'aide du rapport des risques instantanés (« hazard ratio *»). Le* hazard ratio *(HR) est le rapport du risque instantané dans le groupe traité (*h_1*) divisé par le risque dans le groupe contrôle (*h_0*).*

$$hr = \frac{h_0}{h_1}$$

Dans le cas général, le risque instantané est une fonction du temps. Le HR est donc lui aussi une fonction du temps :

$$hr(t) = \frac{h_1(t)}{h_0(t)}$$

Cette fonction décrit l'évolution de l'effet « instantané » du traitement.

Souvent on souhaite caractériser la taille d'un effet par un seul nombre (comme avec le risque relatif par exemple). Avec le hazard ratio, cela nécessite de faire l'hypothèse que ce rapport est constant au cours du temps, même si les risques instantanés varient :

$$\frac{h_0(t)}{h_1(t)} = cte = hr$$

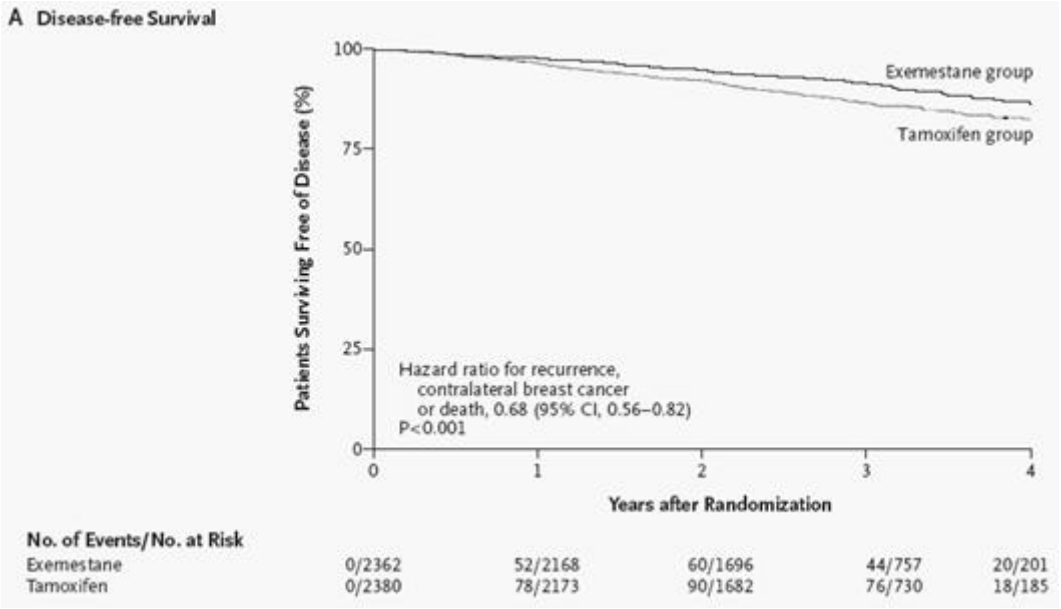
Cette hypothèse est appelée « hypothèse de proportionnalité des risques instantanés » et s'avère assez fréquemment vérifiée. Cela signifie que les courbes de risques instantanés doivent être parallèles au cours du temps. Le risque instantané peut évoluer au cours du temps mais il doit rester proportionnel entre les deux groupes. L'effet du traitement doit donc rester constant au cours du temps.

Le HR peut être interprété comme un risque relatif. C'est le facteur multiplicatif caractérisant l'effet du traitement, mais ce facteur s'applique sur les risques instantanés et non pas sur les risques. Dans les publications, le HR est parfois présenté comme un risque relatif dont il est souvent proche numériquement.

Par exemple, dans un essai de mortalité versus placebo, un hazard ratio de 0,5 signifie que, sous traitement, le risque instantané de décès est seulement la moitié du risque (instantané) sans traitement. En d'autres termes, chaque jour les patients sont exposés à un risque de décès dans la journée réduit de moitié.

Le hazard ratio correspond tout le temps au risque de survenue de l'événement. Un effet bénéfique se traduit donc par une valeur inférieure à 1.

Figure 16 – Exemple d'article rapportant le hazard ratio comme principale mesure de l'effet du traitement. Dans cet essai, le risque (instantané) de récurrence de cancer controlatéral avec Exemestane est 0.68 fois celui sous Tamoxifen (reproduit d'après N Engl J Med 2004;350:1081-92.).



En pratique, à partir des données de survie, le HR est estimé soit à partir de la statistique du logrank, soit en utilisant le modèle de Cox.

Modèle de Cox

Le modèle de Cox est une technique de régression multiple adaptée aux données de survie. Dans un essai thérapeutique, le modèle de Cox est utilisé pour rechercher l'effet traitement en ajustant ou pas sur des covariables. Le modèle de Cox, comme toutes les méthodes multivariées, permet de corriger la mesure le l'effet du traitement des effets qu'auraient pu induire d'éventuelles différences existant au niveau de covariables pronostiques (cf. section : Ajustement *chapitre :* statistiques avancées*). Le modèle de Cox estime le « hazard ratio » lié au traitement.*

Le Tableau 3 rapporte le résultat de la quantification de l'effet de la pravastatine sur la mortalité totale, dans l'essai 4S (4), obtenue par le calcul direct du risque relatif et par l'utilisation d'un modèle de Cox ajusté. En raison de l'importance de l'effectif et du très faible taux de censures avant la fin de l'essai, les valeurs obtenues sont identiques.

Tableau 3 – Calcul de l'effet traitement par le risque relatif ou à l'aide d'un modèle de Cox ajusté dans l'essai 4S (comparaison de la simvastatine au placebo en prévention secondaire).

Mesure de l'effet du traitement sur la mortalité totale	**Estimation (IC 95 %)**
par le risque relatif mesuré en fin d'essai	RR = 0,71 (0,59 ;0,85)
par un modèle de Cox (ajusté sur les caractéristiques de base)	HR = 0,70 (0,58 ;0,85)

Il a été montré qu'il y avait un intérêt à ajuster systématiquement sur les variables pronostiques, même si aucun déséquilibre existe entre les groupes. L'estimation ajustée du HR est toujours plus exacte et plus précise

que celle obtenue sans ajustement (5-7). Les variables utilisées pour l'ajustement doivent cependant être choisie a priori et non pas en fonction des liens observés sur les données (cf. chapitre Statistiques avancées*).*

Le test du Logrank permet aussi d'estimer le rapport HR des risques instantanés. Il procède à la façon d'une méta-analyse, en calculant un HR combiné, non pas à partir d'estimations fournies par différents essais, mais à partir d'estimations effectuées à chaque temps où survient un décès. Comme le HR est considéré comme constant au cours du temps, sa meilleure estimation possible est obtenue en combinant tous les HR mesurés chaque fois que cela est possible, c'est-à-dire à chaque décès. Comme dans une méta-analyse, la combinaison est effectuée en pondérant chaque HR individuel par l'inverse de sa variance.

Les deux modes de lecture d'une courbe de survie

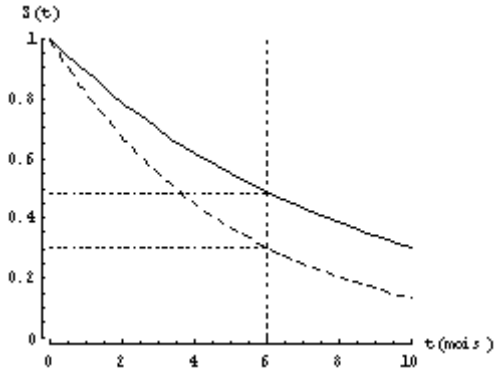
> *Il est abusif de dire qu'un traitement évite 5 décès pour 1000 patients traités durant 5 ans. Il repousse seulement, en moyenne, au-delà de 5 ans le décès de 5 patients pour 1000 patients traités.*

Quand un essai de mortalité d'un nouveau traitement produit un résultat positif, la presse médicale et grand public proclament souvent que « ce nouveau traitement sauve des vies ». L'utilisation des indices comme le NNT conduit à employer des termes comme « nombre de décès évités ». Cette formulation est quelque peu abusive, car tout individu est mortel (8). Les traitements n'évitent pas les décès, ils les retardent simplement. La meilleure façon serait de mesurer leur effet en termes d'augmentation de la durée de survie. Mais, nous avons vu, que cette mesure n'est pas réaliste car très rarement mesurable directement. La difficulté a été contournée en raisonnant sur les probabilités de décès à un moment donné, mais il s'agit que de probabilités (de risque) qui ne peuvent pas être traduites en termes de vies sauvées. Un traitement ***n'évite*** *pas, stricto sensu, 5 décès pour 1000 patients traités durant 5 ans. Il repousse seulement, en moyenne, au-delà de 5 ans le décès de 5 patients pour 1000 patients traités. À la rigueur, il est possible de dire que 5 décès prématurés ont été évités en moyenne sur une période de 5 ans.*

Lecture verticale

La lecture verticale des courbes de survie consiste à mesurer de façon verticale la distance qui sépare deux courbes de survie. Cette mesure s'effectue à un temps donné. La lecture verticale revient à comparer des risques (qui sont égaux à 1 - taux de survie*). C'est donc le mode de lecture implicite réalisé lorsque l'on raisonne en termes d'indices d'efficacité : risque relatif, odds ratio, différence de risque ou NNT.*

Figure 17 – Lecture verticale des courbes de survie



Ce mode de lecture est utilisable même si la durée de suivi n'a pas été suffisante pour enregistrer le décès de tous les patients.

Limitations de la lecture verticale

La Figure 18 représente l'évolution au cours du temps des principaux indices d'efficacité RR, OR et DR. La sous-figure a *représente les deux courbes de survie avec le groupe contrôle en trait pointillé. La sous-figure* c *montre la relation entre le RR et la survie dans le groupe contrôle.*

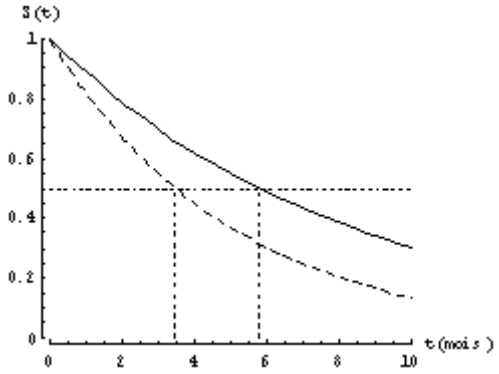
Il apparaît qu'aucun de ces indices n'est constant au cours du temps, bien que le vrai effet traitement représenté par le hasard ratio (rapport des risques instantanés) soit constant. Les différences ne sont cependant très importante que pour des différence de durées de suivi elles aussi très importantes.

Figure 18 – Évolution en fonction de la durée de suivi des indices d'efficacité. a : courbes de survie des deux groupes comparés ;b : évolution du risque relatif au cours du temps ; c : évolution de la différence des risques ; d : évolution de l'odds ratio ; e : représentation du risque relatif en fonction du taux de survie (et non plus du temps).

Lecture horizontale

La lecture horizontale des courbes de survie consiste à mesurer de façon horizontale la distance qui sépare deux courbes de survie. Cette lecture s'effectue à un niveau de survie donné, par exemple, 50%. À ce taux de survie de 50% correspond un temps qui est la médiane des temps de survie. En effet, la durée de survie de la moitié des sujets de la population d'origine a été inférieure à cette valeur, puisqu'à ce temps il n'y a plus que 50% des sujets survivants. Pour refléter l'imprécision des estimations, les médianes de survie doivent être rapportées accompagnées de leur intervalle de confiance.

Figure 19 – Lecture horizontale des courbes de survie



La lecture horizontale permet alors de mesurer l'effet du traitement en terme d'augmentation de la médiane de survie. L'unité de cet effet est l'unité de temps. Par exemple une augmentation de 2,3 mois de la médiane de survie (IC 95% = [1,1 ; 3,7]).

Figure 20 – reproduit d'après N Engl J Med 2004;350:2335-42

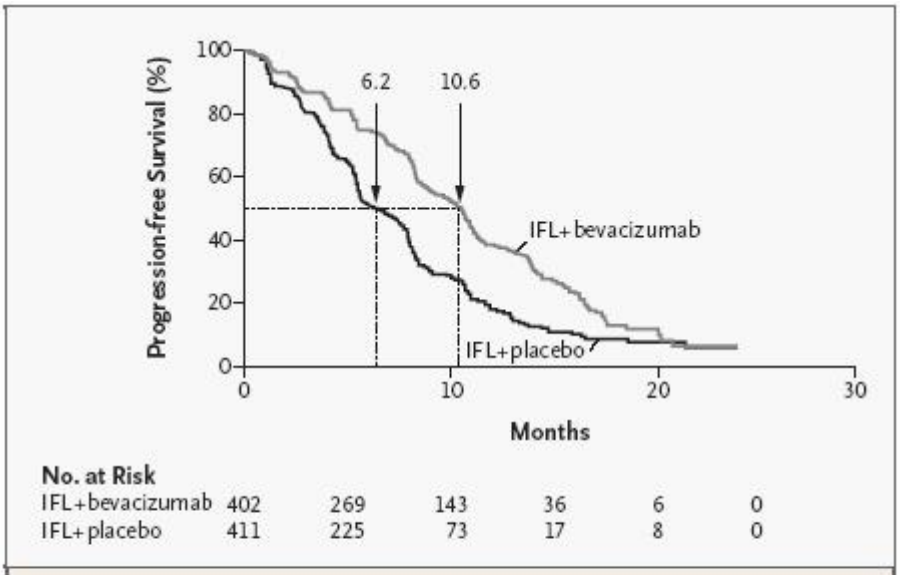
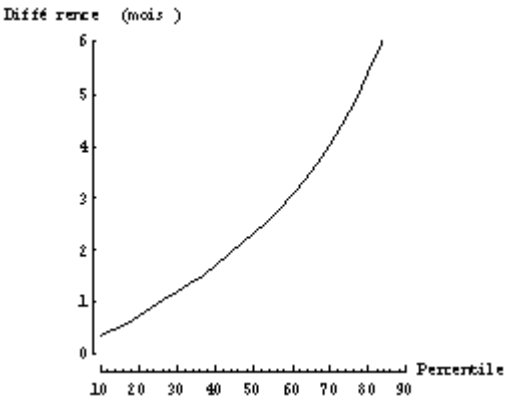


Figure 2. Kaplan–Meier Estimates of Progression-free Survival.
The median duration of progression-free survival (indicated by the dotted lines) was 10.6 months in the group given irinotecan, fluorouracil, and leucovorin (IFL) plus bevacizumab, as compared with 6.2 months in the group given IFL plus placebo, corresponding to a hazard ratio for progression of 0.54 (P<0.001).

Lorsque la durée de suivi a été insuffisante pour atteindre un taux de survie d'au moins 50% dans les deux groupes, l'effet du traitement se mesure en terme d'augmentation d'un autre percentile de survie, par exemple le 75ème percentile. Dans ce cas, la signification de cet effet devient moins intelligible et ne présente pas d'avantage par rapport à la lecture verticale.

De façon similaire à ce qui est observé avec la lecture verticale, la lecture horizontale produit des mesures de l'effet traitement variant en fonction du niveau vertical où est effectuée la lecture. L'augmentation du 90ème percentile de survie est plus importante que l'augmentation de la médiane de survie (cf. Figure 21). Par contre, l'expression sous forme relative de la différence horizontale est constante et ne dépends pas du percentile mesuré (pour un risque instantané constant).

Figure 21 – Évolution de la différence entre les courbes mesurée horizontalement pour différents niveau de percentile. Par contre l'expression sous forme relative de cette différence horizontale est constante (67% dans ce cas).



Exemple de présentation complète des résultats d'une analyse de survie

*"*The median duration of overall survival, the primary end point, was significantly longer in the group given studied treatment than in the group given placebo (20.4 months vs. 15.7 months), which corresponds to a hazard ratio for death of 0.66 (P<0.001) (Table 3 and Fig. 1), or a reduction of 34 percent in the risk of death in the studied treatment group. The one-year survival rate was 74.3 percent in the group given studied treatment and 63.4 percent in the group given placebo (P<0.001)*"*

Table 3. Analysis of Efficacy.*

End Point	IFL plus Placebo	IFL plus Bevacizumab	P Value
Median survival (mo)	15.6	20.3	<0.001
Hazard ratio for death		0.66	
One-year survival rate (%)	63.4	74.3	<0.001

Gain en durée de survie

Durée de survie

En démographie, l'espérance de vie est la moyenne des durées de survie d'un groupe de sujets de même âge. Par exemple, l'espérance de vie des hommes de 40 ans est de 30 ans signifie que la moyenne des durées de survie futures d'un groupe de sujets âgés de 40 ans est de 30 ans.

La durée moyenne de survie est égale à l'aire sous la courbe de survie. Un traitement bénéfique en diminuant le risque instantané augmente l'aire sous la courbe. Le bénéfice apporté par un traitement peut être exprimé en gain de durée moyenne de survie (ou gain en espérance de vie) (9). Ce gain est la surface de la zone comprise en les deux courbes de survie. Souvent ce gain est appelé « gain en espérance de vie ». Le terme espérance de vie est alors utilisé dans une acception différente de celle de l'espérance de vie démographique (cf. rappel) et correspond à la durée de suivi moyenne du groupe considéré.

En pratique, l'utilisation de l'aire sous la courbe se heurte à la même difficulté que le calcul direct de la moyenne des temps de survie : celle de devoir suivre les patients jusqu'au décès de tous. Deux voies sont envisageables pour contourner cette difficulté. Celle d'utiliser un modèle mathématique pour compléter la partie manquante des courbes et celle de calculer un gain en durée de survie sur la durée de l'essai. Ces deux techniques présentent des limites.

Modélisation

L'utilisation de la modélisation permet de compléter la partie manquante des courbes de suivie. Un modèle paramétrique, comme le modèle de Gompertz, est ajusté sur la partie disponible de courbes. Il permet ensuite le calcul de la totalité de l'aire sous la courbe. Cette approche nécessite de faire des hypothèses sur la queue de la distribution des durées de survie, hypothèses qui sont très difficile à vérifier.

L'extension de survie

Un gain en espérance de vie est parfois calculé au terme d'un essai, même si la totalité des durées de survie ne sont pas disponibles (10). Ce calcul n'a cependant pas beaucoup de sens et il est d'interprétation hasardeuse.

Par exemple, le gain en espérance de vie a été calculé pour les 3 grands essais de statines WOSCOPS, 4S et CARE. Dans l'essai de prévention secondaire 4S, au terme du suivi de 5 ans, la simvastatine est associée avec un gain en espérance de vie de 24 jours sur une durée de 5,4 ans (11).

Quoiqu'ils semblent parlants et proches de ce que l'on cherche à faire avec l'utilisation des statines, ces résultats n'ont que peu de valeur. En effet, que signifie une moyenne de durées de survie ne concernant qu'une petite proportion de tous les sujets. Par exemple le résultat avancé pour WOSCOPS doit s'interpréter de la façon suivante : le traitement durant 5 ans par la pravastatine allonge en moyenne de 3 semaines la durée de survie des 4% de patients ***qui décèdent avant 5 ans****. Le gain en espérance de vie des 96% de survivants à plus de 5 ans n'est pas connu.*

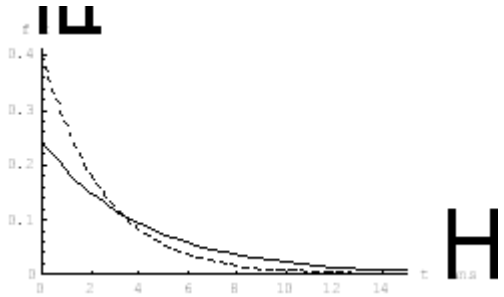
Il apparaît que l'extension de survie partielle mesurée à la fin d'un essai sous-estime fortement le réel gain en espérance de vie. Par exemple, pour WOSCOPS il est possible d'estimer ce gain à environ 2 ans.
Par contre dans les essais comme CONCENSUS où il est possible d'observer la totalité des courbes de survie, le calcul du gain en espérance de vie est possible et représente parfaitement ce que l'on cherche à mesurer (8).

Interprétation d'un gain de survie

Des pièges sont à éviter dans l'interprétation d'un gain en « espérance de vie » (12). Ce gain peut être perçu comme un temps supplémentaire de durée de vie gagné en fin de vie. Ainsi pour certaines personnes, étant donné son caractère lointain, cette augmentation peut paraître de peu d'intérêt : « il est préférable de gagner 1000 euros maintenant que dans 10 ans ». Cependant, un gain en espérance de vie, même minime, traduit en fait un bénéfice immédiat qui est une diminution du risque de décéder à tout moment (risque instantané). Il s'ensuit un décalage dans la distribution des temps de survie (Figure 22) avec une diminution de la fréquence des durées de survie courtes et augmentation de celle des survies prolongées.

De même, il ne faut pas concevoir le gain en espérance de vie comme une prolongation de la durée de survie de chaque sujet. Il est inexact de dire que « sous traitement, la durée de survie de chaque sujet est augmentée de x mois ». En effet, si cette assertion était correcte, cela devrait se traduire par l'apparition d'un plateau initial dans la courbe de survie, prolongeant le taux de survie de 100% sur une durée égale au gain en « espérance de vie ». Le gain en « espérance de vie » est une valeur moyenne : la moyenne du groupe est augmentée, mais en raison de la variabilité, les valeurs individuelles peuvent être inchangées, augmentées ou diminuées.

Figure 22 – Illustration de la modification de la distribution des temps de survie induite par un effet traitement (courbe en trait plein). La courbe en pointillée représente la distribution des temps de survie sans traitement.



Bibliographie

1. *Collet D. Modelling survival data in medical research. London: Chapman & Hall; 1994.*

2. *Hill C, Com-Nougué C, Kramar A, Moreau T, O'Quigley J, Senoussi R, et al. Analyse statistique des données de survie. Paris: Flammarion; 1990.*

3. *Turrisi AT, Kim K, Blum R, Sause WT, Livingston RB, Komaki R, et al. Twice-Daily Compared with Once-Daily Thoracic Radiotherapy in Limited Small-Cell Lung Cancer Treated Concurrently with Cisplatin and Etoposide. NEJM 1999;340(4):265-271.*

4. *Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet 1994;344:1383-89.*

5. *Chastang C, Byar DP, Piantadosi S. A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. Stat Med 1988;7:1243-1255.*

6. *Edwards D. On model prespecification in confirmatory randomized studies. Stat Med 1999;18:771-785.*

7. *Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the cox proportional hazards model. Stat Med 1995;14:735-746.*

8. *Tan LB, Murphy R. Shifts in mortality curves: saving or extending lives? Lancet 1999;354:1378-81.*

9. *Wright JC, Weinstein MC. Gains in life expectancy from medical interventions - standardizing data on outcomes. NEJM 1998;339:380-6.*

10. *Yusuf S, Zucker D, Peduzzi P, Fisher LD, Takaro T, Kennedy JW, et al. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. Lancet 1994;344(8922):563-70.*

11. *Krut LH. On the statins, correcting plasma lipid levels, and preventing the clinical sequelae of atherosclerotic coronary heart disease. Am J Cardiol 1998;81(8):1045-6.*

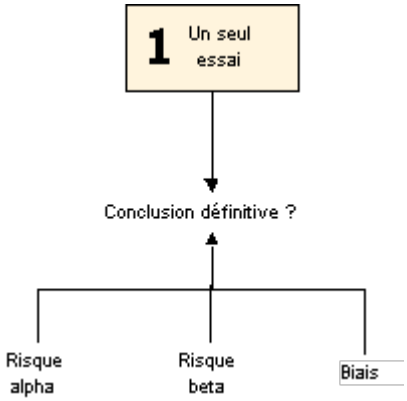
12. *Naimark D, Naglie G, Detsky AS. The meaning of the life expectancy: what is a clinically significant gain? Journal of General Internal Medicine 1994;9:702-707.*

LA META-ANALYSE

Principes généraux

Introduction

L'essai clinique randomisé correctement conçu et réalisé produit les preuves les plus fiables de l'efficacité des traitements. Cependant, le résultat d'un seul essai thérapeutique est-il suffisant pour conclure définitivement à l'efficacité du nouveau traitement ? [1].



Dans un essai, la conclusion sur l'efficacité se base sur un test statistique qui laisse la possibilité d'une erreur statistique. Une différence statistiquement significative peut être le fruit du hasard. Dans ce cas, conclure à l'efficacité du traitement est une conclusion erronée.

De plus, un résultat peut être lié aux conditions de réalisation de l'essai : biais, patients particulièrement bien sélectionnés, environnement de soins particuliers, etc.

En raison de ces points, il est en général indispensable qu'un premier essai clinique positif soit vérifié par un autre. Ce principe n'est pas propre au domaine de l'évaluation clinique et prévaut dans tous les domaines expérimentaux : le résultat d'une expérience doit être vérifié par au moins une autre avant d'être accepté. Dans le domaine réglementaire, il se traduit par la nécessité de disposer des résultats de 2 essais pivots pour délivrer une autorisation de mise sur le marché.

En pratique, on se trouve presque constamment confronté à plusieurs essais qu'il faudra prendre en compte simultanément avant de se prononcer sur l'efficacité du traitement.

Les problèmes liés à la synthèse de plusieurs résultats d'essais cliniques

La synthèse des résultats de plusieurs essais cliniques pose un certain nombre de problèmes. La méta-analyse est la technique qui a été développée pour contourner ces difficultés.

Problèmes statistiques

Dans une série de plusieurs essais, le hasard engendre suivant le cas des résultats faussement positifs (si le traitement est sans efficacité) ou faussement négatifs (si le traitement est efficace). Ces faux résultats conduisent à des décisions erronées s'ils sont interprétés sans tenir compte des autres résultats.

La multiplicité des essais s'accompagne souvent de résultats apparemment discordants et tirer une conclusion globale par la seule logique discursive n'est pas toujours possible. En général, des essais concluants (donnant un résultat statistiquement significatif en faveur de l'efficacité) coexistent avec des résultats non concluants. Ces discordances apparentes donnent la possibilité de soutenir les deux conclusions opposées : celle de l'existence de l'efficacité et celle de son absence.

Schématiquement, les tenants de l'existence de l'efficacité argumenteront à partir des résultats concluants, les résultats non significatifs étant expliqués par un manque de puissance. Les défenseurs de l'absence d'efficacité mettront en avant les résultats non significatifs et expliqueront les résultats positifs par le fait du hasard et du risque de première espèce alpha.

Tableau 1 – Résultats des essais d'angioplastie primaire à la phase aiguë de l'infarctus du myocarde.

	Nb de patients	Mortalité groupe traité	Mortalité groupe contrôle	P
Essai 1	56	6,9%	5,2%	NS
Essai 2	100	6,0%	2,0%	NS
Essai 3	395	2,5%	6,5%	NS
Essai 4	52	4,3%	17,2%	NS
Essai 5	103	4,2%	3,5%	NS
Essai 6	301	1,9%	7,3%	p<0,05

Le Tableau 1 rapporte les résultats de 6 essais qui ont tous comparé l'angioplastie primaire à la fibrinolyse à la phase aiguë de l'infarctus du myocarde. En pratique, il convient de se prononcer à partir de ces données qui représentent toutes les informations non biaisées disponibles sur le sujet. Trois conclusions sont possibles :

1. *Les données disponibles montrent avec un degré de certitude élevé (démontrent) la supériorité de l'angioplastie primaire sur la fibrinolyse. L'angioplastie peut être recommandée en pratique.*
2. *Les données disponibles montrent avec un degré de certitude élevé (démontrent) l'absence de supériorité de l'angioplastie primaire sur la fibrinolyse. L'angioplastie primaire ne doit pas être recommandée en pratique.*
3. *Les données sont insuffisantes pour conclure. Un ou plusieurs nouveaux essais sont nécessaires*

Pour adopter l'une de ces conclusions, ces données doivent être synthétisées. Autrement une analyse individuelle des résultats donne des arguments aussi bien pour soutenir l'existence que l'absence d'efficacité du nouveau traitement. Par exemple, la conclusion à l'absence d'efficacité peut être soutenue en argumentant que la grande majorité des essais a obtenu un résultat non statistiquement significatif (5 essais sur 6). Le résultat significatif de l'essai 6 allant contre cette conclusion est alors mis sur le compte du hasard (risque alpha dans un contexte de comparaisons multiples). Cependant, il est aussi possible de soutenir la conclusion inverse c'est-à-dire celle de l'existence de l'efficacité en argumentant que les résultats non significatifs sont simplement dus à un manque de puissance lié à un effectif insuffisant. Cette efficacité est mise en évidence par l'essai 6 qui a obtenu un résultat significatif du fait d'une puissance correcte (nombre de patients supérieur). Au total, il apparaît difficile de conclure sur l'efficacité ou non de ce nouveau traitement malgré l'existence de plusieurs essais thérapeutiques dont certains de taille conséquente.

Une technique de synthèse satisfaisante devra prendre en compte la possibilité d'un manque de puissance des essais négatifs et d'un risque d'erreur alpha pour les essais positifs. Cette technique de synthèse devra donc faire appel à des méthodes statistiques.

Sélection des essais

Effectuée de manière narrative la synthèse de plusieurs résultats est en général fortement subjective. Classiquement ce type de synthèse s'effectue dans des revues générales de la littérature.

À l'impossibilité de gérer les fluctuations aléatoires s'ajoute celui de la sélection des essais. L'absence de méthodes et de critères définis a priori laisse la possibilité que les essais pris en considération soient sélectionnés. Ranskov a montré que dans les revues de la littérature concernant les traitements hypocholestérolémiants, les essais positifs étaient 5 fois plus cités que les résultats [2]. Ainsi, les revues de la

littérature dans le domaine de l'évaluation des traitements s'apparentent souvent à de simples opinions argumentées par quelques résultats d'essais bien sélectionnés.

C'est principalement pour résoudre ces problèmes et éviter ces travers liés principalement à l'absence de méthode que la méthodologie de la méta-analyse a été développée.

Pollution de la synthèse par les essais biaisés

Les biais des essais eux-mêmes sont la première source de biais d'une synthèse. Si les informations sources sont potentiellement biaisées, le résultat de la synthèse l'est aussi. Seule une sélection appropriée des essais en fonction de la qualité de leur méthode garantit une bonne qualité méthodologique du résultat.

Cependant, le biais introduit par un ou quelques essais biaisés est « dilué » par les essais non biaisés si ces derniers sont majoritaires, en taille et en quantité d'information. La conclusion de la méta-analyse sera moins erronée que celle fondée uniquement sur l'essai biaisé.

La qualité méthodologique d'un essai est difficile à évaluer. De nombreuses échelles de qualité ont été publiées, mais des études empiriques récentes montrent qu'il est possible d'écarter la possibilité de biais avec suffisamment de certitude en utilisant seulement 3 critères [3-5] :

- *le caractère aléatoire de la répartition des patients entre les groupes et plus particulière son imprévisibilité qui garantit que les investigateurs n'ont pas pu déterminer à l'avance dans quel groupe devait aller le patient qu'il souhaitait inclure,*
- *le suivi en double insu quand il était éthiquement possible,*
- *l'absence ou le faible taux des patients randomisés mais exclus de l'analyse.*

Biais de publication

À côté de la répercussion des biais des essais, toute synthèse d'information est sujette à un biais qui lui est propre : le biais de publication (« publication bias *»).*

Les résultats négatifs (essais ne montrant pas de différence significative) sont moins fréquemment publiés que les résultats positifs (essais montrant une différence significative). L'étude du devenir de 285 protocoles soumis au comité d'éthique d'Oxford révèle que 85% des résultats positifs ont été publiés contre seulement 56% des résultats négatifs [6]. Il existe ainsi une publication sélective des résultats positifs au détriment des résultats négatifs. Cela ne veut pas dire que ces derniers ne sont jamais publiés mais plus difficilement et seulement pour une partie d'entre eux.

Les causes de ce phénomène sont certainement nombreuses [7, 8] et impliquent à la fois les comités de lecture des revues, peu séduits par un résultat négatif, et les auteurs qui n'investissent pas dans la rédaction d'un article qui a peu de chance d'être accepté.

Ce phénomène de publication sélective va fausser les conclusions que l'on peut tirer à partir des résultats publiés. C'est le biais de publication. Dans une synthèse, si aucune recherche poussée des essais non publiés n'est entreprise, le risque couru est de ne travailler qu'avec les essais positifs, ce qui conduit à une surestimation de l'efficacité du traitement.

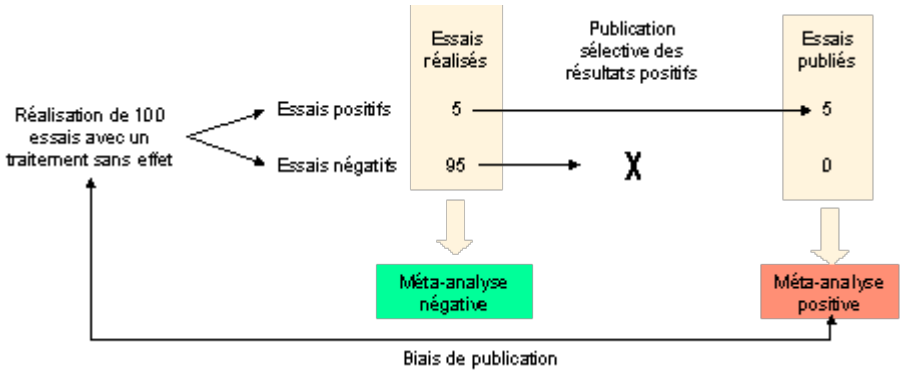


Figure 1 – Si de nombreux essais sont réalisés avec un traitement sans efficacité, certains d'entre eux auront cependant un résultat statistiquement significatif, uniquement du fait du risque d'erreur statistique alpha que l'on consent au niveau du test statistique. Ainsi, si 100 essais sont réalisés, 5 d'entre eux seront positifs à tort, du fait du hasard. Ainsi, si uniquement les essais positifs sont publiés, une synthèse ne portant que sur les résultats publiés donnera une fausse impression d'efficacité du traitement. C'est le biais de publication.

Les conséquences potentiellement dommageables du biais de publication sont illustrées par l'exemple des antiarythmiques de classe 1 en post infarctus avec la non publication en 1980 d'un essai qui montrait une forte augmentation de mortalité avec une molécule de cette classe, la lorcainide.

Une étude empirique (« empirical study ») a montré que l'exclusion, dans les méta-analyses des essais non publiés entraîne en moyenne une surestimation de 15% (IC95% 4% ;28%) de la taille de l'effet [9]. De même, l'exclusion des essais publiés uniquement sous forme « d'abstracts » entraîne en moyenne une surestimation de 33% (IC95% 10% ;60%) de l'effet.

La solution à ce problème serait la mise en place de registres prospectifs d'essais qui en enregistrant les essais à leur mise en place, permettraient, par la suite, de retrouver leur trace même s'ils n'ont jamais été publiés.

Qu'est ce que la méta-analyse

*Une méta-analyse est une synthèse quantifiée, **e**xhaustive, rigoureuse et reproductible des résultats d'essais cliniques répondant à une question thérapeutique donnée.*

La méta-analyse permet de synthétiser les résultats des essais thérapeutiques répondant à une question thérapeutique donnée [10]. Cette synthèse se déroule en suivant une méthodologie rigoureuse qui a pour but d'assurer son impartialité et sa reproductibilité.

La méta-analyse est une synthèse systématique et quantifiée. Elle est systématique car elle implique une recherche exhaustive de tous les essais favorables ou non au traitement étudié, publiés et non publiés. Elle est quantifiée car elle se base sur des calculs statistiques donnant une estimation précise de la taille de l'effet du traitement. L'utilisation de techniques statistiques permet de prendre en compte le fait que les conclusions d'un essai thérapeutique se basent sur des tests statistiques et que les résultats obtenus dans plusieurs essais peuvent être différents, uniquement du fait du hasard.

Les points clés d'une méta-analyse
Une méta-analyse est une **synthèse :** • Exhaustive • Rigoureuse et reproductible • Quantifiée

La méthodologie de la méta-analyse a été conçue pour apporter une solution aux différentes difficultés auxquelles conduit la problématique de la synthèse de plusieurs essais.

Tableau 2 – Problèmes posés par la synthèse de plusieurs essais et solutions apportées par la méta-analyse.

Problèmes posés par la synthèse de plusieurs essais		Solution apportée par la méta-analyse
existence d'un risque d'erreur statistique au niveau des résultats des essais	→	calcul d'un effet traitement commun à partir des données de chaque essai
possibilité d'une sélection arbitraire des essais d'après leurs résultats	→	prise en compte de tous les essais quel que soit leurs résultats
la prise en compte d'essais biaisés qui faussent le résultat de la synthèse	→	sélection des essais dont la qualité méthodologique garantit suffisamment l'absence de biais
biais de publication lié à la non publication des résultats négatifs	→	recherche exhaustive des essais publiés et non publiés

Ainsi les différents principes méthodologiques de la méta-analyse ont pour objet d'apporter une solution satisfaisante aux différents problèmes soulevés par la synthèse de plusieurs essais

Intérêts de la méta-analyse

Liste des intérêts de la méta-analyse

Par rapport à l'analyse d'un seul essai ou de plusieurs pris individuellement, la méta-analyse permet :

- *d'augmenter la puissance statistique (la probabilité de trouver un résultat significatif) de la recherche d'un effet traitement. La méta-analyse est alors utilisée pour mettre en évidence l'effet du traitement dans une situation où les essais déjà réalisés pris individuellement ne permettent pas de conclure car aucun n'a donné de résultat statistiquement significatif,*
- *de réconcilier des résultats apparemment discordants et de lever le doute,*
- *d'augmenter la précision de l'estimation de la taille de l'effet du traitement, en la basant sur une plus grande quantité d'informations, consécutive à l'augmentation du nombre de sujets prenant part à la comparaison,*
- *de synthétiser une somme d'informations parfois très importante,*
- *de tester et augmenter la généralisation d'un résultat à un large éventail de patients. L'estimation issue d'une méta-analyse est ainsi plus proche de l'effet qui sera vraisemblablement obtenu avec l'utilisation "en pratique" du médicament. Pris individuellement, chaque essai a sélectionné avec beaucoup de soin les sujets inclus. En regroupant des essais portant sur des groupes de sujets de caractéristiques différentes, la méta-analyse procure un moyen d'approcher le "patient moyen tout venant" de la population de diffusion,*
- *d'expliquer la variabilité des résultats entre essais (notamment par suite de biais dans certains essais),*
- *de réaliser des analyses en sous-groupes et effectuer une recherche des groupes de patients susceptibles de bénéficier le plus d'un traitement, ou au contraire ne pas en bénéficier. La prise en compte simultanée de plusieurs essais apporte une plus grande variété dans les caractéristiques de base des patients étudiés et aussi des effectifs accrus dans les sous-groupes. Elle permet aussi de vérifier qu'un résultat d'un sous-groupe se retrouve sur l'ensemble des essais,*
- *de mettre un essai en perspective en le confrontant aux autres essais du domaine,*
- *de constater le manque de données fiables dans un domaine et mettre en place un essai,*
- *de répondre à une question non initialement posée par les essais.*

Ainsi, les méta-analyses sont particulièrement utiles quand les essais sont de trop petite taille pour donner des résultats fiables, quand la réalisation d'un essai de grande taille est impossible, quand les essais ont été réalisés mais donnent des résultats discordants ou non concluants ou quand les résultats d'un essai définitif sont attendus.

Les analyses en sous-groupes, avec recherche de différences dans l'effet du traitement entre les sous-groupes (appelées hétérogénéité) évitent une synthèse réductrice qui pourrait noyer dans la masse des essais, des effets spécifiques observés seulement dans certains d'entres eux.

Finalement, pour la recommandation d'un traitement pour la pratique, l'intérêt de la méta-analyse est d'éviter les décisions erronées engendrées par une analyse séparée des essais qui se ferait abuser par les résultats faux positifs (ou faux négatifs) produits par le hasard, ou par des biais méthodologiques, ou par une publication sélective ou par une désinformation.

Les sources de données

Un point important de la méta-analyse est l'exhaustivité. L'obtention de celle-ci nécessite un effort important. La seule utilisation de Medline est insuffisante pour garantir l'exhaustivité de la recherche des essais [11]. En pratique, l'ensemble des sources d'informations disponibles doit être utilisé (Tableau 3). La recherche des essais randomisés ne peut pas se contenter de l'utilisation du seul mot clé « randomised controlled trials », mais doit se baser sur des stratégies de recherche spécifique de bonne sensibilité [12].

Tableau 3 Moyens de recherche des essais à utiliser en méta-analyse dans un but d'exhaustivité

- Bases bibliographiques informatisées : Medline, Embase (plus européenne que Medline), Biosis, Pascal (base française), Lillacs (base americo-latine), bases spécialisées spécifiques du domaine étudié (comme PsyLit, CancerLit, etc.)
- Références des comptes rendus d'essais et des articles de revues et références des références pour obtenir un effet « boule de neige »
- Registre des essais randomisés de la Collaboration Cochrane
- Recherche dans les abstracts des congrès d'essais dont les résultats ont été communiqués oralement, mais qui ne sont pas encore ou qui ne seront jamais publiés
- Registre d'essais thérapeutiques (existant dans certains domaines comme celui de la thrombose ou du cancer)
- Recherche manuelle dans les revues de la spécialité à la recherche d'essais non indexés comme tels dans les bases bibliographiques

Moyens plus spécifiques de la recherche des essais non publiés :

- Recherche auprès des laboratoires pharmaceutiques concernés, auprès des leaders d'opinion du domaine ou des investigateurs potentiels, en particulier pour les essais non publiés
- Recherche dans les abstracts des congrès des essais qui n'auraient donné lieu qu'à une présentation orale non suivie de publication par la suite
- Recherche dans la littérature grise (thèse, rapport interne, revue journalistique, etc.)
- Registre prospectif enregistrant les essais lors de leur mise en place (déclaration aux comités d'éthique, etc.)

Les résultats d'une méta-analyse

Hypothèse fondamentale de la méta-analyse

Une hypothèse fondamentale, l'hypothèse d'homogénéité, est nécessaire pour donner un sens au principe de la méta-analyse. En effet, il n'est possible d'envisager de regrouper plusieurs essais pour estimer l'efficacité d'un traitement, que si l'on considère que la quantité d'effet de ce traitement est une constante, et donc que chaque essai thérapeutique mesure cette même constante (Figure 2). Les variations observées entre plusieurs essais thérapeutiques ne proviennent que des fluctuations aléatoires. Ainsi, il serait possible de concevoir une série d'essais comme une série de mesures d'un même effet traitement, soumises à des fluctuations

d'échantillonnages. Les calculs de méta-analyse cherchent alors la meilleure estimation possible de cet effet traitement commun.

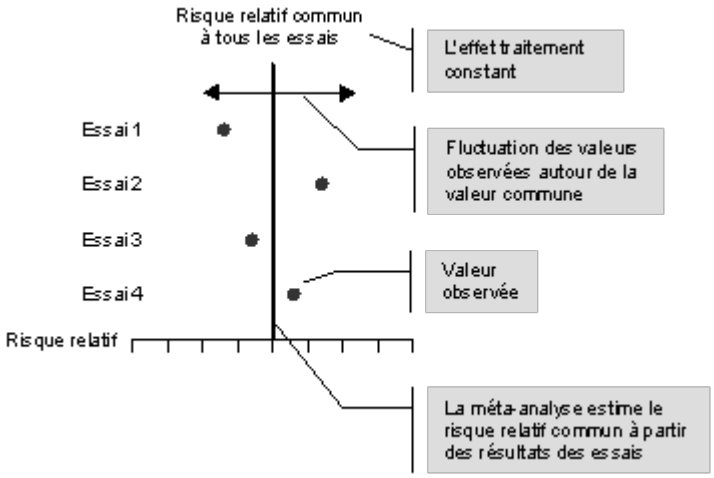


Figure 2 – Hypothèse fondamentale de la méta-analyse

Les calculs de méta-analyse produisent les résultats suivants :

- ***l'estimation de l'effet traitement commun*** *accompagnée de son intervalle de confiance,*
- ***le test d'association*** *qui teste l'existence d'un effet traitement non nul. Si ce test est significatif, il est possible de conclure à l'existence d'un effet traitement non nul.*
- ***le test d'hétérogénéité*** *qui évalue si les résultats de tous les essais peuvent être considérés comme similaires (hypothèse d'homogénéité). Le regroupement de ces essais est alors licite. Si le test d'hétérogénéité est significatif, il existe au moins un essai dont le résultat ne peut pas être considéré comme identique aux autres et le regroupement n'est pas licite. Il convient alors de recourir à des techniques spéciales (modèle d'effet aléatoire).*

Représentation graphique

Les résultats d'une méta-analyse sont fréquemment représentés sous forme graphique (Figure 3). Sur ce graphique, les risques relatifs obtenus au niveau de chaque essai et globalement par la méta-analyse sont représentés encadrés par leur intervalle de confiance (les lignes). Un trait vertical correspondant à la valeur 1 de l'odds ratio matérialise le seuil de non-efficacité. Si l'intervalle de confiance englobe ce repère, le résultat obtenu au niveau de l'essai (ou de la méta-analyse) n'est pas statistiquement significatif. Les risques relatifs supérieurs à 1 témoignent d'un risque supérieur dans le groupe traité par rapport au groupe contrôle. Ce type de schéma permet aussi facilement de positionner chaque essai par rapport au résultat global (position relative de son résultat par rapport à la ligne verticale passant par la valeur globale). L'existence d'un ou plusieurs essais, dont la totalité de l'intervalle de confiance se trouve en dehors de cette ligne, témoigne très probablement d'une hétérogénéité entre les essais.

La Figure 3 représente un des résultats de la méta-analyse des essais des préventions primaires des maladies cardiovasculaires par les traitements hypocholestérolémiants. L'interprétation de ce graphique est la suivante : le regroupement de ces 6 essais représentant 34 000 patients met en évidence une réduction de la fréquence des événements coronariens, caractérisée par un risque relatif de 0,76. Ce résultat est hautement significatif ($p<0,001$). Il n'est pas mis en évidence d'hétérogénéité entre ces résultats (test d'hétérogénéité non significatif). Les fluctuations des résultats autour de l'estimation commune (représentées par le trait vertical en pointillé) surviennent au hasard.

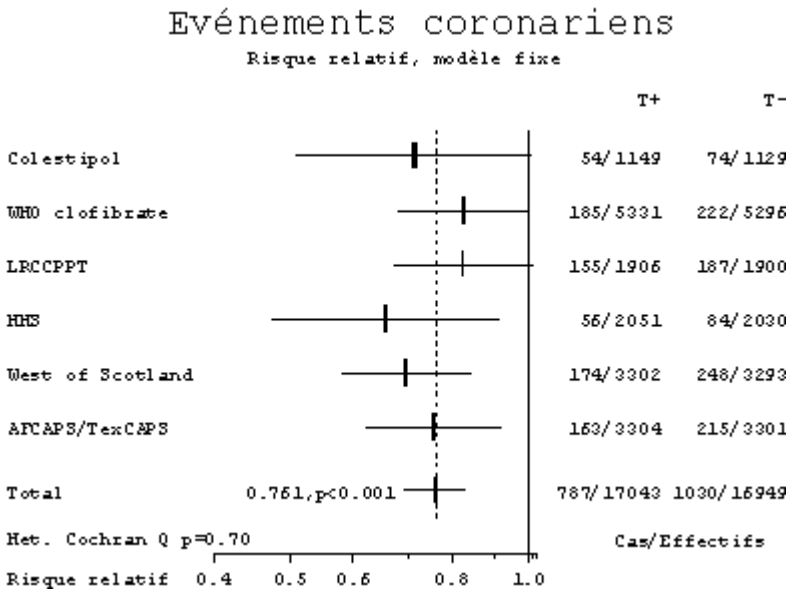


Figure 3 – Graphique typique de méta-analyse. Ce type de graphique représente les risques relatifs des essais entourés de leur intervalle de confiance à 95% ainsi que celui issu des calculs de méta-analyse (« total »). Les deux colonnes numériques de gauche rapportent le nombre d'événements et les effectifs des deux groupes. Le résultat du test d'hétérogénéité (Het. Cochran Q) est aussi présenté.

Hétérogénéité

Il existe une hétérogénéité lorsque les résultats des essais ne peuvent pas être considérés comme similaires. Il existe au moins un essai dont le résultat ne peut pas être considéré comme identique aux autres. L'hétérogénéité est recherchée à l'aide du test du même nom, mais dont la puissance n'est satisfaisante que lorsqu'il y a plusieurs dizaines d'essais. En cas d'hétérogénéité, il est nécessaire de tenter d'identifier le facteur (ou les facteurs) l'expliquant. En cas de succès de cette démarche, la méta-analyse sera réalisée en fonction des sous-groupes définis par ce facteur introduisant l'hétérogénéité. Le but est de monter qu'il n'existe plus d'hétérogénéité à l'intérieur de ces sous-groupes et que, par contre, les résultats des sous groupes sont hétérogènes. Il est alors possible de conclure à l'existence d'une interaction entre l'effet du traitement et le facteur identifié.

Conclusion

La méta-analyse est devenue un outil de routine. Dans Medline, on repère actuellement 600 à 800 méta-analyses par an. Comme tous les outils très performants, elle peut conduire à des aberrations si les conditions d'application ne sont pas respectées. Mais le succès acquis par cette technique provient surtout du fait que la méta-analyse répond à un besoin ressenti par de nombreux acteurs de santé, du médecin prescripteur au décideur de santé publique. La somme des connaissances sur lesquelles doivent se baser maintenant les décisions médicales, et en particulier les choix thérapeutiques, croît sans cesse. Les médecins ont de plus en plus besoin de données synthétiques qui intègrent efficacement l'ensemble des informations existantes pour assurer une base rationnelle à leur décision. En somme la méta-analyse devient l'étape finale indispensable de toute évaluation de l'efficacité clinique d'une thérapeutique (cf. chapitre suivant).

Bibliographie

1. *Cucherat M. Un seul résultat d'essai est-il suffisant? La méta-analyse des essais thérapeutiques. Rev Prat 2000;50:846-50.* PMID:

2. *Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. BMJ 1992;305:15-9.* PMID:

3. *Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408-412.* PMID:

4. *Moher D, Ba'Pham, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-613.* PMID:

5. *Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Controlled Clinical Trials 1990;11:339-.* PMID:

6. *Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991;337:867-872.* PMID:

7. *Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990;263:1385-1389.* PMID:

8. *Dickersin K, Min Y, Meinert CL. Factors influencing publication of research results: follow-up of application submitted to two institutional reviews boards. JAMA 1992;267:374-378.* PMID:

9. *McAuley L, Ba'Pham, Tugwell P, Moher D. Does inclusion of gray literature influence estimates of intervention effectiveness reported in meta-analysis? Lancet 2000;356:1228-1231.* PMID:

10. *Cucherat M, Boissel JP, Leizorovicz A. La méta-analyse des essais thérapeutiques. Paris: Masson; 1997.*

11. *Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Perusing the literature: Comparison of MEDLINE searching with a perinatal trials database. Controlled Clinical Trials 1985;6:306-317.* PMID:

12. *Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994;309:1286-1291.* PMID:

Apports de la méta-analyse

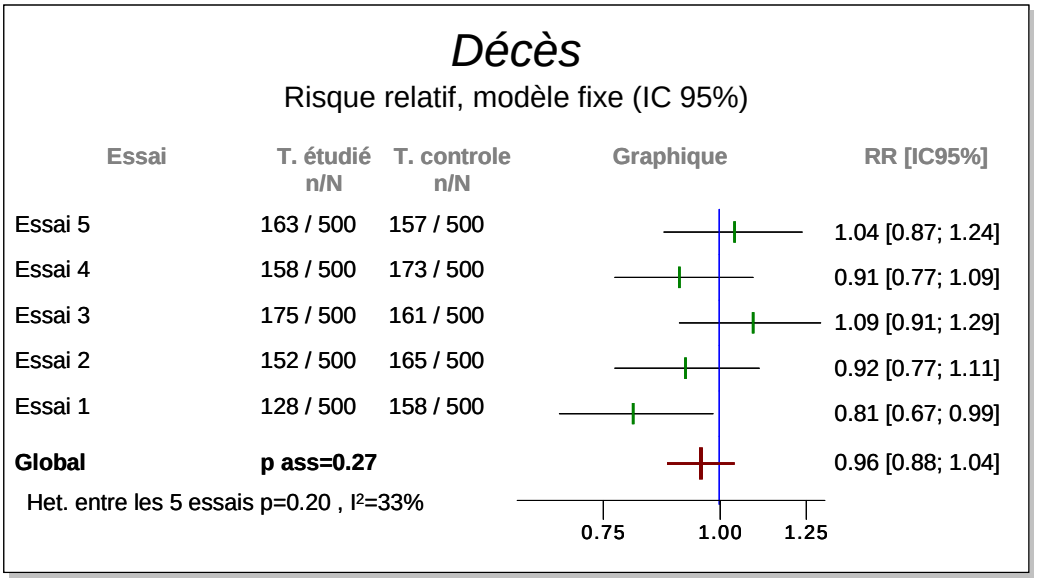
L'interprétation des essais d'un même domaine pris séparément conduit à des conclusions erronées que seule une prise en compte globale de l'ensemble des résultats disponibles par méta-analyse permet d'éviter.
Ces problèmes sont principalement de 3 types.

Mise en avant du résultat favorable

L'énoncé suivant semble satisfaisant et apporter la preuve de l'efficacité du traitement considéré :

> *Dans un essai contrôlé contre placebo, randomisé, en double aveugle, de grande taille (regroupant 1000 patients) le traitement a réduit la mortalité de 19% de manière statistiquement significative (risque relatif 0.81, p<0.05). Cet essai correctement conçu et réalisé démontre ainsi l'efficacité du traitement sur la mortalité.*

En fait la situation réelle est celle dépeinte par la figure suivante.



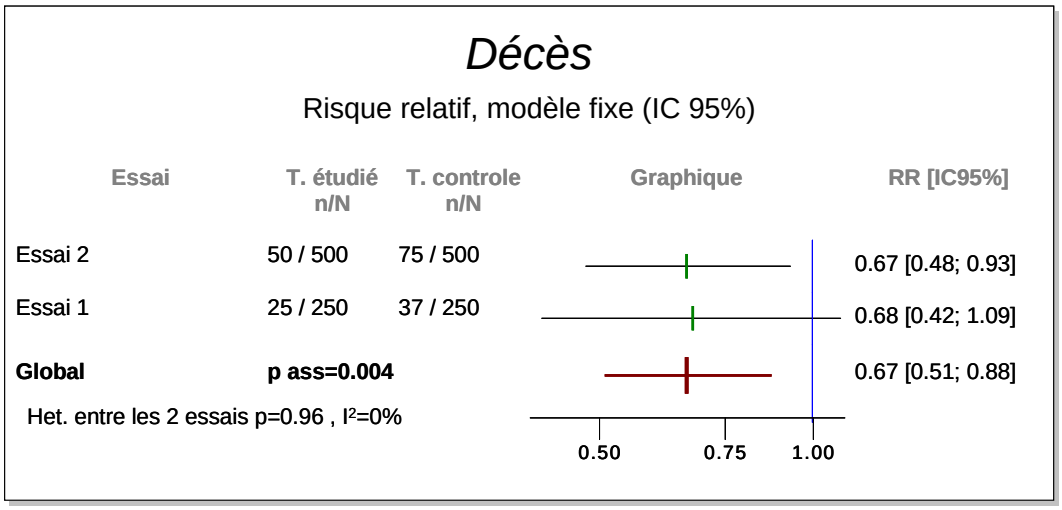
En réalité, 5 essais randomisés de bonne qualité ont été réalisés avec ce traitement. Ces essais sont tous non concluant à l'exception d'un essai l'essai n°1. C'est cet essai qui est mis en avant dans l'énoncé précédent, interprété indépendamment des autres résultats. Souvent d'ailleurs, il est possible de trouver des arguments post hoc pour disqualifier les résultats qui ne nous arrange pas, même si ces mêmes arguments auraient semblé futiles si le résultat de l'essai avait été satisfaisant.
La méta-analyse en faisant la liste exhaustive de tous les résultats à l'abri des biais disponibles évite cet écueil et permet de conclure qu'il n'y a pas de preuve de l'efficacité du traitement considéré (RR=0.96, p=0.27).
Il ya d'ailleurs une absence d'hétérogénéité montrant que le résultat de l'essai n°1 est une simple manifestation de l'erreur statistique alpha d'obtenir par hasard un résultat significatif. En répétant les essais (donc les tests) on répète les occasions de trouver par hasard un argument en faveur de l'effet du traitement. La méta-analyse corrige cette inflation du risque alpha.

Interprétation limitée à la signification statistique

Dans l'énoncé suivant, 2 essais discordants sont interprétés de manière séparée :

Le traitement a été étudié dans 2 essais contrôlé contre placebo, randomisé, en double aveugle. Le premier mené en Europe de l'Est est concluant tandis que le second réalisé aux USA ne l'est pas. C'est 2 etudes montrent donc bien que l'effet du traitement n'est pas le même aux USA et en Europe de l'Est car les contextes de soins sont différents, avec un sous traitement en Europe de l'Est. Les patients étant sous traités, le nouveau traitement peut apporter un bénéfice tandis que chez des patients recevant déjà des traitements efficaces ce nouveau traitement ne présente pas d'intérêt.

En fait la situation réelle est la suivante :



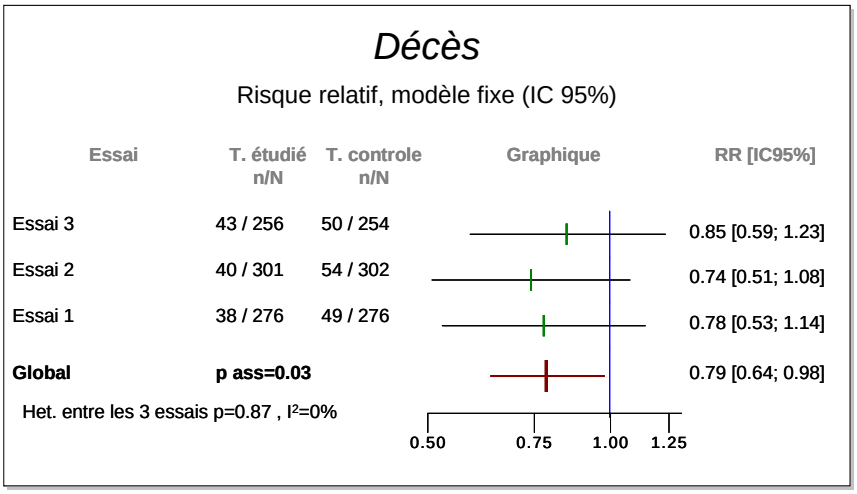
En fait, l'interprétation proposée est erronée car elle se base uniquement sur la dignification statistique des résultats. De ce fait les 2 résultats semblent complètement opposés. En fait lorsque l'on regarde les estimations des tailles d'effet (les risques relatifs), le traitement a eu la même efficacité dans ces 2 essais. Simplement l'essai n° 1 est moins puissant et a donc obtenu un résultat non statistiquement significatif. Cela est évident sur le graphique de la méta-analyse et est confirmé par l'absence complète d'hétérogénéité statistique (p het=0.96).

Absence de prise en compte de la puissance

Dans l'énoncé suivant 3 essais négatifs sont interprétés de manière séparée :

Trois essais contrôlés contre placebo, randomisés, en double aveugle ont été réalisés pour évaluer le même traitement. Ils ont tous non concluants et montrent donc que ce traitement n'est pas efficace.

En fait la situation réelle est la suivante :



Cette interprétation est erronée car elle ne prend pas en compte la puissance des essais. En fait chaque essai est insuffisamment puissant par lui-même pour mettre en évidence l'effet du traitement. La méta-analyse atteint un niveau de puissance permettant de mettre en évidence un effet statistiquement significatif. Les données disponibles ne permettent donc pas de considérer définitivement le traitement contre sans effet. Un nouvel essai de taille suffisante (déterminée à l'aide des résultats de la méta-analyse) est donc souhaitable pour démontrer l'efficacité.

Biais de publication

Le biais de publication a été formalisé à l'occasion du développement des techniques de méta-analyses, mais ce biais n'est pas spécifique de cette approche. Il touche en fait toute action récapitulative des résultats de la recherche (personnelle, revue de la littérature, méta-analyse). Par contre, contrairement aux autres méthodes, la méta-analyse donne des moyens de rechercher le biais de publication : funnel plot ,calcul de la robustesse, etc.

Définition

Toute synthèse d'information est sujette à un biais qui lui est propre : le biais de publication (« publication bias *»).*

Les résultats négatifs (essais ne montrant pas de différence significative) sont moins fréquemment publiés que les résultats positifs (essais montrant une différence significative). L'étude du devenir de 285 protocoles soumis au comité d'éthique d'Oxford révèle que 85% des résultats positifs ont été publiés contre seulement 56% des résultats négatifs [1]. Il existe ainsi une publication sélective des résultats positifs au détriment des résultats négatifs. Cela ne veut pas dire que ces derniers ne sont jamais publiés mais plus difficilement et seulement pour une partie d'entre eux.

Les causes de ce phénomène sont certainement nombreuses [2, 3] et impliquent à la fois les comités de lecture des revues, peu séduits par un résultat négatif, et les auteurs qui n'investissent pas dans la rédaction d'un article qui a peu de chance d'être accepté.

Ce phénomène de publication sélective va fausser les conclusions que l'on peut tirer à partir des résultats publiés. C'est le biais de publication. Dans une synthèse, si aucune recherche poussée des essais non publiés n'est entreprise, le risque couru est de ne travailler qu'avec les essais positifs, ce qui conduit à une surestimation de l'efficacité du traitement.

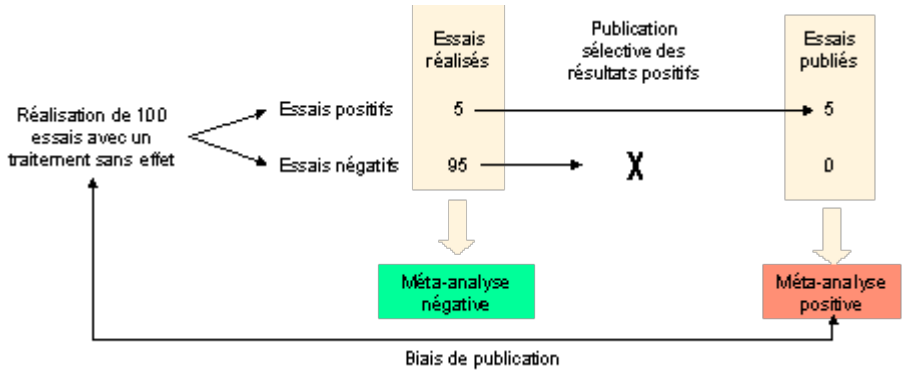


Figure 1 – Si de nombreux essais sont réalisés avec un traitement sans efficacité, certains d'entre eux auront cependant un résultat statistiquement significatif, uniquement du fait du risque d'erreur statistique alpha que l'on consent au niveau du test statistique. Ainsi, si 100 essais sont réalisés, 5 d'entre eux seront positifs à tort, du fait du hasard. Ainsi, si uniquement les essais positifs sont publiés, une synthèse ne portant que sur les résultats publiés donnera une fausse impression d'efficacité du traitement. C'est le biais de publication.

Les conséquences potentiellement dommageables du biais de publication sont illustrées par l'exemple des antiarythmiques de classe 1 en post infarctus avec la non publication en 1980 d'un essai qui montrait une forte augmentation de mortalité avec une molécule de cette classe, la lorcainide.

Une étude empirique (« empirical study ») a montré que l'exclusion, dans les méta-analyses des essais non publiés entraîne en moyenne une surestimation de 15% (IC95% 4% ;28%) de la taille de l'effet [4]. De même, l'exclusion des essais publiés uniquement sous forme « d'abstracts » entraîne en moyenne une surestimation de 33% (IC95% 10% ;60%) de l'effet.

La solution à ce problème serait la mise en place de registres prospectifs d'essais qui en enregistrant les essais à leur mise en place, permettraient, par la suite, de retrouver leur trace même s'ils n'ont jamais été publiés.

Exemple

Une étude récente publiée début 2008 dans le New England Journal of Medicine [5] démontre que le biais de publication (ou plus exactement la publication sélective des essais en fonction de leur résultat) est encore un problème d'actualité.

L'étude a consistée à récupérer auprès de la FDA les résultats de tous les essais de 12 antidépresseurs, soit 74 essais cliniques déclarés au total.

Sur ces 74 essais, 31% (représentant 3449 patients) non pas été publiés. Selon l'analyse FDA, 38 essais ont obtenus des résultats positifs. Ils ont tous été publiés à l'exception d'un seul. A l'exception de 3, les essais considérés par la FDA comme ayant obtenu des résultats négatifs ou questionnable n'ont pas été publiés (22 essais) ou ont été publiés d'uen façon suggérant un résultat plutôt positif selon l'évaluation des auteurs (11 essais).

En ne regardant que la littérature, 94% des essais semblent positifs alors que seulement 51% des essais selon l'analyse FDA ont donné des résultats positifs.

Les registre d'essais

La solution au biais de publication est maintenant apportée par l'impossibilité de publier, dans la majorité des revues, les essais qui n'auraient pas été enregistrés à leur démarrage dans un registre public. Ces registres prospectifs d'essais permettent alors de retrouver la trace de tous les essais entrepris qu'ils soient ou non publiés.

Cette initiative a été prise en septembre 2004 par International Committee of Medical Journal Editors (lien web). Cette association demanda aux journaux de plus publier à partir du 1er juillet 2005 les essais randomisés qui n'aurait pas été déclarés à un registre d'essai lors de leur mise en place. Cette décision permis de solutionner pour tous les nouveaux essais le problème du biais de publication. En effet, à partir de ce moment tout initiateur d'essai ne voulant pas prendre le risque de ne pas pouvoir publier correctement son travail est dans l'obligation d'enregistrer son étude.

Actuellement de nombreux registre d'essais ont été mis en place comme le registre du NIH (www.clinicaltrials.gov), le registre Current Controlled Trials. Un meta-registre est aussi disponible (metaRegister of Controlled Trials).

Dans ce contexte un numéro international standardisé a été proposé, l'ISRCTN. L'OMS a aussi crée son propre registre (the WHO International Clinical Trials Registry Platform). Une liste assez complète des registres existant à travers le monde est disponible à http://www.controlled-trials.com/links/ .

Funnel plot

Principe

Le graphique dit « funnel plot » (dont une traduction possible est « graphe en entonnoir ») consiste à représenter pour chaque étude la valeur estimée de l'effet traitement en fonction de la taille de son échantillon. En l'absence de biais de publication, les différentes estimations de l'effet du traitement vont se répartir autour de la valeur commune. Les estimations dont l'écart type est important car obtenus dans les études de plus faibles effectifs varieront autour de cette valeur centrale avec une plus grande amplitude que celles dont l'écart

type est petit (c'est-à-dire basé sur des plus grands effectifs). Les points se répartissent de façon symétrique de part et d'autre de la valeur centrale et donnent un nuage de points évasé.

Les points situés en bordure de ce nuage correspondent aux résultats statistiquement significatifs. Avec un traitement sans efficacité, seul 5% des points sont dans ce cas (Figure Erreur ***! Signet non défini.****). En cas de biais de publication, la répartition n'est plus homogène. Un déséquilibre apparaît avec disparition des points situés dans la zone correspondant aux résultats non significatifs et le graphique devient creux, d'où son nom de graphique en entonnoir (Figure Erreur* ***! Signet non défini.****). Une autre possibilité de biais de publication est représentée par la non publication des essais allant à l'encontre de l'hypothèse testée (significatifs ou non significatifs). Cette possibilité s'avère même plus fréquente en pratique que la précédente. En effet, les résultats favorables à l'efficacité du traitement sont publiés qu'ils soient significatifs ou non. Par contre, ceux suggérant un effet délétère restent non publiés. Dans ce cas, représenté sur la Figure Erreur* ***! Signet non défini.****, le nuage de point devient asymétrique.*

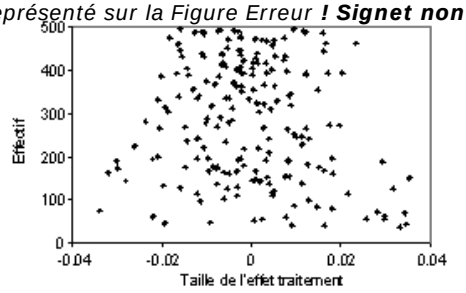


Figure ***Erreur ! Signet non défini.*** *– Funnel plot dans une situation où il n'y a pas de biais de publication*

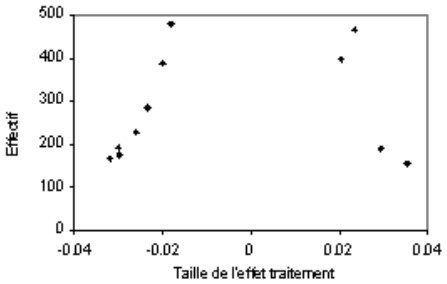


Figure ***Erreur ! Signet non défini.*** *– Funnel plot représentant une situation caricaturale de biais de publication. La totalité des essais non significatifs n'est pas disponible, ce qui donne un aspect creux au graphique.*

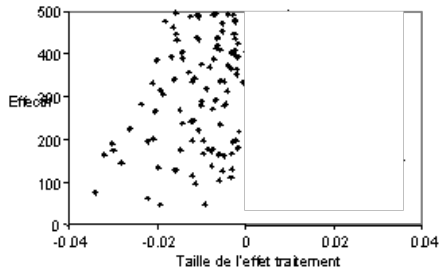


Figure ***Erreur ! Signet non défini.*** *– Traduction sur le funnel plot d'un biais de publication où ce sont les essais suggérant un effet délétère (significatif ou non) sont non publiés.*

L'analyse graphique du « funnel plot » donne ainsi un moyen de vérifier s'il y a lieu de suspecter un biais de publication dans une méta-analyse.

Exemples de funnel plot

La figure 4 donne un exemple d'un funnel plot réel très évocateur d'un biais de publication (dans cette méta-analyse, les effets bénéfiques se traduisent par un effect size positif).

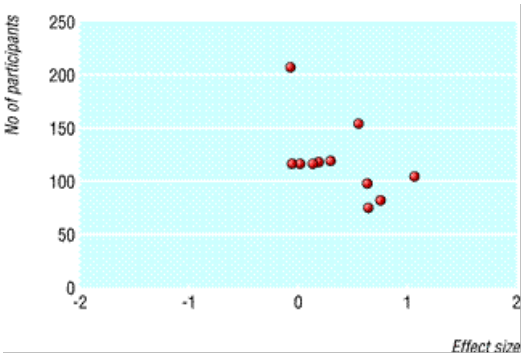


Figure 4 – Funnel plot of randomised controlled trials comparing topical non-steroidal anti-inflammatory drugs with placebo (asymmetry P=0.04). Jinying Lin,Weiya Zhang, Adrian Jones, Michael Doherty. Efficacy of topical non-steroidal anti-inflammatory drugs in the treatment of osteoarthritis: meta-analysis of randomised controlled trials. BMJ, doi:10.1136/bmj.38159.639028.7C (published 30 July 2004)

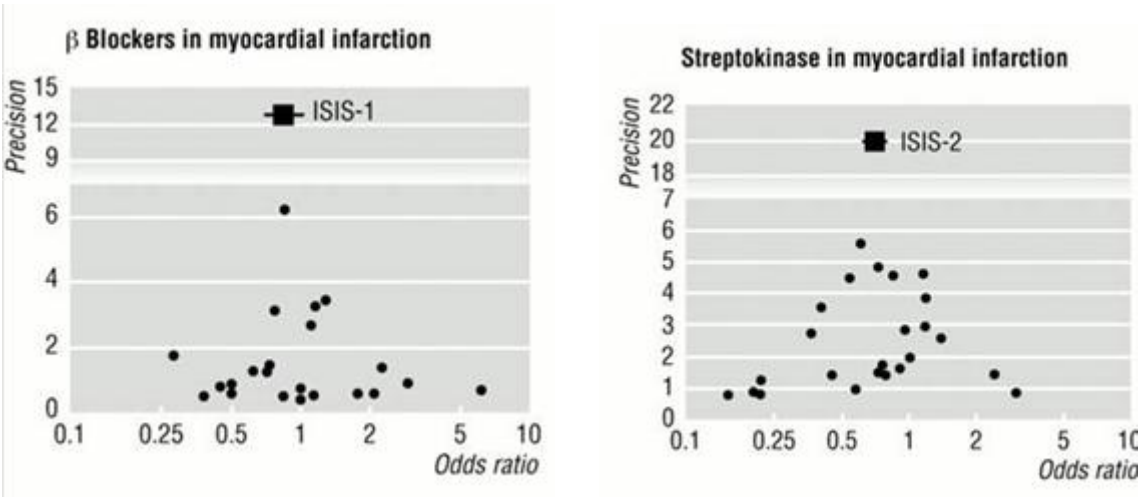


Figure 5 – Exemple de 2 funnel plots permettant d'exclure un biais de publication. En haut du graphique figure le « mega-trial » du domaine.

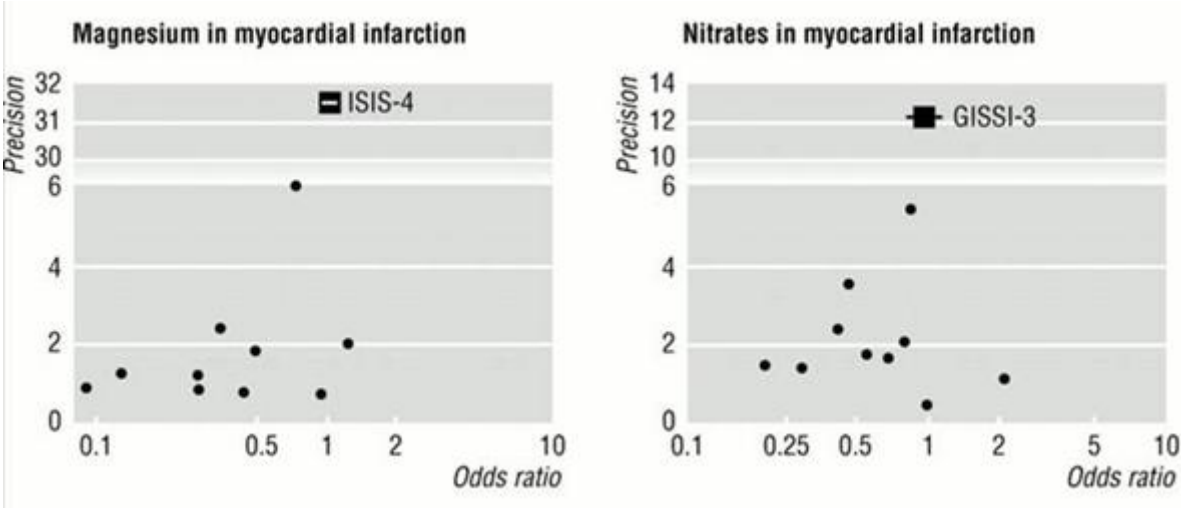


Figure 6 – Exemple de 2 funnel plots évocateur de part leur asymétrie d'un biais de publication. En haut du graphique est représenté le grand essai du domaine montrant l'absence d'efficacité.

Bibliographie

1. *Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991;337:867-872.* PMID:

2. *Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA 1990;263:1385-1389.* PMID:

3. *Dickersin K, Min Y, Meinert CL. Factors influencing publication of research results: follow-up of application submitted to two institutional reviews boards. JAMA 1992;267:374-378.* PMID:

4. *McAuley L, Ba'Pham, Tugwell P, Moher D. Does inclusion of gray literature influence estimates of intervention effectiveness reported in meta-analysis? Lancet 2000;356:1228-1231.* PMID:

5. *Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. N Engl J Med 2008;358(3):252-260.* PMID:

Hétérogénéité

Définition

L'hétérogénéité se définit par le rejet de l'hypothèse d'homogénéité des effets traitement. Le test de cette hypothèse est appelé test d'hétérogénéité ou test d'homogénéité, ce qui induit quelques confusions. Quel que soit son nom, l'hypothèse testée est la même et l'obtention d'un test significatif témoigne d'une hétérogénéité : l'effet d'un essai au moins ne peut pas être considéré comme étant identique à celui des autres essais. L'hypothèse du modèle fixe ne tient pas et la combinaison de tous ces essais devient litigieuse. En effet, la méta-analyse cherche à estimer une valeur qui, par hypothèse, est considérée comme commune à tous les essais.

Le test d'hétérogénéité possède une faible puissance avec le nombre d'essais habituellement rencontré dans les méta-analyses (en général inférieur à une trentaine). A contrario, avec un grand nombre d'essais, une faible hétérogénéité sans pertinence clinique est détectable.

Une hétérogénéité peut être le témoin d'une interaction entre une covariable et l'effet du traitement. Elle peut aussi provenir d'une forte variabilité aléatoire de l'effet sans qu'il soit possible de rattacher ces fluctuations à un ou des facteurs bien précis. L'effet est alors inconstant d'un essai à l'autre et pose la question du bien-fondé du regroupement de ces essais.

Statut de l'hétérogénéité

L'hétérogénéité peut être considérée comme une nuisance que l'on cherchera à éliminer en prenant une méthode adaptée (méthode à effet aléatoire, méthode de Peto). Ces techniques prennent en compte l'hétérogénéité sans chercher à l'expliquer.

A l'opposé, l'hétérogénéité peut être considérée comme informative, témoignant de changement dans l'effet du traitement en fonction des circonstances de sa mesure (profil des patients ou utilisation du traitement).

Statut de l'hétérogénéité	***Approche***
Nuisance	*Prise en compte de l'hétérogénéité (sans l'expliquer) avec un modèle aléatoire*
Information	*Explication de l'hétérogénéité en fonction de covariables*

Que faire devant une hétérogénéité

Recherche des essais induisant l'hétérogénéité

La première chose à faire devant une hétérogénéité est de rechercher le ou les essais qui l'induisent, en s'aidant d'un graphique. Pour confirmer les indications apportées par l'analyse graphique, l'hétérogénéité est recalculée, après suppression des essais suspects, pour s'assurer de sa disparition effective. Après avoir identifié les essais qui induisent l'hétérogénéité, il convient de chercher s'ils diffèrent des autres par l'une de leurs caractéristiques (populations, intervention, qualité méthodologique). Ensuite, si un facteur d'hétérogénéité est suspecté, non seulement le ou les essais induisant l'hétérogénéité devront être exclus de la méta-analyse, mais aussi tous les essais dans lesquels ce facteur est présent .

Recherche d'interaction

Une approche complémentaire consiste à rechercher systématiquement s'il existe une interaction avec une ou plusieurs covariables. Cette recherche d'interaction a pour but de montrer que la taille de l'effet varie en fonction des valeurs prises par une ou plusieurs covariables. Cette démonstration peut être obtenue par plusieurs moyens:

- *les analyses en sous-groupes,*
- *la modélisation de l'effet sur les données résumées par des techniques uni- ou multi-variées,*
- *l'utilisation de modèles uni- ou multi-variés sur les données individuelles (ces techniques sont abordées dans le chapitre consacré aux m éta-analyses sur données individuelles, chapitre 28).*

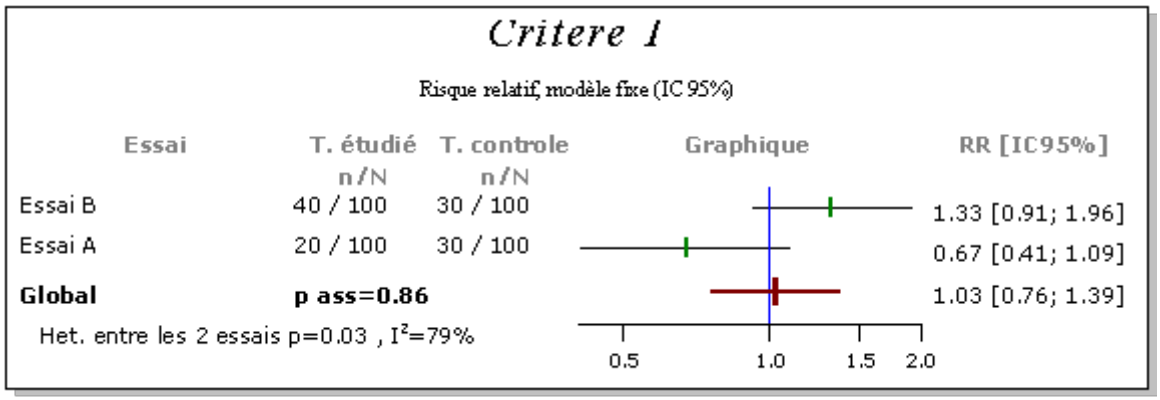
L'existence d'une hétérogénéité n'est pas une condition nécessaire à la recherche des interactions. Celle-ci peut être effectuée de manière systématique, lorsqu'elle correspond, par exemple, à l'un des objectifs de la méta-analyse. Dans ce cas les modalités de la recherche des interactions et les facteurs étudiés sont prévus d'emblée dans le protocole.

Modèle aléatoire

Si l'hétérogénéité observée ne s'explique pas par une interaction, il est possible de recourir à un modèle aléatoire. Ce modèle prend en compte une certaine variabilité aléatoire de l'effet traitement d'un essai à l'autre.

L'estimation obtenue tiendra compte de cette variabilité, l'intervalle de confiance de l'effet combiné sera plus large qu' avec le modèle fixe, et la variance du vrai effet traitement sera estimée (τ2). Les intervalles de confiance sont plus larges car, en plus des fluctuations aléatoires, ils prennent en compte la variabilité du vrai effet traitement.

L'utilisation d'un modèle aléatoire s'accompagne d'un risque de méconnaître l'existence d'une interaction et d'arriver à une conclusion réductrice qui perd une partie de l'information apportée par les essais.



Analyses en sous-groupes

Les analyses en sous-groupes effectuent une recherche de l'interaction de façon univariée, en comparant les résultats obtenus entre deux ou un petit nombre de sous-groupes d'essais. Une interaction est détectée si les résultats des sous-groupes se révèlent statistiquement différents les uns des autres par l'application du test d'hétérogénéité entre les strates. Les sous-groupes sont créés en fonction de la covariable étudiée.

Hétérogénéité entre les sous-groupes

Une interaction entre l'effet du traitement et le facteur définissant les sous-groupes peut être recherchée par une analyse de l'hétérogénéité entre les sous-groupes. Cette analyse consiste en la réalisation de méta-analyses dans chacun des sous-groupes. L'hétérogénéité au sein de chaque sous-groupe est recherchée ainsi que l'hétérogénéité entre les résultats des sous-groupes. Une interaction se manifestera par des essais homogènes au sein de chaque sous-groupe, mais conduisant à une hétérogénéité entre les sous-groupes.

L'effet du traitement est alors significativement différent d'un sous-groupe à l'autre (voir par exemple le cas du délai depuis les symptômes dans la figure 24.3)

Le test d'hétérogénéité des résultats dans chaque sous-groupe utilise la classique statistique Q:

Constitution des sous-groupes

Les sous-groupes pourront être définis de diverses façons :

1. *par le type de traitement. Par exemple, dans la méta-analyse des hypocholestérolémiants il est possible de regrouper les essais en fonction des classes pharmacologiques (fibrates, inhibiteur de l'HMGCoA réductase, résines); en fonction de la molécule (simvastatine, lovastatine, pravastatine); en fonction de la nature de l'intervention (régime, médicaments, chirurgie). Des sous-groupes en fonction de la dose sont aussi envisageables ou en fonction du moment du traitement (par exemple avec les traitements fibrinolytiques dans l'infarctus du myocarde en fonction du délai d'administration par rapport au début des symptômes),*
2. *par le type de mesure du critère de jugement, lorsque plusieurs moyens existent pour mesurer le même critère de jugement. Par exemple, dans les essais de prévention du risque thrombo-embolique par les héparines, différents moyens diagnostiques sont utilisables pour rechercher les phlébites (clinique uniquement, phlébographie, Doppler, fibrinogène marqué),*
3. *par des caractéristiques propres aux patients. Par exemple, en fonction de la tranche d'âge, du sexe, en fonction du pronostic ou du risque de base,*
4. *4. par les conditions de réalisation des essais : essais hospitaliers versus ambulatoires, en fonction de la région ou du pays de réalisation, etc...*
5. *par le type d'essais : essais en double aveugle, en simple aveugle.*

Dans le premier cas, l'opération revient à comparer différents traitements par des comparaisons indirectes. Par contre, dans les autres cas, le but est de rechercher une modification de la taille de l'effet en fonction de facteurs divers, correspondant soit à des caractéristiques des patients, soit à des caractéristiques des essais. Il s'agit alors d'une véritable recherche d'interaction.

La définition des sous-groupes est fixée a priori dans le protocole, sauf dans le cas où le problème se pose après mise en évidence d'une hétérogénéité. Suivant les cas, les sous-groupes peuvent être constitués de deux façons :

1. *Les essais sont répartis entre les différents sous-groupes car au sein d'un essai tous les patients sont identiques vis à vis de la caractéristique définissant les sous-groupes.*
2. *Chaque essai regroupe des patients qui correspondent aux différents sous-groupes. Ces patients doivent donc être répartis entre les différents sous-groupes. Cette situation exige que les données correspondant à chaque type de patients soient rapportées séparément dans le compte rendu de l'essai, c'est à dire que la même analyse en sous-groupe ait été réalisée dans l'essai. Si cette condition n'est pas remplie, seules les données individuelles permettront de recréer les sous-groupes.*

Risque des analyses en sous-groupes

En méta-analyse, les analyses en sous-groupes (aussi appelées stratifiées) font courir, comme dans un essai clinique, le risque de l'inflation non contrôlée de l'erreur de première espèce. La multiplication des tests statistiques (un par sous-groupe) augmente la probabilité d'obtenir un test significatif uniquement par hasard. Un résultat de sous-groupe significatif devient suspect car il est impossible de savoir si ce test révèle une interaction réelle ou s'il s'agit simplement d'un artefact lié à la répétition des tests. Par exemple, dans l'essai ISIS-2, l'aspirine administrée à la phase aiguë de l'infarctus du myocarde produit une réduction significative très importante de la mortalité à 1 mois. Mais, lors de l'analyse en sous-groupe en fonction des signes astrologiques, l'aspirine apparaît inefficace pour les sujets du signe de la vierge ou des gémeaux et plus efficace que la moyenne pour le signe du capricorne. Dans les paradigmes scientifiques actuels, aucune théorie ne permet de penser que ces différences sont réelles!

En effet, il est toujours possible d'obtenir un résultat significatif en multipliant les sous-groupes. Un résultat significatif obtenu dans ces conditions n'a aucune valeur. La réalisation d'analyse en sous-groupes à partir de données issues de plusieurs essais dans une méta-analyse ne résoud pas ce problème (qui est uniquement lié à la répétition des tests).

Génération des hypothèses des sous-groupes

Pour minimiser le risque de résultats significatifs par hasard dans les analyses en sous-groupes, il convient de définir a priori un petit nombre de sous-groupes. Ces analyses s'apparenteront alors à la démarche hypothético-déductive.

Les sous-groupes définis a priori sont retenus de deux manières différentes.

1. *certaines interactions peuvent être recherchées systématiquement avec, par exemple, l' âge, le sexe, la dose, la durée. Ces interactions sont suspectées de façon systématique ou à partir d'un modèle physiopathologique, pharmacologique ou thérapeutique.*
2. *Les interactions peuvent être suspectées à partir de résultats obtenus dans un des essais du domaine. L'hypothèse est générée par les données et il convient d'éviter tous risques de tautologie dans sa confirmation.*

Dans cette dernière situation, la méta-analyse permet de confronter cette hypothèse à d'autres données externes à son processus de génération. Ainsi, il est possible de vérifier si un résultat initialement observé dans un essai se retrouve dans d'autres essais. Dans l'affirmative, les chances que ce résultat soit uniquement le fait du hasard s'amenuisent. A ce niveau, il est possible de discuter l'attitude qui consiste à exclure ou à maintenir l'essai (ou les essais) à l'origine de l'hypothèse dans la méta-analyse. En effet, si le poids de celui-ci est prépondérant dans la méta-analyse (car il comporte à lui tout seul autant de sujets que les autres réunis) le résultat de la méta-analyse ne dépendra presque exclusivement que du résultat de cet essai, conduisant a une confirmation tautologique de l'hypothèse. Une analyse de sensibilité confrontant les résultats obtenus avec et sans les essais à l'origine de l'hypothèse semble être la meilleure solution pour éprouver ce problème potentiel.

A contrario, la pratique de la méta-analyse a fourni des exemples du caractère fallacieux des analyses en sous-groupes dans les essais. Des relations statistiques apparemment fortes, quoique fortuites, entre les variables de base des patients et le résultat observé dans un essai ont été infirmées par la méta-analyse. C'est le cas de la relation entre la topographie de l'infarctus et l'effet des béta-bloquants dans le post-infarctus.

Apport de la méta-analyse par rapport aux analyses en sous-groupes dans les essais

La division de l'échantillon de patients en deux ou plusieurs sous-groupes entraîne une baisse de puissance des comparaisons réalisées dans chaque groupe et du test interaction. La méta-analyse apporte plus de patients et augmente la puissance par rapport à un seul essai.

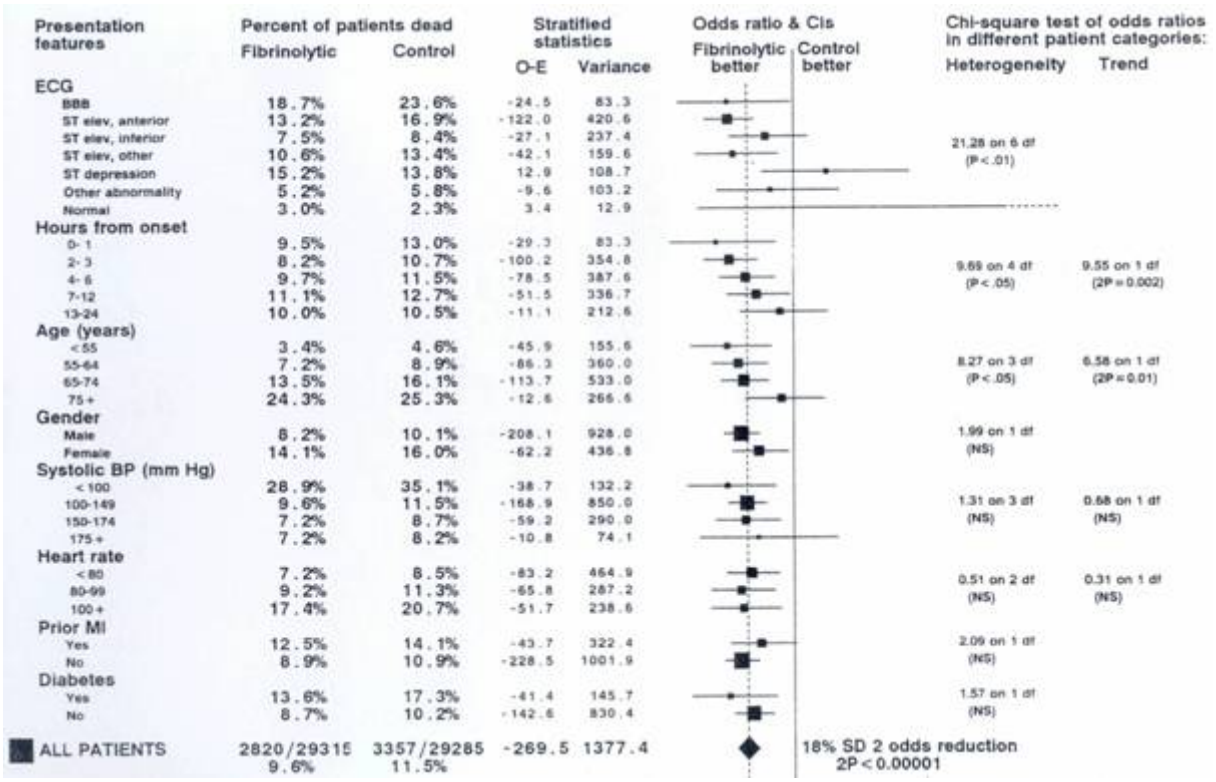
Dans un essai isolé, le nombre de sujets inclus a été calculé pour assurer une puissance suffisante à la comparaison principale. Par rapport à cette comparaison, une recherche d'interaction nécessite plus de patients. Un seul essai est donc en général insuffisant pour garantir une puissance suffisante à une analyse en sous-groupes. La méta-analyse permettra de renforcer les effectifs et d'augmenter la puissance des comparaisons en sous-groupes.

Cependant, même dans le meilleur des cas, la mise en évidence d'une interaction reste du domaine de l'association statistique et ne permet pas de conclure à la relation causale entre le facteur étudié et la variation de la taille de l'effet. En effet, les différentes modalités du facteur étudié ne peuvent pas être contrôlées, ce qui empêche de se prémunir contre l'existence de facteur de confusion.

La méta-analyse classique ne permet que des analyses univariées en sous-groupes. L'ajustement de la recherche d'une interaction, sur d'éventuels facteurs de confusion est donc impossible sans le recours à des techniques multivariées.

Exemple La méta-analyse du « Fibrinolytic Therapy Trialists' Collaborative Group » regroupe 9 essais de grande taille (> 1000 patients) de la fibrinolyse à la phase aiguë du myocarde. Plusieurs sous-groupes ont été étudiés, où une différence d'efficacité de la fibrinolyse était recherchée en fonction de caractéristiques des patients comme les signes ECG, le temps écoulé entre le début des symptômes et la fibrinolyse, l'âge, le sexe, l'existence d'un diabète ou d'antécédents d'infarctus du myocarde. La figure 24.3 représente les résultats de ces différents sous-groupes.

Fig. 24.3. — Exemple d'analyse en sous groupes


Presentation features
Percent of patients dead
Fibrinolytic
Control
Stratified statistics
O-E
Variance
Odds ratio & CIs
Fibrinolytic better
Control better
Chi-square test of odds ratios in different patient categories:
Heterogeneity
Trend
ECG
BBB 18.7% 23.6% -24.5 83.3
ST elev, anterior 13.2% 16.9% -122.0 420.6
ST elev, inferior 7.5% 8.4% -27.1 237.4
ST elev, other 10.6% 13.4% -42.1 159.6
ST depression 15.2% 13.8% 12.9 108.7
Other abnormality 5.2% 5.8% -9.6 103.2
Normal 3.0% 2.3% 3.4 12.9
21.28 on 6 df (P<.01)
Hours from onset
0-1 9.5% 13.0% -29.3 83.3
2-3 8.2% 10.7% -100.2 354.8
4-6 9.7% 11.5% -78.5 387.6
7-12 11.1% 12.7% -51.5 336.7
13-24 10.0% 10.5% -11.1 212.6
9.69 on 4 df (P<.05)
9.55 on 1 df (2P=0.002)
Age (years)
<55 3.4% 4.6% -45.9 155.6
55-64 7.2% 8.9% -86.3 360.0
65-74 13.5% 16.1% -113.7 533.0
75+ 24.3% 25.3% -12.6 266.6
8.27 on 3 df (P<.05)
6.58 on 1 df (2P=0.01)
Gender
Male 8.2% 10.1% -208.1 928.0
Female 14.1% 16.0% -62.2 436.8
1.99 on 1 df (NS)
Systolic BP (mm Hg)
<100 28.9% 35.1% -38.7 132.2
100-149 9.6% 11.5% -168.9 850.0
150-174 7.2% 8.7% -59.2 290.0
175+ 7.2% 8.2% -10.8 74.1
1.31 on 3 df (NS)
0.68 on 1 df (NS)
Heart rate
<80 7.2% 8.5% -83.2 464.9
80-99 9.2% 11.3% -65.8 287.2
100+ 17.4% 20.7% -51.7 238.6
0.51 on 2 df (NS)
0.31 on 1 df (NS)
Prior MI
Yes 12.5% 14.1% -43.7 322.4
No 8.9% 10.9% -228.5 1001.9
2.09 on 1 df (NS)
Diabetes
Yes 13.6% 17.3% -41.4 145.7
No 8.7% 10.2% -142.6 830.4
1.57 on 1 df (NS)
ALL PATIENTS 2820/29315 9.6% 3357/29285 11.5% -269.5 1377.4
18% SD 2 odds reduction
2P<0.00001

Lecture critique d'une méta-analyse

Guide de lecture

Les points à analyser pour déterminer la validité interne (validité méthodologique assurant l'exactitude des résultats) sont les suivants. Ces points seront expliqués et détaillés dans la section 2.

La totalité des essais du domaine a telle été considérée ?

- La recherche des essais publiés a telle été exhaustive ? : recherche dans plusieurs bases bibliographiques informatisées, utilisation des références des articles
- Les essais non-publiés ont-ils été recherchés ? : recherche des abstracts de congrès, prise de contacts avec les promoteurs potentiels et les leaders d'opinion du domaine, etc. ?
- Les essais que je connais et ceux couramment cités sont-ils présents dans la méta-analyse ?
- La date de fin de recherche des essais est-elle récente ? est-il possible que de nouveaux essais n'aient pas été pris en considération ?

Les essais ont-ils été retenus pour la méta-analyse indépendamment de leur résultat ?

- Les raisons d'exclusions des essais sont-elles précisées ? Ces raisons sont-elles justifiées et licites ? (rechercher des critères d'exclusion abusifs cachant une sélection d'après le résultat)
- Les auteurs ont-ils des conflits d'intérêts avec les promoteurs des traitements étudiés

Les essais inclus dans la méta-analyse sont-ils fiables ?

- Les critères de sélection méthodologiques des essais sont-ils pertinents ? et sont-ils à même de ne retenir que des essais potentiellement non biaisés ?
- D'après la description des caractéristiques méthodologiques des essais inclus, ceux-ci semblent-ils suffisamment à l'abri des biais ?

Le résultat agrégatif donné par la méta-analyse a t-il un sens clinique ?

- Les critères de sélection thématiques déterminent des essais qui répondent tous à la même question thérapeutique ? Cette question thérapeutique correspond à des problèmes médicaux rencontrés en pratique ?
- Peut-on conclure à l'absence d'hétérogénéité cliniquement pertinente (au sein des regroupements qui sont proposés au final) ?
- En cas de diversité importante des essais, ses conséquences ont-elles été explorées par des analyses en sous groupes ?
- En cas d'hétérogénéité détectée, celle-ci a t-elle fait l'objet d'une recherche d'explication avant le recours a un modèle aléatoire ?

Le regroupement des résultats a t-il été correctement réalisé ?

- Sont ce, les estimations de l'effet traitement qui ont été regroupés (et non pas les événements et les effectifs) ? c'est à dire le résultat est-il a l'abri du paradoxe de Simpson ?
- Le p du test d'association est-il donné ?
- L'hétérogénéité statistique a telle été recherchée ? (le p d'hétérogénéité est-il disponible ?)
- En cas d'analyse en sous groupe, l'hétérogénéité entre sous groupe est-elle rapportée ? (pour permettre une recherche d'interaction)

Le résultat de la méta-analyse est-il à l'abri d'un biais de publication ?

- Les essais non publiés ont-ils été recherchés de façon adaptée (cf. supra) ?
- La possibilité d'un biais de publication a telle été recherchée à l'aide d'une technique appropriée (Funnel plot) ?

Qu'elle est la pertinence clinique du résultat de la méta-analyse ?

- L'objectif de la méta-analyse est-il cliniquement pertinent ? (problème thérapeutique réel, comparateur validé et pertinent)
- Le critère de jugement principal (ou celui mis en avant) est-il cliniquement pertinent ? (correspond-il directement à un objectif thérapeutique) Existe t-il des critères cliniquement pertinents non envisagés par cette méta-analyse ?
- Les traitements étudiés dans les essais ont-ils étaient utilisés de manière optimale ?
- Les patients étudiés dans les essais correspondent-ils aux patients vus en pratique ?
- La taille de l'effet mis en évidence est-elle suffisante pour être intéressante en pratique ? (attention en raison du gain de puissance que procure le regroupement de plusieurs essais, la méta-analyse peut détecter de manière statistiquement significative des effets de très faible ampleur, insuffisamment importants pour être cliniquement pertinent)

- Le bénéfice absolu du traitement est-il présenté (réduction absolue du risque ou NNT) ?
- La méta-analyse permet t-elle de documenter la balance bénéfice risque ?

Le résultat de la méta-analyse est-il suffisamment probant pour entraîner un changement de pratique ?

- Existe t-il au moins un essai concluant ? ou la seule preuve de l'efficacité est-elle apportée par la méta-analyse ?

L'ESSAI DE NON INFERIORITE

Principe général

Introduction

Les essais de non infériorité (« non-inferiority trial »), parfois appelé par abus de langage essais d'équivalence (« equivalence trial »), deviennent de plus en plus fréquents dans l'évaluation clinique des nouveaux traitements. Ce type d'essais fait appel à une méthodologie et à des techniques statistiques dont le développement est relativement récent [1] et relativement peu connu. De ce fait, des nouveaux traitements peuvent être acceptés sur la base d'essais d'équivalence discutables par méconnaissance des pièges et des spécificités de ce type d'études [2, 3]. En particulier, le processus décisionnel qui leur est attaché nécessite l'introduction d'un seuil d'équivalence choisi arbitrairement. De la valeur de ce seuil dépend grandement le résultat de l'essai.

Malgré son nom, l'essai d'équivalence ne permet pas de conclure que le traitement étudié a une efficacité équivalente à celle du traitement de référence.

Les conclusions de ces essais sont aussi très souvent surinterprétées. Malgré les apparences, l'essai dit « d'équivalence » ne permet pas de conclure que le traitement étudié a une efficacité identique à celle du traitement de référence mais simplement qu'il a une efficacité suffisante. Comme nous le verrons par la suite, les méthodes disponibles permettent seulement de raisonnablement éliminer la possibilité que le traitement étudié soit nettement moins efficace que le traitement de référence. Ces techniques permettent d'exclure que le nouveau traitement entraine une perte d'efficacité supérieure à une certaine limite, fixée a priori et qui est devrait être la plus grande perte d'efficacité cliniquement négligeable.

Ainsi, a l'issu d'un essai de non infériorité concluant, rien ne permet d'exclure que le nouveau traitement soit en réalité moins efficace que le traitement de référence. La seule chose qui soit acquise (avec un risque alpha d'erreur de 5%) est que cette perte d'efficacité est inférieure à la limite que les investigateurs sont prêt à perdre compte tenu des avantages qu'offre le nouveau traitement par ailleurs.

> Note : Les essais d'équivalence clinique ont pour objectif de montrer que deux traitements sont « équivalents » en termes d'efficacité clinique. Ils sont à distinguer des essais de bioéquivalence où l'équivalence ne concerne que des paramètres pharmacocinétiques.

Justification de la recherche de la non infériorité

Dans une pratique fondée sur les preuves, un nouveau traitement n'est adopté que lorsqu'il existe une preuve issue d'essais cliniques qu'il représente un progrès par rapport au traitement de référence (ou par rapport à l'absence de traitement). En général, le progrès thérapeutique est représenté par une efficacité supérieure à celle du traitement de référence. La preuve est apportée par un essai visant à montrer la supériorité du nouveau traitement (essai de supériorité).

Cependant, dans certaines situations, une avancée thérapeutique peut non pas être une efficacité supérieure mais simplement une plus grande facilitée d'utilisation, une meilleure tolérance ou un plus faible coût. Ces avantages pourront être suffisamment intéressants pour justifier l'adoption du nouveau traitement même si son efficacité n'est pas supérieure à celle du traitement de référence, voire est légèrement inférieure. La communauté médicale est prête à accepter de perdre un peu d'efficacité étant donnés les autres avantages. La démonstration de l'intérêt du nouveau traitement sera apportée par un essai cherchant à mettre en évidence la non-infériorité de celui-ci par rapport au traitement de référence.

Par exemple, l'intérêt des héparines de bas poids moléculaires dans le traitement des thromboses veineuses profondes est une plus grande facilité d'utilisation. Il en est de même des changements de modalités d'administration d'un même produit à la recherche d'une plus grande faisabilité. Cette approche a été envisagée pour la fibrinolyse avec l'alteplase à la phase aiguë de l'infarctus (essai COBALT).

Exemple des HBPM dans le traitement des TVP

Jusqu'à récemment, le traitement standard des thromboses veineuses profondes consistait en l'hospitalisation et l'administration intraveineuse continue d'héparine non fractionnée (HNF) durant 5 à 10 jours, suivi par un traitement anticoagulant oral d'au moins 3 mois. L'utilisation de l'héparine non fractionnée nécessite un monitorage biologique pour l'ajustement des doses.

Les héparines de bas poids moléculaires (HBPM) présentent de nombreux avantages par rapport à l'héparine non fractionnée : leur demi-vie plus longue rend possible leur administration en 2, voire 1, prise par jour, l'adaptation de la dose par monitorage biologique n'est pas nécessaire (une adaptation au poids du patient est suffisante). Ces deux points rendent envisageable le traitement de ces patients à domicile.

L'ensemble de ces avantages fait que les HBPM représentent une alternative à l'HNF intéressante en pratique même sans surcroît d'efficacité [4]. En effet, il serait tout à fait justifié d'adopter les HBPM même si elles ne s'avèrent qu'équivalentes en efficacité par rapport au traitement standard. Dans cet exemple ; le progrès thérapeutique réside dans l'amélioration de la praticabilité du traitement, et du confort du patient.

L'évaluation des HBPM dans le traitement de cette maladie s'est donc basée sur des essais de non-infériorité [4].

La désescalade thérapeutique, par exemple en cancérologie, avec un allégement des protocoles de chimiothérapie ou le recours à des traitements chirurgicaux moins délabrants représente aussi une situation où il est facilement justifiable de changer pour de nouveaux protocoles thérapeutiques d'efficacité seulement équivalente aux précédents mais qui préservent mieux la qualité de vie des patients.

> ⚠ *L'approche de non-infériorité ne produit des arguments permettant d'utiliser le nouveau traitement que si celui-ci présente, sur certains points, une supériorité manifeste par rapport au traitement habituel. En leur absence, la perte d'efficacité consentie empêche de conclure que le nouveau traitement représente un progrès thérapeutique par rapport au précédent. Cette remarque est importante car la tentation est grande d'évaluer en équivalence un nouveau traitement non innovant et de vanter ensuite son « efficacité équivalence » pour le faire utiliser en remplacement du précédent. Cette attitude est dangereuse car il n'est pas possible d'exclure qu'elle conduise à remplacer un traitement par un autre, en réalité moins efficace, sans que cette substitution n'apporte un quelconque avantage.*

Ces exemples peuvent se généraliser de la façon suivante : le bénéfice d'un traitement est une notion multifactorielle dans laquelle intervient à la fois l'efficacité vis à vis des critères de jugement clinique mais aussi la tolérance, la faisabilité et le coût. La démonstration de « l'équivalence clinique » d'un nouveau traitement par rapport au traitement de référence est suffisante pour l'adoption de celui-ci chaque fois où le gain obtenu sur les autres dimensions du bénéfice représente un intérêt suffisant pour admettre une efficacité équivalente (c'est-à-dire potentiellement légèrement inférieure). Le Tableau 1 présente quelques situations de ce type.

Cependant cette prise de décision va s'appuyer sur des choix arbitraires qui consistent à décider si les avantages sont « suffisants ». En d'autres termes, quelle diminution de coût, quelle réduction de fréquence des effets indésirables représentent un avantage suffisamment important pour justifier un changement. Toute la difficulté consiste à déterminer la quantité de perte d'efficacité que l'on peut consentir en regard des avantages apportés. Ce choix est le plus souvent arbitraire, dépendant du point de vue et du référentiel choisi. Il constitue la principale difficulté de la prise de décision en équivalence.

Dans certaines situations, l'essai peut servir à montrer l'avantage du nouveau traitement en même temps que sa non-infériorité. Par exemple, une fréquence d'effets indésirables moindre et une efficacité suffisante (non-inférieure). L'essai ne sera concluant que lorsque ces deux hypothèses seront vérifiées simultanément. Une adaptation des tests statistiques aux comparaisons multiples est alors nécessaire.

Tableau 1 – Avantages pouvant justifier une recherche de l'équivalence.

Avantage en terme de tolérance
• *Fréquence des effets secondaires moindre* • *effets secondaires moins graves*
Facilité d'utilisation plus grande :
• *voie d'administration plus simple (par exemple orale par rapport à intraveineuse, bolus à la place*

d'une perfusion)
- *une administration par jour à la place de plusieurs ou dose unique à la place d'un traitement de plusieurs jours*
- *absence d'ajustement de dose*

Inconvénients du traitement plus faibles

- *traitement médical à la place d'un traitement chirurgical*
- *chirurgie moins délabrante*
- *radiothérapie moins prolongée*

Coût plus faible

Absence de différence dans un essai de différence

Lorsque dans un essai de différence, la supériorité n'est pas mise en évidence de façon significative, il peut être tentant de conclure à l'équivalence. Cette conclusion pose plusieurs problèmes.

La puissance est peut-être insuffisante. L'absence de différence significative ne signifie pas qu'il y a absence de différence, mais peut être, simplement que l'essai était insuffisamment puissant pour mettre en évidence la différence qui existe entre deux traitements : « l'absence de preuve n'est pas la preuve de l'absence ».

Conclure à l'équivalence après avoir bâti l'essai pour tester une hypothèse de différence revient à changer d'hypothèse. Le principe de la méthode expérimentale, l'approche « hypothesis testing *» de Fisher n'est pas respecté. L'essai ne peut pas être considéré comme démontrant l'hypothèse d'équivalence étant donné qu'il n'avait pas été conçu pour cela (mais pour démontrer l'hypothèse inverse). Dans cette situation, conclure à la démonstration de l'équivalence est une démarche tautologique. Les données servant à la démonstration sont celles qui ont fait générer l'hypothèse. De plus, se pose des problèmes de qualité méthodologique de l'essai (les contraintes de l'essai d'équivalence sont différentes de celles de l'essai de supériorité) et de fixation post-hoc de la limite d'équivalence (cf. infra).*

Dans un essai incluant 1000 patients par groupe, on observe 30 événements critères de jugement dans chaque groupe. Malgré la stricte identité de ces nombres d'événements, ce résultat est loin de permettre de conclure à l'équivalence d'efficacité. Le risque relatif est bien de 1 mais avec un intervalle de confiance à 95% entre 0,61 et 1,65. Ce qui signifie que ce résultat est compatible avec, en réalité, une augmentation de la fréquence du critère de jugement par le nouveau traitement de 65%. Du fait de cette incertitude statistique, il est donc impossible de conclure à l'équivalence des 2 traitements. En fait, il est toujours impossible de conclure à la stricte équivalence entre 2 traitements car cela nécessiterait un intervalle de confiance de largeur nulle, ce qui est impossible car nécessitant une infinité de patients.

Aspects statistiques de la recherche de l'équivalence

Principe général

La première difficulté que pose l'essai d'équivalence est d'ordre statistique. Les tests statistiques classiques sont construits pour rejeter une hypothèse nulle d'absence de différence afin de pouvoir conclure, avec un risque d'erreur contrôlé, à l'existence d'une différence.

Pour conclure à l'équivalence, c'est-à-dire à l'absence de différence, on pourrait imaginer d'inverser l'hypothèse nulle en cherchant à rejeter une hypothèse d'existence d'une différence. Ceci n'est cependant pas possible car, dans ce cas, l'hypothèse nulle correspond à une infinité de valeurs et il devient impossible de calculer la probabilité d'obtenir la valeur observée sous l'hypothèse nulle. En fait, il est impossible au plan statistique de démontrer que deux traitements sont strictement équivalents.

Cet obstacle est contourné par la recherche d'une équivalence relative, qui consiste à montrer que deux traitements ne sont pas trop différents, et que cette différence reste inférieure à un seuil préalablement fixé. Ce seuil correspond à la quantité d'efficacité que l'on peut consentir de perdre, étant donnés les autres avantages

du nouveau traitement. L'introduction de cette tolérance rend les calculs possibles. La démonstration statistique de l'équivalence relative repose sur un processus fondé sur les intervalles de confiance.

Équivalence ou non-infériorité

Non-infériorité et équivalence sont deux notions très proches. La non-infériorité correspond à une équivalence unilatérale, tandis que l'équivalence vraie est bilatérale.

Pour l'efficacité clinique, la recherche de l'équivalence est, sauf cas exceptionnel, une situation unilatérale. En effet, le nouveau traitement sera intéressant en pratique aussi bien s'il s'avère équivalent que supérieur au traitement de référence, mais ne sera pas utilisé s'il s'avère inférieur. Cette description correspond à une situation unilatérale où l'on cherche à montrer avec un risque d'erreur α contrôlé que le nouveau traitement n'est pas inférieur au traitement contrôle, c'est-à-dire en d'autres termes qu'il est au moins aussi efficace. L'utilisation d'un test bilatéral ferait courir le risque de ne pas conclure dans une situation où l'on ne pourrait pas considérer le nouveau traitement comme équivalent car il serait en fait supérieur. Cette conclusion serait paradoxale et gênante en pratique car elle entraînerait l'impossibilité de recommander un traitement qui, en fait, pourrait être supérieur à l'existant.

Très souvent, par abus de langage, le terme équivalence est employé pour désigner ce qui en réalité est une non-infériorité. Cela vient du fait que dans le domaine de l'évaluation du bénéfice clinique, le nouveau traitement peut être substitué à l'ancien à partir du moment où la non-infériorité est démontrée. La bioéquivalence est une situation d'équivalence vraie bilatérale. Il y a bioéquivalence lorsque les paramètres pharmacocinétiques du nouveau médicament ne sont ni plus ni moins élevés que ceux du traitement de référence.

Seuil de non-infériorité

La décision de conclure à la non-infériorité (du nouveau traitement par rapport au traitement de référence), s'effectuera en comparant la borne supérieure de l'intervalle de confiance avec le seuil de non-infériorité choisi. Si cette borne est inférieure à ce seuil, il est possible de conclure à la non-infériorité avec un risque d'erreur contrôlé. En effet, toutes les vraies valeurs probables de la différence d'efficacité du nouveau traitement par rapport au traitement de référence sont inférieures au seuil préalablement fixé. Par contre si la borne supérieure est supérieure au seuil, il n'est pas possible d'exclure que le nouveau traitement soit moins efficace que le traitement de référence. Un intervalle de confiance unilatéral à 97.5% est utilisé. Cet intervalle de confiance correspond à un risque alpha de 2.5%. Cette valeur a été choisie pour être cohérent avec ce qui se passe dans l'essai de supériorité. En effet avec un test bilatéral et un risque alpha de 5% (test classiquement utilisée pour les essais de supériorité), le risque alpha rattaché à la conclusion de supériorité est de 2.5% (cf. test unilatéraux/bilatéraux). Ainsi dans un essais de supériorité, le risque alpha de conclure à tort à la supériorité est de 2.5% (et le risque de conclure à tort à l'infériorité est aussi de 2.5%, ce qui au total fait 5% pour le risque global de conclure à tort). Avec un intervalle unilatéral à 97.5% dans l'essai de non infériorité, le risque de conclure à tort à la non infériorité est donc aussi de 2.5%, ce qui assure une cohérence entre les 2 approches.

Ce processus de décision est illustré sur la Figure 1.

> *Le seuil de non-infériorité correspond à la plus grande perte d'efficacité par rapport au traitement de référence que l'on peut consentir, compte tenu des autres avantages que présente le traitement.*

La signification du seuil est importante. Il correspond à la plus grande perte d'efficacité par rapport au traitement de référence que l'on consent. Par exemple, un seuil relatif de 10% signifie que l'on considérera le nouveau traitement comme « équivalent » (non inférieur) tant que son efficacité ne sera pas inférieure, en relatif, de 10% à celle du traitement de référence. Au maximum, le nouveau traitement, déclaré comme « équivalent », pourra entraîner une augmentation relative de la fréquence du critère de jugement de 10%.

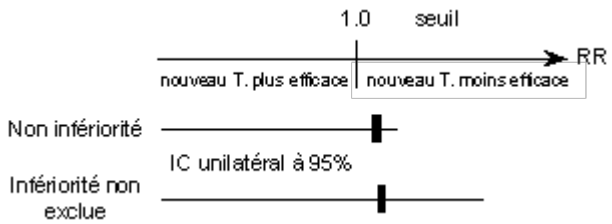


Figure 1 – Illustration du processus de décision de non-infériorité. La borne supérieure du premier intervalle de confiance est inférieure au seuil de non-infériorité choisi. Le nouveau traitement peut être considéré comme non inférieur avec un risque statistique d'erreur de 2.5%. Le second intervalle de confiance englobe le seuil de non-infériorité. Dans ce cas, il n'est pas possible d'exclure que le nouveau traitement soit en fait inférieur (moins efficace) que le traitement de référence.

Apparaît ici l'ambiguïté du terme non inférieur (ou équivalent). Si le seuil choisi correspond à une perte importante d'efficacité, dire que le nouveau traitement est non inférieur est clairement un abus de langage. Il peut être intéressant en pratique étant donné ses autres avantages mais parler de non infériorité, voir d'équivalence est abusif.

En fait, l'essai de non-infériorité ne démontre pas stricto sensu l'équivalence. Il permet simplement d'exclure que, par rapport au traitement de référence, l'efficacité du traitement étudié est inférieure à une certaine limite. On ne peut pas dire que l'équivalence est démontrée car il est possible que le nouveau traitement soit moins efficace que le traitement de référence. Cependant, malgré cela, il peut être acceptable d'utiliser ce nouveau traitement en pratique même s'il n'est pas exclu qu'il soit moins efficace que le traitement de référence car il présente d'autres avantages par ailleurs. Cependant les nuances existant dans l'interprétation de la conclusion disparaissent quand il est conclu, un peu rapidement, que le traitement étudié est équivalent. *Cette formulation, surtout pour le béotien dans le domaine, évoque irrémédiablement l'identité, et amène à penser que les traitements sont interchangeables et conduisent au même résultat.*

Admettre l'équivalence de deux traitements, c'est accepté que le nouveau traitement soit d'une efficacité potentiellement inférieure à celle du traitement de référence.

En fait, un nouveau traitement montré comme équivalent au traitement de référence doit être considéré comme inférieur, jusqu'à preuve du contraire apportée par un essai de supériorité. Dans un classement hiérarchique des traitements par efficacité décroissante, le nouveau traitement arrive en seconde position derrière le traitement de référence.

La valeur du seuil conditionne le nombre de sujets. Plus le seuil est petit, plus l'effectif de l'essai doit être important. Ainsi, il n'est pas réaliste de fixer arbitrairement le seuil à une valeur très petite. La valeur du seuil devra être choisie en fonction de la pathologie, du critère de jugement et de la nature et de l'importance des avantages apportés par le nouveau traitement.

Le choix du seuil de non-infériorité

Introduction

Dans la recherche de la non infériorité, le résultat du test statistique n'est pas aussi absolu que celui d'un test de différence. En non-infériorité, la signification statistique d'un résultat dépend étroitement de la limite d'équivalence choisie. À partir des mêmes données, le test pourra être significatif ou non significatif en fonction de la limite : significatif avec une limite très tolérante et non significatif avec une limite plus stricte.

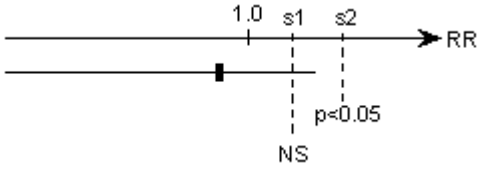


Figure 2 – En fonction de la valeur du seuil choisi, le même résultat peut conduire à un test de non infériorité statistiquement significatif (seuil s2) ou non statistiquement significatif (seuil s1). L'obtention d'un résultat significatif demande d'accepter une plus grande perte d'efficacité.

Le choix du seuil de non infériorité doit permettre d'exclure que le nouveau traitement fasse perdre la totalité du bénéfice apporté par le traitement de référence.

Le choix de la limite est la principale difficulté de l'essai d'équivalence clinique. Quelle perte d'efficacité sur la mortalité justifie une chirurgie moins délabrante en cancérologie ou un traitement fibrinolytique en double bolus dans l'infarctus du myocarde à la place d'une perfusion de 90 minutes ? Le plus souvent le choix est arbitraire, reflète une échelle de valeur, et conduit à des seuils parfois discutables. Ainsi la littérature contient plusieurs exemples dans lesquels la valeur de la limite de non-infériorité s'avère très tolérante et pour le moins discutable. Même si un essai conclut de façon statistiquement significative à la non-infériorité d'un nouveau traitement, il est tout à fait possible de rejeter cette conclusion si l'on considère que le seuil utilisé était trop tolérant. Contrairement au résultat d'un test statistique de différence qui ne peut être remis en cause (zéro est zéro, sans discussion possible).

Exemple

Dans un essai comparant une héparine de bas poids moléculaire (HBPM) à une héparine non fractionnée (HNF), le traitement de référence, dans le traitement des thromboses veineuses profondes symptomatiques. La fréquence du critère de jugement (récidives thromboemboliques) attendue sous HNF était de 7 à 8%. La limite a été fixée à 5% en terme de différence absolue, ce qui signifiait que l'HBPM allait être déclarée équivalente tant qu'elle n'entraînerait pas une fréquence de récidive de 12 à 13%. Cette limite absolue de 5% correspond en fait, à une augmentation relative de 66% de la fréquence attendue dans le groupe HNF. Ainsi les investigateurs étaient prêts à accepter que le nouveau traitement puisse multiplier par 1,66 la fréquence des récidives, parmi lesquelles figure l'embolie pulmonaire. Un autre argument suggère que cette limite absolue de 5% est exagérée. L'HNF entraîne une réduction absolue de l'ordre de 13% par rapport au placebo. Avec la limite de 5%, on accepte une perte de plus d'un tiers (38%) du bénéfice apporté par l'HNF. Heureusement, le résultat de l'essai conduit à un intervalle de confiance dont la borne supérieure est inférieure à cette limite (1,07).

Positionnement par rapport à l'efficacité du traitement de référence

Le choix du seuil de non-infériorité peut être facilité par l'étude du bénéfice qu'apporte le traitement de référence. Le seuil sera fixé de telle façon à ne pas permettre que l'utilisation du nouveau traitement conduise à perdre l'avancée thérapeutique représentée par le traitement de référence. Entre autres, si le traitement de référence a validé son efficacité contre placebo, le seuil choisi garantira que le nouveau traitement ne puisse pas être moins bon que le placebo.

Cette démarche a été utilisée dans l'essai COBALT [5]. L'objectif de cet essai de fibrinolyse à la phase aiguë de l'infarctus du myocarde était de comparer un nouveau traitement représenté par un double bolus d'alteplase au traitement de référence, la perfusion accélérée d'alteplase. La perfusion accélérée est justifiée par les résultats de l'essai GUSTO 1 où ce traitement s'est avéré meilleur que le traitement de référence précédent, la streptokinase. Dans cet essai, la mortalité à 30 jours sous streptokinase était de 7,3% contre 6,3% avec l'alteplase. Le gain apporté par la perfusion accélérée d'alteplase est donc, en différence absolue, de 1% avec un intervalle de confiance bilatéral à 95% de 0,4% à 1,6%. Le vrai effet de l'alteplase se situe entre ces deux bornes et au pire, dans la situation la moins favorable, la différence absolue par rapport à la streptokinase n'est que de 0,4%.

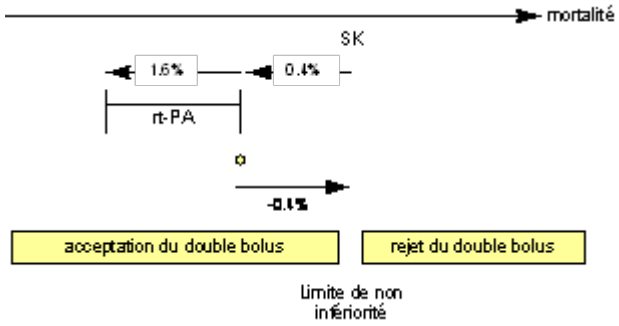


Figure 3 – Illustration du choix de la limite d'équivalence en se basant sur le bénéfice apporté par le traitement de référence. La limite est choisie de telle façon qu'il n'y ait pas de risque de régresser, c'est-à-dire de faire moins bien que le précédent traitement de référence (ici la streptokinase).

En choisissant comme limite cette valeur on est assuré que le nouveau traitement ne pourra pas être considéré comme équivalent alors qu'il fait moins bien que la streptokinase. C'est-à-dire que l'efficacité moindre que l'on pourrait tolérer en regard des avantages pratiques apportés par le double bolus d'alteplase ne conduise pas à perdre tout le bénéfice de la dernière avancée thérapeutique. La démarche de cet essai est exemplaire et montre qu'une définition rigoureuse et argumentée de la non-infériorité est possible. Elle conduit cependant à des valeurs très contraignantes demandant un nombre de sujets de même ordre de grandeur qu'un essai de supériorité.

Dans cet exemple, la perte de 100% de l'efficacité de la perfusion accélérée d'alteplase est acceptable car, même si l'efficacité du double bolus n'est que celle de la streptokinase, le double bolus représente encore un progrès thérapeutique en termes de praticabilité et de meilleure tolérance car il ne présente pas les effets allergiques de la streptokinase.

Un autre exemple du même domaine illustre que le choix de la limite est fréquemment subjectif. L'essai GUSTO 3 [6] comparait un nouveau fibrinolytique, la reteplase qui s'administre en double bolus au traitement de référence qu'est la perfusion accélérée d'alteplase. À l'origine, cet essai était un essai de supériorité dont les résultats ne permettaient pas de conclure à la supériorité de la reteplase. Devant ce résultat, l'objectif de l'étude a été transformé en recherche de l'équivalence clinique. Ce changement post-hoc d'hypothèse pose un problème sur lequel nous reviendrons. Pour l'instant intéressons-nous à la limite de non-infériorité qui a été choisie. Cette limite a été choisie comme dans COBALT par référence à GUSTO 1, mais là où COBALT intégrait l'incertitude statistique en prenant la borne inférieure (0,4%) de l'intervalle de confiance de GUSTO 1, GUSTO 3 considère l'estimation moyenne de 1%. La borne supérieure de l'intervalle obtenue dans GUSTO 3 est de 0,66%, valeur qui autorise de conclure à l'équivalence avec la limite de 1% mais pas avec celle de 0,4%. Les publications de COBALT et GUSTO 3 sont parues dans le même numéro du New England Journal of Medicine et ce contraste entre des choix différents illustre le manque de règles adoptées par tous et les dérives qui peuvent exister.

GUSTO 3 pose aussi la question de savoir s'il est licite de passer d'un objectif de supériorité à un objectif d'équivalence en fonction des résultats. À cause de ce changement post-hoc, le principe de la méthode expérimentale suivant lequel l'hypothèse doit être générée avant le recueil des données (« l'hypothesis testing » de Fisher) n'est pas respecté. L'essai ne peut pas être considéré comme démontrant l'hypothèse d'équivalence étant donné qu'il n'avait pas été conçu pour cela mais pour démontrer l'hypothèse inverse. Dans cette situation, conclure à l'équivalence est une démarche tautologique. Les données servant à la démonstration sont celles qui ont fait générer l'hypothèse. De plus rien n'assure qu'un essai de supériorité répond au critère de qualité de l'essai d'équivalence et que le comparateur soit adéquat.

Ces restrictions méthodologiques sont prises en compte dans la discussion, mais néanmoins la conclusion de GUSTO 3, bien que prudente, suggère fortement l'équivalence.

Cette approche peut être utilisée avec le bénéfice absolu (la différence des risques) ou le bénéfice relatif (risque relatif). Bien que fréquemment utilisée jusqu'à présent, la différence absolue ne tient pas compte du risque de base. Une différence absolue limite fixée à 1% a priori sur l'hypothèse que le risque de base est de 10% correspond à un risque relatif limite de 10%. Si l'essai inclut des patients à faible risque conduisant à un risque de base de 5%, la même limite absolue de 1% correspond alors à un risque relatif limite de 20%. La tendance actuelle est d'utiliser préférentiellement le risque relatif.

Calcul du seuil

En suivant les principes énoncés précédemment, le calcul du seuil se déroule de la façon suivante.

Considérons un essai de non infériorité sur la mortalité du traitement N par rapport au traitement A, dans lequel on souhaite conserver 75% de l'efficacité de A (ainsi l'éventuelle perte d'efficacité acceptable avec N ne doit pas être supérieure à 25% de l'efficacité de A).

En premier il convient de connaitre l'efficacité de A par rapport à son propre contrôle. Ici A a été évalué par rapport au placebo. La méta-analyse des essais de A versus placebo donne comme estimation une différence de risque de -4% avec un intervalle de confiance à 95% entre -2% et -6%.

La plus petite efficacité « garantie » avec A est donc une réduction absolue des décès de 2% (borne péjorative de l'intervalle de confiance). Le seuil de non infériorité, correspondant à la préservation de 75% de cette efficacité, s'obtient par 2%(1-75%) = +0.5%. En effet, 75% de l'efficacité « garantie » de A correspond à une différence absolue de -2%*75%=-1.5%. L'augmentation acceptée de mortalité par rapport à A est donc de -2%-(-1.5%) = +0.5%.*

Pour être considérer comme non inférieur il convient de pouvoir écarter que le nouveau traitement N puisse entrainer une augmentation absolue de plus de 0.5% de la mortalité. La borne péjorative de l'intervalle de confiance de la comparaison N vs A devra donc être inférieure à +0.5%.

Bien que simple, le raisonnement en différence absolue doit être éviter. En effet, l'acceptabilité d'un seuil en différence absolue dépend du risque de départ. Par exemple dans l'essai A versus placebo les risques étaient respectivement de 4% et 8%. Par rapport au risque obtenu sous A (10%) le seuil calculé précédemment correspond à une augmentation relative acceptée de 0.5%/4%=12.5%. Si dans l'essai de non infériorité on obtient un risque sous A de 2% (ce qui assez fréquent car l'essai de non infériorité se déroule quelques années plus tard que l'essai A versus placebo, chez des patients qui bénéficient d'autres avancées thérapeutiques concourant à diminuer encore d'avantage leur risque). Dans cette situation, le seuil de +0.5% représente une augmentation relative de 0.5/2=25%, soit le double ! Pour éviter cette situation, il convient de raisonner directement en risque relatif, ce qui garantira la même augmentation relative consentie quelque soit le risque obtenu avec A dans l'essai de non infériorité.

*Pour l'efficacité de A par rapport au placebo, la méta-analyse donne un risque relatif de 0.85 avec une borne supérieure de l'intervalle de confiance à 95% à 0.92. L'efficacité « garantie » de A par rapport au placebo est donc une réduction relative de risque (RRR) de 8%. Préserver 75% de cette efficacité correspond à une RRR de 75%*8%=6%. Cette RRR de 6% correspond à un risque relatif de 0.94. Le risque relatif seuil est donc 0.94/0.92=1.022.*

Méthodologie

Le but de la méthodologie est d'éviter les biais. Les biais encourus par la recherche de l'équivalence ou de la non-infériorité sont différents de ceux qui peuvent survenir dans les essais de supériorité. La méthodologie doit donc être adaptée et elle diffère des principes classiques de l'essai de supériorité.

La conclusion d'un essai de non-infériorité peut être biaisée si l'efficacité développée par le traitement de référence a été moindre que ce qu'elle aurait du être. Dans ce cas, un nouveau traitement pourtant nettement moins efficace que le traitement de référence apparaîtra à tort équivalent. Le traitement intrinsèquement le plus efficace a été ramené au niveau du moins efficace.

Tableau 2 – Liste des biais possibles

1) L'efficacité du traitement de référence est altérée
a) Le traitement de référence n'est pas administré correctement : dose trop faible ou trop forte, durée du traitement trop courte, etc.
b) Le traitement de référence est administré à des patients chez lesquels il est moins ou pas efficace
c) Le traitement de référence n'est pas le meilleur traitement possible (problème d'interprétation plus que biais)
d) Le traitement de référence est facilement arrêté pour effet indésirables
2) L'efficacité du nouveau traitement est renforcée par des traitements concomitants

La méthodologie de l'essai de non-infériorité doit donc veiller à ce que le traitement de référence développe correctement toute son efficacité (administration correcte) et que l'estimation de son effet reflète bien sa

véritable efficacité (sensibilité et spécificité correctes de la mesure du critère). La confirmation de l'absence d'un biais à ce niveau pourrait être obtenue en incluant dans l'essai un bras placebo pour s'assurer que l'efficacité du traitement de référence est bien celle attendue. Cependant, l'emploi d'un placebo dans une situation où il existe un traitement de référence est rarement possible. Une validation externe est nécessaire.

L'utilisation de traitements concomitants peut aussi entraîner un biais. Si une forte proportion des patients des deux groupes reçoit des traitements concomitants efficaces, l'efficacité observée sera identique dans les deux groupes. Mais il ne s'agira pas de l'efficacité propre des traitements testés mais celle des traitements concomitants. Une équivalence sera observée même si le nouveau traitement est moins efficace.

Dans l'essai de non infériorité, l'analyse en intention de traiter favorise l'hypothèse testée. Par contre, l'analyse per-protocole est conservatrice.

Dans un essai de non-infériorité, l'analyse en intention de traité est moins à l'abri de biais que l'analyse per-protocole. Plusieurs composantes de l'analyse en intention de traiter sont susceptibles de réduire la mesure de l'efficacité des traitements dans les deux groupes et en particulier celle du traitement de référence. La différence entre les groupes tend donc à diminuer, ce qui, dans le cas de l'essai d'équivalence, favorise l'hypothèse testée. Il en résulte donc un biais.

Contrairement à l'essai de supériorité, l'analyse potentiellement la moins biaisée dans l'essai de non-infériorité est l'analyse en per-protocole où seuls les patients traités en stricte conformité avec le protocole sont maintenus dans l'analyse. En pratique, une conclusion sûre n'est possible que lorsque les analyses en intention de traiter et per-protocole donnent des résultats similaires.

L'analyse en intention de traiter évalue l'équivalence des stratégies thérapeutiques.
L'analyse per-protocole évalue l'équivalence des traitements à l'intérieur des stratégies thérapeutiques.

Tableau 3 – Situations introduisant un biais vers l'absence d'effet à travers l'analyse en intention de traiter. En fait, toutes ces situations conduisent aux mêmes conséquences : la dilution et la convergence des effets.

Situations	**Effet sur la différence entre les deux groupes**
Patients qui n'ont pas reçu le traitement alloué	Égalisation des traitements reçus dans les 2 groupes
Arrêt prématuré du traitement de l'étude	
Déviation au protocole, administration de traitement concomitant interdit	Le groupe du traitement testé reçoit des traitements aussi efficaces que le traitement de référence
Patients inclus à tort	Patients ne pouvant pas répondre au traitement car insensibles aux différences de traitement reçu entre les 2 groupes
Prise en compte des perdus de vue comme des valeurs manquantes	Dilution et convergence des effets des traitements

Par exemple, si le nouveau traitement est inférieur, le recours aux traitements concomitants pour échec du traitement testé sera plus fréquent dans le groupe du nouveau traitement que dans celui du traitement de

référence. L'efficacité du nouveau traitement sera renforcée par celle des autres traitements et il pourra apparaître équivalent dans une analyse en intention de traiter. Par contre, dans une analyse per-protocole où les patients qui ont arrêté précocement le traitement testé ont été exclus, l'insuffisance d'efficacité du nouveau traitement apparaîtra.

Au total, l'essai de non-infériorité est extrêmement sensible à sa qualité méthodologique. Un fort taux de perdus de vue, d'écarts au protocole ou d'arrêts prématurés des traitements risque d'égaliser l'efficacité dans les deux groupes et de biaiser le résultat.

Bibliographie

1. *Blackwelder WC. "Proving the null hypothesis" in clinical trials. Controlled Clinical Trials 1982;3:345-53.* PMID:

2. *Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996;313:36-9.* PMID:

3. *Makuch R, Johnson M. Issues in planning and interpreting active control equivalence studies. J Clin Epidemiol 1989;42:503-11.* PMID:

4. *Koopman MM, Prandoni P, Piovella F, Ockelford PA, Brandjes DP, van der Meer J, et al. Treatment of venous thrombosis with intravenous unfractionated heparin administered in the hospital as compared with subcutaneous low- molecular-weight heparin administered at home. The Tasman Study Group. NEJM 1996;334(11):682-7.* PMID:

5. *The Continuous Infusion versus Double-bolus Administration of Alteplase (COBALT) Investigators. A comparison of continuous infusion of alteplase with double-bolus administration for acute myocardial infarction. NEJM 1997;337:1124-30.* PMID:

6. *The Global Use of Strategies to Open Occluded Coronary (GUSTO 3) Investigators. A comparison of reteplase with alteplase for acute myocardial infraction. NEJM 1997;337:1118-23.* PMID:

Utilisation du placebo putatif

Principe

Le principe de cette approche est de calculer à partir des résultats de deux essais, l'un comparant le traitement à A au placebo et l'autre le traitement B au traitement A, l'effet du traitement B par rapport au placebo.

L'interprétation du résultat d'un essai de non-infériorité peut se ramener à un problème de comparaison indirecte. Cette approche de l'équivalence consiste à extrapoler l'efficacité du nouveau traitement par rapport au placebo (putatif) à partir d'un essai contre un comparateur actif. Ce calcul s'effectue à partir de l'estimation de l'efficacité du traitement de référence par rapport au placebo et de celle du nouveau traitement par rapport au traitement de référence. Cette efficacité extrapolée permet ensuite de s'assurer que le nouveau traitement fait mieux de le placebo. Elle peut ensuite être comparée avec celle du traitement de référence. Le problème de la détermination d'une limite de perte d'efficacité consentie se transforme alors en une comparaison des niveaux d'efficacité des deux traitements concurrents, discutée en fonction des autres avantages que procure le nouveau traitement.

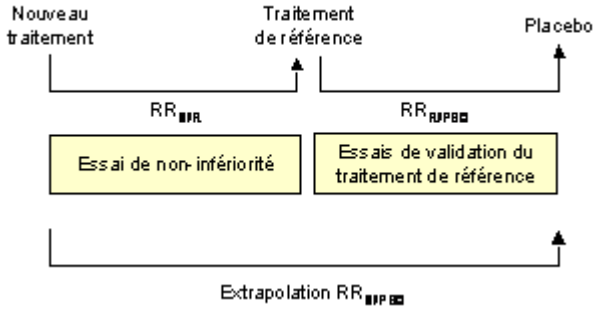


Figure 4 – Illustration du processus d'extrapolation

Calcul

Le risque relatif extrapolé de la comparaison du nouveau traitement versus placebo est obtenu par :

$$RR_{N1PBO} = RR_{C1PBO} RR_{N1C}$$

où $RR_{C/PBO}$ désigne le risque relatif de la comparaison du traitement de référence au placebo (estimé dans un essai ou par la méta-analyse de plusieurs) et $RR_{N/C}$ celui de la comparaison du nouveau traitement au traitement de référence (issu de l'essai de non infériorité).
Le calcul de l'intervalle de confiance nécessite de passer par les logarithmes et de calculer les variances des log des risques relatifs $RR_{C/PBO}$ et $RR_{N/C}$ à partir de leur intervalle de confiance. La variance du logarithme du risque relatif extrapolé ($RR_{N/PBO}$) est alors obtenue par :

$$\mathrm{var}\left[\log\left(RR_{N1PBO}\right)\right] = \mathrm{var}\left[\log\left(RR_{C1PBO}\right)\right] + \mathrm{var}\left[\log\left(RR_{N1C}\right)\right]$$

La variance du logarithme du risque relatif s'obtient à partir des bornes de l'intervalle de confiance par :

$$\mathrm{var}\log RR = \left(\frac{\log(\overline{RR}) - \log(\underline{RR})}{2 \times 1.96}\right)^2$$

Où la notation $\overline{x}, \underline{x}$ désigne respectivement la borne inférieure et la borne supérieure de l'intervalle de confiance.

À partir de cette variance, il est simple de calculer les bornes de l'intervalle de confiance de la comparaison extrapolée :

$$\overline{RR_{N1PBO}}, \underline{RR_{N1PBO}} = exp\left[\log\left(RR_{N1PBO}\right) \pm 1.96 \sqrt{\mathrm{var}\left(\log RR_{N1PBO}\right)}\right]$$

Tableau 4 – Exemple de calculs d'extrapolation d'un risque relatif.

	RR (IC à 95%)	log(RR)	var(log RR)	
Compa (C vs P				149
Nouvea compar				066
Nouvea placebo (extrap				214

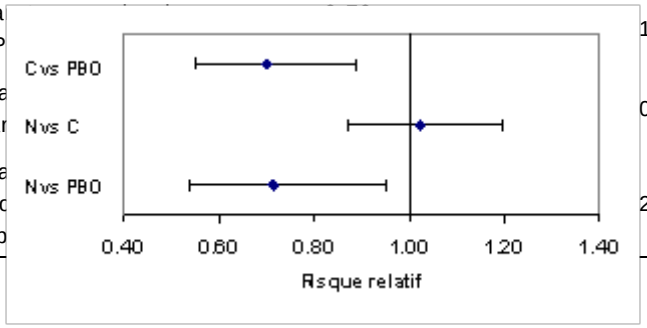
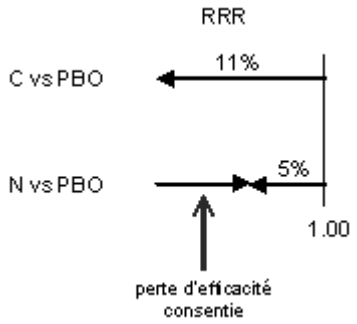


Figure 5 – Exemple de calculs d'extrapolation d'un risque relatif.

Ce graphique montre ainsi, que le nouveau traitement est supérieur au placebo. Son efficacité est très certainement proche de celle du traitement de référence. Mais il n'est pas possible d'exclure avec certitude une efficacité moindre, visualisée par une borne supérieure de l'IC du nouveau traitement plus élevée que celle du traitement de référence (0,95 à la place de 0,89).

Calcul de la perte réelle d'efficacité possible

A partir de l'estimation de l'efficacité du nouveau traitement par rapport au placebo putatif, il est possible de calculer quelle est la réelle perte d'efficacité possible avec le nouveau traitement. Il ne s'agit donc plus de la perte d'efficacité maximale que l'on était prêt à perdre a priori (limite de non infériorité du protocole) mais d'une estimation tenant compte du résultat observé dans l'essai de non infériorité.
Ce calcul s'effectue à partir de la borne péjorative de l'intervalle de confiance de l'estimation extrapolée par rapport au placebo putatif et de la borne péjorative de l'efficacité du traitement de référence par rapport au placebo.
Avec les données de l'exemple ci-dessus, cela revient à calculer quelle perte d'efficacité représente un RR de 0,95 par rapport au RR de 0,89.
Un risque relatif de 0,95 correspond à une réduction relative de risque de 5%, tandis que 0,89 donne une RRR de 11%.



La RRR du nouveau traitement (5%) représente 45% de la RRR du comparateur (11%) (45%=5/11*100), soit une perte de 55% de l'efficacité du comparateur. Il n'est pas possible de raisonnablement exclure que le nouveau traitement ne développe que 45% de l'efficacité du traitement de référence.

Lecture critique de l'essai de non infériorité

La lecture critique des essais de non infériorité comporte quelques particularités par rapport à celle des essais de supériorité. Sur certain point, les critères d'acceptabilité sont même complètement opposés à ceux de l'essai de supériorité.

Validité interne

Le biais majeur auquel est confronté l'essai de non infériorité est de conclure à la non infériorité d'un traitement alors qu'en réalité celui-ci est inférieur au traitement de référence. Toutes les situations qui auront tendance à faire disparaître une réelle différence qui existe entre les 2 traitements sont susceptibles d'engendrer ce type de biais.

1) Évaluation des aspects méthodologiques de la validité interne
    a) Respect de l'approche hypothético-déductive
        i) Pour respecter l'approche hypothético-déductive, l'hypothèse de non infériorité (et le choix du seuil) doit avoir été formulée a priori (avant la prise de connaissance des résultats) et non de façon post hoc après avoir constater l'impossibilité de conclure à la supériorité dans un essai conçu initialement pour montrer la supériorité du nouveau traitement.
    b) Recherche d'un biais de sélection
        i) Un biais de sélection sera évoqué devant toutes situations asymétriques défavorisant le traitement de référence (par exemple, les patients du groupe nouveau traitement sont plus à risque que ceux du groupe du traitement de référence)
        ii) Un biais de sélection sera aussi évoqué si les patients inclus sont à faible risque. dans ce cas, le critère de jugement va mesurer un taux résiduel d'événements qui sera identique dans les deux groupes quel que soient l'efficacité des traitements.
    c) Recherche d'un biais de mesure
        i) Un biais de mesure peut être évoqué lorsque la mesure du critère de jugement manque de sensibilité et/ou de spécificité. Les résultats obtenus dans les 2 groupes sont alors déconnectés de la réalité et mesurent des valeurs résiduelles (bruits de fond) qui ont tendance à être égale dans les 2 groupes.
    d) Recherche d'un biais de réalisation
        i) Un biais de réalisation est évoqué en cas d'administration non optimale du traitement de référence : dose insuffisante, schéma d'administration non optimal, dose effectivement reçue par les patients non optimale.
        ii) Un biais de réalisation survient aussi en cas d'administration de traitements concomitants aux 2 groupes masquant ainsi la différence existant entre les 2 traitements
    e) Recherche d'un biais d'attrition
        i) Un biais est possible avec l'analyse en intention de traiter.
2) Réalité statistique
        i) La décision de non infériorité doit reposer une un intervalle de confiance unilatéral à 5%
        ii) (Il s'agit bien d'une démarche de non infériorité et non pas une conclusion à l'équivalence à partir d'un résultat NS)

Cohérence externe

La cohérence externe d'un résultat d'essai de non infériorité est assurée quand :

- le traitement de référence est validé, sur le même critère de jugement (ou équivalent)
- le résultat est cohérent avec les autres résultats du domaine : pharmaco, épidémio, autres essais
- si l'essai analysé ne le démontre pas, il existe d'autres études démontrant parfaitement la supériorité du nouveau traitement sur le traitement de référence sur d'autres plans que l'efficacité

Pertinence clinique

L'évaluation de la pertinence clinique repose sur les mêmes éléments que ceux utilisés pour l'essai de supériorité auxquels s'ajoutent les points suivants.

1) Pertinence de la limite de non infériorité
    i) La perte d'efficacité consentie est inférieure à l'efficacité du traitement de référence et cette perte est acceptable
2) Pertinence de l'efficacité effectivement montrée par placebo putatif et comparaison indirecte
    i) quel que soit la limite de non infériorité utilisé, est-ce que le nouveau traitement à une efficacité suffisante par rapport au placebo putatif et par rapport aux autres traitement possibles

PERTINENCE CLINIQUE

Indices d'efficacité pour les critères binaires

Introduction

Les indices d'efficacité pour critères binaires quantifient l'efficacité d'un traitement à partir des modifications observées dans la fréquence de survenue d'un événement clinique utilisé comme critère de jugement. Si, par exemple, le critère est le décès, ces indices quantifient la réduction de la mortalité (c'est-à-dire la réduction de la fréquence des décès) provoquée par le traitement.
Plusieurs indices sont utilisables avec ces critères de jugement binaires : risque relatif, odds ratio, différence des risques, NNT.

Données nécessaires aux calculs

Ces indices sont calculés à partir de la fréquence de survenue (risque) du critère de jugement dans les deux groupes expérimental et contrôle. Le terme risque est synonyme de fréquence, il est dérivé du domaine de l'épidémiologie. Dans la suite, ces deux termes seront utilisés indifféremment. Dans un essai, le risque correspond à l'incidence du critère de jugement. Ces risques sont calculés à partir des effectifs (« *size, sample size, effective* ») et du nombre d'événements observés dans chacun des deux groupes. Les données nécessaires sont celles du Tableau 1.

Tableau 1 – Données aux calculs des indices d'efficacités avec les critères de jugement binaires.

Groupe	**Effectif**	**Événements**	**Risque**
traitement étudié	n1	x1	$r_1 = \frac{x_1}{n_1}$
traitement contrôle	n0	x0	$r_0 = \frac{x_0}{n_0}$

Le risque r_0 qui correspond au risque du groupe contrôle est dénommé risque de base (car il correspond en quelque sorte au risque spontanée des patients). Il est aussi appelé risque sans traitement dans les essais contre placebo.

Les indices mesurent en quelque sorte la « distance » qui sépare les risques observés entre le groupe expérimental et le groupe contrôle suivant différente métrique.

Dans la suite, les calculs des différents indices seront illustrés à l'aide de l'exemple suivant :

Tableau 2 – Exemple de résultat apporté par un essai avec un critère de jugement binaire

	Effectif	Événements	Risque
traitement étudié	250	21	0.08 (8%)
traitement contrôle	246	36	0.15 (15%)

Le risque relatif

Définition et généralités

Le risque relatif (« *relative risk* ») est le rapport du risque r_1 obtenu sous traitement divisé par le risque de base r_0 :

$$RR = \frac{r_1}{r_0}$$

Dans notre exemple, le risque relatif vaut RR = 0,08 / 0,15 = 0,53
La réduction relative de risque (RRR) est assez fréquemment utilisée à la place du risque relatif :

$$RRR = (1 - RR)\ 100\%$$

Pour l'exemple, RRR = (1 - 0,53)×100% = 47%. Le traitement entraîne une réduction relative de la fréquence de l'événement (le risque) de 47%.
Le risque relatif (ou la réduction relative des risques) mesure le bénéfice relatif et appartient à la classe des mesures d'effet multiplicatives.

TABLE 3. INCIDENCE OF THE PRIMARY OUTCOME AND OF DEATHS FROM ANY CAUSE.

OUTCOME	RAMIPRIL GROUP (N=4645)	PLACEBO GROUP (N=4652)	RELATIVE RISK (95% CI)*	Z STATISTIC	P VALUE†
	no. (%)				
Myocardial infarction, stroke, or death from cardiovascular causes‡	651 (14.0)	826 (17.8)	0.78 (0.70–0.86)	−4.87	<0.001
Death from cardiovascular causes§	282 (6.1)	377 (8.1)	0.74 (0.64–0.87)	−3.78	<0.001
Myocardial infarction§	459 (9.9)	570 (12.3)	0.80 (0.70–0.90)	−3.63	<0.001
Stroke§	156 (3.4)	226 (4.9)	0.68 (0.56–0.84)	−3.69	<0.001
Death from noncardiovascular causes	200 (4.3)	192 (4.1)	1.03 (0.85–1.26)	0.33	0.74
Death from any cause	482 (10.4)	569 (12.2)	0.84 (0.75–0.95)	−2.79	0.005

*CI denotes confidence interval.

Figure 1 – Exemple d'essai quantifiant les effets à l'aide de risque relatif

	Perindopril (n=6110)	Placebo (n=6108)	Relative risk reduction (95% CI)	p
Cardiovascular mortality, MI, or cardiac arrest	488 (8·0%)	603 (9·9%)	20% (9 to 29)	0·0003
Cardiovascular mortality	215 (3·5%)	249 (4·1%)	14% (–3 to 28)	0·107
Non-fatal MI	295 (4·8%)	378 (6·2%)	22% (10 to 33)	0·001
Cardiac arrest	6 (0·1%)	11 (0·2%)	46% (–47 to 80)	0·22
Total mortality, non-fatal MI, unstable angina, cardiac arrest	904 (14·8%)	1043 (17·1%)	14% (6 to 21)	0·0009
Total mortality	375 (6·1%)	420 (6·9%)	11% (–2 to 23)	0·1

Table 3: **Frequency of primary and selected secondary outcomes**

Figure 2 – Exemple de présentation sous forme de réduction relative du risque

Interprétation du risque relatif

Le risque relatif est une mesure relative. Il exprime l'effet relativement à la fréquence de base de l'événement. Le but d'une mesure relative est de réaliser un ajustement sur la valeur initiale et donc d'obtenir une mesure indépendante de cette-ci. L'objectif est d'obtenir une même valeur que le risque de base soit faible ou important.

Selon les conventions habituelles, un RR<1 témoigne d'un effet bénéfique tandis que RR=1 marque un traitement sans efficacité. Un RR>1 signale un effet délétère (Quand le critère de jugement a valeur d'échec du traitement, ce qui est le cas le plus fréquemment).

La valeur 1 est importante pour l'interprétation du risque relatif. Dans le cas le plus courant où le critère de jugement a valeur d'échec du traitement, par exemple le décès ; un risque relatif inférieur à 1 est obtenu quand la fréquence de l'événement sous traitement est inférieure à celle sans traitement. Le traitement réduit donc la fréquence de l'événement. Un RR<1 signale donc un traitement bénéfique.
Un risque relatif de 1 traduit un traitement sans efficacité : la fréquence de l'événement est la même avec ou sans traitement. Un RR>1 témoigne d'un traitement délétère qui augmente la fréquence de l'événement.
L'interprétation est inversée si le critère de jugement est un événement favorable comme la survenue d'une grossesse dans un traitement de l'hypofertilité. Un risque relatif >1 témoigne alors d'un effet bénéfique (augmentation de la fréquence de survenue d'une grossesse).
Dans un essai contre traitement de référence, un risque relatif inférieur à 1 indique que le traitement testé est supérieur au traitement contrôle tandis qu'un RR>1 témoigne d'une plus grande efficacité du traitement contrôle.
La valeur du risque relatif est le coefficient par lequel il faut multiplier le risque sans traitement pour obtenir celui sous traitement $r_1 = r_0 RR$. Par exemple, un risque relatif de 0,7 signifie que le risque sous traitement est 0,7 fois celui sans traitement (éventuellement, il est possible de dire qu'avec le traitement, le risque est 70% du risque de base). Avec un risque relatif de 0,5, le risque est divisé par 2. Quand RR=4, le risque est multiplié par 4.
Dans le domaine des effets bénéfiques, plus le risque relatif est proche de zéro, plus le bénéfice apporté par le traitement est important. Un risque relatif de 0,3 témoigne d'un effet plus important qu'un RR de 0,7. En effet, les réductions relatives des risques sont respectivement de 70% et de 30%. Pour les effets délétères, plus le risque relatif est supérieur à un, plus l'effet délétère est important en taille.

Calcul de l'intervalle de confiance du risque relatif

Le calcul de l'intervalle de confiance du risque relatif passe par celui de l'intervalle de confiance du logarithme du risque relatif, car le logarithme (népérien) du risque relatif est distribué selon une loi normale et il est possible d'approcher sa variance.

La variance du logarithme du risque relatif est :

$$\mathrm{var}\left(\log RR\right) = \frac{1}{x_1} - \frac{1}{n_1} + \frac{1}{x_0} - \frac{1}{n_0}$$

Les bornes inférieure (*bi*) et supérieure (*bs*) de l'intervalle de confiance à 95% du logarithme du risque relatif sont obtenues par :

$$bi, bs\left(\log RR\right) = \log\left(RR\right) \pm 1,96\sqrt{\mathrm{var}\left(\log RR\right)}$$

Les bornes de l'IC à 95% du risque relatif s'obtiennent par :

$$bi\left(RR\right) = \exp\left[bi\left(\log RR\right)\right] \text{ et } \quad bs\left(RR\right) = \exp\left[bs\left(\log RR\right)\right]$$

L'odds ratio ou rapport des cotes

Définition

Odds

L'odds ratio (dont une traduction littérale en français peut être « rapport des cotes ») est le rapport de l'odds de l'événement (sa cote, il s'agit de la cote des parieurs, comme par exemple la cote d'un cheval) dans le

groupe traité divisé par l'odds de l'événement dans le groupe contrôle. L'odds est égale à c = r /(1 – r) *où* r *est la fréquence de l'événement. Ainsi un odds est le rapport du nombre de patients présentant l'événement,* r×n, *divisé par le nombre de patients ne présentant pas l'événement,* (1-r)×n. *Par exemple, un odds de 0,25 correspond au rapport 2/8 et signifie que pour 2 patients présentant l'événement, 8 ne le présentent pas (0,25=2/8=*r/(1-r) *). Dans la même situation, le risque est 0,20 (2/10).*

Un odds peut aussi être interprétée de la façon suivante : dans un groupe, pour 100 patients ne présentant pas l'événement étudié, 100*c *le présentent.*

L'odds ratio

L'odds ratio s'obtient par

$$OR = \frac{c_1}{c_0} = \frac{r_1/(1-r_1)}{r_0/(1-r_0)}$$

où c_1 *et* c_0 *sont respectivement les odds dans le groupe traité et dans le groupe contrôle.*

Avec l'exemple précédent :

$$OR = \frac{0{,}08/(1-0{,}08)}{0{,}15/(1-0{,}15)} = \frac{0{,}08/0{,}92}{0{,}15/0{,}85} = \frac{0{,}087}{0{,}176} = 0{,}49$$

L'odds ratio peut être aussi exprimé en réduction relative des odds (« relative change in odds *») obtenu par :*

$RRO = (1-OR)\times 100\%$ *. Ainsi avec l'exemple,* RRO = (1-0,49)×100% = 51%.

En épidémiologie, l'odds ratio est utilisé avec les études cas-témoin pour appréhender le risque relatif qui ne peut pas être calculé directement. En effet, dans ces études, les nombres de cas et de témoins sont fixés par l'investigateur « sans que l'on connaisse la population dont-ils sont issues ». On ne peut pas le risque, et par conséquent le risque relatif, car on ne connaît pas le dénominateur. Dans certaines publications, le terme risque relatif est d'ailleurs improprement utilisé pour présenter les résultats obtenus avec un odds ratio.

L'odds ratio est assez peu utilisé dans l'essai thérapeutique. Par contre il est fréquent en méta-analyse (Figure 3).

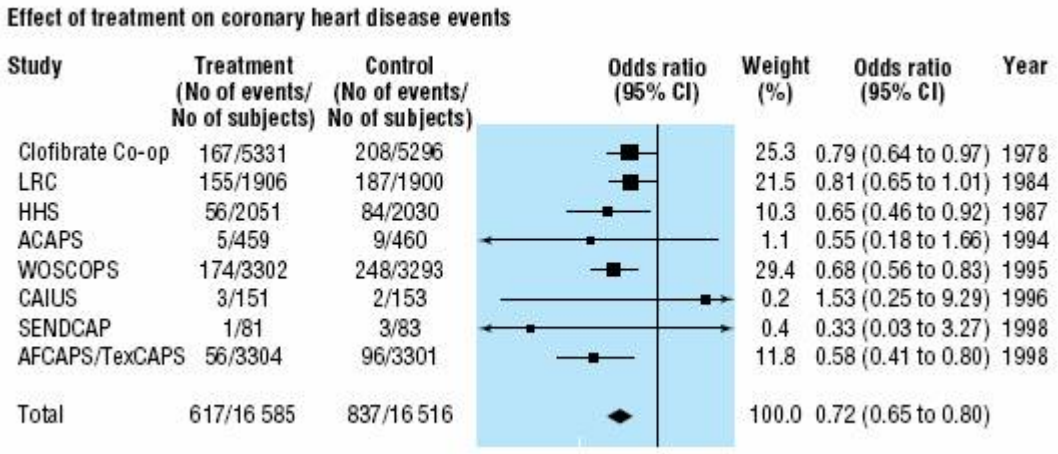


Figure 3 – Exemple de résultats rapportés sous la forme d'odds ratio dans une méta-analyse.

Interprétation

L'odds ratio s'interprète de façon similaire au risque relatif (nous allons d'ailleurs voir que l'odds ratio est proche du risque relatif). Un odds ratio de 1 correspond à l'absence d'effet. En cas d'effet bénéfique, l'odds ratio est inférieur à 1 et il est supérieur à 1 en cas d'effet délétère. Plus l'odds ratio est éloigné de 1, plus l'effet est important.

Comparaison mathématique du risque relatif et de l'odds ratio

L'odds ratio peut aussi se calculer directement à partir de la table de contingence. En considérant les notations suivantes :

	Événement	Non événement	*Effectifs*
traitement étudié	A	B	*n1*
traitement contrôle	C	D	*n0*

l'odds ratio se calcule par :

$$OR = \frac{A\,D}{B\,C}$$

tandis que le risque relatif est :

$$RR = \frac{A(C+D)}{C(A+B)}$$

Ainsi quand la fréquence de l'événement est faible, c'est-à-dire quand A est petit comparé à B et C petit par rapport à D, C+D est assimilable à D et A+B à B. La formule du risque relatif devient alors similaire à celle de l'odds ratio.

Relation entre l'odds ratio et le risque relatif

L'odds ratio est une estimation du risque relatif lorsque la fréquence de l'événement est faible.

Sur un même jeu de données, l'odds ratio est en général assez proche du risque relatif. Ainsi avec notre exemple, OR=0,49 et RR=0,53.

L'odds ratio est d'autant plus proche du risque relatif que le risque de base dans le groupe contrôle est faible. Le Tableau 3 compare la valeur des risques relatifs et des odds ratio calculés à partir des mêmes données, dans différentes situations où le risque de base est plus ou moins grand. La

figure 4 représente graphiquement cette évolution de l'odds ratio en fonction du risque de base.

Tableau 3. Relation entre l'odds ratio (OR) et le risque de base (R0) dans une situation ou le risque relatif (RR) est constant et égal à 0,8.

R0	RR	R1=R0*RR	OR
0.02	0.80	0.02	0.80
0.10	0.80	0.08	0.78
0.20	0.80	0.16	0.76
0.40	0.80	0.32	0.71

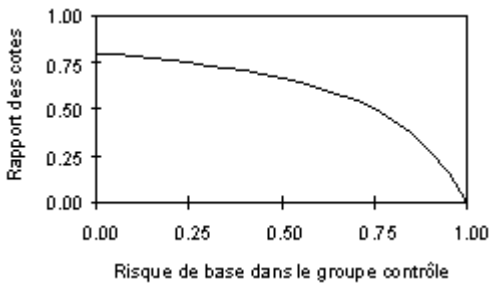


Figure 4 – Évolution de l'odds ratio en fonction du risque de base pour un risque relatif constant de 0,8.

Quand le risque de base est faible, l'odds ratio est donc une bonne approximation du risque relatif. À l'inverse, l'odds ratio tend vers zéro quand le risque de base est très important (voisin de 1) quelle que soit la valeur du risque relatif.

En termes, d'efficacité d'un traitement, l'odds ratio aura tendance à surestimer l'effet du traitement quand le risque de base est élevé (au dessus de 25% environ). Dans ce cas, les résultats présentés avec un odds ratio seront plus favorables au traitement que ceux basés sur le risque relatif et un odds ratio ne peut plus être interprété comme un risque relatif. L'odds ratio ne peut plus être utilisé pour estimer la réduction relative des risques ; et il doit être interprété en termes de réduction des odds, qui est plus importante que la réduction des risques. Par exemple, avec les données de la dernière ligne du Tableau 3, l'odds ratio de 0,71 ne doit pas être interprété comme une réduction du risque de 29% puisque celle-ci n'est que de 20% par construction, mais bien comme une réduction relative des odds de 29%. Dans ce cas, l'effet du traitement exprimé en réduction des odds semble plus important que lorsqu'il est exprimé en réduction de risque.

Bien que l'odds ratio soit quelque peu décrié à l'heure actuelle comme mesure de l'effet d'un traitement [1, 2], il présente néanmoins différents intérêts (cf. § Avantages et inconvénients du risque relatif par rapport à l'. L'odds ratio est aussi l'indice naturel de paramètrisation d'une table 2x2. De ce fait, il intervient dans de nombreuses méthodes de calculs statistiques (régression logistique).

Calcul de l'intervalle de confiance de l'odds ratio

Comme pour le risque relatif, le calcul de l'intervalle de confiance de l'odds ratio se fait par l'intermédiaire de l'intervalle de confiance de son logarithme, car le logarithme (logarithme népérien) de l'odds ratio est distribué selon une loi normale et il est possible d'approcher sa variance.

La variance du logarithme de l'odds ratio est :

$$\mathrm{var}\left(\log RC\right) = \frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_0} + \frac{1}{n_0 - n_1}$$

Les bornes inférieure (*bi*) et supérieure (*bs*) de l'intervalle de confiance à 95% du logarithme de l'odds ratio sont obtenues par :

$$bi, bs\left(\log RC\right) = \log\left(RC\right) \pm 1,96\sqrt{\mathrm{var}\left(\log RC\right)}$$

Les bornes de l'intervalle de confiance à 95% de l'odds ratio s'obtiennent alors par :

$bi\left(RC\right) = \exp\left[bi\left(\log RC\right)\right]$ et $bs\left(RC\right) = \exp\left[bs\left(\log RC\right)\right]$

Surestimation de l'effet traitement par l'odds ratio. Exemple des unités de soins intensifs pour AVC

Une méta-analyse réalisée par le Cochrane Stroke Review Group [3] a évalué l'efficacité en terme de mortalité totale ou de dépendance à la fin de la période de suivi de la prise en charge en unités neuro-vasculaires spécialisées des patients ayant présenté un accident vasculaire cérébral constitué ou transitoire, quelle qu'en soit l'ancienneté.

Cette méta-analyse conclu à un bénéfice important apporté par les unités de soins intensifs sur le risque de mortalité totale ou de dépendance, caractérisé par un odds ratio de 0,75 IC95%=[0,65 ;0,87]. Cet odds ratio ne doit cependant pas être interprété comme une réduction relative du risque de 25% car le risque de base varie entre 39,1 % et 92,3 % en fonction des essais. Ce résultat se situe donc dans une zone de risque de base où l'odds ratio sous-estime fortement le risque relatif. En refaisant les calculs de la méta-analyse directement avec le risque relatif on obtient RR=0,90 IC95%=[0,85 ;0,95], résultat toujours en faveur d'un effet du traitement mais qui apparaît alors moins important (réduction relative du risque de 10%). Dans cet exemple, l'odds ratio traduit la réduction relative des odds apporté par le traitement mais ne permet pas d'estimer la réduction relative de risque. Interpréter l'odds ratio comme un risque relatif revient à surestimer fortement (de 15%) le bénéfice de l'intervention : réduction relative de risque de 25% à la place de 10%.

Avantages et inconvénients du risque relatif par rapport à l'odds ratio

Le risque relatif est la mesure préférentielle du bénéfice relatif apporté par un traitement. Il présente cependant certaines limites.

Deux indices mesurent le bénéfice relatif. Sont-ils cependant identiques ? En pratique, le risque relatif est préféré à l'odds ratio car :

- *il est d'interprétation intuitive,*
- *c'est un indice qui vient naturellement à l'esprit pour mesurer l'effet d'un traitement,*
- *il donne une valeur similaire à celle de l'odds ratio car dans une grande majorité des situations le risque de base est peu important.*

Cependant, le risque relatif présente certaines limites. Il ne peut pas prendre la même plage de valeur en fonction du risque de base. Imaginons, que dans une population où la fréquence de base de l'événement est de 0,2, un traitement délétère entraîne un risque relatif de 4 conduisant à une fréquence sous traitement de 0,8. La mesure de l'effet de ce même traitement dans une population où la fréquence de base est supérieure à 0,25 aurait forcément conduit à un risque relatif plus petit. En effet au-dessus de r_0>0,25, *le produit* r_0 rr *qui est la fréquence sous traitement* r_1 *est supérieur à 1 ce qui est une valeur impossible. Pour un risque de base* r_0*=0,3, le risque relatif maximum qui correspond à un risque sous traitement de* r_1=1 *est de 1/0,3 = 3,33 inférieur à 4. Ainsi, le risque relatif ne peut pas être considéré comme étant une caractéristique universelle d'un effet traitement étant donné que sa valeur est contrainte en partie par le risque de base de la population dans laquelle est étudié l'effet. L'odds ratio (cf. infra) ne présente pas cet inconvénient. Une même valeur d'odds ratio est compatible avec tous les risques de base.*

L'autre limite de cet indice provient de son asymétrie. Le risque relatif n'est pas symétrique dans l'opposition présence / absence de l'événement. Si un traitement entraîne une réduction relative du risque de 20% sur la mortalité il n'entraîne, par exemple pour un risque de base de 10%., qu'une augmentation de 2% de la survie. La survie est pourtant symétrique à la mortalité : une réduction de mortalité entraîne une augmentation de la survie, mais les tailles d'effet ne sont pas conservées. Si la fréquence de présence de l'événement est r_1 *dans le groupe traité et* r_0 *dans le groupe contrôle, la fréquence de l'absence de l'événement est respectivement dans ces deux groupes* 1-r_1 *et* 1-r_0*. Le risque relatif rattaché à l'absence d'événement n'est pas l'inverse de celui rattaché à la présence de l'événement.*

*Avec l'exemple initial, la fréquence de l'absence de l'événement est de 1-8%=92% dans le groupe traité et de 1-15%=85% dans le groupe contrôle. Nous avons vu que le traitement entraîne une réduction relative de 47% de la fréquence de l'événement. Par contre le traitement n'augmente que de 8% la fréquence des patients exempts de l'événement ((1-0,92 / 0,85)*100%) et non pas de 47%. Le traitement pourrait paraître moins efficace avec le critère de jugement de prévention de l'événement par rapport au critère d'échec du traitement. Il s'agit pourtant des deux façons symétriques de voir le problème. L'odds ratio ne présente pas cet inconvénient.*

La différence des risques et NNT

La différence des risques

Définition et calcul

La différence des risques (« risk difference *»), appelée aussi différence absolue ou bénéfice absolu, est égale à la différence entre le risque sous traitement (*r_1*) et le risque sans traitement (*r_0*). Elle se calcule par :*

$$DR = r_1 - r_0$$

Le calcul de la différence des risques à partir des données de l'exemple donne :

$$DR = 0,08 - 0,15 = 0,07 \text{ soit } -7\%.$$

La différence des risques donne la taille de l'effet non ajustée sur la valeur initiale.

*Il est possible de rapporter la différence des risque sous forme de réduction absolue des risques RAR («*absolute reduction in risk*») qui est égale à :*

$$RAR = (1 - DR) \times 100\%$$

TABLE 4. ANALYSIS OF THE RATES AND RISKS OF DEATH FROM ANY CAUSE AT 28 DAYS.*

VARIABLE	PLACEBO GROUP	DROTRECOGIN ALFA ACTIVATED GROUP	P VALUE†	RELATIVE RISK OF DEATH (95% CI)‡	ABSOLUTE REDUCTION IN RISK (95% CI)§
	no./total no. (%)				%
Treated patients					
Nonstratified analysis	259/840 (30.8)	210/850 (24.7)	0.005	0.80 (0.69 to 0.94)	6.1 (1.9 to 10.4)
Stratified analysis¶			0.005	0.81 (0.70 to 0.93)	6.2 (1.6 to 10.8)
Protein C deficiency					
Yes	215/670 (32.1)	182/709 (25.7)	0.009	0.80 (0.68 to 0.95)	6.4 (1.6 to 11.2)
No	28/105 (26.7)	14/90 (15.6)	0.06	0.58 (0.33 to 1.04)	11.1 (−0.4 to 22.6)
Unknown	16/65 (24.6)	14/51 (27.5)	0.73	1.12 (0.60 to 2.07)	−2.8 (−19.0 to 13.4)
Randomized patients‖					
Nonstratified analysis	268/857 (31.3)	216/871 (24.8)	0.003	0.79 (0.68 to 0.92)	6.5 (2.2 to 10.7)

*Patients were analyzed in the treatment group to which they were assigned at randomization. CI denotes confidence interval.

†Two-sided P values for the nonstratified and subgroup analyses are based on Pearson's chi-square tests, and the P value for the primary stratified analysis is based on the Cochran–Mantel–Haenszel test.

‡The relative risk of death is calculated as the mortality rate in the drotrecogin alfa activated group divided by the mortality rate in the placebo group.

Figure 5 – Exemple d'essai quantifiant les effets à la fois avec un risque relatif et la réduction absolue des risques (1-DR)

Interprétation

En l'absence d'effet du traitement, la différence est nulle. Un effet bénéfique se traduit par une différence des risques négative et un effet délétère par une valeur positive. Plus la valeur absolue de la différence de risque est importante plus l'effet est grand. Une différence des risques de -7% signifie que le traitement évite la survenue de 7 événements pour 100 patients traités.

Parfois, la différence des risques est calculée comme étant la différence entre le risque sans traitement et celui avec traitement : r_0-r_1 *(avec les données de l'exemple cela donne 0,015-0,08 = 0,07). Dans ce cas, la signification de la différence des risques est inversée : valeur positive pour un effet bénéfique, valeur négative pour une effet délétère. La première définition présente l'avantage de localiser sur un graphique les valeurs correspondantes à un effet bénéfique à gauche de la valeur témoignant l'absence d'effet, de la même façon qu'avec le risque relatif ou l'odds ratio.*

Calcul de l'intervalle de confiance

Les bornes de l'intervalle de confiance à 95% de la différence des risques s'obtiennent par :

$$bi,bs(DR) = DR \pm 1{,}96\sqrt{\text{var}(DR)}$$

où var(*DR*) désigne la variance de la différence des risques dont une approximation est :

$$\text{var}(DR) = \frac{r_1(1-r_1)}{n_1} + \frac{r_0(1-r_0)}{n_0}$$

Le NNT

Cet indice nommé par l'abréviation NNT pour Nombre de sujet Nécessaire de Traiter (ou en anglais « Number Needed to Treat *») correspond au nombre moyen de sujets qu'il est nécessaire de traiter pour éviter 1 événement. Il est égal à l'inverse de la différence des risques :*

$$NNT = \frac{1}{DR} = \frac{1}{r_1 - r_0}$$

Dans notre exemple, le NNT est égal à NNT = 1/-0,07 = -14. Le signe moins témoigne d'un effet bénéfique et l'interprétation en fonction du signe est identique à celle de la différence des risques.

Attention, pour le calcul du NNT, la différence des risques ne doit pas être exprimée en pourcentage. En effet, 1/7=0,14 n'est pas la bonne valeur du NNT (rappel 7%=0,07).

Un NNT de 14 signifie qu'il faut traiter en moyenne 14 patients pour éviter un événement. En effet, sans traitement le nombre d'événements attendu chez 14 sujets est de 14×0,15=2,1 tandis que sous traitement ce nombre est de 14×0,08=1,1, ce qui correspond bien à un patient de moins. En moyenne, tous les NNT patients traités, un événement est évité.

Exemple

"Treatment with atorvastatin 10 mg daily for 4 years in 1000 such patients would prevent 37 first major cardiovascular events … . One major first cardiovascular event would be avoided for every 27 patients treated for 4 years. We have expressed these measures of absolute benefit at our median duration of follow-up of 4 years" D'après ref. [4].

Différence des risques, NNT et durée de suivi

Le NNT et la différence des risques, dépendent de la durée de suivi [5]. Celle-ci doit donc toujours être précisée quand on rapporte ces indices. Le Tableau 4 présente l'évolution au cours du temps des risques absolus avec et sans traitement, de la différence des risques et du NNT dans une situation où le risque annuel est de 5% et le risque relatif constant et égal à 0,80. Plus la durée de suivi est longue, plus le risque sans traitement est élevé. La différence des risques augmente aussi avec la durée de suivi alors que le NNT diminue. Cette relation entre NNT et différence de risque et durée de suivi doit être prise en compte lorsque l'on compare différents traitements.

Tableau 4 – Évolution en fonction de la durée de suivi de la différence des risques et du NNT

Suivi (ans)	r0	r1	DR	NNT
1	5%	4%	-1.0%	- 100
2	10%	8%	-2.0%	- 51
3	14%	11%	-2.9%	- 35
4	19%	15%	-3.7%	- 27
5	23%	18%	-4.5%	- 22
10	26%	21%	-5.3%	- 19

NNT et différence de risque

Différence des risques et NNT véhiculent la même information et quantifient le nombre d'événements évités par nombres de patients traités. En effet, cette information peut être exprimée :

- *soit en nombre d'événements évités pour 100 (ou 1000) patients traités,*
- *soit en nombre de patients à traiter pour éviter 1 événement.*

Par exemple, une différence des risques de 3/100 signifie que le traitement de 100 sujets durant la période de suivi de l'essai a permis d'éviter la survenue de 3 événements. La même situation correspond à un NNT de 33 signifiant qu'il faut traiter 33 patients durant la période de suivi de l'essai pour éviter 1 événement.

Interprétation

Depuis son introduction par Sackett [6], cet indice obtient un grand succès, car il semble très « parlant » pour les praticiens. Par exemple, à partir du nombre de patients souffrant de la maladie étudiée présent dans la clientèle d'un médecin, il estime le nombre d'événements que le traitement pourrait éviter. Il mesure en quelque sorte « l'énergie » moyenne qu'il faut dépenser pour obtenir un succès thérapeutique.

Son expression en termes d'un événement évité pour NNT patients traités a cependant tendance à faire croire qu'il existe réellement un individu qui ne présentera pas l'événement avec le traitement alors qu'il l'aurait présenté sans traitement. Il y a individualisation du bénéfice, mais cette interprétation est fausse*. Cet indice est issu d'un développement valable en moyenne, et l'interprétation la plus probable est que tous les patients bénéficient un peu du traitement (il y a réduction de leur probabilité de faire un événement) et que cette*

réduction se traduit en moyenne par une différence de 1 du nombre moyen d'événements attendus sans et avec traitement dans un groupe de NNT patients.

Lorsque l'on dit qu'il faut traiter, par exemple, 30 patients pour éviter un événement, la tentation est forte de traduire cela en : un patient bénéficie du traitement tous les 30 patients traités. Par extension, sur 30 patients traités, 29 n'en bénéficient pas. Cette interprétation est abusive en personnalisant trop le bénéfice. Par exemple, il se peut que tous les patients bénéficient un peu du traitement en ayant par exemple leur espérance de vie augmentée de quelques mois. Donc aucun patient ne bénéficie plus que les autres.

Exemple de sur-interprétation du NNT
"There's a lot of expensive technology. Some of it provides benefit for only a tiny fraction of patients and at great cost" to everybody else, Sox says. With breast cancer, for example, 800 women must be screened for 13 years to prevent a single death. Yet, women aren't going to eschew mammography because of the cost to society of the test. (http://www.healthscout.com/template.asp?page=newsdetail&ap=1&id=507477 visité le 24/04/05).

Pour un même domaine thérapeutique, le NNT (comme le bénéfice absolu) varie plus d'un essai à l'autre que le bénéfice relatif (mesuré par un risque relatif ou odds ratio) car le risque de base varie souvent d'un essai à l'autre. Ce point a d'ailleurs été objectivé empiriquement [7]. Le Tableau 5 compare les risques relatifs et les NNT obtenus dans 4 essais comparant une statine au placebo dans la prévention des maladies cardiovasculaires [8].

Pour cette raison, les NNT ne doivent pas être utilisés pour comparer deux essais, encore moins pour comparer deux traitements concurrents pour une même pathologie. Des différences de NNT pouvant témoigner aussi bien d'une différence d'efficacité des traitements que de différences de risques de bases des patients entre les essais. La comparaison des NNT de deux traitements nécessite de se baser sur des NNT correspondant à un même risque base [9]. Cette standardisation est obtenue en calculant les NNT à partir des risques relatifs et d'un risque de base de référence.

Tableau 5. Comparaison des risques relatifs et des NNT concernant les événements coronariens observés dans 4 essais de prévention du risque cardiovasculaire par les statines.

	Risque relatif	NNT (DR)	Risque de base
WOSCOPS	0,70	44 (23 ‰)	7,5%
CARE	0,77	33 (30 ‰)	13,1%
LIPID	0,77	28 (35 ‰)	15,9%
4S	0,70	15 (67 ‰)	22,6%

Pour un risque de base fixé R_c*, le NNT s'obtient à partir du risque relatif* RR *par la formule :*

$$NNT = \frac{1}{R_c(RR-1)}$$

*Par exemple, avec un traitement caractérisé par un risque relatif de 0,8, le NNT correspondant à un risque de base de 10% est égal à 1/(0,1*0,2)=50. En effet, sous traitement, le risque est 0,1*0,8=8% ce qui donne une différence des risques de 10%-8%=2% soit un NNT=1/0,02=50.*

La comparaison des bénéfices absolus de différents traitements doit s'effectuer à risque de base identique.

Le Tableau 6 illustre les pièges tendus par la comparaison directe des NNT de deux traitements évalués dans des essais différents. À partir des NNT obtenus par les essais, on serait amené à conclure que le traitement A apporte un plus grand bénéfice absolu que le traitement B. Cependant, cette conclusion est erronée car le risque de base de l'essai évaluant le traitement B est plus petit que celui de l'essai du traitement A. En utilisant un NNT standardisé, par exemple pour un risque de base de 10% (qui se situe au milieu des risques des essais), il apparaît que le traitement B apporte un plus fort bénéfice absolu que le traitement A.

Tableau 6. Comparaison des NNT de deux traitements évalués dans des essais différents.

	NNT observé dans l'essai	Risque relatif	Risque de base	NNT standardisé pour un Rc=10%
Essai A (traitement A)	35	0,81	0,15	52
Essai B (traitement B)	53	0,69	0,06	32

NNH

La notion de NNT peut être étendue aux événements délétères. Il s'agit alors du « Number needed to harm (NNH) *» qui quantifie le nombre de patients qu'il faut exposer au traitement pour observer, en moyenne, un effet indésirable du traitement (cf. chapitre* Rapport bénéfice risque*).*

Table 4 Risk of side effects with latanoprost and timolol

Side effects	No of trials	Crude event rate		RR (95% CI)	RD (95% CI)	NNH (95% CI)
		Latanoprost	Timolol			
Withdrawals due to adverse effects	2	9/277	14/285	0.70 (0.30, 1.66)	−0.02 (−0.05, 0.16)	NA
Hyperaemia	6	51/586	20/503	2.20 (1.33, 3.65)**	0.05 (0.02, 0.07)**	21 (14, 42)**
Conjunctivitis	3	7/419	5/320	0.80 (0.25, 2.53)	0.006 (−0.001, 0.02)	NA
Increased pigmentation	4	21/478	0/387†	8.01 (1.87, 34.30)*	0.03 (0.01, 0.04)**	36 (22, 91)**
Hypotension	1	0/149†	2/145	0.19 (0.01, 4.02)	−0.01 (−0.04, 0.01)	NA
Bradycardia	1	0/87†	2/91	0.21 (0.01, 4.29)	−0.02 (−0.06, 0.02)	NA

RR = relative risk; RD = risk difference or attribute risk; NNH = number needed to harm; CI = confidence interval.
†For trials with event rate of zero, 0.5 was added to each cell of the individual 2 × 2 table to calculate RR, RD, or NNH.
*p≤0.05, **p≤0.01.

Figure 6 – Exemple de quantification des événements indésirables faisant appel au NNH. D'après ref. [10].

Calcul de l'intervalle de confiance d'un NNT

Il n'existe pas de méthode de calcul satisfaisante de l'intervalle de confiance du NNT

Un autre problème posé par le NNT est que le calcul de son intervalle de confiance n'est pas direct.

La première idée qui vient à l'esprit est de prendre comme borne de l'intervalle de confiance du NNT l'inverse des bornes de l'intervalle de confiance de la différence des risques. Ainsi à partir d'une différence des risques de 0,07 avec un IC95% = [0,02 ;0,12], on obtient un NNT de 14,3 avec comme borne de son intervalle de confiance 1/0,02=50 et 1/0,12=8,3 soit comme IC95% [8,3 ;50].

Un problème survient avec cette méthode lorsque les bornes de l'intervalle de confiance de la différence des risques sont de part et d'autre de zéro. En effet, avec une différence de risque de 0,01 dont l'IC95% est [–0,04 ;0,06] le calcul précédent donne comme borne 1/-0,04=-25 et 1/0,06=16,6 soit un IC95%=[-25 ;16,6] pour un NNT de 1/0,01=100. L'estimation ponctuelle (100) est située en dehors de l'intervalle de confiance. Ce phénomène provient de la nature hyperbolique de la transformation 1/x. Une méthode de calcul non entièrement satisfaisante a été proposée par Doug Altman [11], et une autre par R Bender [12].

Comparaison des différents indices, Bénéfice relatif — bénéfice absolu

L'efficacité d'un traitement est envisageable suivant deux axes : le bénéfice relatif et le bénéfice absolu

Les différents indices d'efficacité que nous venons de voir, risque relatif, odds-ratio, différence des risques, et nombre de sujets qu'il faut traiter pour éviter un événement (NNT), ne véhiculent pas exactement la même information clinique. Ainsi, ces mesures ne donnent pas exactement les mêmes renseignements sur la pertinence clinique d'un effet. Les deux premières (odds ratio, risque relatif) sont des mesures relatives et estiment un bénéfice relatif, tandis que les deux dernières (différence des risques et NNT) mesurent un bénéfice absolu.
Le bénéfice relatif est plutôt une information explicative. Il est le reflet direct de l'efficacité du traitement. Il est en général constant d'une population à l'autre et sa valeur caractérise le traitement pour un large éventail de situations. C'est une information qui intéresse le chercheur. C'est aussi un indice qui permet de comparer différents traitements de la même maladie. En cela il intéresse le prescripteur et le décideur de santé publique.
Le bénéfice absolu reflète plus les conséquences apportées par un traitement au niveau d'une population. Il est spécifique d'une situation particulière : traitement caractérisé par son risque relatif, type de patients conditionnant le niveau de risque de base, durée de traitement ou de suivi. Il est plus pertinent en termes de santé publique.

Tableau 7 – Comparaison du bénéfice relatif et du bénéfice absolu

Bénéfice relatif	Bénéfice absolu
(risque relatif, odds ratio)	(différence des risques, NNT)
indice explicatif indice caractéristique du traitement intéresse le chercheur intéresse le médecin et le décideur de santé publique pour apprécier le bénéfice absolu	indice pragmatique indice caractéristique d'une situation (traitement, patient, durée) reflète l'importance du bénéfice par rapport au risque de base intéresse le médecin et le décideur de santé publique

Une réduction relative de risque de 30% est déjà une réduction conséquente, qui, d'ailleurs, n'est que rarement observée. Malgré cela, la pertinence clinique de cet effet dépend du risque de base. En effet, réduire en relatif de 30% un événement fréquent est bien plus intéressant que de réduire dans la même proportion un événement rare. Si le risque de base est de 50%, sous l'effet d'une réduction de 30%, il devient 35%, donnant une différence de risque de 15%. Avec un risque initial de 5%, la même réduction relative aboutit à un risque sous traitement de 3,5%, correspondant à une différence absolue de 1,5%. En termes d'événements évités pour 1000 sujets traités, le premier cas de figure correspond à 150 événements évités, tandis que le second à seulement 15. Aussi bien du point de vue de la santé publique, que du point de vue individuel, la première situation est plus intéressante que la seconde.
En pratique, l'intensité de l'efficacité d'un traitement est quantifiée par le risque relatif car le risque relatif se révèle relativement constant d'une situation clinique à une autre [7], il caractérise le traitement comme une sorte de « constante universelle ». Cependant, ce risque relatif ne permet pas de juger directement de la pertinence clinique de l'utilisation de ce traitement sans une situation clinique particulière. En effet, cette pertinence clinique va dépendre du bénéfice absolu apporté par le traitement dans cette situation et qui dépend du risque relatif et du risque de base. Ainsi, en fonction du niveau de risque des patients, un même traitement pourra apporter un bénéfice plus où moins intéressant. En règle générale, le traitement de patients à très bas risque n'a que peu d'intérêt pour la collectivité comme pour le patient.

Guide d'appréciation du bénéfice apporté par un traitement

D'une manière générale, pour apprécier le bénéfice apporté par un traitement, il va être nécessaire de vérifier que le risque relatif est constant à travers les essais (aux fluctuations d'échantillonnages près) puis d'estimer ce risque relatif. Ceci est effectué par la méta-analyse en utilisant les outils de recherche de l'hétérogénéité et l'estimation du risque relatif commun.
Ensuite le bénéfice absolu apporté par ce traitement est estimé pour un ou plusieurs risques de bases correspondant aux patients plus ou moins à risque de la population cible. La détermination de la valeur du

risque de base peut s'effectuer à partir des groupes contrôles (placebo) des essais. La plage de variation des risques de base est déterminée, puis le risque moyen ou médian. Dans les situations où les niveaux de risque des patients sont plus ou moins élevés entre les essais, il est alors possible de déterminer la valeur de risque caractérisant le mieux les patients à faible risque, à risque moyen et à haut risque, conduisant à la détermination de trois valeurs de bénéfice absolu. Les études épidémiologiques sont aussi utilisables pour cerner le risque de bases.
Pour les essais contre traitement actif, le risque de base a utilisé est celui qui est obtenu avec ce traitement. Le calcul donne le surcroît de bénéfice absolu apporté par le nouveau traitement vis-à-vis du précédent.

Influence de la présentation des résultats sur la perception de l'intensité de l'efficacité

Plusieurs études ont montré que le choix de l'indice utilisé pour rapporter les résultats d'un essai influe sur la perception de l'importance de l'efficacité d'un traitement. Les résultats d'un essai rapportés sous forme de bénéfice relatif suggèrent spontanément une efficacité plus importante que les mêmes résultats rapportés sous forme de bénéfice absolu [13].
Ainsi une réduction relative du risque de 50% est plus frappante qu'une réduction du risque de 5 pour 1000 (0,5%) (de 1% à 0,5%) ou qu'un NNT de 200.
Pour ne pas se laisser abuser par ce phénomène, il est nécessaire d'analyser systématiquement le bénéfice relatif et sa traduction en bénéfice absolu pour les risques de bases couramment rencontrés avec la pathologie considérée.

Bénéfice faible et maladie fréquente

Les bénéfices absolus de petite taille ne sont pas forcément sans intérêt. Lorsque la maladie est très fréquente, un petit bénéfice aura pour conséquence un grand nombre d'événements évités au niveau de la population tout entière. Si l'événement prévenu est sérieux (décès), un traitement de ce type présente un intérêt en termes de santé publique.
Par exemple, environ 100 000 infarctus du myocarde surviennent chaque année en France. Une réduction de la mortalité de 1% (bénéfice supplémentaire apportée par la perfusion accélérée d'alteplase par rapport à la streptokinase) évite 1 000 décès en phase précoce.

Autres indices pour critères binaires

Le rapport des taux et la différence des taux

La « fréquence de survenue » des critères binaires est parfois exprimée sous la forme de taux d'incidence plutôt qu'en termes de risque. Par rapport au risque, un taux est calculé en utilisant comme dénominateur la durée cumulée de suivi. Le taux d'incidence s'exprime donc en nombre d'événements par personnes-années de suivi. Cette approche est particulière utile quand les individus étudiés ont eu des durées de suivi très différentes. En effet, plus le temps de surveillance d'un sujet est important, plus la probabilité qu'il présente l'événement est élevé. L'utilisation de la durée cumulée de suivi permet donc de corriger le calcul du risque de l'hétérogénéité des durées de suivi. En effet, entre deux groupes de même taille et regroupant des sujets identiques, un plus grand nombre d'événements sera observé dans celui qui aura été observé sur une plus longue période. Le risque sera alors plus important dans ce groupe que dans l'autre. Par contre, les taux d'incidence seront identiques.

Tableau 8 – Expression des résultats d'un essai en termes de taux d'incidence

	Durée de suivi cumulée	Nombre d'événements	Risque	Taux d'incidence
Groupe 1 (n=100)	190 PA[†]	19	19/100=19%	19/190=10 pour 100 PA
Groupe 2 (n=100)	110 PA[†]	11	11/100=11%	11/110=10 pour 100 PA

[†] PA : personnes années

Le taux d'incidence est en quelque sorte la vitesse de survenue de l'événement. Les courbes de survie sont une autre approche permettant aussi de prendre en compte la vitesse de survenue des événements.
Le taux d'incidence ne peut pas être obtenu en divisant le risque par la durée de la période d'observation, même si tous les sujets ont été suivi durant cette période. En effet, la durée de suivi cumulée intègre le temps de survenue des décès et elle est donc inférieure à la durée de la période d'observation. Pour les patients décédés, le temps de suivi est inférieur à la durée de la période d'observation.
Lorsque les taux d'incidence sont utilisées, le rapport des taux mesure le bénéfice relatif du traitement et la différence des taux mesure le bénéfice absolu.

Exemple

L'essai Syst-Eur [14] compare un traitement actif au placebo dans l'hypertension artérielle du sujet âgé. 2398 sujets ont été inclus dans le groupe nitrendipine et 2297 dans le groupe placebo (Figure 7). Le nombre de patients années dans le groupe traité est de 5995, il est de 5709 dans le groupe placebo. Les durées de suivi varient entre 1 et 97 mois avec une médiane de 2 ans. 123 décès sont survenus dans le groupe traité et 137 dans le groupe contrôle, conduisant à des risques de mortalité de respectivement 123/2398=5,13% et de 137/2297=5,96%. Les taux de mortalité dans les groupes traités et contrôle sont respectivement de 20.5 et 24.0 pour 1000 personnes années, ce qui donne une différence des taux de –14 avec un IC95% de [-33 ;9] (p=0,22)

Cause of death	Rate per 1000 patient-years (number of deaths)		Difference (active minus placebo)	
	Placebo (n=2297)	Active (n=2398)	% rate (95% CI)	p
All causes	24·0 (137)	20·5 (123)	–14 (–33 to 9)	0·22
Unknown cause	0·4 (2)	0·7 (4)	..	..
Cardiovascular				
All cardiovascular	13·5 (77)	9·8 (59)	–27 (–48 to 2)	0·07
Stroke	3·7 (21)	2·7 (16)	–27 (–62 to 39)	0·33
Cardiac mortality*	9·1 (52)	6·7 (40)	–27 (–51 to 11)	0·14
Heart failure	1·8 (10)	1·3 (8)	–24 (–70 to 93)	0·57
Coronary mortality†	7·4 (42)	5·3 (32)	–27 (–54 to 15)	0·17
Myocardial infarction	2·6 (15)	1·2 (7)	–56 (–82 to 9)	0·08
Sudden death	4·7 (27)	4·2 (25)	–12 (–49 to 52)	0·65
Dissecting aortic aneurysm	0·4 (2)	0·2 (1)	..	..
Pulmonary embolism	0·2 (1)	0·3 (2)	..	..
Peripheral arterial disease	0·2 (1)	0 (0)	..	..
Non-cardiovascular				
Total	10·2 (58)	10·0 (60)	–1 (–31 to 41)	0·95
Cancer	4·4 (25)	3·0 (18)	–31 (–63 to 26)	0·22

*Included deaths from heart failure and coronary mortality.
†Consisted of fatal myocardial infarction and sudden death.

Table 3: **Mortality by treatment group**

Figure 7 – Résultats de l'essai Syst Eur rapportés sous forme de taux d'incidence (pour 1000 patients années) avec quantification de l'efficacité par la différence des taux.

Bibliographie

1. Sackett DL, Deeks JJ, Altman DG. Down with the odds ratio! Evidence Based Medicine 1996;1:164-6. *PMID:*
2. Bland JM, Altman DG. The odds ratio. BMJ 2000;320:1468. *PMID:*
3. Organised inpatient (stroke unit) care for stroke. Stroke Unit Trialists' Collaboration. Cochrane Database Syst Rev 2000(2):CD000197. *PMID:*
4. Randomised trial of efficacy and safety of inhaled zanamivir in treatment of influenza A and B virus infections. The MIST (Management of Influenza in the Southern Hemisphere Trialists) Study Group. Lancet 1998;352(9144):1877-81. *PMID: 9863784*.
5. Lubsen J, Hoes A, Grobbee D. Implications of trial results: the potentially misleading notions of number needed to treat and average duration of life gained. Lancet 2000;356:1757-59. *PMID:*
6. Sackett DL. On some clinically useful measures of the effects of treatment. Evidence Based Medecine 1996;1:37-38. *PMID:*
7. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med 1998;17:1923-42. *PMID:*
8. Cucherat M, Lièvre M, Gueyffier F. Clinical benefits of cholesterol lowering treatments. Meta-analysis of randomized therapeutic trials. Presse Med 2000;29:965-76. *PMID:*
9. McQuay HJ, Moore AR. Using numerical results from systematic review in clinical practice. Anns of Internal Medicine 1997;126:712-20. *PMID:*
10. Zhang WY, Po AL, Dua HS, Azuara-Blanco A. Meta-analysis of randomised controlled trials comparing latanoprost with timolol in the treatment of patients with open angle glaucoma or ocular hypertension. Br J Ophthalmol 2001;85(8):983-90. *PMID: 11466259*.
11. Altman DG. Confidence intervals for the number needed to treat. BMJ 1998;317(7168):1309-12. *PMID:*
12. Bender R, Lange S. Adjusting for multiple testing - when and how? J Clin Epidemiol 2001;54(4):343-9. *PMID:*
13. McGettigan P, O'Connell D. The effect of information framing on the practice of physicians. A systematic review of the published literature. Journal of General Internal Medicine 1999;14:633-642. *PMID:*
14. Staessen JA, Fagard R, Thijs L, Celis H, Arabidze GG, Birkenhager WH, et al. Randomised double-blind comparison of placebo and active treatment for older patients with isolated systolic hypertension. The Systolic Hypertension in Europe (Syst-Eur) Trial Investigators. Lancet 1997;350(9080):757-64. *PMID:*

Les indices d'efficacité pour les critères quantitatifs

Indices courants

L'effet d'un traitement sur un critère quantitatif (appelé aussi critère continu) s'exprime généralement en termes de différence des valeurs moyennes obtenues sous traitement et sans traitement. Cette différence peut être soit une différence absolue soit une différence relative. La précision de l'estimation de l'effet est donnée par l'intervalle de confiance de ces différences, qui visualise le plus petit effet que l'on ne peut pas raisonnablement exclure.

Dans un essai mesurant l'effet sur la pression artérielle d'un traitement anti-hypertenseur, les résultats suivants ont été obtenus :

Groupe	PAS à l'inclusion	PAS en fin d'essai	Changement
Traitement étudié	x_1=140 mmHg	y_1=120 mmHg	c_1=-20 mmHg
Placebo	x_0=141 mmHg	y_0=135 mmHg	c_0=-6 mmHg

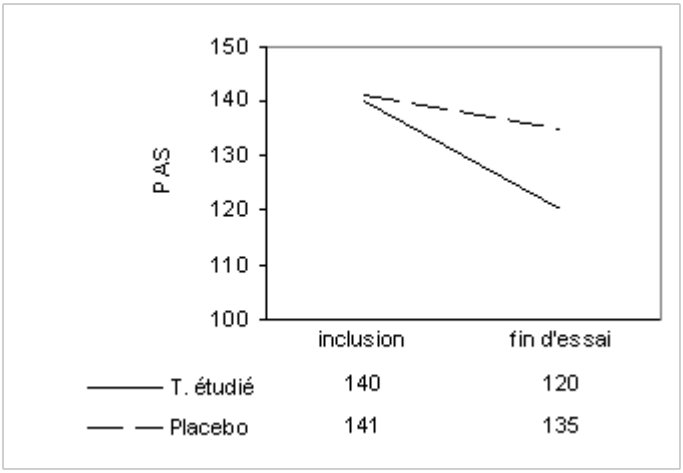


Figure 1 – Résultat d'un essai utilisant un critère de jugement quantitatif mesuré avant et après la période de traitement

Plusieurs possibilités s'offrent à nous pour calculer la taille de l'effet du traitement antihypertenseur :

- *La différence absolue des les PAS moyennes mesurées en fin d'essai : 120-135=-15 mmHg, qui permet de dire que le traitement entraîne une baisse absolue moyenne de 15 mmHg de la PAS*
- *La différence relative des PAS moyennes mesurées en fin d'essai :* $\frac{120-135}{135}=0,11=11\%$*, le traitement entraîne une baisse relative moyenne de 11% de la valeur de PAS*
- *L'effet du traitement peut aussi être estimé par la différence des changements de PAS entre le début et la fin de l'essai. Cette « double différence » permet de prendre en compte une différence initiale entre les 2 groupes (qui dans l'exemple est de 1 mmHg). En utilisant la différence absolue des changements, l'effet est de : 20 - 6 = 14 mmHg.*
- *Cette différence des changements peut aussi être exprimée sous forme relative. Se pose alors la question du dénominateur à utiliser qui peut être : la valeur sous placebo après traitement, la valeur initiale sous placebo, la moyenne des 2 valeurs initiales (cf. tableau ci-dessous).*
- *La différence de changement peut être rapportée au changement observé sous placebo pour exprimer l'effet propre du traitement en multiple de la baisse spontanée. Étant donné que cette baisse spontanée est un phénomène très dépendant des circonstances (variabilité des mesures, durée de la période d'observation, etc.) cet indice n'a que peu d'intérêt.*

Tableau 1 – Différentes façon de mesurer l'effet traitement avec un critère de jugement quantitatif

Différence absolue fin d'essai	y_1-y_0	120-135 = -15 mmHg
Différence relative fin essai	$(y_1-y_0)/y_0 \times 100\%$	-15/135 = 11%
Différence des changements	$(y_1-x_1)-(y_0-x_0)$ = c_1 - c_0	-14 mmHg
Différence relative des changements	$(c_1-c_0)/y_0$	(20-6)/135 = 10,4%
	$(c_1-c_0)/x_0$	(20-6)/141 = 10%
	$(c_1-c_0)/[(x_0+x_1)/2]$	14/140,5 = 10%

	Active	Placebo	Difference	p
Last 3 days of each treatment period				
Symptom score	5·32 (0·49)	5·69 (0·49)	−0·38 (0·29)	0·21
Reliever (puffs/day)	5·04 (0·82)	4·76 (0·64)	0·29 (0·34)	0·40
Morning PEF (L/min)	288·4 (15·2)	282·5 (13·9)	5·89 (5·68)	0·37
Evening PEF (L/min)	300·8 (15·1)	299·3 (13·2)	1·52 (6·64)	0·88
PEF variability	12·2% (1·8)	13·4% (2·5)	−1·0% (0·1)	0·32
Last 7 days of each treatment period				
Symptom score	5·67 (0·47)	5·62 (0·46)	0·05 (0·22)	0·73
Reliever (puffs/day)	5·06 (0·72)	4·65 (0·59)	0·41 (0·24)	0·06
Morning PEF (L/min)	285 (14·8)	284 (13·9)	1·18 (3·41)	0·82
Evening PEF (L/min)	300 (14·5)	300 (13·3)	−0·5 (3·71)	0·81
PEF variability	13·2% (1·9)	13·4% (2·2)	0% (0·1)	0·86

Values are means (SE) for the last 3 or 7 days of each treatment period. There were no significant period or carryover effects. Symptoms were scored 0–3 for each of six variables (total out of 18).

Table 2: **Diary data from last 3 days and last 7 days of each treatment period**

Figure 2 – Exemple de quantification des effets d'un traitement sur des variables quantitatives à l'aide de la différence des moyennes.

	Pravastatin (n = 249)	Atorvastatin (n = 253)	*P* Value Between Groups*
	Atheroma Volume, mm^3		
Baseline			
Mean (SD)	194.5 (114.8)	184.4 (115.7)	
Median (IQR)	168.6 (117.4 to 246.2)	161.9 (111.0 to 228.2)	.20
Follow-up			
Mean (SD)	199.6 (112.3)	183.9 (108.8)	
Median (IQR)	180.0 (125.5 to 255.3)	160.9 (107.4 to 240.3)	.05
Nominal change			
Mean (SD)	5.1 (31.4)	−0.4 (31.8)	
Median (95% CI)	4.4 (0.1 to 6.0)	−0.9 (−3.5 to 1.6)	.02†
P value compared with baseline‡	.01	.72	
Percentage change, %			
Mean (SD)	5.4 (20.1)	4.1 (29.6)	
Median (95% CI)	2.7 (0.2 to 4.7)	−0.4 (−2.4 to 1.5)	.02§
P value compared with baseline‡	.001	.98	

Figure 3 – Exemple d'un tableau regroupant les différentes façons d'exprimer les résultats obtenus au niveau d'une variable quantitative. Bien entendu il existe un problème de multiplicité des tests statistiques et ces résultats sont purement exploratoires ou explicatifs.

Effet standardisé

Une possibilité pour apprécier la taille d'un effet obtenue sur une variable continue est de l'exprimer en terme d'effet standardisé (« effect size »).

Définition

Par définition, l'effet standardisé est la différence des moyennes standardisées sur l'écart-type commun aux deux groupes. De ce fait, l'écart-type de cette nouvelle variable est égal à 1. L'estimation de l'effet standardisé à partir de valeurs observées est obtenue par :

$$\delta = \frac{\mu 1 - \mu 0}{\sigma}$$

où $\mu 1$ *et* $\mu 0$ *sont respectivement les moyennes observées des groupes traité et contrôle, et* s *une estimation de l'écart-type. Plusieurs choix sont possibles pour* s *: l'écart type du groupe contrôle, l'écart-type du groupe traité, ou un écart type commun* s *qui combine ces deux écarts types.*

L'effet standardisé est une valeur sans dimension. La différence des moyennes est divisée par une grandeur de même unité, l'écart-type. Le rapport « perd » l'unité de la grandeur initiale. En partie à cause de ce point, tous les effets standardisés, même issus de mesures différentes, sont comparables. Les variables initiales étant supposées distribuées selon une loi gaussienne, l'effet standardisé l'est aussi.

Application à la comparaison de deux traitements

L'effet standardisé peut être conçu comme étant la différence qui existe entre la distribution des valeurs du critère de jugement dans le groupe contrôle et celle des valeurs dans le groupe expérimental. Par hypothèse, ces deux distributions sont gaussiennes et de même écart-type. Graphiquement, l'effet standardisé est la distance qui sépare les deux modes de ces deux distributions. La Figure 4 illustre un effet standardisé de 1. Le patient moyen sous traitement (symbolisé par le trait vertical d'abscisse 1) a la même valeur du critère de jugement que le patient du groupe contrôle situé au 84eme percentile de sa distribution. C'est-à-dire qu'il occupe un rang où seulement 16% des sujets ont spontanément, avant traitement, une valeur du critère de jugement supérieure. Si, par exemple, le critère de jugement est le périmètre de marche chez des patients artéritiques, le patient moyen aura une amélioration de ces performances qui l'amènera à un niveau où seulement 16% des sujets ont spontanément un périmètre de marche supérieur.

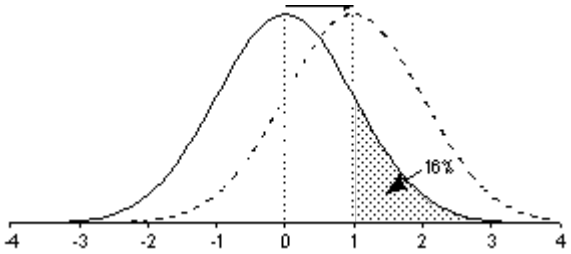


Figure 4 – Interprétation d'un effet standardisé de 1.

Cet indice ne donne cependant pas une vision directe de la pertinence clinique de l'effet. En effet, si le score initial est peu discriminant (faible variabilité), un effet traitement, ne modifiant que peu cliniquement l'état des patients, peut se traduire par un effet standardisé important.

Exemple

Par exemple, avec les données de l'essai MAST concernant le score de handicap, l'écart type des observations est de 5,4 dans le groupe streptokinase (n=81) et de 6,8 dans le groupe placebo (n=94). L'écart type global est de (80×5,4+93×6,8)/(80+93)=6,15. L'effect size correspondant à la différence de 1,8 points entre les deux groupes est : 1,8 / 6,15 = 0,29. Cet effect size correspond à une avancée du patient moyen au niveau du 61ème percentile. Cela veut dire que sous traitement 50% des patients ont un niveau de performance supérieur à celui que seulement 39% des patients avaient spontanément.

Appréciation de la taille de l'effet

La représentation graphique du décalage des distributions d'une variable continue entre le groupe contrôle et traité visualise la taille de l'effet traitement. La Figure 5 représente ce décalage pour différentes tailles d'effet exprimées en terme d'effect size et d'odds ratio. Les effets traitement de tailles courantes (effects size de l'ordre de 0,5 ou odds-ratio proche de 0,60) correspondent à des décalages de faible ampleur. Pour que sous traitement, la moitié des patients soit « mieux » que la totalité des patients sans traitement, l'effet traitement doit correspondre à un effect size de 2 ou à un odds ratio de 0,14. Des effets traitements aussi important sont rarement rencontrés.

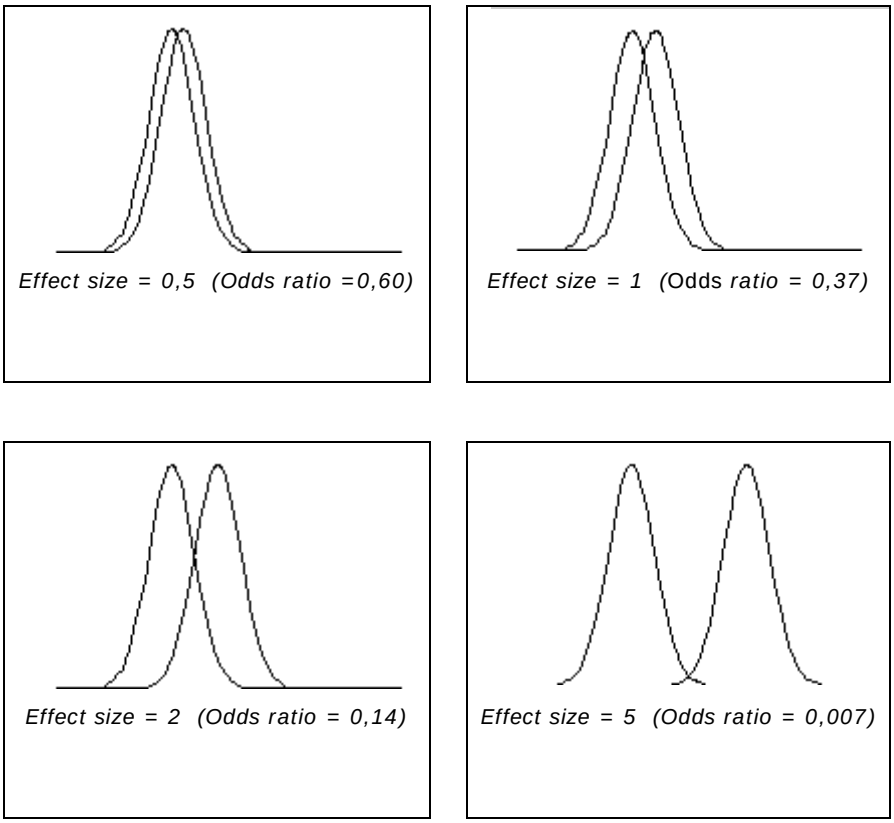


Figure 5 – Représentation graphique en terme de distance entre les distributions d'effets de tailles différentes.

PERTINENCE DES CRITERES DE JUGEMENT

Les critères de jugement

Les critères de jugement sont les critères sur lesquels est jugée l'efficacité des traitements.

Les critères de jugement (« *endpoints* » ou « *outcomes* ») sont des variables dont la valeur numérique est susceptible de se modifier sous l'effet du traitement. Ils permettent une mise en évidence « quantitative » de l'effet du traitement étudié par une différence de valeur entre les deux groupes de l'essai. Ces critères peuvent être de nature très variée : durée d'une maladie, fréquence de survenue d'un événement, taux de mortalité, etc. Par exemple, un critère de jugement utilisable pour mettre en évidence le bénéfice apporté par l'enalapril dans l'insuffisance cardiaque est la survenue d'un décès durant une période de suivi d'un an. L'effet du traitement sera ainsi mesuré en comparant le taux de mortalité à un an du groupe enalapril à celui du groupe placebo, qui sont respectivement 16% et 23%, témoignant ainsi numériquement d'un effet du traitement (cette différence est statistiquement significative).
Dans une situation donnée, de nombreuses variables peuvent être utilisées comme critère de jugement. Cependant toutes ces variables n'ont pas la même pertinence clinique et conduisent à des résultats qui n'ont pas tous la même valeur médicale vis-à-vis de la validation de l'efficacité des traitements.

Les différents types de critères

Événements cliniques

L'utilisation des événements cliniques comme critère de jugement produit une variable binaire : présence/absence de l'événement. Cette variable se traduit au niveau d'un groupe par le nombre de patients présentant le critère, ce qui permet d'estimer une fréquence de survenue de l'événement aussi appelée risque. Par exemple, la survenue d'un infarctus du myocarde est un critère de jugement fréquemment utilisé dans l'évaluation d'un traitement de l'hypertension ou de l'hypercholestérolémie. Dans le traitement chirurgical des éventrations sur cicatrice, la récidive est l'événement clinique employé comme critère clinique.
La façon dont sont définis les événements cliniques est importante. Suivant la définition utilisée, le même vocable ne recouvrira pas exactement les mêmes entités nosologiques. La pertinence clinique du critère pourra en être influencée.
Par exemple, un infarctus du myocarde défini uniquement par une élévation enzymatique n'a pas la même signification que s'il est défini par des signes cliniques et électriques. Dans le premier cas, des élévations isolées des enzymes cardiaques pourront être étiquetées « infarctus » sans en avoir la même valeur pronostique péjorative.
La performance diagnostique (sensibilité et spécificité) de la procédure utilisée pour détecter la survenue d'un événement clinique influence la mesure de l'effet du traitement. Cette détection doit être la plus spécifique possible pour éviter le biais de mesure et la plus sensible possible pour apporter de la puissance à la comparaison (en garantissant un risque de base dans le groupe contrôle élevé) (cf. infra). Schématiquement,

avec un test peu spécifique l'effet d'un traitement est dilué par la présence des faux positifs. Avec un test peu sensible, les nombreux faux négatifs empêchent la matérialisation de l'effet traitement.
Ce type de critère peut être analysé de nombreuses manières (décrites dans le chapitre sur les indices d'efficacités). D'une manière générale il s'agit de comparer des pourcentages entre les groupes (Figure 1).

TABLE 2. CARDIOVASCULAR EVENTS ACCORDING TO TREATMENT GROUP.

EVENT*	PLACEBO (N=4502)	PRAVASTATIN (N=4512)	REDUCTION IN RISK (95% CI)†	P VALUE‡
	no. (%)		%	
Death due to CHD	373 (8.3)	287 (6.4)	24 (12–35)	<0.001
Death due to CVD	433 (9.6)	331 (7.3)	25 (13–35)	<0.001
Death from any cause	633 (14.1)	498 (11.0)	22 (13–31)	<0.001
Death due to CHD or nonfatal MI	715 (15.9)	557 (12.3)	24 (15–32)	<0.001
Any MI	463 (10.3)	336 (7.4)	29 (18–38)	<0.001
CABG	520 (11.6)	415 (9.2)	22 (11–31)	<0.001
PTCA	253 (5.6)	210 (4.7)	19 (3–33)	0.024
CABG or PTCA	708 (15.7)	585 (13.0)	20 (10–28)	<0.001
Hospitalization for unstable angina	1106 (24.6)	1005 (22.3)	12 (4–19)	0.005
Any stroke	204 (4.5)	169 (3.7)	19 (0–34)	0.048

*CHD denotes coronary heart disease, CVD cardiovascular disease, MI myocardial infarction, CABG coronary-artery bypass surgery, and PTCA percutaneous transluminal coronary angioplasty.

†Relative reductions in risk are for the pravastatin group as compared with the placebo group and have been estimated on the basis of the hazard ratio in a Cox regression analysis. CI denotes confidence interval.

‡P values were derived with the stratified log-rank test.

Figure 1 – Exemple de présentation de résultat avec des événements cliniques. L'analyse s'effectue en comparant les fréquences de survenue entre les groupes (ce qui revient à faire une comparaison de proportion ou de pourcentage) et en quantifiant l'effet du traitement à l'aide d'une réduction relative du risque.

Critères composites

Définition

Un critère composite (« composite endpoint *») est un critère qui prend en considération simultanément plusieurs événements cliniques (*

Tableau 5). Le critère « événements coronariens mortels ou non mortels » est un critère composite formé du regroupement des événements coronariens non mortels (infarctus principalement) et des décès d'origine coronarienne. Un patient présente le critère composite à partir du moment où il est victime de l'un des deux événements.

La survenue successive de plusieurs des composantes du critère composite ne donne lieu qu'à une seule occurrence du critère composite. Par exemple, un patient présentant successivement un infarctus non mortel, puis décédant d'une récidive ne sera comptée qu'une seule fois pour le critère composite « événements coronariens mortels ou non mortels ». De ce fait, le nombre d'occurrences du critère composite n'est en général pas la somme des nombres d'occurrences de chacune de ses composantes. Par exemple dans le Tableau 4, le nombre d'événements coronariens mortels et non-mortels est inférieur à la somme des décès coronariens et des infarctus non mortels car 4 patients sont décédés après avoir présentés antérieurement un infarctus non-mortel.

Tableau 4 – Exemple où le nombre d'occurrences du critère composite n'est pas la somme des nombres d'occurrences des événements qui le compose.

Critère	***n***
Infarctus non mortels	*20*
Décès coronariens	*22*
Critère composite : événements mortels ou non mortels	*38*

L'utilisation des critères composites présente plusieurs intérêts : sensibiliser la recherche d'un effet, mesurer directement le rapport bénéfice/risque ou regrouper des équivalents du même phénomène clinique. Cependant, ils posent assez fréquemment des problèmes d'interprétation, particulièrement quand les composantes sont de gravités cliniques différentes.

Tableau 5 – Exemple de critères composites utilisés dans différents domaines.

Critère composite	***Domaine d'utilisation***
Événements coronariens majeurs (« MACE Major Coronary Events ») : décès, infarctus, nécessité de revascularisation	*Essai d'angioplastie ou de stent coronarien*
Survie sans progression : durée jusqu'au critère composite décès ou progression de la maladie nécessitant un traitement (chirurgie)	*Cancérologie*

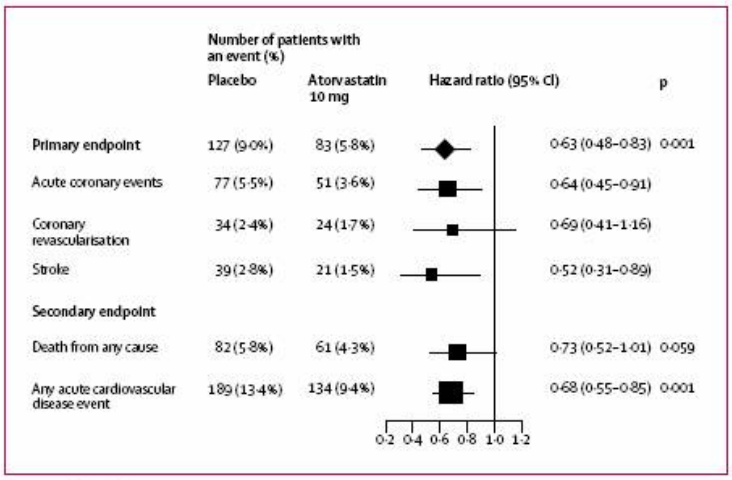


Figure 3: Effect of treatment on primary and secondary endpoints
Total number of acute coronary events, coronary revascularisations, and strokes separately do not equal the total number of primary events shown above, because only the first of these events is included in the primary endpoint. Thus, an individual who has had a stroke and a revascularisation will be counted only once in the primary endpoint but will appear in both separate totals for revascularisation and stroke. Symbol size is proportional to amount of statistical information.

Figure 2 – Exemple présentation graphique des résultats obtenus avec un critère composite. En dessous du résultat du critère composite (primary endpoint) apparaissent ses composantes (acute coronary events, coronary revascularization, stroke).

TABLE 3. INCIDENCE OF THE PRIMARY OUTCOME AND OF DEATHS FROM ANY CAUSE.

OUTCOME	RAMIPRIL GROUP (N=4645)	PLACEBO GROUP (N=4652)	RELATIVE RISK (95% CI)*	Z STATISTIC	P VALUE†
	no. (%)				
Myocardial infarction, stroke, or death from cardiovascular causes‡	651 (14.0)	826 (17.8)	0.78 (0.70–0.86)	−4.87	<0.001
Death from cardiovascular causes§	282 (6.1)	377 (8.1)	0.74 (0.64–0.87)	−3.78	<0.001
Myocardial infarction§	459 (9.9)	570 (12.3)	0.80 (0.70–0.90)	−3.63	<0.001
Stroke§	156 (3.4)	226 (4.9)	0.68 (0.56–0.84)	−3.69	<0.001
Death from noncardiovascular causes	200 (4.3)	192 (4.1)	1.03 (0.85–1.26)	0.33	0.74
Death from any cause	482 (10.4)	569 (12.2)	0.84 (0.75–0.95)	−2.79	0.005

*CI denotes confidence interval.

†P values were calculated with use of the log-rank test.

‡In the substudy, 34 of 244 patients (13.9 percent) assigned to take a low dose of ramipril (2.5 mg per day) reached the composite end point, as compared with 31 of 244 assigned to take 10 mg of ramipril per day (12.7 percent) and 41 of 244 assigned to placebo (16.8 percent). The inclusion of the data from the low-dose group did not change the overall results (relative risk of the primary outcome, 0.78; 95 percent confidence interval, 0.70 to 0.86).

§All patients with this outcome are included.

Figure 3 – Exemple de présentation tabulaire des résultats obtenus avec un critère de jugement composite. Comme avec les événements cliniques « purs », l'analyse s'effectue en comparant les pourcentages correspondant à la fréquence de survenu des événements et en quantifiant l'effet à l'aide du risque relatif (ou d'un autre indice).

Intérêts des critères composites

Regrouper des équivalents du même phénomène clinique

Dans certaines situations, différents types d'événements cliniques peuvent être dus au même processus morbide ; phénomène que cherche à enrayer le traitement étudié. Dans ce cas, la survenue de l'un de ces événements a la valeur d'échec du traitement. Vis-à-vis de l'action du traitement étudié, ces différents événements sont donc équivalents. Leur regroupement permet de prendre en compte la diversité des manifestations d'un processus pathologique.

Par exemple, dans un essai de prévention du risque coronarien par un hypocholestérolémiant, la survenue d'un décès coronarien ou d'un infarctus non mortel a la même valeur : celle de l'échec de la prévention. Le fait qu'un événement s'avère mortel ou non dépend de nombreux facteurs indépendants du traitement préventif : délai

de prise en charge, nature des traitements de la phase aiguë reçus par le patient, localisation et gravité de l'infarctus, etc. Le critère composite formé par le regroupement de ces deux événements a donc un sens et évalue l'impact de ce traitement sur sa cible directe : le risque coronarien.

Mesurer de la balance bénéfice-risque

Le regroupement au sein d'un même critère des événements influencés favorablement par le traitement et de ces effets indésirables donne une mesure de la balance bénéfice risque.

Par exemple, les décès et les accidents vasculaires cérébraux invalidant ont été regroupés dans les essais de fibrinolyse à la phase aiguë de l'infarctus, afin de pondérer le gain en survie par les complications hémorragiques graves induites par la fibrinolyse. Ce critère composite appréhende la survie sans invalidité.

Cette approche donne cependant le même poids aux événements prévenus et aux événements induits : un AVC équivaut à un décès évité. Nous discuterons dans le chapitre consacré à l'appréciation de la balance bénéfice-risque des problèmes posés par cette pondération implicite et de l'influence de la gravité des événements regroupés.

Sensibiliser la recherche d'un effet

La fréquence de survenue d'un critère composite est supérieure à la fréquence de chacune de ses composantes. De ce fait, les critères composites augmentent la puissance de la recherche de l'effet traitement. Le nombre de sujets nécessaires est moindre avec un critère composite qu'avec une seule de ses composantes.

Cette utilisation pose cependant plusieurs problèmes. En cas de regroupement d'événement de pertinence clinique variable, le critère global va être principalement le reflet des événements de moindre importance si ces derniers sont prépondérants en fréquence. Ainsi, lorsque le critère clinique pertinent n'est que l'une des composantes d'un critère combiné , la mise en évidence d'un effet sur le critère composite ne permet souvent pas d'inférer un effet sur le critère clinique pertinent. Dans un essai d'antiagrégants lors de l'angioplastie, l'observation d'une réduction significative de la fréquence du critère « décès+revascularisation+stent » ne démontre pas l'aptitude du traitement à réduire la mortalité. Car la mortalité ne représente qu'une petite part du critère composite. Bien que plus facile à mettre en évidence, un effet sur un critère composite n'équivaut pas à une démonstration de l'efficacité sur le critère clinique le plus pertinent.

Un effet délétère sur la mortalité peut être tamponné dans un critère clinique par un effet favorable sur un critère bien plus fréquent. L'utilisation du critère composite fait passer à côté de cet effet indésirable et de l'absence de bénéfice clinique du traitement.

Exemple

Dans les essais de prévention secondaire du risque cardiovasculaire par les statines, comme par exemple LIPID (7), le risque de décès coronariens est de 8,3% tandis que la fréquence des événements coronariens mortels et non mortels est de 15,9%. La mise en évidence d'un effet sur le critère composite demanderait bien moins de sujets qu'avec le critère de mortalité.

Analyse statistique

Une analyse stratifiée sur les composantes d'un critère composite augmente la puissance du test statistique, en utilisant, par exemple, un test du logrank stratifié avec les données de survie (7).

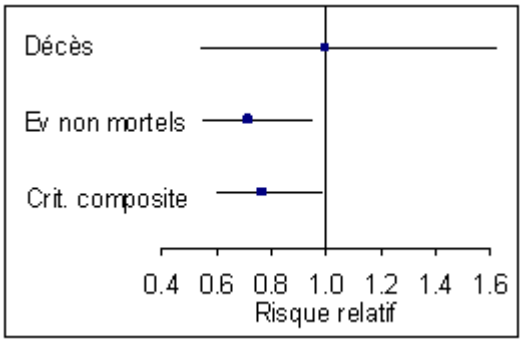
Difficultés d'interprétation des critères composites

Les critères composites posent assez fréquemment des problèmes d'interprétation, principalement quand l'effet du traitement n'est pas uniforme sur toutes les composantes.

Dans les exemples suivants, un critère composite est utilisé pour appréhender l'effet d'un traitement sur la morbi-mortalité, ce critère regroupant décès et événements non mortels. L'utilisation de ce critère combiné peut conduire à des conclusions paradoxales dans certaines situations.

Situation n°1

Dans cet exemple, le résultat obtenu au niveau du critère composite pourrait s'interpréter comme une réduction signification de morbi-mortalité. Pourtant la mortalité n'est pas modifiée, mais comme son poids dans le critère composite est faible (17% du critère composite) elle influence peu le total. La réduction obtenue au niveau du critère principal est uniquement due à un effet sur une des composantes. La conclusion à une réduction de morbi-mortalité est excessive, car aucun effet sur la mortalité n'a été enregistrée.

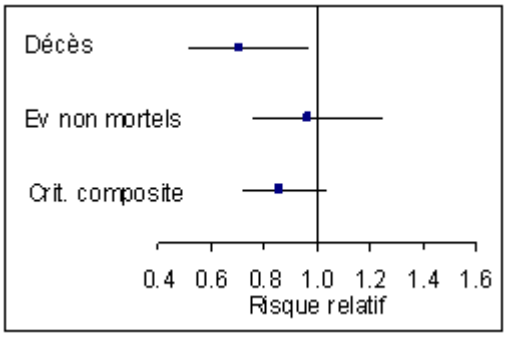


	T+	*T-*	*RR (IC 95%)*	*p*
Décès	*20/500 (4%)*	*20/500 (4%)*	*1 (0.54;1.84)*	*1.00*
Événements non mortels	*72/500 (14.4%)*	*100/500 (20%)*	*0.72 (0.55;0.95)*	*0.02*
Critère composite	*92/500 (18.4%)*	*120/500 (24%)*	*0.77 (0.6;0.98)*	*0.03*

Figure 4 – Situation problématique d'utilisation d'un critère composite

Situation n°2

Dans cet exemple, le traitement entraîne une réduction de la mortalité mais pas de la fréquence des événements non mortels. Par un phénomène de dilution, le critère composite ne permet pas de conclure à l'efficacité du traitement. L'effet sur la mortalité est dilué par l'autre composante car elle ne représente que 44% du critère composite. Le résultat obtenu au niveau du critère composite tendrait à faire conclure qu'aucune efficacité n'a été mise en évidence sur la morbi-mortalité, alors qu'il existe une réduction de la composante la plus pertinente du critère composite, la mortalité.

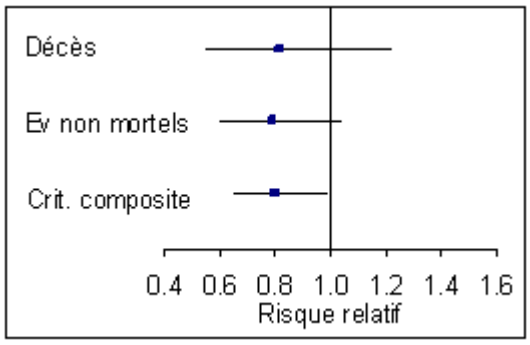


	T+	*T-*	*RR (IC 95%)*	*p*
Décès	*57/500 (11.4%)*	*80/500 (16%)*	*0.71 (0.52;0.97)*	*0.03*
Événements non mortels	*97/500 (19.4%)*	*100/500 (20%)*	*0.97 (0.76;1.25)*	*0.81*
Critère composite	*154/500 (30.8%)*	*180/500 (36%)*	*0.86 (0.72;1.03)*	*0.09*

Figure 5 – Situation problématique d'utilisation d'un critère composite

Situation n°3

Cette situation est la situation idéale. Les deux composantes sont modifiées de la même façon par le traitement, mais aucun des effets n'est significatif au niveau des composantes. Par contre, leur regroupement met en évidence un effet statistiquement significatif. La conclusion que le traitement réduit la mortalité et la fréquence des événements non mortels est assez plausible dans ce cas. C'est par exemple le cas dans les résultats rapportés par la Figure 2 et Figure 3.



	T+	*T-*	*RR (IC 95%)*	*P*
Décès	*41/500 (8.2%)*	*50/500 (10%)*	*0.82 (0.55;1.22)*	*0.32*
Événements non mortels	*79/500 (15.8%)*	*100/500 (20%)*	*0.79 (0.6;1.03)*	*0.08*
Critère composite	*120/500 (24%)*	*150/500 (30%)*	*0.8 (0.65;0.98)*	*0.03*

Figure 6 – Situation où un critère de jugement composite ne pose pas de problème d'interprétation

Exemple

L'essai Benestent 2 a comparé la pose d'un stent à l'angioplastie simple chez des patients ayant un angor stable ou stabilisé (8). Le critère de jugement principal était composite : décès, infarctus du myocarde, chirurgie de pontage coronarien ou geste de revascularisation transcutané réitératif.

	Angioplasty n=410	***Stent n=413***	***Relative risk (95%)***
Death	1	2	
Q-wave myocardial infarction	7	5	
Non Q-wave myocardial infarction	6	10	
Coronary artery bypass surgery	6	6	
Repeat PTCA	56	33	
Composite endpoint	*79*	*53*	*0,67 (0,48;0,92)*

La différence observable au niveau du critère composite a été essentiellement obtenue sur la répétition des gestes de revascularisation transcutanée. Il serait alors abusif de déclarer qu'il existe une différence au niveau des événements coronariens décès, infarctus ou revascularisation.

Les échelles et scores

Les échelles (« *scale* ») et les scores permettent de mesurer l'intensité d'un phénomène clinique comme une gêne fonctionnelle, l'intensité d'un symptôme, l'extension d'une maladie, son stade évolutif, etc. Les échelles et

les scores (la distinction terminologique entre échelles et scores n'est pas universelle. Les scores sont très souvent appelés échelles) sont très prisés en médecine car ils permettent de quantifier numériquement des phénomènes qui ne se caractérisent pas par une dimension physique.

Les échelles

Les échelles sont obtenues en découpant en différents stades (que l'on appelle aussi classe, grade, etc.) le continuum de gravité de la maladie étudiée. Chaque classe est caractérisée par un chiffre ou un adverbe matérialisant la relation d'ordre existant entre ces classes. Par exemple, une échelle de mesure de l'intensité d'une douleur peut-être : 0 : absente, 1 : modérée, 2 : importante, 3 : insoutenable.
Le résultat d'une échelle n'est pas assimilable à une variable continue lorsque le nombre de valeurs possibles est faible. Le recours aux outils statistiques spécifiques aux variables continues (moyenne, test de comparaison de moyennes) pose un certain nombre de problèmes. En réalité, il s'agit d'une variable qualitative ordinale dont l'analyse repose sur la description de la répartition des valeurs et sur des comparaisons à l'aide du test du chi-2. Cette remarque peut aussi concerner les scores.

Tableau 6 – Exemples d'échelles

Intensité d'un symptôme	• Stade NYHA de dyspnée • Souffle cardiaque gradé de 0 à 6
Gravité	• Stade de gravité de l'asthme (bénin, moyen, sévère, aggravé) • Stade d'encéphalopathie hépatique (I,II,III,IV)

Échelle de Rankin : handicap après AVC
0 = Absence de symptômes
1 = Symptômes mineurs sans retentissement sur la vie quotidienne
2 = Symptôme ou handicap mineur qui conduit à certaines restrictions dans le mode de vie, mais qui n'interfère pas avec la capacité du patient à se prendre en charge
3 = Handicap modéré qui restreint significativement le mode de vie et/ou empêche une existence totalement indépendante
4 = Handicap modérément sévère qui empêche clairement une existence indépendante bien que nécessitant pas une attention constante
5 = Handicap sévère entraînant une dépendance totale et nécessitant une attention jour et nuit

Les scores

Les scores permettent de mesurer des phénomènes multidimensionnels. Le score se calcule en cotant un certain nombre d'items analysant les différentes composantes du processus étudié puis en faisant la somme des note attribuées afin d'obtenir un score global. le but du score est de refléter en un seul nombre la totalité des dimensions envisagées.
Par exemple le score d'Apgar qui évalue la gravité des troubles respiratoires et neurologiques à la naissance d'après certains signes cliniques. Les nombres de points correspondants à chaque critère sont additionnés en un score global. Plus le score est bas, plus l'état du nourrisson est préoccupant.

Tableau 7 – Calcul du score d'Apgar

Critères	**Nombre de points**		
	0	**1**	**2**
Couleur	Cyanosée ou pale	Corps rose, extrémités bleues	Complètement rose
Rythme cardiaque	Absent	<100	>100
Respiration	Absente	Irrégulière, lente	Bonne,

Réponse réflexe au cathéter nasal	Sans	Grimace	cri vigoureux Éternuement, toux
Tonus musculaire	Hypotonique	Légère flexion des extrémités	Actif et tonique

Le plus souvent ces scores ont été établis à partir d'études pronostiques. Les items du score sont en fait les facteurs retrouvés associés avec le pronostique et le nombre de points de chaque item est une pondération proportionnelle à son importance dans le pronostic. Ce sont donc en fait des outils simplifiés de prédiction du risque d'évolution favorable (décès, survenue d'une complication, etc.)

Score de Barthel : évaluation du handicap après AVC

	Avec aide	Indépendant
1.Alimentation (si les aliments doivent être coupés = aide)	1	2
2.Déplacement de la chaise roulante au lit et retour	1-2	3
3.Toilette personnelle	0	1
4.Aller et revenir des toilettes	1	2
5.Se baigner seul	0	1
6.Marche sur un sol plat	2	3
7.Monter ou descendre des escaliers	1	2
8.Habillage (comprenant laçage des chaussures, boutonnage)	1	2
9.Continence anale	1	2
10.Continence vésicale	1	2
Total	_____	_____

Ce score prend des valeurs entre 0 et 20. Chaque item se noie dans les autres et une même valeur de score peut être obtenue avec des altérations fonctionnelles différentes. L'aspect multidimensionnel du handicap disparaît. À la fin, un changement de score de 1 n'a plus de signification clinique précise.
L'analyse statistique des scores repose souvent sur la comparaison des scores moyens de chaque groupe (moyenne des scores de chaque patient). En général, les distributions ne sont pas symétriques et il est plus adapté de comparer les médianes.

Problème d'interprétation des scores et des échelles

À coté des questions de qualité métrologique des échelles et des scores (reproductibilité, exactitude, homogénéité) que nous n'aborderons pas ici, les scores et les échelles posent différents problèmes d'interprétation.
La comparaison s'effectue en calculant le score moyen dans chaque groupe (cf. Figure 8). La moyenne est susceptible de prendre des valeurs que ne prennent pas les scores ou les échelles elles mêmes. Par exemple, des valeurs fractionnaires comme 5,68 ou 4,2 alors le score ne prend que des valeurs entières entre 1 et 10. Le patient moyen est donc affublé d'un score qui n'existe pas. Ainsi que signifie une différence de 0,9 points de l'échelle de handicap ? L'utilisation de la médiane pour décrire la position centrale de la population sur l'échelle ou sur le score ne conduit pas à ce problème.
Un autre point est la proportionnalité de la métrique. Est-ce qu'un changement de 1 point représente la même modification dans le phénomène étudié quel que soit le niveau de départ. En d'autres termes, le score mesure-t-il, par le même changement de valeur, un même effet chez des sujets de valeurs initiales différentes.

	Pallidotomy group (n=18)			Control group (n=16)			p*
	Baseline	6 months	Change	Baseline	6 months	Change	
Primary outcome							
UPDRS 3	47 (24–81)	32·5 (16–66)	15 (−13 to 27)	52·5 (23–82)	56·5 (19–91)	−2 (−15 to 9)	0·0004
Secondary outcome							
Pain VAS (mm)	27 (2–100)	14 (0–69)	3·5 (−20 to 77)	15·5 (0–87)	22 (0–84)	−0·5 (−23 to 45)	0·13
Barthell index	10·5 (4–20)	18 (6–20)	2·5 (−2 to 11)	11·5 (3–19)	8 (4–19)	−0·5 (−7 to 3)	0·004
UPDRS 2	30 (11–41)	21 (8–38)	7 (−8 to 20)	32 (14–45)	35 (15–46)	−2 (−11 to 6)	0·002
Schwab and England scale	35 (20–80)	70 (20–90)	15 (−10 to 40)	35 (10–80)	30 (10–80)	−5 (−30 to 10)	0·0009

A positive change score signifies improvement.
*Mann-Whitney *U* test.
Table 3: **Primary and secondary outcomes—median (range) scores of clinical rating scales for defined off phase assessment**

Figure 8 – Exemple de résultats obtenus avec des scores (UPDRS3, Barthell, UPDRS 2) ou une échelle visuelle analogique (Pain VAS, Schwab and England scale).

Exemples

Exemple 1 - L'essai MAST-E comparait la streptokinase au placebo dans le traitement des accidents vasculaires cérébraux. Un des critères de jugement était la mesure du niveau de handicap à l'aide du score de Barthel. Six mois après l'AVC, la moyenne (± erreur standard) de ce score était 13,0±0,7 dans le groupe placebo et de 14,8±0,6 dans le groupe streptokinase. La différence est à la limite de la signification statistique : p=0,06. Étant donnée la construction du score de Barthel, la signification clinique d'une différence de 1,8 points n'est pas simple à appréhender et il n'est pas aisé de dire si cet effet représente une véritable amélioration de l'état des patients.

Exemple 2 - Retour sur l'exemple des inhibiteurs de la phosphodiestérase
L'exemple des agents inotropes inhibiteurs de la phosphodiestérase présenté précédement procure aussi l'occasion de discuter des problèmes d'interprétation de la pertinence clinique d'un effet observé sur une échelle de score et de sa confrontation à un effet sur un critère clinique.
La question qui se pose est de savoir si l'amélioration de la qualité de vie ou de la symptomatologie est suffisamment importante pour éventuellement rendre acceptable un surcroît de mortalité. Avant d'envisager le problème éthique d'une réduction des chances de survie sous prétexte d'une amélioration fonctionnelle, il convient de pouvoir confronter la pertinence clinique des tailles des effets obtenus respectivement sur la mortalité et sur les signes fonctionnels.
Ce n'est pas parce qu'il y a détection d'un effet statistiquement significatif sur la qualité de vie (Avec les critères de jugement continus, des effets de petite taille ne peuvent s'avérer statistiquement significatifs, en particulier si la variabilité est faible. Il peut donc y avoir une dissociation forte entre signification statistique et pertinence clinique), que celui-ci est notable et intéressant pour les patients, et suffisamment important pour constituer une amélioration substantielle pouvant éventuellement justifier l'acceptation d'une surmortalité. Par exemple, dans l'essai vesnarinone (9), les effets étaient recherchés sur le changement médian du score de qualité de vie entre l'entrée dans l'essai et le moment de la mesure. Numériquement l'effet était faible. Initialement le score médian étaient de 56 points. A 8 semaines, le score de qualité de vie (« *Minnesota Living with Heart Failure Questionnaire* ») s'améliorait de 7 points dans le groupe vesnarinone 60mg contre une amélioration médiane de seulement 5 points dans le groupe placebo. Du fait de l'effectif important cette différence était hautement significative (p<0,001) mais il convient de s'interroger sur la pertinence clinique d'un tel effet qui ne représente qu'un surcroît d'amélioration de 2 points sur un échelle allant de 0 à 105.
En d'autres termes, la surmortalité observée représente-t-elle un coût acceptable en regard du bénéfice obtenu sur les symptômes. Il est crucial dans cette situation de pouvoir traduire en terme clinique (évaluer la pertinence clinique) la différence de score de qualité de vie en des termes qui la rende comparable à la surmortalité. Une binarisation du score en utilisant un seuil exigeant est l'un de ces moyens.

Les durées

Une durée peut être utilisée comme critère de jugement ([1][1] Les durées sont souvent définies par la survenue d'un événement clinique. Il existe donc une parenté forte entre ces deux types de critères de jugement. Les durées permettent d'appréhender une part de la dynamique de la survenue des événements cliniques). C'est par exemple la durée de survie ou de délai de survenue d'un événement. Le but du traitement étant d'accroître

ce type de durée. Dans d'autres situations, il s'agit de la durée des symptômes ou de la durée de la maladie (cf. Figure 7). Dans ce cas, le but du traitement est de réduire cette durée.

	Intention-to-treat population			Influenza infected		
	Placebo (n=235)	Oseltamivir 75 mg (n=241)	Oseltamivir 150 mg (n=243)	Placebo (n=161)	Oseltamivir 75 mg (n=158)	Oseltamivir 150 mg bid (n=156)
All patients						
Median (95% CI) duration of illness (h)	116·1 (99·8–129·5)	97·6 (79·1–115·3)	89·4 (79·1–103·7)	116·5 (101·5–137·8)	87·4 (73·3–104·7)	81·8 (68·2–100·0)
p	..	0·05*	0·03*	..	0·02*	0·01*
Median (range) total symptom score AUC	916·6 (0–5996·0)	851·3 (0–6069·0)	708·5 (0–5811·0)	943·0 (0–5408·0)	773·3 (0–3793·0)	708·5 (0–4797·0)
p	..	0·1†	0·03†	..	0·01†	0·003†
Median duration for return to normal sleep quality (h)	249 (204–367)	199 (158–252)	204 (156–247)	204 (179–275)	170 (151–223)	159 (147–221)
p	..	0·03‡	0·02‡	..	0·02‡	0·01‡
Median (range) scale score AUC						
Health	735 (105–1564)	804 (108–3530)	796 (48–3309)	746 (141–1411)	809 (108–3530)	806 (48–3309)
p	..	0·002†	0·007†	..	0·003†	0·0008†
Activity	690 (69–1595)	769 (0–3163)	762 (48–3237)	703 (69–1585)	787 (0–3163)	793 (108–3237)
p	..	0·008†	0·002†	..	0·02†	0·002†
Median (95% CI) time to alleviation of cough (h)	62 (46–72)	34 (23–48)	32 (22–42)	72 (62–99)	48 (30–61)	35 (22–44)
p	..	0·02	0·006	..	0·007	0·0001
Median (95% CI) time to become afebrile (≤37·2°C)	59 (48–68)	45 (38–52)	41 (33–46)	67 (56–74)	39 (31–45)	37 (31–44)
p	..	0·2	0·003	..	0·002	<0·0001
Median overall paracetamol (g)	3·0	3·0	2·0	3·0	2·5	2·0
Patients treated within 24 h						
Total	127 (54%)	130 (54%)	118 (49%)	83 (52%)	77 (49%)	66 (42%)
Median (95% CI) duration of illness (h)	114·6 (94·3–135·2)	76·8 (67·5–104·4)	80·0 (63·5–100·0)	117 (103·0–143·8)	74·5 (68·2–98·0)	70·7 (54·0–89·4)
p	..	0·03*	0·05*	..	0·02*	0·01*

AUC=area under curve. *Weighted Mantel-Haenszel test, adjusted for multiple comparisons. †Extended Wilcoxon's rank sum test. ‡Weighted Mantel-Haenszel test.

Table 2: **Duration and severity of illness in intention-to-treat population and influenza-infected patients**

Figure 7 – Exemple de résultats obtenus sur des durées de maladie. Ce type de critère s'analyse en comparant les médianes plutôt que les moyennes en raison du caractère asymétrique des distributions observées.

En termes d'interprétation, les durées sont plus informatives que la simple mention de la survenue d'un événement au cours d'une période d'observation. L'analyse statistique des durées pose cependant des problèmes particuliers (asymétrie des distributions, durée de suivi variable, censure) qui sont détaillés dans le chapitre : Les courbes de survie*. Il apparaitra que durée et fréquence de survenue des événements sont intimement liées (lecture horizontale ou verticale des courbes de survie). L'analyse des durées étant difficile, elle est avantageusement remplacée par l'analyse des courbes de survie (ou de taux cumulée des évènements).*

De plus, en pratique, il n'est pas toujours simple de déterminer le moment de survenue d'un événement.

Par exemple, dans les maladies bénignes, la détermination de la durée de la maladie nécessite de connaître le moment de survenue de la guérison avec suffisamment de précision et d'exactitude. Pour cela, il serait nécessaire d'examiner régulièrement les patients afin de ne pas rater le moment où survient la guérison. En effet, le recueil rétrospectif de cette information, lors d'un interrogatoire tardif, est insuffisamment fiable. Un manque d'exactitude à ce niveau augmente la variabilité des mesures et réduit la puissance statistique. Un manque de précision réduit la puissance métrologique. Une précision de l'ordre d'une journée pour une maladie qui guérit en une semaine ne permet pas de discriminer des différences de durée de l'ordre de quelques heures.

En pratique, il s'avère plus simple et donc plus fiable de mesurer la proportion de guérison obtenue avec un recul donné. Tous les patients sont revus en consultation à cette date pour déterminer s'ils sont guéris ou non. En pratique, la fréquence de guérison évaluée à un moment donné s'avère bien plus facile à mesurer de façon fiable que la durée d'évolution de la maladie.

En fait, les durées ne sont faciles à déterminer qu'avec les événements qui laissent une trace fiable et facilement accessible de manière systématique, comme les hospitalisations ou le décès. Lorsqu'un tel système de recueil de l'information n'existe pas, l'information doit être recueillie spécifiquement pour l'essai, ce qui s'avère en général réalisable qu'une seule fois ou un nombre limité de fois. Il devient alors nécessaire de raisonner en termes de fréquence à une date donnée.

Par exemple, dans les essais de prévention cardiovasculaire par les hypocholestérolémiants qui durent de 5 à 7 ans, les patients sont revus régulièrement, en général annuellement. Si un événement survient entre deux

visites, il est parfois difficile de déterminer sa date exacte, sauf si l'événement a été suffisamment grave pour conclure à une hospitalisation où à l'intervention d'un service d'urgence.

Les échelles et scores

Les échelles (« scale ») et les scores permettent de mesurer l'intensité d'un phénomène clinique comme une gêne fonctionnelle, l'intensité d'un symptôme, l'extension d'une maladie, son stade évolutif, etc. Les échelles et les scores (la distinction terminologique entre échelles et scores n'est pas universelle. Les scores sont très souvent appelés échelles) sont très prisés en médecine car ils permettent de quantifier numériquement des phénomènes qui ne se caractérisent pas par une dimension physique.

Les échelles

Les échelles sont obtenues en découpant en différents stades (que l'on appelle aussi classe, grade, etc.) le continuum de gravité de la maladie étudiée. Chaque classe est caractérisée par un chiffre ou un adverbe matérialisant la relation d'ordre existant entre ces classes. Par exemple, une échelle de mesure de l'intensité d'une douleur peut-être : 0 : absente, 1 : modérée, 2 : importante, 3 : insoutenable.

Le résultat d'une échelle n'est pas assimilable à une variable continue lorsque le nombre de valeurs possibles est faible. Le recours aux outils statistiques spécifiques aux variables continues (moyenne, test de comparaison de moyennes) pose un certain nombre de problèmes. En réalité, il s'agit d'une variable qualitative ordinale dont l'analyse repose sur la description de la répartition des valeurs et sur des comparaisons à l'aide du test du chi-2. Cette remarque peut aussi concerner les scores.

Tableau 6 – Exemples d'échelles

Intensité d'un symptôme	• *Stade NYHA de dyspnée* • *Souffle cardiaque gradé de 0 à 6*
Gravité	• *Stade de gravité de l'asthme (bénin, moyen, sévère, aggravé)* • *Stade d'encéphalopathie hépatique (I,II,III,IV)*

Échelle de Rankin : handicap après AVC

0 = Absence de symptômes
1 = Symptômes mineurs sans retentissement sur la vie quotidienne
2 = Symptôme ou handicap mineur qui conduit à certaines restrictions dans le mode de vie, mais qui n'interfère pas avec la capacité du patient à se prendre en charge
3 = Handicap modéré qui restreint significativement le mode de vie et/ou empêche une existence totalement indépendante
4 = Handicap modérément sévère qui empêche clairement une existence indépendante bien que nécessitant pas une attention constante
5 = Handicap sévère entraînant une dépendance totale et nécessitant une attention jour et nuit

Les scores

Les scores permettent de mesurer des phénomènes multidimensionnels. Le score se calcule en cotant un certain nombre d'items analysant les différentes composantes du processus étudié puis en faisant la somme des note attribuées afin d'obtenir un score global. le but du score est de refléter en un seul nombre la totalité des dimensions envisagées.

Par exemple le score d'Apgar qui évalue la gravité des troubles respiratoires et neurologiques à la naissance d'après certains signes cliniques. Les nombres de points correspondants à chaque critère sont additionnés en un score global. Plus le score est bas, plus l'état du nourrisson est préoccupant.

Tableau 7 – Calcul du score d'Apgar

Critères	***Nombre de points***		
	0	***1***	***2***
Couleur	*Cyanosée ou pale*	*Corps rose, extrémités bleues*	*Complètement rose*
Rythme cardiaque	*Absent*	*<100*	*>100*
Respiration	*Absente*	*Irrégulière, lente*	*Bonne, cri vigoureux*
Réponse réflexe au cathéter nasal	*Sans*	*Grimace*	*Éternuement, toux*
Tonus musculaire	*Hypotonique*	*Légère flexion des extrémités*	*Actif et tonique*

Le plus souvent ces scores ont été établis à partir d'études pronostiques. Les items du score sont en fait les facteurs retrouvés associés avec le pronostique et le nombre de points de chaque item est une pondération proportionnelle à son importance dans le pronostic. Ce sont donc en fait des outils simplifiés de prédiction du risque d'évolution favorable (décès, survenue d'une complication, etc.)

Score de Barthel : évaluation du handicap après AVC

	Avec aide	*Indépendant*
1.Alimentation (si les aliments doivent être coupés = aide)	*1*	*2*
2.Déplacement de la chaise roulante au lit et retour	*1-2*	*3*
3.Toilette personnelle	*0*	*1*
4.Aller et revenir des toilettes	*1*	*2*
5.Se baigner seul	*0*	*1*
6.Marche sur un sol plat	*2*	*3*
7.Monter ou descendre des escaliers	*1*	*2*
8.Habillage (comprenant laçage des chaussures, boutonnage)	*1*	*2*
9.Continence anale	*1*	*2*
10.Continence vésicale	*1*	*2*
Total	_____	_____

Ce score prend des valeurs entre 0 et 20. Chaque item se noie dans les autres et une même valeur de score peut être obtenue avec des altérations fonctionnelles différentes. L'aspect multidimensionnel du handicap disparaît. À la fin, un changement de score de 1 n'a plus de signification clinique précise.

L'analyse statistique des scores repose souvent sur la comparaison des scores moyens de chaque groupe (moyenne des scores de chaque patient). En général, les distributions ne sont pas symétriques et il est plus adapté de comparer les médianes.

Problème d'interprétation des scores et des échelles

À coté des questions de qualité métrologique des échelles et des scores (reproductibilité, exactitude, homogénéité) que nous n'aborderons pas ici, les scores et les échelles posent différents problèmes d'interprétation.

La comparaison s'effectue en calculant le score moyen dans chaque groupe (cf. Figure 8). La moyenne est susceptible de prendre des valeurs que ne prennent pas les scores ou les échelles elles mêmes. Par exemple, des valeurs fractionnaires comme 5,68 ou 4,2 alors le score ne prend que des valeurs entières entre 1 et 10. Le patient moyen est donc affublé d'un score qui n'existe pas. Ainsi que signifie une différence de 0,9 points de l'échelle de handicap ? L'utilisation de la médiane pour décrire la position centrale de la population sur l'échelle ou sur le score ne conduit pas à ce problème.

Un autre point est la proportionnalité de la métrique. Est-ce qu'un changement de 1 point représente la même modification dans le phénomène étudié quel que soit le niveau de départ. En d'autres termes, le score mesure-t-il, par le même changement de valeur, un même effet chez des sujets de valeurs initiales différentes.

	Pallidotomy group (n=18)			Control group (n=16)			P*
	Baseline	6 months	Change	Baseline	6 months	Change	
Primary outcome							
UPDRS 3	47 (24–81)	32·5 (16–66)	15 (−13 to 27)	52·5 (23–82)	56·5 (19–91)	−2 (−15 to 9)	0·0004
Secondary outcome							
Pain VAS (mm)	27 (2–100)	14 (0–69)	3·5 (−20 to 77)	15·5 (0–87)	22 (0–84)	−0·5 (−23 to 45)	0·13
Barthell index	10·5 (4–20)	18 (6–20)	2·5 (−2 to 11)	11·5 (3–19)	8 (4–19)	−0·5 (−7 to 3)	0·004
UPDRS 2	30 (11–41)	21 (8–38)	7 (−8 to 20)	32 (14–45)	35 (15–46)	−2 (−11 to 6)	0·002
Schwab and England scale	35 (20–80)	70 (20–90)	15 (−10 to 40)	35 (10–80)	30 (10–80)	−5 (−30 to 10)	0·0009

A positive change score signifies improvement.
*Mann-Whitney *U* test.

Table 3: **Primary and secondary outcomes—median (range) scores of clinical rating scales for defined off phase assessment**

Figure 8 – Exemple de résultats obtenus avec des scores (UPDRS3, Barthell, UPDRS 2) ou une échelle visuelle analogique (Pain VAS, Schwab and England scale).

Exemples

Exemple 1 *- L'essai MAST-E comparait la streptokinase au placebo dans le traitement des accidents vasculaires cérébraux. Un des critères de jugement était la mesure du niveau de handicap à l'aide du score de Barthel. Six mois après l'AVC, la moyenne (± erreur standard) de ce score était 13,0±0,7 dans le groupe placebo et de 14,8±0,6 dans le groupe streptokinase. La différence est à la limite de la signification statistique : p=0,06. Étant donnée la construction du score de Barthel, la signification clinique d'une différence de 1,8 points n'est pas simple à appréhender et il n'est pas aisé de dire si cet effet représente une véritable amélioration de l'état des patients.*

Exemple 2 - Retour sur l'exemple des inhibiteurs de la phosphodiestérase

L'exemple des agents inotropes inhibiteurs de la phosphodiestérase présenté précédement procure aussi l'occasion de discuter des problèmes d'interprétation de la pertinence clinique d'un effet observé sur une échelle de score et de sa confrontation à un effet sur un critère clinique.

La question qui se pose est de savoir si l'amélioration de la qualité de vie ou de la symptomatologie est suffisamment importante pour éventuellement rendre acceptable un surcroît de mortalité. Avant d'envisager le problème éthique d'une réduction des chances de survie sous prétexte d'une amélioration fonctionnelle, il convient de pouvoir confronter la pertinence clinique des tailles des effets obtenus respectivement sur la mortalité et sur les signes fonctionnels.

Ce n'est pas parce qu'il y a détection d'un effet statistiquement significatif sur la qualité de vie (Avec les critères de jugement continus, des effets de petite taille ne peuvent s'avérer statistiquement significatifs, en particulier si la variabilité est faible. Il peut donc y avoir une dissociation forte entre signification statistique et pertinence clinique), que celui-ci est notable et intéressant pour les patients, et suffisamment important pour constituer une amélioration substantielle pouvant éventuellement justifier l'acceptation d'une surmortalité. Par exemple, dans l'essai vesnarinone (9), les effets étaient recherchés sur le changement médian du score de qualité de vie entre l'entrée dans l'essai et le moment de la mesure. Numériquement l'effet était faible. Initialement le score médian étaient de 56 points. A 8 semaines, le score de qualité de vie (« Minnesota Living with Heart Failure Questionnaire *») s'améliorait de 7 points dans le groupe vesnarinone 60mg contre une amélioration médiane de seulement 5 points dans le groupe placebo. Du fait de l'effectif important cette différence était hautement significative (p<0,001) mais il convient de s'interroger sur la pertinence clinique d'un tel effet qui ne représente qu'un surcroît d'amélioration de 2 points sur un échelle allant de 0 à 105.*

En d'autres termes, la surmortalité observée représente-t-elle un coût acceptable en regard du bénéfice obtenu sur les symptômes. Il est crucial dans cette situation de pouvoir traduire en terme clinique (évaluer la pertinence clinique) la différence de score de qualité de vie en des termes qui la rende comparable à la surmortalité. Une binarisation du score en utilisant un seuil exigeant est l'un de ces moyens.

Les mesures quantitatives

Des variables quantitatives liées directement à un processus physiologique ou biologique peuvent être utilisées comme critère de jugement, comme la valeur d'un paramètre biologique, le résultat d'une épreuve fonctionnelle, etc. (Figure 9).

Tableau 8 – Exemple de critères de jugement basés sur des paramètres quantitatifs

Paramètre biologique	• *Charge virale (SIDA)* • *Taux de CD4 (SIDA)* • *Température (infections)*
Épreuve fonctionnelle	• *Périmètre de marche (artériopathie des membres inférieurs)* • *Débit expiratoire forcé (asthme)*
Autres types	• *Échelle visuelle analogique (douleur)*

L'avantage des mesures quantitatives est d'être relativement sensible et de demander un petit nombre de patients afin de mettre en évidence les effets d'un traitement.

	Active	Placebo	Difference	p
Last 3 days of each treatment period				
Symptom score	5·32 (0·49)	5·69 (0·49)	−0·38 (0·29)	0·21
Reliever (puffs/day)	5·04 (0·82)	4·76 (0·64)	0·29 (0·34)	0·40
Morning PEF (L/min)	288·4 (15·2)	282·5 (13·9)	5·89 (5·68)	0·37
Evening PEF (L/min)	300·8 (15·1)	299·3 (13·2)	1·52 (6·64)	0·88
PEF variability	12·2% (1·8)	13·4% (2·5)	−1·0% (0·1)	0·32
Last 7 days of each treatment period				
Symptom score	5·67 (0·47)	5·62 (0·46)	0·05 (0·22)	0·73
Reliever (puffs/day)	5·06 (0·72)	4·65 (0·59)	0·41 (0·24)	0·06
Morning PEF (L/min)	285 (14·8)	284 (13·9)	1·18 (3·41)	0·82
Evening PEF (L/min)	300 (14·5)	300 (13·3)	−0·5 (3·71)	0·81
PEF variability	13·2% (1·9)	13·4% (2·2)	0% (0·1)	0·86

Values are means (SE) for the last 3 or 7 days of each treatment period. There were no significant period or carryover effects. Symptoms were scored 0–3 for each of six variables (total out of 18).

Table 2: **Diary data from last 3 days and last 7 days of each treatment period**

Figure 9 – Exemple de tableau rapportant les résultats obtenus au niveau de critères de jugement quantitatif.

Expression de l'effet traitement

L'effet du traitement peut être mesuré de différentes façons (cf. chapitre Indices d'efficacités*) :*

- *différence des valeurs moyennes mesurées en fin d'essai entre les deux groupes. Cette mesure n'est pas ajustée sur les valeurs initiales et sous-entend donc une comparabilité initiale forte des groupes (mêmes moyennes à l'inclusion).*
- *différence entre les deux groupes des changements moyens observés dans chaque groupe entre le début et la fin de l'essai.*

La charge virale

La charge virale est couramment utilisée dans les essais de traitement de l'infection à VIH où elle est parfois considérée, seule ou en association avec le taux de CD4, comme un critère de substitution.
Les modifications de charge virale peuvent être exprimées de plusieurs façon.
a) **Logarithme de la réduction obtenue en fin de l'essai**. Pour chaque patient, le logarithme décimal de la valeur initiale de charge virale est soustrait au logarithme de la valeur mesurée en fin d'essais. La moyenne ou la médiane de ces résultats individuels sont ensuite calculées dans chaque groupe de l'essai. Le fait de travailler sur une échelle logarithme équivaut à calculer le rapport de la valeurs initiale sur la valeur finale de la charge virale. Par exemple, avec un dosage dont le seuil de détection est de 400 copies d'ARN/mL et un traitement qui rend indétectable la charge virale entraînera un changement de 3 log chez un patient dont la valeur initiale est de 400 000 copies/mL, et seulement un changement de 1 log chez un autre patient dont la valeur initiale est de 4 000 copies/mL. Cette mesure est donc sensible au seuil de détection du dosage utilisé. Chez un même patient, plus une méthode est sensible plus elle montrera un changement important. Il est donc important de connaître le dosage utilisé pour comparer les résultats de plusieurs essais.
b) **Logarithme des pics de réduction**. Avec cette méthode, la valeur de base est soustraite à la plus faible valeur de charge virale observée durant le suivi, quel que soit le moment de survenue de ce minimum. Cette différence est ensuite exprimée en terme logarithmique. Cette mesure est à proscrire car, avec un traitement sans effet, elle prend les fluctuations aléatoires vers le bas comme baisse engendrée.
c) **Aire sous la courbe**. L'aire sous la courbe est calculée comme une moyenne des mesures pondérées par les intervalles de temps entre mesure. Cette mesure intègre à la fois l'intensité des réductions observées, la précocité de leur survenue et leur durée.

Pertinence clinique

L'interprétation d'une différence induite par un traitement sur une mesure quantitative et la détermination de sa signification clinique n'est pas toujours évidente.

Comme avec les scores et les échelles, la pertinence clinique d'un effet traitement observé au niveau d'une mesure quantitative n'est pas toujours simple à établir. En plus, les mesures quantitatives font courir le risque de donner assez facilement des différences statistiquement significatives mais de faible ampleur et donc sans signification clinique. Les variables continues sont aussi plus souvent en rapport avec un critère intermédiaire qu'avec un critère clinique.

Par exemple, quel est l'intérêt clinique d'un traitement vasodilatateur augmentant de 20m en moyenne le périmètre de marche de patients ayant une artériopathie des membres inférieurs de stade 2 ? Le traitement augmente effectivement le périmètre de marche des patients mais est-ce qu'une augmentation moyenne de 20 mètres modifie la vie des patients et représente pour eux une amélioration notable ? De plus il convient de faire attention à l'interprétation d'un effet moyen (cf. chapitre Statistiques avancées*) et de ne pas penser que tous les patients ont leur propre périmètre de marche augmenté systématiquement de 20 mètres.*

Exemple

Dans l'incontinence urinaire féminine, un critère utilisé consiste à compter le nombre de fuites urinaires liées à un besoin impérieux sur une certaine période de temps, en général une semaine. Chez des femmes ayant en moyenne 20,7 fuites par semaine, un traitement apporte une réduction supplémentaire par rapport au placebo du nombre de fuites par semaine de 4,3 (le placebo étant associé avec une réduction de 8 par rapport aux valeurs initiales) (10). Ce qui ramène sous traitement la fréquence moyenne des fuites à 8,4 par semaine.

Cet effet hautement statistiquement signification ($p<0.0001$) est cependant très peu pertinent car ce qui gêne ces femmes, entre autres, c'est la nécessité de porter une protection. Avoir 8,4 fuites par semaine à la place de 20,7 jours ne solutionne pas leur problème car elles continuent à devoir porter des protections. Un critère plus pertinent consisterait à dénombrer les succès thérapeutiques : c'est-à-dire les femmes n'ayant plus de fuite sur une période de temps significative, ou éventuellement, dans les cas sévères, le nombre de protections utilisées par jour.

Cet exemple illustre donc le fait que les critères quantitatifs sont fréquemment sans pertinence clinique, non pas parce qu'ils correspondent à une entité nosologique sans intérêt, mais plutôt parce qu'ils peuvent facilement mettre en évidence des effets de petite taille et/ou parce que la gêne subie par les patients n'est pas proportionnelle à la valeur moyenne du critère.

Une solution est de binariser les mesures quantitatives (cf. infra) pour créer une variable binaire de succès-échec thérapeutique. Par exemple, chez les patients artériopathiques, un objectif thérapeutique pourrait être de rendre le périmètre de marche supérieur à 500m. Le bénéfice du traitement s'évaluera alors en termes de nombre de patients atteignant cet objectif.

Mesures répétées

La mesure des critères continus est souvent répétée au cours du suivi d'un essai afin d'appréhender la dynamique de l'effet thérapeutique. L'analyse statistique présente cependant un certain nombre de difficulté (11, 12). La réalisation de plusieurs comparaison temps par temps se heurte au problèmes des comparaisons statistiques multiples. Des techniques statistiques pour mesures répétées sont disponibles mais nécessitent de faire des hypothèses fortes sur les données.

Effet permanent. Après avoir s'être établi, l'effet du traitement se maintient durant toute la période d'observation.

Effet temporaire. Il n'existe pas de différence entre les deux groupes à la fin de l'essai, mais un effet traitement a existé temporairement.

Il n'existe pas de différence en fin d'essai, mais la réduction a été obtenue plus rapidement dans un groupe que dans l'autre.

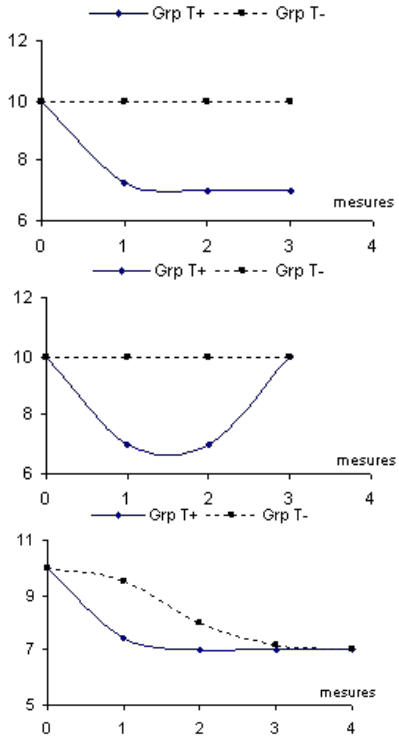


Figure 10 – Différents cas de figures d'effet traitement possible avec des mesures répétées d'un critère de jugement quantitatif

Le moment de mesure de l'effet est donc un point important avec les critères de jugements continus.

Analyse statistique

Au niveau statistique, les mesures répétées (appelées aussi mesures longitudinales) donnent lieu à de nombreuses comparaisons statistiques conduisant à une inflation du risque alpha. Pour éviter cela, les données peuvent être comparées en bloc, avec, par exemple, une analyse de variance pour mesures répétées qui, si elle s'avère statistiquement significative, permet de dire qu'il existe un effet du traitement à au moins un moment de mesure. L'identification de ce ou ces moments pose ensuite toute une série de problèmes statistiques non entièrement résolus (non indépendance des mesures, répétition des tests statistiques, etc.).

Pour contourner ces difficultés, les mesures répétées peuvent être résumées par leur moyenne, par l'aire sous la courbe ou par un modèle d'évolution. La comparaison porte ensuite sur les paramètres résumés ce qui réduit à un le nombre de comparaison statistiques.

Exemple : évaluation des antalgiques

L'évaluation des antalgiques fait appel à des mesures quantitatives de l'intensité de la douleur (13). L'outil le plus fréquemment utilisé est l'échelle visuelle analogique. On demande au sujet d'indiquer l'intensité de sa douleur par une marque sur une ligne horizontal de 10 cm qui à son extrémité gauche porte la mention « absence de douleur » et à son autre extrémité la mention « douleur maximale imaginable ». D'autres approches se basent sur des échelles verbales simples ou des échelles numériques simples où la douleur est cotée par un nombre allant de 0 (absence de douleur) à 5 (douleur maximale imaginable).

Dans un essai, ces mesures sont répétées au cours du temps afin d'étudier l'évolution chronologique de l'effet de l'antalgique. Afin d'exploiter l'ensemble de l'information apportée par la répétition des mesures, il est souhaitable d'intégrer ces multiples mesures afin de juger globalement de l'effet sur une période de temps donnée. Ces indices intégratifs sont calculés dans chaque groupe de l'essai et l'effet du traitement est déterminé par la différence existant entre le groupe expérimental et le groupe contrôle. Plusieurs types d'intégration sont possibles.

À chaque temps de mesure est calculé le PID (« pain intensity difference *») qui est la différence d'intensité de la douleur par rapport à la valeur basale. Ces valeurs permettent de déterminer le max PID qui est la valeur maximale de PID obtenue durant la période de temps ou Tmax PID qui est le temps d'obtention de l'effet antalgique maximum depuis la prise.*

L'aire sous la courbe est appréciée par le SPID « sum of PID *» qui est la somme de tous les PID.*

D'autres indices comme le PAR (« pain relief *») sont basés, non plus sur l'intensité de la douleur mais sur son soulagement. Pour cela, l'échelle visuelle analogique utilisée porte à son extrémité gauche la mention « aucun soulagement » et son extrémité droite « soulagement complet ». Là aussi on peut établir un max PAR ou un Tmax PAR. Un score global, le TOTPAR («* Total of Pain relief *»), est calculé en utilisant l'aire sous la courbe pour une période de temps définie. Le TOTPAR s'exprime aussi en pourcentage du TOTPAR maximum qui serait obtenue par un traitement qui donnerait un soulagement complet durant la période d'observation.*

Binarisation

Une solution aux problèmes d'interprétation clinique des effets mesurés par des échelles, des scores ou des variables continues est de les transformer en une variable binaire en fonction d'une valeur seuil. Cette transformation exprime le résultat en termes de succès thérapeutiques. L'effet du traitement est alors mesuré, par exemple, par la proportion de patients ayant un score supérieur (ou inférieur) à la valeur seuil..

En général, la binarisation est construite de façon à séparer les patients sévèrement atteints de ceux qui le sont moins : score de dépression important, périmètre de marche très limité.

Une autre façon est de choisir un seuil qui correspond à une guérison ou à une quasi-guérison. Dans ce cas, la variable binaire évalue la proportion de patients guéris par le traitement. C'est par exemple le cas avec l'hypertension artérielle où il est possible de créer une variable binaire : normalisation de la pression artérielle (Cette variable ne correspond cependant pas à un critère clinique : une variable binaire ne correspond pas automatiquement à un critère clinique). La binarisation peut aussi être basée sur un seuil de changement de la variable continue : par exemple une amélioration de plus de 30% du périmètre de marche. Cette définition du succès thérapeutique peut manquer de pertinence clinique si l'on prend un seuil d'amélioration peu ambitieux. D'autres définitions sont possibles faisant intervenir la durée de l'amélioration (maintient au-dessus du seuil durant au moins x jours, etc…), ou un objectif fixé par le patient lui même.

Exemple de l'ACR20

L'American College of Rheumatology a défini un critère pour mesurer l'efficacité des traitements de fonds de la polyarthrite rhumatoïde (antirhumatismaux d'action lente), l'ACR20. Ce critère est défini d'une part par une amélioration d'au moins 20% à la fois du nombre d'articulations douloureuses et du nombre d'articulations tuméfiées, et d'autre part par une amélioration d'au moins 20% du score des critères suivants : évaluation globale par le malade, évaluation globale par le médecin, score à l'échelle HAQ (Health Assessment Questionnaries) ou sa version modifiée M-HAQ remplie par le malade, et soit l'amélioration de la vitesse de sédimentation, soit du taux de protéine C réactive. L'effet d'un traitement est évalué en comparant le pourcentage de patients satisfaisant ce critère.

Le taux de patient ACR 20% est actuellement largement utilisé mais sa signification clinique est controversée (14Ê, 15). Par exemple, une réduction de 20% peut simplement traduire que les articulations douloureuses ou tuméfiées est passé de 15 à 12, ou de 5 à 4.

D'autres critères, comme ACR 50 ou ACR 70 sont obtenus de la même façon qu'ACR 20 en exigeant des taux d'amélioration de 50% ou 70%. Comparé avec ces deux indices, l'ACR 20 conduit plus facilement à des différences statistiquement significatives entre traitement et placebo, mais sa pertinence clinique est discutée. Malgré cela, la plupart des essais sont actuellement réalisés avec la mesure de ce critère au bout de 6 mois.

Les constituants de l'ACR 20 sont en fait des mesures de l'activité inflammatoire réversible. Ces mesures sont utilisées comme critère intermédiaire à la place de l'évaluation des dégâts articulaires comme l'érosion radiographique, de la déformation articulaire ou de toute autre évolution au long terme.

De ce fait, des essais au long cours restent nécessaires pour mieux évaluer la progression de la PR sur des critères pertinents : déformations des articulations, handicap fonctionnel, nécessité d'intervention chirurgicale articulaire, manifestations extra articulaires, mortalité.

Problèmes divers posés par la binarisation

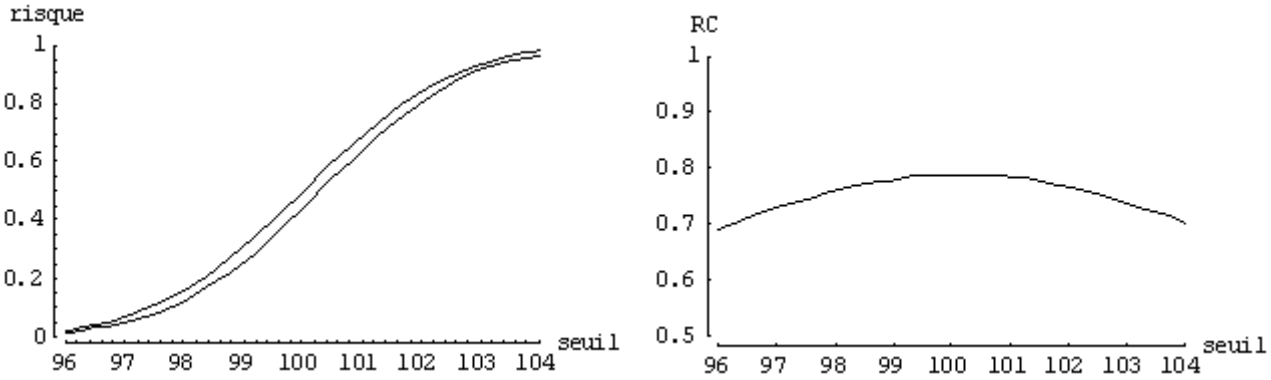
Bien que la binarisation permette de mieux appréhender la pertinence d'un effet, celle-ci n'est que rarement effectuée. De plus, lorsqu'elle est faite, il est rare que le même seuil soit utilisé d'un essai à l'autre. Il semble alors difficile de comparer les résultats des essais entre eux. En fait, il est possible de démontrer que sous certaines conditions, l'odds ratio calculé après binarisation est assez indépendant de la valeur du seuil et dépend surtout de la différence entre les deux distributions (c'est-à-dire entre les moyennes par rapport à leur variance). Ainsi, la comparaison d'études est possible même si des seuils différents ont été utilisés.

Les mêmes propriétés donnent le moyen de calculer un odds ratio même si aucune binarisation n'est disponible. Ainsi, il devient possible d'exprimer l'effet du traitement en termes d'odds ratio, comme si une binarisation avait été effectuée, et si l'on compare plusieurs essais, comme si le même seuil de binarisation avait été utilisé.

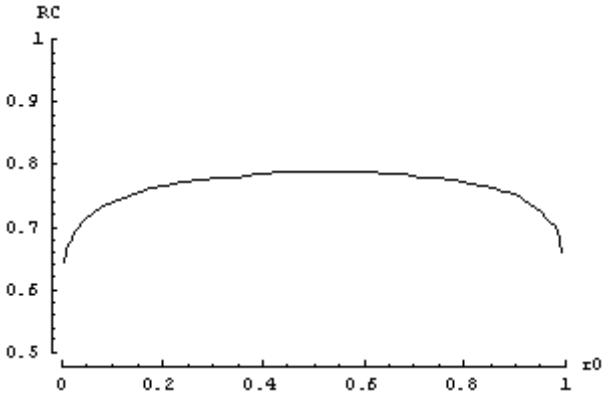
Bien entendu, ces calculs sont des extrapolations qui ne sont exactes que si les distributions des variables continues suivent une distribution particulière, la distribution logistique, qui est proche de la distribution normale. La méthode est relativement robuste à une déviation de la loi normale à conditions qu'elle reste symétrique.

Calcul d'un odds ratio à partir d'une différence de moyenne

La première des figures ci-dessous représente l'évolution de la fréquence du critère binaire dans les 2 groupes en fonction du seuil de binarisation utilisé (loi normale). La seconde figure représente l'évolution de l'odds ratio (RC) en fonction de ce seuil de binarisation.



Lorsque l'on représente l'odds ratio en fonction de la fréquence du critère binaire dans le groupe contrôle (r0), il apparaît que l'odds ratio reste presque inchangé pour une large étendue de risque de base (de 15 à 85%).



Ainsi, lorsque la distribution de la variable continue est proche d'une distribution normale, l'odds ratio obtenu après binarisation est quasiment indépendant de la valeur du seuil choisie pour la binarisation, tant que cette binarisation ne conduit pas à une fréquence du critère proche de 0% ou de 100%. De ce fait, un odds ratio extrapolé à partir de la différence des moyennes représente correctement celui qui aurait été obtenu lors d'une binarisation.
L'extrapolation d'un odds ratio à partir d'une différence de moyenne se calcule de la façon suivante.
On fait l'hypothèse que la distribution de la variable continue est normale (au pire symétrique) et de même variance dans les deux groupes. L'odds ratio est alors obtenu par :

$$Log(OR) = \frac{\pi}{\sqrt{3}} \frac{\bar{x}_0 - \bar{x}_1}{s}$$

où $\bar{x}_1$ et $\bar{x}_0$ représente les deux moyennes et

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}}$$

s_1^2 et s_0^2 désignent les variances et n1 et n0 les effectifs des deux groupes.
La variance du logarithme de l'odds ratio est :

$$\text{var}(LogOR) = \frac{\pi^2}{3}\left[\frac{n_1 n_0}{n_1 + n_0} + \frac{\frac{\bar{x}_1 - \bar{x}_0}{s}}{2(n_1 + n_0) - 2}\right]$$

Divers

Nombre d'épisodes durant une période de temps

Dans certaines situations, le critère de jugement est susceptible de se produire plusieurs fois chez le même patient durant le suivi. Par exemple, chez l'insuffisant rénal dialysé, les patients ressentent parfois des crampes musculaires douloureuses. Un critère de jugement opportun pour évaluer un traitement préventif est la survenue des crampes lors d'une séance de dialyse, survenue qui a valeur d'échec thérapeutique. Les séances de dialyse se succédant rapidement il est possible dans un essai de suivre les patients durant un nombre important de séances.

> *Traiter des observations multiples en provenance du même patient comme des données indépendantes est une erreur (16, 17).*

Malgré les apparences, ce critère n'est pas un critère binaire. En effet, au niveau d'un patient, le critère de jugement n'est pas la survenue d'une crampe lors d'une séance de dialyse mais la fréquence de survenue de crampes par séance (4/12 dans le cas du patient précédent). Il s'agit alors d'un critère continu dont on calcule la moyenne dans chaque groupe de traitement. Il serait erroné de calculer la fréquence de survenue des crampes en divisant le nombre total de crampes par le nombre total de séances (analyse basée sur les séances). L'unité statistique serait la séance de dialyse et il n'y aurait pas indépendance des unités statistiques. En effet, raisonner sur les séances ne prend pas en compte le fait que les séances d'un même patient sont corrélées entre elles. Par exemple le patient n°4 est à faible risque de crampes. De ce fait, l'essai a été

prolongé chez lui durant plus de séances que les autres patients qui présentaient plus de crampes et chez lesquels le recours à un autre traitement a été plus rapide. Ce patient aurait donc un poids supérieur aux autres dans le calcul de la fréquence. Par contre, dans une analyse basée sur les patients, les fréquences de survenue de crampes par séance calculées pour chaque patient sont indépendantes entre-elles. Le calcul de leur moyenne ne pose aucun problème.

Tableau 9 – Analyse basée sur les patients ou basée sur les séances de la fréquence de survenue de crampes lors de séances de dialyse.

Patients	*Nombre de séances de dialyse suivies*	*Nombre de séances avec crampes*	*Fréquence des séances avec crampes*
1	*10*	*5*	*5/10 = 0,50*
2	*12*	*9*	*9/12 = 0,75*
3	*5*	*4*	*4/5 = 0,80*
4	*24*	*2*	*2/24 = 0,80*
5	*9*	*12*	*9/12 = 0,75*
Total	*60*	*32*	*analyse basée sur les*
analyse basée sur les séances		*32 / 60 = 0,53*	*patients = 0,72*

L'analyse correcte de plusieurs mesures par patients fait appel à des techniques statistiques adaptées (modèle hiérarchique) mais il s'avère en pratique que cette approche n'entraîne pas de diminution importante du nombre de sujets nécessaires.

Exemple

La possibilité de faire régresser (ou de stopper la progression) les plaques athéromes coronariennes par les traitements hypocholestérolémiants a été recherchée par les modifications de diamètre des sténoses survenant chez des patients entre deux coronarographies, l'une avant et l'autre après 1 an de traitement. La présence de plusieurs sténoses par patient est fréquente ce qui permet de mesurer plus d'un changement de diamètre chez un même patient.

L'analyse correcte de ces données consiste à prendre comme unité statistique le patient (et non pas la sténose) et à ne retenir qu'une valeur de changement de diamètre par patient. Pour cela une lésion index doit être fixée lors du premier examen et non pas choisie a posteriori. En effet, il ne serait pas raisonnable de choisir la lésion sur laquelle le plus fort changement a été observé.

La prise en compte de plusieurs lésions par patient entraîne une augmentation artificielle de l'effectif et considère comme indépendantes (en probabilité) des valeurs qui sont mesurées chez le même patient.

Conséquences de la performance de la méthode diagnostique

Le niveau de performance diagnostique de la méthode utilisée pour mesurer le critère de jugement influence la mesure de l'effet traitement. Une méthode peu sensible et/ou peu spécifique entraîne une sous-estimation de l'effet traitement dans un essai thérapeutique (18).

Théorie

Le nombre d'événements observés est la somme des vrais positifs (« true positive *») et des faux positifs («* false positive *»). Le taux de vrais positifs dépend de la sensibilité* se *et de la vraie fréquence de l'événement* i*. Le taux de faux positifs dépend de la spécificité* sp *et de la vraie fréquence de l'absence d'événements* 1-i*.*

La fréquence observée du critère de jugement sans traitement (dans le groupe contrôle d'un essai) est :

$$r_0 = i \quad se + (1-i)(1-sp)$$

Sous traitement la vraie fréquence de l'événement est modifiée par le vrai risque relatif θ *caractérisant l'effet traitement :*

$$r_1 = i \quad \theta \quad se + (1-\theta \quad i)(1-sp)$$

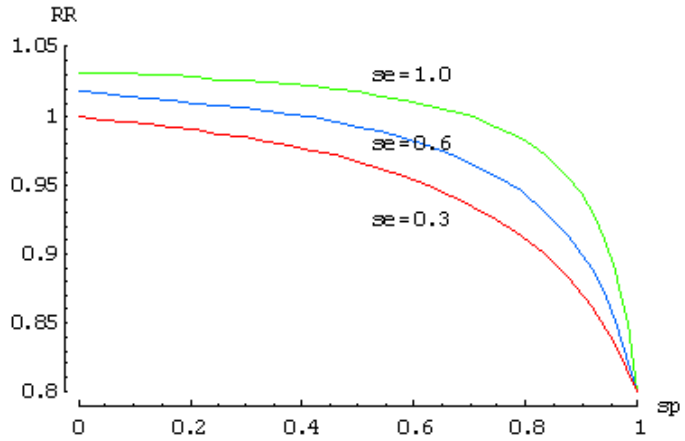
Finalement le risque relatif rr *observé dans l'essai est :*

$$rr = \frac{r_1}{r_0} = \frac{i\ \theta\ se + (1-\theta\ i)(1-sp)}{i\ se + (1-\ i)(1-sp)}$$

Représentations graphiques

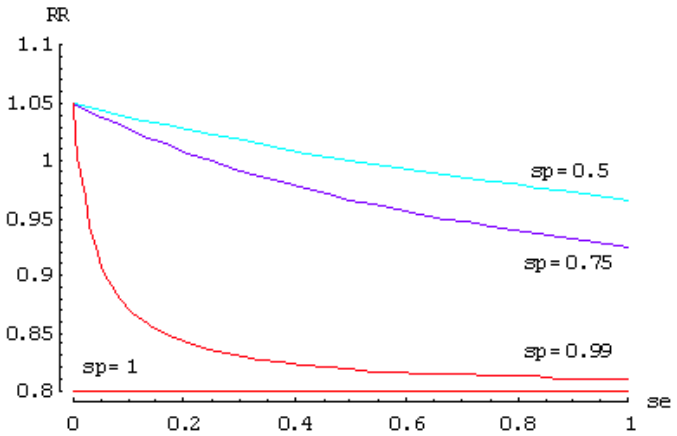
La Figure 11 représente l'évolution de l'estimation d'un vrai risque relatif de 0,8 en fonction de la spécificité pour 3 valeurs différentes de sensibilité. L'incidence est de 0,2. Quelle que soit la sensibilité, le risque relatif tend vers 1 ou plus lorsque la spécificité diminue. Le risque relatif est sensible à la valeur de la spécificité y compris pour une sensibilité parfaite de 1.

Figure 11 – Évolution du risque relatif en fonction de la spécificité pour différentes valeurs de sensibilité. La vraie valeur du risque relatif est 0,8.



La Figure 12 représente l'évolution du risque relatif estimée en fonction de la sensibilité pour différentes valeurs de spécificité. Excepté lorsque la spécificité est égale à 1, le risque relatif tend vers 1 et plus quand la sensibilité diminue. Lorsque la spécificité est de 1, l'estimation du risque relatif est insensible à la sensibilité.

Figure 12 – Évolution du risque relatif en fonction de la sensibilité pour différentes valeurs de spécificité. La vraie valeur du risque relatif est 0,8.



Bibliographie

1. Gerin P, Dazord A, Boissel JP, Hanauer MT, Moleur P, Chauvin F. L'évaluation de la qualité de vie dans les essais thérapeutiques. Thérapie 1989;44:355-64.
2. Schraub S, Mercier M, Arveux P. Mesure de la qualité de vie en cancérologie. Presse Med 2000;29:310-18.
3. Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? Ann Intern Med 1996;125:605-613.
4. Greenhalgh T. How to read a paper: papers that report drug trials. BMJ 1997;315:480-483.
5. Advanced colorectal cancer meta-analysis project. Modulation of fluorouracil by leucovorin in patients with advanced colorectal cancer: evidence in terms of response rate. J Clin Oncol 1992;10:896-903.
6. Riggs BL, Hodgson SF, O'Fallon WM, Chao EY, Wahner HW, Muhs JM. Effect of floride treatment on teh fracture rate in postmenopausal women with osteoporosis. NEJM 1990;322:802-809.
7. The Long Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. NEJM 1998;339:1349-57.
8. Serruys PW, van Hout B, Bonnier H, Legrand V, Garcia E, Macaya C, et al. Randomised comparison of implantation of heparin-coated stents with balloon angioplasty in selected patients with coronary artery disease (Benestent II). Lancet 1998;352(9129):673-81.
9. Cohn JNG, S.O. A dose-dependant increase in mortality with vesnarinone among patient with severe heart failure. Vesnarinone Trial Investigators. NEJM 1998;339:1810-6.
10. Khullar V, Hill S, Laval KU, Schiotz HA, Jonas U, Versi E. Treatment of urge-predominant mixed urinary incontinence with tolterodine extended release: a randomized, placebo-controlled trial. Urology 2004;64(2):269-74; discussion 274-5.
11. Liu C, Li Wan Po A, Blumhardt LD. "Summary measure" statistics for assessing the outcome of treatment trials in relapsing-remitting multiple sclerosis. J Neurol Neurosurg Psychiatry 1998;64:726-729.
12. Senn S, Stevens L, Chaturvedi N. Repeated measures in clinical trials: simple strategies for analysis using summary measures. Stat Med 2000;19:861-877.
13. Dubray C. Etude clinique des médicaments antalgiques. Thérapie 1999;54:135-145.
14. Pincus T, Stein CM. ACR20: clinical or statistical significance? Arthritis Rheum 1999;42:1572-1576.
15. van Gestel A, van Riel L. Improvement criteria - clinical and statistical significance: comment on the article by Pincus and Stein. Arthritis Rheum 2000;43:1658-1659.
16. Bolton S. Independence and statistical inference in clinical trial designs: a tutorial review. J Clin Pharmacol 1998;38(5):408-12.
17. Altman DG, Bland JM. Statistics notes. Units of analysis. BMJ 1997;314(7098):1874.
18. Rodgers A, MacMahon S. Systematic underestimation of treatment effects as a result of diagnostic test inaccuracy: implications for the interpretation and design of thromboprophylaxis trials. Thrombosis and Haemostasis 1995;73:167-71.

Pertinence clinique – Sélection des patients et population incluse

Représentativité des patients inclus dans l'essai

Introduction

La représentativité des patients est acquise quand aucun type de patients ciblés n'a été systématiquement évincé de l'essai

L'analyse de la population étudiée vérifie que celle-ci est constituée de patients représentatifs de ceux vus en pratique et qu'elle n'est pas composée de patients sélectionnés. Il s'agit de déterminer si la population étudiée est proche de la population cible.

Ce point est jugé d'après la définition des critères d'inclusion et d'exclusion (est-ce que les critères de sélection sont trop étroits ou trop larges ?) et d'après la description des patients effectivement inclus (ces patients ont-ils en moyenne les mêmes caractéristiques que la population cible ?). Les critères de sélection déterminent la population qui était ciblée.

La description de la population incluse confirme dans quelles mesures cette cible a effectivement été atteinte. En effet, il n'est pas impossible qu'un essai n'arrive pas exactement à inclure les patients qu'il recherche et ceci pour diverses raisons. Par exemple, les investigateurs hésitent à inclure certains types de patients, les patients sont pris en charge d'une façon qui exclut leur inclusion dans l'essai, etc.

La présence de certains types de patients en une faible proportion ne signifie pas obligatoirement que ces patients ont été systématiquement évincés du recrutement. Ces patients ne représentent peut-être qu'une faible proportion de la population cible et il est donc normal qu'ils soient peu représentés dans l'échantillon de l'essai.

Le but est d'obtenir un échantillon relativement représentatif de la population ciblée. Les différents types de patients devraient se retrouver dans l'échantillon de l'essai dans les mêmes proportions que celles de leur répartition dans la population ciblée. Comme l'échantillon d'un essai n'est pas constitué par tirage aléatoire dans la population cible, cette représentativité n'est pas acquise de fait. Pour l'atteindre, le recrutement des patients doit éviter d'exclure systématiquement des types de patients couramment rencontrés en pratique.

Extrapolation des résultats

La question de l'extrapolabilité du résultat est la suivante : est-ce que l'efficacité obtenue sur des patients particuliers est informative vis à vis de l'efficacité du traitement chez des patients présentant d'autres caractéristiques ?

Un manque de représentativité de la population étudiée n'est gênant que si l'effet du traitement varie substantiellement en fonction des caractéristiques des patients. Dans ce cas, l'efficacité observée chez certains types de patients n'est pas le reflet de l'efficacité chez d'autres types de patients ou dans la population cible de la thérapeutique.

Un scénario fâcheux serait un traitement inefficace chez la majorité des patients de la population cible mais seulement efficace chez des sujets très particuliers et chez lesquels l'essai aurait été conduit. Toutefois, il existe peu d'exemple de ce type.

Cependant l'accent est souvent mis sur une éventuelle possibilité de ce type, conduisant à déclarer tout résultat d'essai non extrapolable sous le prétexte que, presque toujours, les patients inclus dans les essais ne sont pas représentatif des patients tout venant. Cette attitude, bien que couramment enseignée, est certainement exagérée pour deux raisons. La première est que, dans les essais qui prennent garde à ce point, les patients recrutés sont très proches de ceux de la population cible. L'autre raison est qu'il existe peu d'exemples où l'efficacité du traitement change de façon notable d'un type de patient à l'autre. Ainsi, même si les patients inclus dans un essai ne sont pas strictement représentatifs de la population générale, le résultat peut être cependant retenu. La remise en cause de l'extrapolabilité du résultat ne sera fait que s'il existe des arguments positifs forts faisant craindre un changement d'efficacité important en dehors des patients de l'essai.

Bien entendu tout ceci est une affaire de nuance. Un résultat obtenu sur des patients notablement hyper-sélectionnés n'est pas aussi convainquant qu'un résultat obtenu sur une large variété de patients, même si dans ce dernier cas la population de l'essai n'est pas totalement représentative.

Très souvent la population dépend de la façon dont est posée la question de recherche comme le montre le tableau suivant :

Question	Population de l'essai
Effet chez les patients souffrant d'un angor y compris chez les diabétiques	Diabétiques présents dans la même proportion que leur part relative dans la population générale des patients angoreux. Cette population ne permet pas de répondre à la question « que fait le traitement chez les diabétiques ? »
Effet chez les patients angoreux diabétiques	Échantillon d'angoreux diabétiques uniquement

Patients représentatifs des patients vus en pratique

Pour être représentative de la pratique médicale de tous les jours, l'inclusion des patients doit être basée sur des critères larges, peu sélectifs tels qu'utilisés en pratique pour définir la maladie cible. L'essai a comme but de documenter la pratique médicale telle qu'elle sera mise en œuvre avec ce traitement. C'est un essai pragmatique dont le but est de savoir si l'utilisation du traitement permet en pratique d'atteindre les objectifs thérapeutiques [192].

La méthode de l'essai thérapeutique peut aussi être utilisée avec une finalité plus cognitive. Ces essais explicatifs ont alors pour but d'expliquer des mécanismes et se rapprochent de la recherche fondamentale. L'objectif de ces essais est d'étudier les mécanismes d'effet des traitements (justification théorique) mais pas d'apporter la justification factuelle de leur utilisation. Il est alors nécessaire de sélectionner un groupe de patients le plus homogène possible, afin de se placer dans les meilleures conditions expérimentales possibles. Étant donné que les patients de ces essais ne sont pas représentatifs de ceux qui sont vus en pratique, les résultats ont une faible pertinence clinique.

Ces principes généraux conduisent aux oppositions présentées dans le Tableau 2.

Au total, un essai pragmatique n'utilise des critères de sélection que si ces critères sont indispensables pour définir la population cible du traitement. De ce fait ces critères seront ensuite utilisés dans la pratique médicale. Les critères diagnostiques issus des essais de la fibrinolyse illustrent ce principe.

Tableau 1 – Différences au niveau de la population ciblée et des critères d'inclusion entre essai pragmatique et essai explicatif

	Essai pragmatique	Essai explicatif
Objectif	*Documenter le bénéfice qu'apportera le traitement aux patients qui seront traités en pratique*	*Connaître l'effet du traitement dans des conditions très précises dans un but d'explication*
Finalité	*Pratique : valider une pratique médicale*	*Cognitive : connaissance et explication*
Population visée	Représentative des futurs patients qui seront traités	La plus homogène possible afin de contrôler au maximum l'influence d'autres facteurs
Critères	Peu restrictif. Correspondant au critère utilisé en pratique pour décider un traitement et/ou pour caractériser un risque. Faisant appel à des examens de routine	Précis définissant parfaitement les patients recherchés. Faisant appel à des techniques sophistiquées

Exemple

Les grands essais qui ont démontré l'intérêt de la fibrinolyse à la phase aiguë de l'infarctus du myocarde (ISIS 2 [121], GISSI [109]) ont inclus les patients porteurs d'une suspicion de nécrose en utilisant des critères différents de ceux définissant classiquement l'infarctus du myocarde constitué (douleur, élévation enzymatique et apparition d'onde Q de nécrose à l'ECG). Dans les essais les patients étaient éligibles quand ils présentaient une douleur trinitro-résistante prolongée et des signes électriques d'ischémie aiguë (sus décalage de ST de plus de 2mm dans au moins 2 dérivations).

Le bénéfice sur la mortalité totale des fibrinolytiques étant démontré chez ces patients, les mêmes critères ont été ensuite utilisés en pratique médicale courante.

Détermination de la population cible

Dans l'approche pragmatique, tous les patients porteurs de la maladie appartiennent potentiellement à la population cible de la thérapeutique, car, si en pratique le traitement confirme son intérêt, tous ces patients seront susceptibles d'être traités. De ce fait il convient de valider le traitement sur un échantillon des patients chez lesquels il sera utilisé dans la pratique médicale future. Cette démarche globale n'est cependant envisageable que s'il n'existe a priori aucune raison de penser que l'effet du traitement puisse être nettement différent chez certains types de patients. Dans le cas contraire, où il existe, a priori, des arguments pour penser que le traitement sera chez certains types de patients sans efficacité ou délétère, il n'est plus opportun de regrouper ces patients avec les autres dans un même essai. En effet, un effet délétère chez une faible proportion des patients peut être masqué par le bénéfice obtenu chez les autres patients.

Tableau 2 – Oppositions entre essais pragmatiques et essais explicatifs.

Essai pragmatique	**Essai explicatif**
Pas d'age limite (excepté >18 ans pour des raisons légales)	Plage d'age parfaitement définie et limitée
Pas de sélection sur le pronostic	Sélection de patients à bon pronostic pour limiter le risque de données manquantes liées au décès du patient.
Pas d'exclusion de comorbidité	Exclusion de toute comorbidité ne faisant pas partie de l'objectif de l'essai
Aucune limitation de l'utilisation des traitements validés sauf raisons particulières	Exclusion des traitements concomitants
Définition courante de la maladie (attention parfois les essais font changer la définition des maladies, exemple IDM) *Définition clinique et biologique (sérologie) d'une maladie virale*	Définition hautement spécialisée *Diagnostic et typage virologique dans une maladie virale*
Utilisation de méthode diagnostique simple et disponible en routine *Infarctus du myocarde diagnostiqué par la clinique et l'ECG*	Utilisation de méthode hautement spécialisée (de plus haute sensibilité et spécificité) *Infarctus diagnostiqué par la scintigraphie*

Dans ce cas, il est nécessaire de disposer de deux essais : un répondant à la question générale avec exclusion de la sous-population de patients répondant potentiellement de façon différente au traitement. L'autre essai répond à la question spécifique chez ces patients. Ces deux essais peuvent être réalisés sous la forme d'un seul essai stratifié.

La difficulté de cette approche est de déterminer, a priori, s'il y a lieu de suspecter que l'effet du traitement va être différent chez certains types de patients. La tendance spontanée est d'ailleurs de surestimer la variabilité de l'effet. Il existe relativement peu de situations où une interaction forte a été mise en évidence entre l'effet et les caractéristique des patients. Par exemple, il est fréquemment avancé que les patients à haut risque ne doivent pas être étudiés avec des patients à bas risque dans un même essai. Il n'existe pourtant pas de relation directe entre le risque et le bénéfice relatif. En effet, il s'avère fréquemment que le risque relatif est identique quel que soit le risque de base (cf. ***Erreur ! Source du renvoi introuvable.****). Cependant, pour les traitements*

s'accompagnant d'effet indésirables fréquents et/ou graves la balance bénéfice-risque peut être conditionné par le risque de base des patients.

Le Tableau 3 récapitule quelques situations pour lesquelles il est licite d'envisager des variations dans l'effet traitement et la séparation de la population cible en plusieurs sous-populations étudiées séparément.

Tableau 3 – Situations pouvant faire suspecter des variations dans l'effet du traitement

- *Variations dans la pharmacocinétique du médicament entraînant suivant les cas une diminution de l'effet (en cas de diminution des concentrations) ou une augmentation des effets indésirables (en cas d'élévation de la concentration) : insuffisance rénale, inhibition ou induction enzymatique, etc…*
- *Modification, voire disparition, de la cible d'action du traitement. Par exemple, à la phase aiguë de l'accident vasculaire cérébral ischémique, une reperfusion tardive apportera probablement moins de bénéfice qu'une précoce en raison de l'évolution rapide de l'ischémie réversible vers la nécrose irréversible.*
- *Trait génétique conduisant à une plus grande susceptibilité aux effets indésirables (par l'intermédiaire de la pharmacocinétique ou de la pharmacodynamie) ou à une modification du mécanisme d'action ou de la cible du traitement (modification des récepteurs par exemple).*
- *Risque de base faible laissant la possibilité aux effets indésirables de contrebalancer le bénéfice apporté.*
- *Stade d'irréversibilité des lésions ou des processus morbides.*
- *etc.*

Au total, la population incluse dans un essai doit être représentative des patients pour lesquels existent en pratique des questions non résolues sur leur traitement ou un espoir d'amélioration de l'efficacité.

Exemple de relation entre le risque de base et l'effet du traitement

Les analyse en sous-groupe de la méta-analyse de la fibrinolytique à la phase aiguë de l'infarctus [94] fourni deux exemples illustrant parfaitement l'absence de systématisation dans la relation entre taille de l'effet et risque de base.

Aucune relation n'est trouvée entre l'efficacité de la fibrinolyse sur la mortalité à 35 jours (mesuré par l'odds ratio, OR) et la fréquence cardiaque (FC) qui est pourtant un facteur pronostique conditionnant fortement le risque de base (r_0).

FC	r_0	OR (IC95%)
<80	8,50%	0,84 (0,76 ;0,92)
80-99	11,30%	0,80 (0,71 ;0,89)
100+	20,70%	0,81 (0,71 ;0,91)
Test hétérogénéité : NS		

Par contre, l'efficacité de la fibrinolyse dépend hautement du délai de prise en charge médicale de depuis le début des symptômes qui est sans influence sur le risque de base (r_0).

Délai	r_0	OR (IC95%)
0-1	13,00%	0,70 (0,57 ;0,87)
2-6	10,70%	0,75 (0,68 ;0,84)
4-6	11,50%	0,82 (0,74 ;0,90)
7-12	12,70%	0,86 (0,77 ;0,95)
13-24	10,50%	0,95 (0,83 ;1,09)
Test hétérogénéité : p<0,05 ; tendance p=0,002		

Liste de contrôle 1 – Point à vérifier pour établir la pertinence clinique de la population incluse.

- Le risque d'événements est proche de celui habituellement rencontré chez ce type de patients.
- Les critères de sélection utilisent des méthodes couramment disponibles en routine.
- Pas de limite d'age supérieure.
- Définition opérationnelle de la maladie telle qu'utilisée en pratique et qui est à même de prendre en compte les variations de pratique.

Conséquences du regroupement de patients trop hétérogènes

Le mélange de patients différents dans un même essai n'est pas gênant tant qu'il n'existe pas de fortes modifications de l'effet du traitement entre les différents types de patients. Ce qui est le plus gênant ce ne sont pas les variations dans le risque de base mais celles dans la taille, voir le sens de l'effet du traitement (interaction). En effet, les différences de risques de bases auront simplement une répercussion sur la puissance de l'essai. Par contre l'existence d'une interaction conduit à mesurer avec un seul indicateur un effet variable d'un type de patient à l'autre

Le mélange de types de patients ne bénéficiant pas du traitement de la même manière peut rester cependant intéressant en rendant compte de l'effet "moyen" sur les patients traités en pratique de la même façon.

Populations particulières

Femmes

Les femmes sont souvent exclues des essais en raison d'éventuel risque tératologique des traitements testés. Afin d'éviter l'exposition d'éventuelle grossesse à ce type de risque, les femmes sont soit systématiquement exclues soit incluses à la conditions qu'elles n'aient pas de potentiel de procréation (ménopause, ou contraception de haut niveau de fiabilité). Ainsi, les femmes sont en général sous représentées et les risques tératologiques des traitements ne peuvent être connus qu'à travers l'observation rétrospective (pharmacovigilance ou équivalent pour les traitements non médicamenteux).

Personnes âgées

Chez la personne âgée de nombreux facteurs peuvent entraîner une modification de l'efficacité d'un traitement ainsi que de sa sécurité : altération de la pharmacocinétique, modification des mécanismes physiopathologiques, polymédicamentation, plus grande susceptibilité aux effets secondaires, etc…

Tableau 4 – Conséquences des risques compétitifs chez la personnes agées

	Sans traitement	Sous traitement
Personne âgée		
Risque cible du traitement	20%	10% (RR=0,5)
Risques compétitifs	20%	20%
Risque total	40%	30%
Réduction relative de la mortalité totale	15% (1 – 0,3/0,4)	
Personne d'age moyen		
Risque cible du traitement	20%	10% (RR=0,5)
Risques compétitifs	5%	5%
Risque total	25%	15%
Réduction relative de la mortalité totale	40% (1 – 0,15/0,25)	

De plus, les nombreux risques compétitifs existant à cet âge minorent le bénéfice d'un traitement réduisant une mortalité spécifique. Réduire par un traitement une cause de mortalité parmi d'autres n'aura pas beaucoup de conséquences sur la mortalité totale. Ainsi un traitement bénéfique chez le sujet d'âge moyen, n'aura peut-être pas les mêmes conséquences chez les sujets âgés. Le Tableau 4 illustre ce point avec un exemple où la cause de décès cible de la thérapeutique engendre le même risque de 20% quelque soit l'âge. Chez la personne âgée, les autres risques compétitifs diluent davantage le retentissement au niveau de la mortalité totale de l'effet du traitement que chez la personne d'âge moyen.

Le traitement semble apporter un bénéfice moins important chez la personne âgées que chez les sujets plus jeunes, se traduisant par une réduction relative de risque 2,6 fois plus petite. Il convient cependant de remarquer que le bénéfice absolu est le même chez les personnes âgées que chez le sujet d'âge moyen. Le traitement de 100 patients évite la survenue de 10 événements quel que soit l'âge de patients. Ainsi, il apparaît que l'augmentation du risque de base rencontrée chez la personne âgée est susceptible de compenser une baisse de l'efficacité relative du traitement. À l'opposé, l'augmentation du risque de base peut, dans d'autres cas, rendre intéressant chez les personnes âgées un traitement sans grande utilité chez le sujet d'âge moyen. Ces possibilités d'interaction mettent l'accent sur la nécessité de disposer d'essais incluant les personnes âgées.

Origines ethniques et/ou géographiques

Les essais multicentriques recrutent maintenant des patients dans de nombreux pays à travers le monde, et regroupent donc des patients issus d'origine géographique et ethnique variée, traités dans des contextes de soins différents. Ce cosmopolitisme présente de nombreux avantages : recrutement d'effectifs très importants, validation du traitement à travers un grand nombre de pratique de soins et de types de patients.

Cependant il existe des situations où l'inclusion au sein d'un même essai de patients trop différents les uns des autres altère la signification du résultat. C'est, par exemple, le cas avec des traitements pour lesquels des interactions génétiques sont connues ou fortement probables. Il en est de même quand les contextes de soins sont extrêmement différents. La recherche de variations de l'effet du traitement en fonction du pays est couramment réalisée à la recherche de ce type de problèmes mais cette approche se heurte aux limites des analyses en sous groupes. Une littérature abondante existe sur ce sujet.

Balance bénéfice risque

Introduction

Les conséquences sur l'organisme d'un traitement, médicamenteux ou d'autre nature, ne sont jamais exclusivement bénéfiques, mais s'accompagnent d'effets indésirables, plus ou moins sévères, plus ou moins intenses ou fréquents. Il n'est guère d'exemple de traitement largement usité, qui n'ait, un jour ou l'autre, provoqué un effet fâcheux, parfois sérieux. La prise de décision doit donc mettre en balance les effets bénéfiques et les effets négatifs. Les questions qui surgissent à ce niveau sont : « Est-ce que le risque lié aux effets indésirables est acceptable compte tenu de la maladie traitée ? », « les effets indésirables ne contrebalancent-ils pas la totalité du bénéfice ? ».

Classiquement cette problématique s'aborde par l'étude du rapport bénéfice/risque. Mais, malgré ce que pourrait faire penser sa dénomination, ce rapport est difficilement quantifiable. La plupart du temps, son évaluation est discursive, même si elle est basée sur des données numériques, car bénéfice et risque ne sont pas de la même nature, et il n'est pas possible de les confronter directement.

Tableau 1 – Quelques situations où bénéfice et effets indésirables sont de natures différentes

Situations	Bénéfice	Effets indésirables
Oméprazole dans ulcère duodénale	Augmentation du taux de cicatrisation	Troubles neuropsychiatrique, rares troubles hématologiques
Statines dans la prévention secondaire des maladies cardiovasculaire	Réduction de mortalité et de morbi-mortalité	Myalgies parfois associées à une rhabdomyolyse, se compliquant exceptionnellement d'insuffisance rénale aiguë
IEC insuffisance cardiaque	Réduction de mortalité et de morbi-mortalité	Toux, rare œdème de Quincke, augmentation modérée de la créatinine

À ce niveau, le but de l'interprétation des résultats est de déterminer si la balance bénéfice risque est favorable ou non au traitement. Le raisonnement et les éléments d'interprétation vont être différents en fonction de la gravité relative des événements indésirables par rapports à la maladie ou à la nature du bénéfice apporté.

Effets indésirables de gravité supérieure à la maladie

Lorsque les événements indésirables sont de gravité bien supérieure à la pathologie traitée, la balance bénéfice/risque sera défavorable lorsque le risque encouru est disproportionné par rapport au bénéfice apporté. C'est le cas, par exemple, de la survenue de syndromes de Lyell potentiellement mortels avec la prise d'AINS pour soulager des traumatismes bénins. Par contre, avec des chimiothérapies anticancéreuses, un risque faible de syndrome de Lyell est acceptable compte tenu de la gravité de la maladie et du bénéfice attendu sur la mortalité.

En fait, la gravité des événements indésirables est à confronter, non pas à la gravité de la maladie, mais à la nature du bénéfice. Ainsi, la survenue de syndromes de Lyell avec un antiémétique est peu acceptable même dans un contexte oncologique.

Tableau 2 – Quelques situations où le risque lié aux effets indésirables a été jugé excessif par rapport au bénéfice apporté.

Situation	Risque lié aux effets indésirables	Bénéfice prouvé
Alosetron dans le colon irritable [1]	Infarctus mésentérique et occlusion intestinale (dont certains cas mortels). Ce traitement a été retiré de la commercialisation 9 mois après son introduction sur le marché aux USA à la suite du décès de 5 patients	Diminution de l'inconfort et des douleurs abdominales
Troglitazone dans le diabète de type 2	Insuffisance hépatique conduisant dans certains cas au décès ou à la transplantation. Le troglitazone a été retiré du marché 3 ans après sa commercialisation à la suite de 90 cas d'insuffisance hépatique dont 70 mortels ou ayant nécessités une transplantation.	Diminution des taux d'hémoglobine glycosylée (pas d'effet prouvé sur les complications du diabète)

Cette évaluation repose sur un jugement de valeur et va donc dépendre de l'appréciation qui est faite de la gravité des événements indésirables et de l'importance de leur fréquence de survenue. Cette appréciation repose sur des échelles de valeur qui peuvent être différentes suivants les médecins, les patients et qui dépendent des choix sociétaux. Ces éléments expliquent les différences d'appréciation des mêmes données qui se rencontrent parfois dans l'évaluation de la balance bénéfice risque.

Effets indésirables de même nature que la maladie

Certains événements indésirables sont de même nature que les événements que le traitement cherche à prévenir. Ils sont donc susceptibles de contrebalancer en partie ou en totalité le bénéfice apporté par le traitement (cf. Tableau 3).

C'est, par exemple, le cas des décès par hémorragie cérébrale induit par l'utilisation des fibrinolytiques à la phase aiguë de l'infarctus du myocarde. Dans ce cas le critère de jugement utilisé pour l'efficacité, le décès, prends aussi en compte les effets délétères et évalue directement la balance bénéfice risque. Par exemple, avec la streptokinase comparée au placebo, la mortalité totale est réduite de 23% ce qui permet de dire que, même si il existe des effets délétères mortels, ceux-ci sont quantitativement minoritaire à coté des effets bénéfiques du traitement. Cette situation est la plus propice à une tentative de quantification du rapport bénéfice risque (cf. section suivante).

Fréquemment, les événements indésirables graves s'accompagne d'un cortège d'événements plus bénins mais qui entraînent une gêne substantielle des patients. Ces deux dimensions doivent être prise en compte dans la décision.

Tableau 3 – Exemples de situations où les effets indésirables sont de même nature que le bénéfice apporté par le traitement

Situation	Effets indésirables	Bénéfice
Aspirine en prévention des maladies coronariennes ischémiques (American Physicians' Health study) [2]	Augmentation de la fréquence des accidents cérébraux hémorragiques	Diminution de la fréquence des événements coronariens (infarctus)
Œstrogènes et hypercholestérolémie (CDP) [3]	Augmentation de la fréquence des thromboses veineuses	Diminution de la fréquence des événements coronariens (infarctus)

Les effets indésirables bénins posent des problèmes identiques lorsque le bénéfice apporté par le traitement est de même nature. Par exemple, dans le syndrome d'instabilité vésicale, un traitement qui améliorerait la symptomatologie en diminuant le nombre de mictions, mais qui s'accompagnerait d'effets indésirables fréquents de type anti-cholinergique (sécheresse buccale et oculaire, nervosité, somnolence, etc.) aurait une balance bénéfice risque défavorable. Pour un grand nombre de patients, le désagrément induit par le traitement serait analogue à celui qui est soulagé par le traitement. Bien entendu cette appréciation est susceptible de varier en fonction du patient, suivant son vécu de la symptomatologie urinaire.

Synthèse

Au total, la balance bénéfice risque est défavorable dans les situations suivantes dont les principales caractéristiques sont récapitulées dans le Tableau 5 :

- la gravité des événements indésirables fait courir un risque disproportionné par rapport à la gravité de la maladie ou par rapport à la nature du bénéfice apporté,
- la fréquence des événements indésirables contrebalance le bénéfice.

Des facteurs externes rentrent aussi en ligne de compte dans l'appréciation de la balance bénéfice risque (Tableau 4), comme :

- les risques inhérents aux autres traitements déjà existant : la balance bénéfice/risque d'un nouveau traitement sera jugé défavorable s'il est moins bien toléré que les traitements déjà disponibles, même si dans l'absolu, la balance bénéfice risque pourrait être acceptable ou si l'efficacité du nouveau traitement est légèrement supérieure à celle des autres,
- la différence de perception entre la personne exposée au risque, la victime potentielle, et le décideur (cf. tableau ci-dessous). La perception du risque par les patients vient donc s'interposer entre l'évaluation de la balance bénéfice risque et la décision qui doit tenir compte de ce filtre déformant.
-

Tableau 4 – Différence de perception de la nature du risque entre le décideur et la victime

	Analyse du décideur	Analyse de la victime
Nature du bénéfice	Social, donc général	Particulier, donc individuel
Nature du détriment	Risque, donc probabilité	Danger, donc intolérable

Mesure du rapport bénéfice/risque

Plusieurs approches sont possibles pour essayer d'appréhender numériquement la balance bénéfice-risque, c'est-à-dire de calculer le rapport bénéfice/risque. Leur objectif est de fournir un seul indice numérique intégrant les aspects positifs et négatifs d'un traitement sur lequel pourrait se baser la décision. Quelle que soit la voie suivie, l'obtention d'un tel indice se heurte au même problème : celui de pondérer les effets positifs et

négatifs en fonction de la gravité des événements qu'ils concernent. Cette pondération débouche sur des choix liés à des échelles de valeur, dépendant de nombreux paramètres externes (gravité, fréquence, réversibilité, coûts induits, etc…). Ces choix s'avèrent donc arbitraires et dépendants d'un contexte particulier.

Tableau 5 – Les deux situations dans lesquelles les effets indésirables compromettent la balance bénéfice risque (BBR).

Événements graves et non spécifiques de la maladie	**Événements fréquents peu spécifiques**
Très difficile à détecter au cours des essais cliniques	Mise en évidence par **comparaison à un groupe contrôle** (analyse de sécurité d'un essai d'efficacité).
Détectable en pharmacovigilance, avec quantification à l'aide **d'études d'observation.**	Indétectable en pharmacovigilance (en raison d'un bruit de fond important)
Mécanisme souvent inconnu	soit conséquence de l'effet thérapeutique (ex AVC et streptokinase), soit effet secondaire de mécanisme plus ou moins connu (ex AINS et effet indésirables gastro intestinaux)
Rendent la BBR défavorable quand ils font courir un risque disproportionné par rapport à la gravité de la maladie ou au bénéfice apporté, ou quand ils sont plus fréquents qu'avec les médicaments concurrents de même efficacité	Rendent la BBR défavorable quand : - le produit fréquence × gravité annule le bénéfice - une fréquence excessive compromet l'observance (ex constipation avec dose efficace de cholestyramine)

Critères composites

Le principe des critères composites pour la mesure du rapport bénéfice/risque est de regrouper les événements positifs et les événements négatifs au sein d'un même critère.

À la phase aiguë de l'infarctus, la comparaison de l'alteplase (t-PA) au fibrinolytique de référence, la streptokinase, a fait appel à un critère combiné regroupant les décès et les accidents vasculaires cérébraux (AVC) invalidants (Tableau 6) [4]. Ce critère prend en compte simultanément le bénéfice du traitement, c'est-à-dire la réduction du nombre de décès, et les effets délétères les plus graves représentés par les AVC consécutifs aux hémorragies cérébrales induites par la fibrinolyse. Ce critère intègre à la fois la diminution du nombre de décès et l'augmentation du nombre d'AVC induits. Les résultats obtenus sont les suivants.

Tableau 6 – Exemple de critère composite regroupant à la fois une dimension d'efficacité (décès) et de sécurité (AVC invalidant).

	Alteplase n=10 344	streptokinase n=20 173	p
Décès ou AVC invalidant	7,8%	6,9%	p=0,006
Décès	7,3%	6,3%	p=0,04
AVC invalidant	0,5%	0,6%	NS

Une réduction de ce critère composite s'interprète comme une réduction du nombre de décès sans AVC invalidants. Le surcroît d'AVC invalidants ne contrebalance pas totalement la réduction de mortalité.

L'addition des effets positifs et négatifs sous-entend une pondération de 1 pour 1 : un événement délétère a la même valeur qu'un événement bénéfique. Le bénéfice apporté par un événement évité est entièrement effacé par la survenue d'un événement délétère.

Dans l'exemple de la fibrinolyse, il a donc été considéré que la gravité d'un AVC invalidant est la même que celle d'un décès. S'il survient autant d'AVC invalidants que de décès évités, le traitement sera déclaré sans intérêt. Cette pondération reflète une certaine échelle de valeur : être victime d'un AVC invalidant équivaut à décéder. Cette échelle de valeur va aussi transparaître dans la définition de ce qu'est un AVC invalidant.

L'obtention d'une réduction importante du critère de jugement composite rend moins critique l'arbitraire de la pondération, le résultat s'avérant peu sensible à la pondération utilisée.

Rapport des "bénéfices" absolus

Il a été proposé de recourir au bénéfice absolu (et au NNT) pour effectuer une confrontation numérique du bénéfice et du risque lorsque les effets indésirables et les événements cibles de la thérapeutique sont de nature similaire [5].

Dans l'exemple présenté dans le tableau suivant, le traitement de 1000 patients permet d'éviter 36 événements cibles tandis qu'il engendre 12 événements indésirables. Au total, le rapport bénéfice/risque se traduit par une réduction net de 36-12=24 événements cibles sans événements indésirables.

	Groupe traité n=1 000	**Groupe contrôle n=1 000**	**Bénéfice / maléfice absolu**
Effet + événement cible	156	192	-3,6 %
Effet – événement indésirable	35	23	+1,2 %

De plus il est possible de calculer le rapport NNH/NNT qui quantifie en moyenne le nombre d'événements évités pour un événement indésirable induit. Il s'agit de l'importance relative des effets indésirables par rapport aux effets bénéfiques. Avec les résultats de l'exemple, le calcul donne les résultats suivants :

Effet + (événement cible)	*NNT = 27,7*	*NNH/NNT = 3*	*En moyenne, un événement indésirable survient pour 3 événements cibles évités par le traitement.*
Effet – (événement indésirable)	*NNH = 83,3*		

Exemple

Dans ISIS 2, la fréquence des hémorragies majeures durant l'hospitalisation passe de 18/8491 (0,21%) sous placebo à 46/8490 (0,54%) sous streptokinase, soit un surcroît de 0,33% en différence absolue. Le NNH est de 1/0,0033=303. En moyenne, une hémorragie supplémentaire survient durant le séjour hospitalier tous les 303 patients traités. En comparaison, la réduction de mortalité est de 2,77% en absolu (9,2% vs 12,0%) ce qui correspond à un NNT=36. En moyenne, une hémorragie majeure survient tous les 303/36=8,4 décès évités.

Cette approche de mesure du rapport bénéfice/risque est en fait proche de la précédente basée sur l'utilisation d'un critère composite. Les bénéfices absolus obtenus sur chacune des composantes sont additionnés à la place des événements dans un même critère composite.

De plus, contrairement à l'utilisation d'un critère composite, la confrontation des estimations du NNT et du NNH n'intègre pas l'imprécision de ces estimations et conduit à raisonner sans tenir compte de leur intervalle de confiance.

Estimation du rapport bénéfice/risque en fonction du risque de base

Interaction arithmétique

Le bénéfice absolu dépend du risque de base.

Pour un traitement donné, le risque relatif s'avère plutôt constant à travers les essais, même s'ils ont été réalisés dans différentes populations [6]. En méta-analyse, cette constance du risque relatif donne un sens au regroupement de plusieurs essais. D'une manière générale, c'est aussi en raison de cette relative stabilité des effets des traitements qu'il est justifié d'extrapoler le résultat des essais aux patients futurs.

Arithmétiquement le bénéfice absolu apporté par un traitement est d'autant plus important que le risque de base est élevé.

Un des paramètres qui distinguent les populations de patients de différents essais est le risque de base. Des essais ont inclu des patients à haut risque et d'autres des patients à faible risque. De ce fait, même si le bénéfice relatif du traitement est le même dans tous ces essais, les conséquences du traitement, appréhendées, par exemple, par le bénéfice absolu, vont être différentes. En effet, le bénéfice absolu sera d'autant plus grand que le risque de base est élevé.

La différence de risque DR *se calcule à partir du risque relatif* RR *et du risque de base* r_0 *par la formule :*

$$DR = r_0 - r_0 \times RR$$
$$= r_0 (1 - RR)$$

La différence de risque est donc directement proportionnelle au risque de base. Ainsi, le bénéfice absolu apporté par un traitement varie en fonction du risque de base en dehors de toute variation de l'effet du traitement lui-même. Sans aucune variation de l'effet du traitement, les sujets à haut risque tirent plus de bénéfice d'un traitement que les sujets à bas risque.

Cette situation où le risque relatif est constant quel que soit le risque de base se retrouve fréquemment. Dans une analyse de 112 méta-analyses, Schmid et al. trouve que le risque relatif ne varie avec le risque de base que dans 13% des cas tandis qu'une relation est trouvée avec la différence des risques dans 31% des cas [6]. L'hypothèse d'un risque relatif constant est donc en général raisonnable et peut être testée en méta-analyse grâce à des outils de méta-régression [7-9].

Il existe cependant quelques exemples où le risque relatif est différent chez le patient à haut ou à faible risque. L'effet des inhibiteurs de l'enzyme de conversion dans le traitement de l'insuffisance cardiaque congestive semble plus important chez les patients avec un risque annuel de mortalité supérieur à 15% (RR=0,64, IC95% =[0,51 ;0,81]) que chez les sujets à bas risque (RR=0,88, IC95%=[0,80 ;0,97]). Ce résultat est obtenu par la méta-analyse de 39 essais dont 28 ont inclus des patients à bas risque et 11 des patients à haut risque [10].

Conséquence pour la décision thérapeutique

Cette interaction arithmétique a des répercussions au niveau de la décision de traiter les patients [11]. Avec un traitement ayant démontré sont efficacité, une abstention thérapeutique pourra être justifiée chez les patients à très faible risque en raison du très faible bénéfice absolu attendu chez eux. La petitesse du bénéfice à attendre du traitement doit alors être confrontée aux désagréments subit par le patient (astreinte à suivre un traitement, etc.), au coût du traitement et, surtout, au risque d'effets indésirables. Les patients à très faible risque sont exposés au risque d'effets indésirables pour un bénéfice attendu quasi nul.

Ceci étant d'autant moins acceptable que les effets indésirables sont potentiellement graves.

Par exemple, une complication grave de la fibrillation auriculaire, non due à une pathologie rhumatismale, est l'embolie cérébrale qui provoque un accident vasculaire cérébral plus ou moins invalidant et entraînant assez fréquemment le décès. Les anticoagulants oraux réduisent le risque de survenue de ces embolies, mais au prix d'un risque induit d'hémorragies cérébrales. Des essais thérapeutiques ont évalué l'efficacité de cette approche thérapeutique. Dans ces essais, le risque de base d'accidents cérébraux vasculaires ischémiques est variable et réduit de façon proportionnelle par le traitement. Par contre, le risque sous traitement d'hémorragies cérébrales est relativement fixe d'un essai à l'autre. Ainsi, le risque global d'accidents vasculaires cérébraux (ischémiques et hémorragiques) se décompose en deux composantes : ischémique modifiée de façon multiplicative par le traitement et iatrogène dont le risque sous traitement est constant quel que soit le risque de base d'accidents vasculaires cérébraux (en presque totalité d'origine ischémique chez ces patients). Ces différences dans la forme de ces deux modèles d'effet s'expliquent par le fait que ces deux effets ont des mécanismes et des sites d'action différents. L'effet sur les accidents cérébraux ischémiques proviendrait d'une action sur les processus

thrombo-emboliques au niveau de l'atrium gauche, processus qui conditionnent directement le risque de base de ces malades. Par contre, la composante iatrogène découlerait d'une action sur un site différent (paroi vasculaire cérébrale) et le risque de survenue de ces complications hémorragiques est indépendant du risque thromboembolique cardiaque.

Il est possible de représenter graphiquement ces différents effets du traitement anticoagulant oral sur un graphe où le risque spontané est en abscisse et le risque sous traitement en ordonnée. Ce type de graphe est appelé graphique de modèle d'effet [12] ou graphique de L'Abbé.

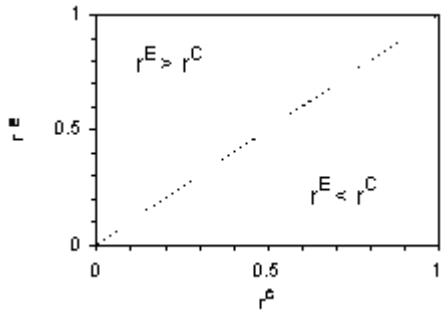


Figure 1 – Graphique de modèle d'effet.

Ce graphe est scindé en deux zones par la première bissectrice. Dans la zone située en dessous de la bissectrice, le risque sous traitement est inférieur au risque de base $r_1 < r_0$*, et correspond à un effet bénéfique. Les points situés sur la bissectrice sont tels que* $r_0 = r_1$ *et représentent l'absence d'effet traitement, tandis que la zone au-dessus de la bissectrice matérialise les effets délétères (où* $r_1 > r_0$*). Sur ce graphe, les résultats des essais peuvent être représentés par des points de coordonnées (*$x = r_0$, $y = r_1$*).*

L'action des anticoagulants oraux sur la survenue des accidents vasculaires ischémiques diminue leur risque de survenue de façon proportionnelle (risque relatif). Cet effet est représenté graphiquement par une droite issue de l'origine est dont la pente est égale au risque relatif.

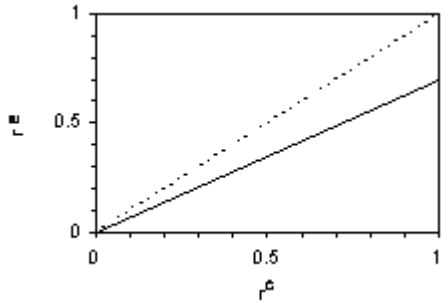


Figure 2 – Modèle d'effet multiplicatif.

Par contre, le risque délétère d'accident vasculaire cérébral d'origine hémorragique se traduit par une droite parallèle à la bissectrice et située au-dessus d'elle. En effet, ce risque délétère est constant quel que soit le risque de base d'AVC ischémique et s'ajoute donc à celui-ci.

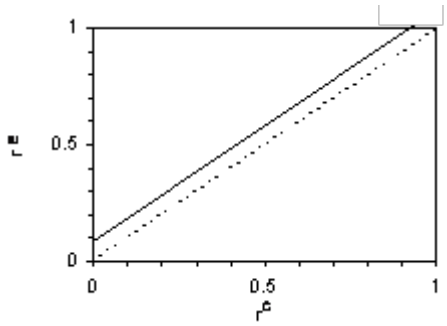


Figure 3 – Modèle d'effet additif.

Au total, l'effet net des anticoagulants oraux est la somme des effets bénéfiques et des effets délétères. Il se traduit par une droite coupant la bissectrice. Cette intersection définit alors un seuil de risque de base délimitant deux régions où l'effet du traitement anticoagulant oral sur le risque d'AVC est totalement différent.

En dessous de ce seuil, le traitement est globalement délétère car les effets indésirables du traitement sont plus importants que le bénéfice qui est limité du fait d'un risque spontané faible. Au-dessus du seuil, le bénéfice devient plus important que les effets délétères et globalement le traitement réduit la fréquence des AVC.

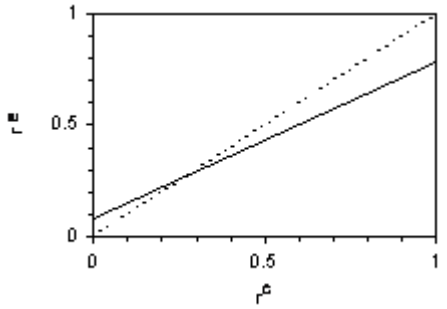


Figure 4- Modèle d'effet mixte.

Cette propriété est fondamentale, car elle conduit au fait qu'un même traitement peut être à la fois bénéfique ou délétère suivant le risque de base rattaché à la situation dans laquelle il est utilisé. Ainsi, disparaît tout manichéisme dans l'effet d'un traitement. Cette propriété débouche aussi sur le constat dans une telle situation de l'insuffisance des mesures simples qui peuvent ne refléter qu'un aspect de l'effet du traitement. Ainsi, il n'est plus possible de se contenter d'un seul indice pour caractériser l'effet du traitement mais de deux. De même, l'effet dont pourra bénéficier un patient donné n'est plus absolu mais dépend de son risque initial.

L'utilisation de la warfarine pour la prévention des accidents vasculaires cérébraux dans la fibrillation auriculaire produit une réduction du risque d'AVC de 73% mais s'accompagne d'une augmentation absolue du risque d'hémorragies cérébrales mortelles de 3%[11]. Les données de suivi de SPAF permet aussi de déterminer le risque annuel d'AVC en fonction du nombre de facteurs de risque que présentent les patients : hypertension, insuffisance cardiaque, antécédent thromboembolique, dysfonction ventriculaire gauche, taille de l'atrium augmentée [13].

Le Tableau 7 compare le bénéfice absolu au risque hémorragique en fonction du nombre de facteurs de risque. Pour un patient ne présentant aucun facteur de risque, la warfarine évite annuellement en moyenne 4,4 AVC pour 1000 patients traités, mais induit 30 hémorragies cérébrales mortelles. La balance bénéfice risque est donc très déséquilibrée. Par contre, pour les patients présentant deux ou trois facteurs de risques, la warfarine permet annuellement en moyenne d'éviter 91 AVC pour 1000 patients en engendrant 30 hémorragies. La balance bénéfice risque s'avère alors positive avec 3 AVC évités pour une hémorragie induite.

Tableau 7 – Comparaison du bénéfice absolu au risque hémorragique en fonction du nombre de facteurs de risque.

Nombre de facteurs de risque	Risque de base d'AVC	Bénéfice absolu AVC	Risque absolu d'hémorragies
0	0,6%	4,4/1000	30/1000
1	4,8%	35/1000	30/1000
2 ou 3	12,5%	91/1000	30/1000

1. *Camilleri M, Nothcutt AR, Kong S, Dukes GE, McSorley D, Mangel AW. Efficacy and safety of alosetron in women with irritable bowel syndrome: a randomised, placebo-controlled trial. Lancet 2000;355:1035-40.* PMID:

2. *Final report on the aspirin component of the ongoing Physicians' Health Study. Steering Committee of the Physicians' Health Study Research Group. NEJM 1989;321(3):129-35.* PMID:

3. *The Coronary Drug Project. Findings leading to discontinuation of the 2.5-mg day estrogen group. The coronary Drug Project Research Group. JAMA 1973;226(6):652-7.* PMID:

4. *The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. NEJM 1993;329:673-682.* PMID:

5. *McQuay HJ, Moore AR. Using numerical results from systematic review in clinical practice. Anns of Internal Medicine 1997;126:712-20.* PMID:

6. *Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med 1998;17:1923-42.* PMID:

7. *Walter DD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. Stat Med 1997;16:2883-2900.* PMID:

8. *McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. Stat Med 1996;15:1713-1728.* PMID:

9. *Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. BMJ 1996;313:735-738.* PMID:

10. *North of England Evidence-based Guideline Development Project. ACE inhibitors in teh primary management of adults with symptomatic heart failure. Newcastle-upon-Tyne: Centre fro Healthh Services Research; 1997.*

11. *Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. BMJ 1995;311:1356-1359.* PMID:

12. *Cucherat M, Boissel JP. Modèles d'effet et méta-analyse. Thérapie 1997;52:13-18.* PMID:

13. *SPAF investigators. Warfarin versus aspirin for prevention of thromboembolism in atrial fibrillation: stroke prevention in atrial fibrillation II study. Lancet 1994;343:687-691.* PMID: