Press therm climat 2005;142:75-99

COMMENT PEUT-ON JUGER DE LA VALIDITÉ D'UN ESSAI THÉRAPEUTIQUE ?

Revue de la littérature et réflexions sur la conception pratique des grilles de lecture dans les traitements physiques[1]

R FORESTIER[2], A FRANÇON[2], B GRABER-DUVERNAY[2]

Résumé – L'objectif de ce travail est de déterminer, à l'aide d'une revue de la littérature, les fondements de la validité des essais thérapeutiques et des grilles de lecture qui servent à les juger dans les revues systématiques et les méta-analyses.

Méthode : Recherche à l'aide de mots-clefs et par nom d'auteurs sur la base de données Medline puis par analyse de la bibliographie des articles identifiés comme pertinents. *Résultats :* Dans certains domaines : le recueil du consentement éclairé, la comparaison de la publication avec le protocole, la procédure de sélection des patients, la procédure de randomisation, le calcul du nombre de sujets nécessaires, les abandons et les perdus de vue, l'insu des patients et des examinateurs, une partie des modes de comparaison (à un placebo et à un traitement de référence), l'ajustement du seuil de probabilité (au nombre de mesures et de critères) et l'analyse en intention de traiter une influence est démontrée sur les résultats des essais thérapeutiques. Dans d'autres domaines, nous n'avons pas trouvé de démonstration scientifique, même si leur effet est très probable : les contraintes des traitements non médicamenteux sont particulières.

Conclusion : La validité des grilles de lecture quantitatives est faible en l'absence de démonstration scientifique probante de leur pertinence. Il est probable que des grilles spécifiques sont nécessaires pour juger les essais non médicamenteux.

Mots-clés : Méthodologie, validité, essais thérapeutiques, grilles de lecture

HOW CAN WE JUDGE CLINICAL TRIAL VALIDITY. Review and consequences on the practical conception of scale in physical therapy.

Key-Words: Methodology, validity, clinical trial, checklist

[1] Les éléments de ce travail seront retrouvés dans un article des mêmes auteurs paru dans les Annales de Réadaptation et de médecine physique 2005;48:250-8 sous le titre : Les paramètres de validité d'un essai thérapeutique et leur influence sur l'élaboration d'une médecine fondée sur les preuves

[2] Centre de recherche rhumatologique et thermal, BP 234, 73100 Aix-les-Bains Cedex

Introduction

À l'ère de la médecine fondée sur les preuves, les autorités sanitaires et les experts scientifiques fondent de plus en plus leur jugement sur les revues systématiques et les méta-analyses d'essais thérapeutiques.

Ces travaux scientifiques, extrêmement longs à mener et dont la méthodologie comporte de grands efforts de rigueur, utilisent, pour sélectionner les essais thérapeutiques, des grilles de lecture fondées sur les connaissances que nous avons des différents biais méthodologiques. Les plus anciennes remontent au début des années soixante et se présentent comme des listes de cases à cocher ou check-lists. Ce sont les grilles qualitatives. On voit maintenant apparaître, depuis le début des années quatre-vingt, des grilles quantitatives qui donnent une note à chaque essai qu'elles sont chargées de juger [1]. Elles ont été utilisées dans de nombreux travaux scientifiques, y compris sur les traitements physiques [2-8]. La pratique devient tellement courante qu'il existe maintenant des grilles de lecture pour vérifier la validité des méta-analyses et des revues systématiques [9].

Si le principe de leur utilisation n'est guère contestable, il n'est pas certain qu'en pratique ces grilles soient toutes fondées sur des démarches scientifiques aussi rigoureuses que les études qu'elles sont chargées d'évaluer. En outre, les différentes thérapeutiques ne comportent pas toujours les mêmes contraintes d'évaluation. Il pourrait donc apparaître une dérive qui consisterait à faire un amalgame entre la valeur des essais thérapeutiques dans un domaine donné et la valeur de la thérapeutique elle-même. Ceci pourrait avoir des conséquences particulières en rhumatologie où nous utilisons couramment des méthodes thérapeutiques non médicamenteuses dont les contraintes d'évaluation semblent assez différentes des traitements chimiques et qui, pour cette raison, pourraient être moins bien « notées ».

Le but de ce travail est d'examiner, à l'aide d'une revue de la littérature, comment il est possible de juger la validité d'un essai thérapeutique par la qualité de sa présentation, par les options méthodologiques qu'ont choisies ses concepteurs et par les conditions de son déroulement. Une attention particulière sera portée dans les domaines où les essais médicamenteux et non médicamenteux diffèrent. Nous examinerons ensuite quelles sont les conséquences de ces constatations sur la conception des grilles de lecture.

Nous espérons que cette réflexion permettra de porter un jugement plus pertinent sur les essais thérapeutiques et qu'elle aidera à concevoir des grilles de lecture qui prennent en compte les contraintes particulières qui s'exercent dans les essais non médicamenteux.

Nous partirons du principe que le but idéal d'un essai thérapeutique est de déterminer ce que le traitement évalué apporte à des patients dans des conditions courantes d'utilisation pour son prescripteur.

Matériel et méthode

La recherche bibliographique a utilisé la base de données Medline avec les mots clés : méthodologie, grilles de lecture, revue systématique, biais, validité interne et validité

externe, analyse en intention de traiter. Nous avons ensuite lu les titres des différentes études sur la méthodologie afin de voir si elles étaient en rapport avec notre recherche. Lorsqu'ils étaient disponibles, les articles ont été analysés en totalité. Dans le cas contraire, nous avons simplement utilisé le résumé. Nous avons alors signalé les références par un * dans la bibliographie. Nous avons ensuite étendu la recherche à divers noms d'auteurs : Chalmers, Altman, Moher, Jadad, Sackett, Guyatt, Oxmann, Colditz, Tugwell, Jüni, Schulz. L'analyse de la bibliographie des articles identifiés a permis d'ajouter quelques références pertinentes.

Résultats

Pour plus de clarté, nous avons suivi le plan habituel de rédaction des articles scientifiques portant sur l'évaluation d'un traitement.

Pour les lecteurs non avertis, rappelons que la validité interne d'un essai correspond à l'intensité de la relation de cause à effet entre le traitement étudié et l'amélioration observée chez les patients. Elle pourrait être assimilée à la rigueur de la démonstration scientifique. La validité externe traduit les possibilités d'extrapolation à la population générale [10]. On a déjà observé depuis longtemps qu'elles étaient difficiles à concilier et qu'augmenter l'une conduisait souvent à diminuer l'autre [11].

La qualité globale d'un essai a une influence variable sur le résultat de celui-ci. Ainsi, dans un recueil de 5 méta-analyses comparant des essais de bonne qualité à des essais de faible qualité, on constate que l'effet du traitement étudié était sous-estimé deux fois sur cinq et surestimé une fois sur cinq. Les interventions avaient un effet similaire une fois sur cinq et dans un cas, aucune conclusion n'était possible [12]. Curieusement, un autre auteur constatait, à la même époque, que les essais de faible qualité surestimaient les effets du traitement de 34% [13]. Enfin, un dernier auteur a étudié la corrélation entre la qualité méthodologique des essais et leur résultat : elle n'était pas significative [14]. Il y a donc des données contradictoires sur la façon dont la validité globale d'une étude peut influer sur son résultat.

Les domaines pour lesquels une influence est démontrée sur le résultat de l'étude sont résumés dans le tableau 1. Une liste non limitative des domaines qui restent à explorer est proposée dans le tableau 2.

Rubriques

« Titre et résumé »

S'il s'agit d'une étude comparative, et si les groupes ont été obtenus par randomisation, il est important que le mot figure en toutes lettres dans le titre. Ceci permet au lecteur d'identifier immédiatement l'article comme un essai thérapeutique randomisé.

Le résumé devra en principe refléter l'ensemble des constatations avec honnêteté sans privilégier les résultats les plus significatifs afin de permettre au lecteur de se faire une idée exacte de l'effet du traitement évalué [15].

Nous n'avons pas trouvé de travaux méthodologiques évaluant l'impact d'un résumé dont les résultats seraient discordants avec les résultats de l'étude. Il est probable qu'un tel résumé est susceptible de modifier l'attitude thérapeutique d'un lecteur qui ne lirait pas l'article en totalité.

« Introduction »

Il est d'usage de détailler les différentes hypothèses et la justification de l'étude. Il est également normal de donner les raisons du choix de la dose ou du mode d'administration du traitement.

« Méthode »

Consentement éclairé et comité d'éthique

Il est souhaitable et, de toute façon, obligatoire depuis la déclaration d'Helsinki, de déposer un protocole auprès d'un comité d'éthique lorsqu'on réalise un essai comparatif. Le recueil d'un consentement éclairé et de l'avis d'un comité d'éthique est de plus en plus souvent rapporté dans les articles mais cela reste inconstant même dans les travaux les plus récents [16]. Un auteur a observé que le recueil d'un consentement éclairé, avant un essai randomisé, diminuait la différence entre le placebo et le produit actif [17]. Lorsque de nombreux patients éligibles pour un essai thérapeutique refusent d'être randomisés, il risque d'apparaître un biais de non-consentement qui limite la validité externe de l'étude. Pour le mesurer, un auteur a proposé de continuer à évaluer, en parallèle, les patients qui ont refusé de se soumettre à la randomisation et ceux qui l'ont acceptée [18]. Ainsi, il pense pouvoir déceler une éventuelle différence dans les valeurs initiales et dans l'évolution qui donnerait une meilleure idée de la validité externe de l'étude.

Il serait par ailleurs souhaitable que le comité d'éthique assure le suivi des protocoles qui lui ont été soumis afin de lutter contre la non-publication de données défavorables aux traitements évalués lorsque les essais sont financés par leurs fabricants. La mise à la disposition du public d'un registre des protocoles déposés auprès des comités d'éthique, permettrait également de diminuer le biais de publication qui fausse les résultats des méta-analyses. À l'aide d'une méthode empirique très simple, un auteur a estimé que certaines méta-analyses Cochrane (réputées pour être les plus rigoureuses) ne recueillaient pas toutes les études pertinentes (26 sur 48). Dans dix cas sur 48, le nombre d'études manquantes était significatif. Dans quatre cas sur 48, ce biais avait modifié l'effet de l'intervention [19]. Un autre a comparé les essais de 135 méta-analyses dont 32 incluaient la « littérature grise » (essais non-publiés ou publiés avec une diffusion limitée) [20]. Les essais publiés mesuraient un effet traitement de 15% supérieur à celui de la littérature grise (ratio des odds-ratio : 1,15 ; intervalle de confiance : 1,04 à 1,28). Il concluait que l'exclusion de la littérature grise conduisait à surestimer l'effet des interventions thérapeutiques.

Populations étudiées : sélection, restrictions et homogénéité des patients

Les conditions de recrutement sont déterminantes pour la validité externe d'un essai.

• Description des critères d'inclusion

La population étudiée (avec une certaine pathologie ou symptôme) est définie par des critères d'inclusion et d'exclusion.

Il est souhaitable que les critères d'inclusion soient validés par des études préalables ou, à défaut, qu'ils aient fait l'objet d'un consensus d'experts. En dernier recours on se contentera des critères d'inclusion consacrés par l'usage.

Ces critères d'inclusion doivent éviter de constituer une sélection de patients présentant uniquement la forme typique de la maladie. On ne validerait alors aucun traitement pour les formes atypiques ou limites qui sont fréquentes en pratique. Que penser, en effet, d'un traitement dont l'efficacité a été validée sur seulement 1% des sujets potentiellement éligibles en raison de critères d'inclusion et exclusion trop restrictifs ?

Pour sélectionner une population représentative, l'idéal serait d'inclure la totalité des patients présentant la maladie ou symptôme (inclusion de sujets consécutifs). Si cela est impossible en pratique, on peut procéder à un tirage au sort, mais il est alors indispensable de décrire avec précision la procédure de tirage au sort afin de pouvoir vérifier qu'elle est réellement aléatoire.

Il est démontré que l'exclusion d'une part importante des patients dans une étude conduit à des populations non-représentatives [21-22]. Dans une revue sur les interventions de santé publique, un auteur a constaté que les études rapportaient de façon inconstante la représentativité de la population sur laquelle avait été effectuée l'intervention [23].

Une méta-analyse, réalisée sur 172 essais cliniques randomisés publiés dans quatre revues scientifiques à *impact factor* élevé, a étudié le recrutement des sujets [24]. On ne pouvait calculer la proportion de patients enrôlés (sur le nombre total de patients éligibles) que dans 49 essais. Cette proportion était de 64,6% en moyenne (41 à 82%). La raison la plus souvent observée de non-inclusion était le refus de participer (86%), les abandons (5,8%), l'aggravation clinique. Sur les patients enrôlés, la proportion de patients réellement recrutés pour l'étude était de 54% (32 à 77%). Au total, le nombre moyen de patients à examiner pour en recruter un, allait de 1,8 à 68 selon le type d'essais… À l'opposé, 20 études déclaraient avoir recruté 100% des patients éligibles mais celles-ci excluaient de leurs calculs les patients qui avaient refusé de participer...

• Critères d'exclusion

Dans une optique pragmatique, les critères d'exclusion doivent être limités aux facteurs qui peuvent faire attribuer à tort l'amélioration des patients à un facteur de confusion plutôt qu'au traitement. Certains d'entre eux ne peuvent être exclus malgré tout car ils correspondent à des facteurs pronostiques importants et fréquents (association à une anxiété dans la dépression, dépression associée dans la lombalgie chronique…). De ce fait, l'élimination des patients présentant ces facteurs de confusion conduirait à faire l'étude sur une population non représentative.

Un auteur a étudié, rétrospectivement, la totalité des femmes consultant pour ostéoporose. Il a constaté que seulement 3,3% à 20% des patientes auraient pu être incluses dans un essai thérapeutique. Les principales raisons de non-inclusion étaient : l'âge, la

présence de co-morbidité ou la présence de traitements associés [25].

Une étude réalisée par nos soins, de validité limitée par le petit nombre de sujets, avait constaté que les patients présentant tous les critères d'inclusion et d'exclusion pour un essai clinique avaient une réponse au traitement supérieure à ceux ne présentant pas ces critères [26]. Il est donc possible que certains critères d'inclusion et d'exclusion conduisent à sélectionner des populations de répondeurs.

• Restrictions à une population homogène de patients

Cette restriction, recommandée par la plupart des grilles de lecture, permet de diminuer le risque de facteurs de confusion méconnus en utilisant des patients très similaires. Dans une méta-analyse, elle permet de rassembler des données issues d'études différentes. Elle pourrait avoir comme avantage de diminuer les effectifs de patients nécessaires pour obtenir une différence significative puisqu'ils sont plus homogènes. Elle a par contre l'inconvénient de conduire à une population non représentative qui limite la validité externe.

Taille de l'étude : calcul préalable du nombre de sujets nécessaires

Il s'agit du nombre de patients inclus dans l'étude aptes à la randomisation. Le calcul du nombre de sujets nécessaires permet la prise en compte de l'erreur bêta (risque de conclure à tort à l'absence d'effet par manque de puissance statistique), du bénéfice minimum cliniquement pertinent et de la variabilité de la réponse au traitement. L'erreur bêta n'est pas toujours prise en compte dans les études. Ainsi, un auteur a constaté sur 383 essais randomisés que 102 avaient un résultat négatif. Selon les années, seulement 16% à 36% avaient une puissance statistique suffisante pour affirmer une différence de 50% ou de 25% [33]. Un autre auteur, en 2001, a fait les mêmes constatations sur les essais cliniques randomisés en chirurgie [34].

Randomisation

Un auteur a recensé huit méta-analyses lui permettant de comparer des essais randomisés et non randomisés du même traitement [12]. Sur ces huit études, cinq surestimaient l'effet, une rapportait des effets similaires et deux sous-estimaient l'effet du traitement évalué. Lorsque les méta-analyses comparaient le résultat des essais randomisés et non randomisés des mêmes traitements, 2/3 ne pouvaient donner aucune conclusion et 1/3 concluait à des effets similaires. L'absence de randomisation conduisait parfois à une surestimation de l'effet mais aussi, plus rarement, à une sous-estimation.

Un auteur propose d'utiliser un facteur de correction en fonction ou non de l'existence d'une randomisation et de l'insu du patient : si la répartition des groupes n'est pas randomisée, il faut diminuer l'effet du nouveau traitement de 15% ; si l'aveugle des patients n'est pas préservé mais que l'étude est randomisée, il faut diminuer l'effet du nouveau traitement de 11% [27]. Ces valeurs ont été calculées sur des essais de traitements médicaux. Sur les traitements chirurgicaux [28], il faut distinguer les actes à visée préventive et curative. Lorsque le geste est à visée curative, le nouveau traitement apporte un gain de 56% en moyenne si l'essai est randomisé, de 62% s'il est contrôlé en groupes parallèles (comparaison simultanée), 63% lors des comparaisons à des séries historiques. En prévention secondaire, le gain moyen est de 53% pour les essais randomisés, 58% pour les essais non

randomisés. Un autre auteur, à partir de recueil d'essais utilisés dans des méta-analyses, a constaté que l'absence de randomisation conduit à surestimer les *odds ratio* de 17% [29].

Ainsi, la randomisation en plusieurs groupes semble avoir des effets variables selon les circonstances.

• Procédure de randomisation adéquate

Pour de multiples auteurs et rédacteurs de grilles de lecture, la randomisation utilisée pour répartir les traitements doit réellement être basée sur le hasard [8]. Il peut s'agir d'une table de nombres aléatoires, le jet de pièce de monnaie ou le mélange de jeu de cartes. Les mauvaises procédures d'allocation utilisent le numéro de dossier, la date d'admission, la date de naissance, une répartition alternée [30]... Elles introduisent une logique qui peut être devinée par les différents intervenants. Elles conduisent à terme à une rupture de l'insu de l'évaluateur qui peut conduire à une rupture de tous les insus. La randomisation par blocs est une solution acceptable, mais la taille des blocs doit être suffisante et si possible variable (en décidant la taille des blocs de façon aléatoire). Un auteur considère que la taille des blocs doit être au minimum de 4 sous peine de provoquer une rupture de l'insu de l'évaluateur [31].

• Allocation en insu

La procédure de randomisation utilisée pour répartir les patients entre les groupes devrait être réalisée en aveugle (c'est-à-dire sans que quiconque sache à quel groupe il a été attribué). L'absence de randomisation en aveugle peut conduire à surestimer les effets. Ainsi, un auteur, à partir d'un recueil d'essais utilisés dans une méta-analyse, a considéré que l'absence de randomisation en insu augmentait les *odds ratio* de 41% et qu'un insu inadéquat ou peu clair les augmentait de 30% [12].

D'autres auteurs ont estimé que l'absence de randomisation en aveugle surestimait les effets de 35% [13]. L'insu de la randomisation lui paraissait avoir un effet potentiel supérieur à celui du traitement examiné.

En pratique, il semble que l'absence de randomisation en insu conduise, elle aussi, à une rupture de tous les insus. Elle débouche également sur un biais de sélection des patients qui tendent à choisir le traitement qu'ils préfèrent ou des médecins qui tendent à choisir, pour un patient supposé mauvais répondeur, le placebo plutôt que le produit actif. Il compromet la validité interne et externe de l'étude [32]. Certains proposent une randomisation centralisée, réalisée à distance du site où est réalisée l'étude afin de favoriser l'obtention de l'insu.

Le domaine de la randomisation est probablement un des plus étudiés. Il est probablement un des rares dont l'influence a été chiffrée.

• Comparabilité du pronostic

Dans la mesure où le but de la randomisation est de constituer des groupes comparables, même pour les variables inconnues, il est important de vérifier que les groupes sont comparables pour les variables d'évaluation et les variables pronostiques les plus importantes. Lorsqu'il existe une différence dans les facteurs pronostiques, on suspecte l'exis-

tence d'un biais de sélection des patients (avec le risque alpha que cette différence soit liée au hasard). Nous n'avons pas trouvé d'études sur l'influence de ce biais. Il faut noter que, si la randomisation a réussi à constituer des groupes comparables, elle ne signifie pas que ceux-ci sont équivalents (il faudrait pour cela calculer préalablement un nombre de sujets suffisant à déterminer un risque bêta acceptable pour déclarer l'équivalence). Il est probable que les critères pronostiques sont d'autant meilleurs qu'ils ont été validés par des études préalables mais nous n'avons pas trouvé d'études sur l'influence de cette validation. La préférence des patients pour le traitement évalué est un facteur pronostique important. Il retentit à plusieurs échelons de la validité. Lors de la randomisation, les patients exerceront une pression plus importante pour aller dans le groupe de traitement qu'ils préfèrent. Au cours du suivi, le nombre de perdus de vue risque d'être différent si la majorité des patients préfère un traitement à l'autre (ce qui avantage souvent le traitement jugé le plus « moderne », le plus efficace ou le moins risqué).

Les écarts au protocole

Les « écarts au protocole » comprennent les « perdus de vue » et les « écarts au protocole de traitement ». Ces derniers sont les écarts à l'administration du traitement telle qu'elle a été codifiée dans le protocole : absence totale ou partielle de traitement (abandons), administration de traitements non autorisés, modification de la posologie prévue, voire administration du traitement de l'autre groupe, etc.

• Ecarts au protocole de traitement

L'absence d' « écart au protocole de traitement » est, bien entendu, souhaitable. Lorsqu'il y a des « écarts au protocole de traitement », il est important que leur nombre soit donné séparément dans chaque groupe et que leurs motifs soient comparables. Les « écarts au protocole de traitement » sont une source importante de biais mais l'influence réelle de celui-ci est vraisemblablement proportionnelle à leur nombre. S'ils sont très importants, l'applicabilité du traitement sera mise en doute et l'on pourra être éventuellement conduit à ne pas analyser l'essai. Exclure les « écarts au protocole de traitement » expose à un biais potentiel. Certains auteurs insistent pour garder dans leur groupe de départ les sujets qui ont reçu le traitement prévu pour l'autre groupe. Cet aspect sera détaillé au chapitre sur l'analyse en intention de traiter.

• Perdus de vue

Les perdus de vue sont les patients pour lesquels on ne dispose pas du critère de jugement. Ils sont une source de biais importante et il est souhaitable qu'ils soient les moins nombreux possibles. Il est également souhaitable que le nombre et les motifs des pertes de vue soient comparables dans les groupes évalués. Les perdus de vue peuvent ne pas être pris en compte dans l'analyse mais le test effectué sera moins puissant et leur exclusion peut biaiser la comparaison. Exclure les perdus de vue équivaut en effet à considérer qu'ils se comportent de la même façon que les autres sujets de leur groupe. Si cette hypothèse n'est pas admissible, il est possible, en cas de critère de jugement dichotomique, de les considérer tous comme des succès ou comme des échecs. On peut choisir également l'« hypothèse du biais maximum » qui revient à se placer dans les conditions les

plus défavorables pour le traitement à évaluer : en cas de critère de jugement qualitatif, on considérera tous les perdus du groupe traité comme des échecs et tous ceux du groupe témoin comme des succès ; en cas de critère de jugement quantitatif, on affectera à tous les perdus de vue du groupe traité la valeur la plus mauvaise des patients non perdus de vue du groupe traité, et pour les perdus de vue du groupe témoin la valeur la meilleure des patients non perdus de vue du groupe non traité. Dans ces conditions, la mise en évidence d'une différence en faveur du traitement à l'étude existe a fortiori dans tous les cas de figure.

Lorsque que le nombre de données manquantes est trop important, toute conclusion valide d'un essai devient impossible (facteur de biais majeur).

Insu (aveugle) du patient

L'insu du patient intervient à plusieurs niveaux dans la validité de l'étude.

On a coutume de dire qu'une frontière existe entre les traitements médicamenteux et les autres. Pour l'insu du patient, il nous semble que la frontière est plutôt située entre les thérapeutiques industrielles et non-industrielles.

Les traitements qui reposent sur une technologie industrielle sont les médicaments, mais aussi différentes méthodes de physiothérapie (infrarouges, stimulation électrique transcutanée, champs électromagnétiques). Ils ont en commun de pouvoir être administrés en insu (un appareil électrique déconnecté, une pilule placebo). Ils peuvent également être totalement standardisés (il est possible de donner exactement la même dose pour les patients d'une étude), ce qui sera discuté dans le chapitre sur la mesure des effets.

En chirurgie, un insu est possible et a déjà permis, par exemple, de constater que l'arthroscopie n'était pas supérieure à une pseudo intervention dans l'arthrose du genou [35]. L'insu du patient en chirurgie pose malgré tout d'importants problèmes éthiques et n'est certainement pas possible dans tous les cas, en particulier dans les gestes engageant le pronostic vital. La place de la chirurgie est donc particulière puisqu'elle est en principe simulable mais non-technologique.

Les études portant sur des substances chimiques accordent beaucoup d'importance à l'aveugle du patient. Nous aborderons ce problème, ainsi que son influence potentielle sur le résultat de l'étude, plus longuement au chapitre suivant. Il est évident que l'insu du patient est rarement possible dans les traitements physiques puisque, dans un aveugle complet, les patients ne connaissent pas les traitements reçus. Pour éviter cet inconvénient, on conseille parfois de recruter des patients qui ne sont pas familiers avec l'intervention évaluée (on parle alors de patients complètement naïfs). Pour des raisons pratiques, on recrute parfois des patients chez qui l'intervention expérimentale n'a pas été donnée durant l'année précédant l'évaluation (on parle alors de patients partiellement naïfs). Lorsque l'étude n'évalue pas le traitement sur ceux qui l'ont déjà reçu dans le passé, la validité interne est plus importante mais la validité externe est plus faible.
Il est souhaitable d'évaluer l'aveugle réel des patients au cours de l'étude et les conditions dans lesquelles celui-ci a éventuellement été rompu.

« Traitement »

La partie traitement doit décrire avec précision l'intervention expérimentale et l'intervention contrôle de façon à ce que n'importe qui puisse la répéter.

L'observance du traitement a, bien entendu, une influence sur le résultat de l'étude. Elle dépend beaucoup des conditions de son déroulement. Dans les traitements physiques, l'expérience du thérapeute a également une importance sur le résultat de l'étude.

Types de comparaison

La nature du comparateur conditionne en partie le résultat de l'étude.
Le traitement évalué peut-être comparé à une intervention placebo mais celle-ci est difficile à mettre en œuvre dans les traitements physiques. Il peut parfois paraître préférable de comparer celui-ci à un traitement de référence. Dans ce cas, la fiabilité de l'étude n'est pas la même s'il s'agit d'un traitement validé par des études préalables ou seulement consacré par l'usage.

• Comparaison à un placebo

L'effet placebo a fait l'objet de nombreuses publications.

Dans une expérimentation sur la caféine, un premier auteur a montré que le fait de faire croire au patient qu'on lui donnait un produit inactif diminuait la réaction normale à la caféine [36]. Il en conclut que le consentement demandé aux patients, dans les études en double insu, diminue la validité externe d'une étude. Un deuxième a retrouvé un effet identique mais considère que l'information du patient de la possibilité d'un placebo n'a qu'une influence limitée sur le résultat [37]. Un troisième auteur a essayé de quantifier l'effet placebo dans une méta-analyse comparant celui-ci à l'absence de traitement. Il a constaté que le placebo n'était supérieur à l'absence de traitement que pour les critères subjectifs et dans le cas de variables continues [38]. Pour les variables discontinues et pour les critères objectifs rapportés sous forme de variables continues, le placebo n'avait aucun effet significatif. L'analyse en sous-groupe (qui était prévue dans le protocole) des essais sur la douleur montre, à l'opposé, que le placebo a un effet significatif qui est de l'ordre de 6,5 mm à l'échelle visuelle analogique.

La comparaison à un placebo pose d'autres problèmes théoriques. Ainsi Kirsch a comparé deux groupes de patients qui recevaient le même placebo [39]. Un des groupes avait des instructions correspondant à un double insu (placebo versus caféine). L'autre avait des informations « décevantes » l'informant qu'il recevait un placebo. Pour certaines des variables mesurées (tension artérielle systolique, vigilance) les réactions des deux groupes étaient opposées.
La comparaison à un placebo augmente indiscutablement la validité interne d'une étude. Il est probable qu'elle diminue parallèlement sa validité externe.

• Comparaison à un traitement de référence

Si le traitement évalué est comparé à un traitement de référence il est important de savoir s'il s'agit d'un traitement consacré par l'usage ou si son effet a été validé. Dans ce dernier cas, il faudra citer une référence de validation dans l'article.

Un auteur a constaté que les essais comparant deux traitements avaient tendance à avantager le plus récent au détriment du plus ancien [40]. Même s'il montre que cette différence est en partie imputable à la présentation de l'article qui avantage le nouveau produit, on ne peut s'empêcher de penser qu'un certain nombre de patients (ceux qui acceptent de rentrer dans un essai par exemple…) ont l'espoir de trouver dans le nouveau produit un effet supplémentaire à celui des produits classiques. Ce facteur est particulièrement important quand une intervention placebo n'est pas réalisable. Elle avantage le traitement que les patients préfèrent.

• Comparaison à l'absence de traitement ou à la poursuite du traitement habituel

L'absence de placebo, valable dans la plupart des traitements physiques, pourrait conduire à revenir à un schéma plus simple avec comparaison à l'absence de traitement.

Ce type de comparaison est difficile à mener en pratique car le groupe contrôle ne reste pas volontiers dans l'étude puisqu'il n'en retire aucun bénéfice. Ceci est susceptible d'être source de déception, augmente le nombre de perdus de vue (particulièrement chez les patients les plus graves) et introduit donc un biais dans le groupe témoin.

De surcroît, il conduit à inclure l'effet placebo en plus de l'effet spécifique éventuel du traitement évalué, ce qui complique l'interprétation du résultat.

Plusieurs auteurs ont pensé à comparer le traitement physique avec la poursuite du traitement habituel. Afin de limiter le nombre de perdus de vue, ils proposent aux patients du groupe témoin un traitement à la fin de la période de surveillance. Cette méthode a l'inconvénient potentiel de majorer la gêne fonctionnelle des sujets témoins qui pourraient redouter de ne pas recevoir le traitement proposé s'ils sont trop améliorés [41]. Comme dans le cas précédent, certains patients du groupe contrôle pourraient avoir un ressentiment justifié par le fait que leur entrée dans l'étude correspondait à un moment où le « traitement habituel » s'avérait inefficace.

Traitements associés

La validité interne d'un essai est plus importante en l'absence de tout traitement associé. Elle reste difficile à proposer sur le plan éthique surtout lorsque l'essai envisage une longue période de surveillance. Elle risque aussi d'induire un biais de sélection de patients non-représentatifs ayant, par exemple, une gêne fonctionnelle moins importante et pour qui la possibilité de ne rien pouvoir adjoindre au traitement évalué représente un risque acceptable.

En pratique, autoriser les traitements associés représente une situation plus proche de la réalité. Il est nécessaire de vérifier que ceux-ci sont comparables dans les deux groupes afin de ne pas attribuer à tort l'amélioration observée à ce facteur de confusion.

Insu du thérapeute.

Afin de s'approcher le plus possible de l'effet réel du traitement, il pourrait être souhaitable que le thérapeute ne connaisse pas l'appartenance de son patient à un essai thérapeutique. L'insu du thérapeute, s'il est recherché, doit être lui-même évalué à la fin de l'étude et il faut donc décrire les conditions dans lesquelles il a été réalisé. A notre

connaissance l'influence de l'insu du thérapeute n'a jamais été évaluée.

« Résultats »

Un organigramme pour la répartition des patients a été proposé par le CONSORT statement (Consolidated Report of Randomized Clinical Trial) [42]. Il permet de vérifier quel type d'analyse a été menée sur les sujets de l'étude. Il devrait être rapporté dans tous les essais thérapeutiques (figure 1).

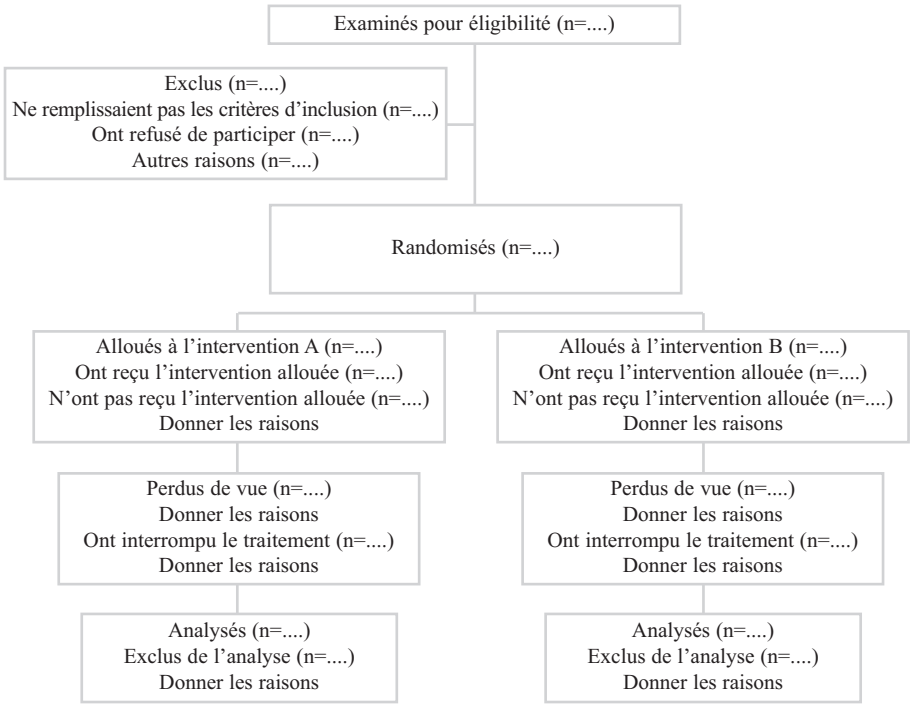

Figure 1 : Diagramme de flux proposé par le CONSORT statement (2001). Il permet au lecteur de l'article de connaître la façon dont les sujets ont été recrutés et leur devenir au cours de l'étude d'après [42].

Mesure des effets

• Critères de jugement

Les résultats des mesures des différents critères de jugement doivent être rapportés explicitement (moyenne et déviations standard, ou médianes et quartile, ou moyenne et intervalle de confiance). Ils doivent être donnés pour tous les critères de jugement importants et aux moments les plus importants. Il est nécessaire de préciser par qui a été mesurée la variable (le patient, le thérapeute ou un évaluateur indépendant). Il faut aussi préciser l'expérience de l'évaluateur et si la façon dont les mesures ont été réalisées est pertinente.

La pertinence des critères est déterminante pour la qualité de l'étude. Il est souvent préférable de choisir un critère de jugement principal et, pour celui-ci, d'opter pour un critère qualitatif dont l'amélioration est perceptible par le patient. La pertinence de l'effet a autant d'importance que sa signification statistique.

Certains essais élaborent des critères composites dont la validité n'a pas toujours été démontrée.

Comme nous l'avons dit, la standardisation de la séquence thérapeutique est d'usage très répandu dans les traitements médicamenteux. À l'opposé, la plupart des traitements qui reposent sur l'intervention d'un thérapeute (kinésithérapeute, médecin, chirurgien, psychothérapeute) sont modulables en fonction du déroulement de la séquence thérapeutique. Nous pensons qu'il est illusoire, et de surcroît non souhaitable, que ce dernier type d'intervention soit standardisé pour les besoins de l'évaluation. En effet, cette standardisation ne correspond pas à une situation clinique vraisemblable en pratique. Elle fait perdre un des avantages que ces traitements peuvent avoir en s'adaptant aux circonstances en temps réel. L'absence de standardisation a, par contre, l'inconvénient d'entraîner une plus grande variabilité de la réponse qui nécessite l'inclusion d'un plus grand nombre de sujets.
La confrontation du protocole déposé auprès du comité d'éthique avec la présentation de l'étude permettrait, dans l'idéal, d'éviter la manipulation ou la publication partielle des données les plus avantageuses en faveur du traitement étudié. Cela semble s'être produit récemment dans un essai qui a eu un retentissement notable sur les dépenses de santé puisqu'il portait sur un nouvel anti-inflammatoire non stéroïdien [43].

• Période de suivi

La durée de la période de suivi influe sur la validité interne et externe de l'étude. Lorsqu'elle est de courte durée, la validité interne est élevée puisque le risque de perdus de vue est plus faible mais la validité externe peut être limitée s'il s'agit d'une pathologie chronique dont les traitements doivent être mesurés sur le long terme.

À l'opposé, une longue période de suivi expose à un plus grand nombre de perdus de vue qui diminue la validité interne. Elle rend aussi plus probable l'apparition d'un facteur de confusion méconnu. Elle a par contre l'avantage de donner une meilleure idée de l'effet d'un traitement proposé dans une pathologie chronique.

Le timing du traitement doit être le même dans les deux groupes sous peine d'exposer l'essai aux biais rencontrés dans les études non randomisées. Si le traitement a été réalisé à un moment différent, les patients des deux groupes ont pu évoluer différemment et ne plus être comparables.

• Effets indésirables

Les effets indésirables doivent être décrits avec précision dans chaque groupe ; leurs relations éventuelles avec les abandons du traitement et les perdus de vue également.

Dans une étude évaluant l'efficacité d'une thérapeutique, le fait que les effets indésirables soient comparables entre les deux groupes ne signifie nullement qu'ils sont équivalents.

En effet, certains traitements ont des effets indésirables graves mais rares. Ils ne peuvent être décelés que par des études portant sur des effectifs de patients très importants.

• Insu (aveugle) de l'évaluateur.

L'influence que peut exercer l'évaluateur sur le résultat de l'étude est déterminante. Son insu doit toujours être recherché mais, dans les traitements physiques, où l'insu des patients est impossible, il est important d'évaluer la façon dont il a été préservé.
L'indépendance de l'évaluateur par rapport au traitement évalué sera discutée dans le chapitre consacré au conflit d'intérêts.

Méthode d'analyse statistique

Règles générales de l'analyse

Le calcul de l'intervalle de confiance est l'objectif du calcul statistique. Un auteur rappelait que la formule de base était d'une grande simplicité [44] :

$$I.\ de\ confiance = \frac{Signal}{Bruit}\ x\ \sqrt{Taille\ de\ l'étude}$$

En pratique, il est bien entendu important que les tests statistiques soient adaptés à la distribution des données. Lorsque celles-ci sont distribuées de façon normale, les tests paramétriques sont adaptés. Dans tous les autres cas, il faut utiliser les tests non paramétriques. Il est donc indispensable de vérifier le type de distribution des données avant de commencer l'analyse.
Il est préférable de rapporter la valeur exacte de la probabilité obtenue plutôt que des seuils de valeurs arbitraires (comme p inférieur à 0,05 ou 0,01). Dans la comparaison statistique, il faut préciser l'intervalle de confiance et l'amplitude de la différence [15].
Lorsque les auteurs choisissent d'utiliser un critère principal, la valeur de $p<0,05$ est consacrée par l'usage. Si chaque critère de jugement est mesuré plusieurs fois et que plusieurs critères de jugement sont rapportés, cette valeur doit être abaissée à l'aide d'un facteur de correction [45]. Dans le cas contraire, on s'expose à une erreur de type alpha qui est la probabilité de conclure, à tort, à l'activité d'un traitement inactif.
Le calcul de la différence entre les groupes, pour le critère principal, est particulièrement important. Il reste quand même intéressant de calculer l'évolution par rapport à l'état de départ qui renseigne sur l'effet global que pourrait avoir le traitement même s'il est probable que la participation à un essai clinique renforce l'effet placebo.
Le calcul de l'effet taille (variation du critère étudié/écart type de la variation) donne une indication sur l'importance de l'amélioration clinique. Il permet aussi de comparer des critères de jugement entre eux [59].

Méthode d'analyse : en « intention de traiter »

Dans ce type d'analyse, les résultats de tous les patients randomisés sont reportés y compris ceux qui n'ont pas terminé le traitement, de même que les autres écarts au protocole de traitement. L'analyse en intention de traiter est typiquement pragmatique ; elle permet d'éviter un biais de sélection des patients à posteriori, après obtention des résultats de l'étude. Bien entendu, la validité de l'étude reste moins bonne si le nombre de

perdus de vue est élevé, rendant l'analyse en intention de traiter sans intérêt. Un auteur recommande une variante intitulée : « analyse en intention de traiter modifiée » [46]. L'analyse exclut les patients qui n'ont jamais reçu le traitement ou ceux qui n'ont jamais été évalués après avoir reçu le traitement. Cette variante élimine un facteur de confusion mais diminue la représentativité de la population. Il est également utile de faire « une analyse des *completers* » qui renseigne sur ce que le traitement évalué apporterait dans l'absolu, c'est-à-dire sans tenir compte des problèmes d'observance et de tolérance.
Même dans les publications récentes des revues à *impact factor* élevé, l'analyse en intention de traiter n'est rapportée que dans la moitié des essais cliniques randomisés [47]. Dans ces derniers, la majorité ne précisait pas nettement comment ils avaient pu intégrer les déviations par rapport à la randomisation, les inclusions dans le mauvais groupe et les données manquantes. Certains articles excluaient de l'analyse les patients n'ayant jamais débuté l'intervention étudiée. D'autres continuaient à exclure jusqu'à 5% des patients inclus sous des prétextes variés.

En pratique, l'analyse en intention de traiter doit être considérée comme une stratégie complète pour conduire un essai thérapeutique plutôt que comme une simple méthode d'analyse statistique. Dans la méthodologie, elle est essentielle pour les études pragmatiques. Elle nécessite que toutes les inclusions et exclusions soient justifiées. Dans la conduite de l'étude, il est essentiel que le nombre de perdus de vue soit le plus limité possible. Il faut continuer à évaluer la totalité des sujets qui abandonnent au cours du traitement. L'analyse doit porter sur la totalité des sujets dans le groupe où ils ont été randomisés et évaluer l'effet potentiel des données manquantes. Dans la publication, il faut préciser si une analyse en intention de traiter a été réalisée, décrire avec exactitude les effets potentiels des réponses manquantes et baser les conclusions sur celle-ci [47].

Analyse en sous-groupe

Elle doit être, dans l'idéal, limitée à un nombre prédéfini d'hypothèses énoncées dans le protocole initial. Il est souvent intéressant d'envisager les interactions entre le traitement évalué et un facteur pronostique (par exemple ancienneté de la lombalgie ou existence d'un accident de travail dans la lombalgie chronique). L'interprétation des résultats doit être faite avec prudence et être présentée comme des découvertes à explorer plutôt que comme une démonstration.
Le nombre d'analyses en sous-groupe, réalisées à posteriori, est par définition infini. Si l'on se réfère au seuil de probabilité de 0,05, le chercheur qui ferait 100 analyses en sous-groupe à posteriori, aurait donc cinq chances sur cent d'en trouver de positives par les simples lois du hasard. C'est pourquoi il est particulièrement important que les auteurs précisent quelles analyses en sous-groupe ont été prévues au protocole et lesquelles sont des découvertes exploratoires. Ceci peut également être vérifié en comparant le protocole initial avec la présentation des résultats

Insu (aveugle) du statisticien

Même si cet insu est rarement mentionné dans les essais et n'est pas couramment retrouvé dans les grilles de lecture, il a une influence potentielle sur le résultat de l'étude. Cette influence n'a jamais été évaluée à notre connaissance, mais nous pouvons supposer

qu'un statisticien intéressé par le nouveau traitement tendrait à multiplier les tests statistiques pour prouver son impression (ce qui reviendrait à la situation du paragraphe précédent). Une planification de l'analyse préalablement au déroulement de l'étude permettrait de limiter les conséquences de l'absence d'insu du statisticien.

Corrections

En cas de données manquantes, ou de perdus de vue, ou lorsqu'il y a une différence significative entre les variables lors de l'évaluation initiale, il peut être préférable d'apporter des corrections plutôt que d'utiliser les résultats bruts.

Concernant les perdus de vue, on peut considérer, selon les circonstances, que tous ont un mauvais résultat (hypothèse du biais maximum), tous un bon résultat, ou bien leur attribuer la variation moyenne de la population. On peut enfin leur attribuer le résultat de la dernière mesure connue. Compte tenu de la complexité de leur mise en œuvre, les corrections doivent être effectuées par un statisticien.

« Discussion »

La discussion devrait en principe aborder les hypothèses de l'étude, les sources de biais potentiels, la généralisation possible des résultats et la situation des conclusions de l'étude par rapport aux connaissances actuelles.

Plus globalement, elle pourrait tenter de déterminer quel est le service médical rendu par le traitement évalué et ce qu'apporte l'étude par rapport à lui : traitement symptomatique ou thérapeutique de fond ; complément ou substitut à d'autres traitements ; traitement indispensable, simple adjuvant ou ultime recours ; intérêt médico-économique ; à qui le traitement évalué rend-il service (indications précises, formes cliniques définies) ?

« Conclusion »

Dans la conclusion, les résultats de l'étude doivent être présentés avec sincérité. Ce n'est pas toujours le cas : ainsi un auteur a constaté, dans une revue de la littérature, que 76% des articles avaient une conclusion douteuse ou invalide [40]. Il a constaté notamment, une déclaration d'efficacité en l'absence de groupe placebo, des déclarations d'équivalence en l'absence de calcul du nombre de sujets nécessaires, une sous-estimation des biais qui favorisaient systématiquement la nouvelle drogue, une structure des essais sans concordance avec les résultats présentés.

« Conflit d'intérêt potentiel »

L'origine du financement d'une étude est susceptible d'avoir une influence sur la façon dont sont présentés ses résultats. Il en va de même lorsqu'un des auteurs de l'article tire une partie de ses revenus du traitement évalué. Ainsi, dans une analyse critique des traitements de l'asthme, un auteur a constaté que la totalité des revues sponsorisées par un laboratoire pharmaceutique présentaient des insuffisances méthodologiques qui avantageaient le produit fabriqué par la firme [48]. La présentation partielle des résultats et le changement à posteriori du critère de jugement principal qui avantage le nouvel anti-

inflammatoire dans l'étude CLASS a peut-être été favorisée par un conflit d'intérêt (l'étude était financée par le fabricant) [43]. Ces changements fondamentaux dans la structure de l'étude n'avaient en tout cas pas été signalés dans la publication originale.

En pratique, il est difficile de connaître exactement l'ampleur du phénomène mais un auteur a observé, dans une enquête où le taux de réponse était faible (63% des auteurs américains ont refusé de répondre), que 87% des chercheurs qui avaient répondu, avaient une « interaction » significative avec l'industrie pharmaceutique. Il a observé que cette interaction concernait 100% des auteurs de recommandations pour la pratique clinique dans des domaines aussi variés que les arrêts cardiaques, la dépression, le diabète, l'ulcère peptique, l'hypercholestérolémie et l'arthrose [49].

Certains ont proposé différents niveaux d'interaction pouvant influer sur le résultat de l'étude [50]. La situation la plus favorable pour une étude est une indépendance totale vis-à-vis du traitement étudié et, par exemple, l'absence de support par un laboratoire. S'il y a un support affiché par un laboratoire, si un salarié du laboratoire est cité dans les auteurs, si le médicament est fourni par le fabricant, si l'article est publié dans un journal sponsorisé par un laboratoire pharmaceutique, si un au moins des auteurs de l'essai est payé par un laboratoire : il y a un conflit d'intérêt potentiel. On a proposé comme conflit d'intérêt significatif entre un auteur et le financement d'une recherche la somme de 10 000 dollars [51]. Cette somme, peut-être valable pour les Etats-Unis, est certainement à adapter au niveau de vie du pays où réside l'auteur.

Nous pensons qu'une influence du financement sur le résultat de l'étude pourrait exister même pour des sommes bien inférieures.

Commentaires

Même si certains facteurs importants ont été déjà explorés, les connaissances ne nous semblent pas suffisantes pour estimer l'influence de la validité globale d'une étude sur son résultat final. Il nous semble que ceci hypothèque la pertinence des grilles de lecture quantitatives utilisées dans certaines revues systématiques et méta-analyses. De surcroît, ces grilles traitent indistinctement les essais médicamenteux et non-médicamenteux alors que nous avons vu que les contraintes méthodologiques étaient parfois différentes.

Il nous semble, par contre, tout à fait licite de noter la qualité de la présentation en tenant compte des déterminants de la validité qui sont connus ou supposés. Le score obtenu permet de savoir s'il est possible de déterminer la validité de l'étude mais ne préjuge pas de la validité elle-même (même si les deux phénomènes sont souvent liés en pratique).

Nous avons bien conscience du caractère non-exhaustif de cette revue. La bibliographie par Medline est connue pour représenter un peu moins de la moitié des publications et elle privilégie nettement la langue anglaise. Nous ne sommes pas remontés avant les années 80 car, passée cette date, les résumés des articles ne sont plus disponibles et les articles eux-mêmes sont plus difficiles à obtenir. Il est donc tout à fait probable que nous n'avons pas identifié tous les domaines susceptibles de retentir sur le résultat d'un essai,

ni leur validation. Nous pensons que ceci n'est pas suffisant pour hypothéquer la validité de notre conclusion et espérons que d'autres, mieux documentés, poursuivront la réflexion que nous avons entreprise.

Le retentissement de la validité d'un essai thérapeutique sur son résultat a déjà été estimé dans certains domaines : le recueil du consentement éclairé, la comparaison de la publication avec le protocole, la procédure de sélection des patients, la procédure de randomisation, le calcul du nombre de sujets nécessaires, les abandons et les perdus de vue, l'insu des patients, une partie des modes de comparaison (à un placebo et à un traitement de référence), l'ajustement du seuil de probabilité (au nombre de mesures et de critères) et l'analyse en intention de traiter. Il faut remarquer que tous ces aspects sont souvent intriqués et que leur effet sur le résultat de l'essai peut être contradictoire. Dans d'autres domaines, une influence sur le résultat de l'étude est suspectée mais n'a pas été chiffrée. Pour finir, il existe obligatoirement des domaines méthodologiques encore inconnus mais dont l'influence est importante sur le résultat des essais.

La validité des grilles quantitatives est donc limitée par les connaissances incomplètes que nous possédons sur les différents biais. C'est pourquoi il n'est pas étonnant que plusieurs auteurs aient relevé des contradictions dans les revues et méta-analyses [52-56]. La comparaison de 25 scores de qualité a été faite par l'un d'entre eux pour la comparaison des héparines classiques avec les héparines de bas poids moléculaires. Les résultats étaient discordants selon la grille utilisée et le degré de rigueur des essais jugés [53]. Un autre a réalisé un travail similaire sur l'exercice physique dans la lombalgie chronique [54]. Il a remarqué que la conclusion d'une revue systématique pouvait varier en utilisant 25 scores de qualité différents sur les mêmes essais thérapeutiques. Le coefficient de corrélation des scores entre les différentes grilles variait de 0,49 à 0,94. La reproductibilité inter-observateur était également variable mais supérieure à 0,7 pour seulement 10 échelles. Cette reproductibilité était de surcroît différente selon la qualité de l'essai (elle était moins bonne pour les essais de faible qualité). Finalement, avec 6 grilles (dont la grille Cochrane), il y avait une évidence forte que l'exercice physique paraît supérieur au traitement habituel. Avec 4 grilles, il existait une évidence modérée d'efficacité supérieure. Les auteurs remarquaient également que leur cotation avec la grille utilisée par la collaboration Cochrane les avait menés à une conclusion différente. Un auteur enfin [52] a remarqué que les grilles désavantageaient systématiquement les essais chirurgicaux. De surcroît, il a remarqué que la plupart des grilles (96%) reposaient sur des critères « acceptés par la communauté scientifique » et qu'ils n'étaient donc pas validés.
La méthode DELPHI, utilisée par certains pour valider les critères, a l'avantage de représenter un consensus d'experts à un moment donné [8]. Elle a l'inconvénient potentiel d'utiliser le plus petit dénominateur commun entre les experts. Cette méthode pourrait donc en théorie éliminer des items pertinents mais non consensuels.

Beaucoup de ces grilles donnent des références de validation. Celles-ci portent surtout sur la reproductibilité [57]. Nous n'avons pas trouvé de travaux permettant de démontrer à partir de quelle note un essai thérapeutique était invalide. La différence de score minimal entre deux études pour considérer que l'une est supérieure à l'autre n'est pas connue.

Une auteure a comparé la rigueur méthodologique de l'évaluation des traitements non pharmacologiques et pharmacologiques dans l'arthrose des membres [58]. Elle a constaté une différence en faveur de l'évaluation pharmacologique. Cette différence n'était pas due à une moindre rigueur dans la méthode de randomisation ni dans l'analyse en intention de traiter. Elle était principalement provoquée par l'absence d'utilisation d'un placebo et des insus (patients, soignants et évaluateurs) qui sont beaucoup plus difficiles à mettre en œuvre dans les traitements non médicamenteux.

Notre revue, et les constatations de ces derniers auteurs, incitent à penser que ni la sensibilité, ni la spécificité des grilles de lectures quantitatives ne sont connues avec précision. Ces grilles ne tiennent pas compte du type de traitement évalué alors que nous avons vu que les contraintes pouvaient être différentes. Il faut donc considérer avec réserves les conclusions qu'elles permettent de formuler.

Conclusions

Nous sommes encore au fond de la caverne de Platon, le dos tourné à la lumière. Les connaissances accumulées à ce jour, permettent d'avoir une idée de certains facteurs qui influencent potentiellement le résultat des essais thérapeutiques. Elles ne sont pas suffisantes pour connaître l'amplitude exacte de ceux-ci dans une étude donnée.

Ces lacunes théoriques expliquent certainement une grande partie des différences de note observées d'une grille à l'autre.

Malgré l'attrait que peut représenter leur usage, nous pensons qu'il faut renoncer aux grilles de lectures quantitatives. Il nous semble préférable de revenir aux grilles qualitatives (check-lists en anglais) en attendant que les progrès des connaissances permettent d'établir un score qui reflète la valeur réelle de l'étude.

Des recherches sont encore nécessaires pour juger de l'influence réelle des différents biais et autres facteurs qui modifient les résultats des essais médicamenteux et non-médicamenteux. Il est probable que ces derniers pourraient faire l'objet de grilles de lecture spécifiques en raison de leurs contraintes méthodologiques, et que la comparaison avec les traitements médicamenteux est hasardeuse.

Références

1. Chalmers TC, Smith H, Blackburn B, Sylverman B, Scroeder B, Reitman D, Ambroz A. Method for assessing the quality of randomised trials. *Controlled clinical trial* 1981;2:31-49.*
2. Verhagen AP, De Wet HCW, De Bie RA, Kessels AGH, Boers M, Knipschild PG. Taking Bath. The efficacity of balneotherapy in patients with arthritis: a systematic review. *J Rheumatol* 1997;24:1964-71.
3. Koes BW et coll. Spinal manipulation for low back pain. An updates systematic review of randomized clinical trial. *Spine* 1996;21(24):2860-71.
4. Van Tulder MW, Koes BW, Bouter LM. Conservative treatment of acute and chronic nonspecific low back pain. A systematic review of randomized controlled trial of the most common intervention. *Spine* 1997;22(18):2128-2156.

5. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factor of non specific neck pain: a systematic review. *Pain* 1998;77(1):1-13.
6. Hill CL, La Valley MP, Felson DT. Secular changes in the quality of published randomized clinical trial in rheumatology. *Arthritis Rheum* 2002;46(3):779-784.
7. Van Der Heidjen G, Beurskens A, Koes BW, Assendelft JJ, De Vet HC, Bouter LM. The efficacy of traction for back and neck pain : a systematic blinded review of randomized clinical trial method. *Phys Ther* 1995;75(2):93-104.
8. Verhagen AP, De Vet HCW, De Bies RA, Kessels AGH, Boers M, Bouter LM, Knipschild PG. The DELPHI list : A criteria list for quality assessment of randomized clinical trial developed by Delphi consensus. *J Clin Epid* 1998;1235-41.
9. Oxman AD Guyatt GH, validation of an index of the quality of review articles. *J clin Epidemiol* 1991;44:1271-8.
10. Slack MK, Draugalis JR. Establishing the internal and external validity of experimental studies. *Am J Health Syst Pharm* 2001 Nov 15;58(22):2173-81.
11. Fisher AA, Carlaw RW. Family planning field research projects: balancing internal against external validity. *Stud Fam Plann* 1983;14(1):3-8.
12. Kunz R, Oxman AD. The unpredictability paradox: Review of empirical comparison of randomized and non randomized trials. *BMJ* 1998;317:1185-90.
13. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen T. Does quality of report of randomized trials affects estimates of intervention efficacy reported in meta-analysis ? *Lancet* 1998;352:609-13.
14. Verhagen AP, De Vet HC, Vermeer F, Widdershoven JW, de Bie RA, Kessels AG, Boers M, Van den Brandt PA. The influence of methodological quality on the conclusion of a landmark meta-analysis on thrombolyic therapy. *Int J Technol Assess Health Care* 2002;18(1):11-23.*
15. Poccock SJ, Hugues MD, Lee RJ. Statistical problems in the reporting of clinical trials. A surveys of three medical journals. *N Engl J Med* 1987;317:426-32.
16. Yanks V, Rennie D. Reporting informed consent and ethic committee approval in clinical trial. *JAMA* 2002;287:2835-2838.
17. Kleijnen J, de Craen AJM, van Everdingen J, Krol L. Placebo effect in double blind clinical trial: a review of interaction with medications. *Lancet* 1994;344:1347-1349.
18. Marcus SM. Assessing non-consent bias with parallel randomized and non randomized clinical trial. *J Clin Epidemiol* 1997;50(7):823-828.
19. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of publication bias on meta-analyses. *BMJ* 2000;320:1574-7.
20. Mc Auley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimate of intervention effectiveness reported in meta-analyses ? *Lancet* 2000 ;356 (9237):1228-1231.*
21. Zimmermann M, Mattia JI, Posternak MA. Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice ? *Am J Psychiatry* 2002;159(3).*
22. Licht WR, Gouliaev G, Vestergaard P, Frydenberg M. Generalisability of result from randomised drug trials. A trial on antimanic treatment. *Br J Psychiatry* 1997,170:264-7.
23. Glasgow RE, Bull SS, Gillette C Klesges LM, Dzewaltowski DA. Behavior change intervention research in healthcare settings. A review of recent report with emphasis on external validity. *Am J Prev Med* 2002;23(1):62-9.*
24. Gross CP, Mallory R, Heiat A, Krumholz HM. Reporting the recruitment process in clinical trial: who are these patients and how did they get there ? *Ann Intern Med* 2002;137:10-16.
25. Dowd R, Recker RR, Heaney RP. Study subjects and ordinary patients. *Osteoporosis int* 2000;11:533-536.

26. Forestier R, Françon A. Le biais de sélection des patients dans les études randomisées. *Rev Rhum* [Ed Fr.] 2001;68(10-11):97.
27. Colditz GA Miller GN, Mosteller F. How study design affect outcome in comparison of therapy I : *Medical. Stat Med* 1989;8(4):441-54.*
28. Miller JN, Coldiz GA, Mosteller F. How study design affect outcome in comparison of therapy II : Surgical. *Stat Med* 1989;8(4):455-56.*
29. Schulz K, Chalmer I, Hayes R, Altman D. Empirical dimensions of bias. Dimension of methodological quality associated with estimate of treatment effects in controlled trials. *JAMA* 1995;273(5):408-412.
30. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, no choice. *Lancet* 2002;359(9305):515-9.*
31. Bouvenot G, Vray M. Essai cliniques : cherchez les biais ! *Rev Rhum* [Ed Fr.]1993 60(6):412-415.
32. Berger VW. Detecting selection bias in randomized clinical trial. *Control Clin Trial* 1999;20(4):319-27.*
33. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trial. *JAMA* 1994;272(2):122-4.
34. Dimick JB, Diener-West M, Lipsett PA. Negative result of randomized clinical trials published in the surgical literature : equivalency or error ? *Arch Surg* 2001;136(7):796-800.*
35. Mooseley B, O'Malley K, Petersen NJ et coll. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *N Engl J Med* 2002;347(2):81-88.
36. Kirsch I, Rosadino MJ. Do double blind studies with informed consent yeld externally validity result? An empirical test. *Psychopharmacology* 1993;110(4):437-42.*
37. Nash JM, Holroyd K, Rokicki L, Kvaal, Penzien D. The influence of placebo awareness on stimulant drug response in a double blind trial. *Psychopharmacology* 2002;161(3):213-21.
38. Hrobjartsson A, Gotche PC. Is the placebo powerless? An analysis of clinical trial comparing placebo with no treatment. *N Engl J Med* 2001;344:1594-602.
39. Kirsch I, Weixel LJ. Double blind versus deceptive administration of a placebo. *Behav Neurosci* 1988;102(2):319-23.*
40. Gotzche P. Methodology and overt and hidden bias in report of 196 double blind trial of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. *Control Clin Tral* 1989 10:31-56.
41. Hadler NM. Spa therapy was effective in spa therapy. *ACP J Club*. 1994 Jul-Aug;121 Suppl 1:14
42. Moher D, Schulz KF, Altman DG. The consort statement: revised recommendations for improving the quality of report parallel group randomised trials. *Lancet* 2001;357:1191-94.
43. Jüni P, Rutjes AW, Dieppe P. Are selective COX 2 inhibitors superior to traditional non steroidal anti-inflammatory drugs ? Adequate analysis of the CLASS trial indicates that this may not be the case. *BMJ* 2002;324:1287-8.
44. Sacckett DL. Why randomised controlled trial fail but needn't: failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or to understand!). *CMAJ* 2001;165(9):1226-1237.
45. Schwarz D. *Méthodes statistiques à l'usage des médecins et des biologistes*. Flammarion Paris 1987 4è ed.
46. Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. *CMAJ*. 1997 May 15;156(10):1411-6.
47. Hollis S, Campbell F. What is meant by intention to treat analysis? Surveys of published randomized trials. *BMJ* 1999;319:670-4.

48. Jadad AR, Moher M, Browman GP, Booker L, Singouin C, Fuentes M, Stevens R. Systematic review and meta-analyses on treatment of asthma: critical evaluation. *BMJ* 2000;320:537-40.
49. Choudhry NK, Stelfox HT, Detsky AS. Relationship between author of clinical practice guidelines and the pharmaceutical industry. *JAMA* 2002;287(5):612-7.
50. Rochon PA, Guwiz JH, Simms RW et coll. A study of manufacturer-supported trials of non-steroidal drugs in the treatment of arthritis. *Arch Intern Med* 1994;154:157-63.
51. Drazen JM, Curfman GD. Financial association of authors. *N Engl J Med* 2002;346(24):1901-2.
52. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh. Assessing the quality of randomized clinical trial: an annotated bibliography of scales and checklists. *Control Clinical Trial* 1995;16(1):62-73.
53. Jüni P, Wischi A, Bloch R, Egger M. The hazards of scoring quality of clinical trial for meta-analysis. *JAMA* 1999;282(11):1054-60.*
54. Colle F, Rannou F, Revel M, Fermanian J, Poiraudeau S. Impact of quality scales on level of evidence inferred from a systematic review of exercice therapy and low back pain. *Arch Phys Rehabil* 2002;83(12):1745-52.
55. Furlan AD, Clarke J, Esmail R, Sinclair S, Irvin E, Bombardier C. A critical review of the reviews on treatment of chronic low back pain. *Spine* 2001;26(7):E155-E162.
56. De Vet H, De Bie R, Van der Heijden G, Verhagen AP, Sijkpe P, Knipschild P. Systematic review on the basis of methodological criteria. *Physiotherapy* 1997; 83(6):284-8.
57. Verhagen AP, de Vet HC, de Bie, Kessels AG, Boers M, Knipschild PG. Balneotherapy and quality assessment: interobserver reliability of the maastricht criteria list and the need for blinded quality assesment. *J Clin Epidemiol* 1998;51(4):335-41.
58. Boutron I, Tubach F, Giraudeau B, Ravaud P. Methodological difference in clinical trial evaluating nonpharmacological and pharmacological treatments of hip and knee osteoarthritis. *JAMA* 2003;290(8):1062-70.
59. Liang MH, Fossel AH, Larson MG. Comparison of five Health status instruments for orthopedic evaluation. *Med Care* 1990;28(7):632-642.

Tableau I : Influence mesurée de la méthodologie des essais sur le résultat. Les domaines étudiés ne couvrent pas tous les aspects d'un essai thérapeutique. Dans un certain nombre de cas, l'influence du domaine étudié n'est pas rapportée par les auteurs de l'article ou n'est pas disponible.

Thème de l'article	Étude	Plan expérimental	Conclusion de l'article	Importance de l'effet [IC 95%]	Signification statistique
Non-publication des articles	Mc Auley 2000[20]	Articles publiés versus (vs) non-publiés	publication préférentielle des essais positifs surestime l'effet du traitement	1,15[1,04 à 1,28]	
Qualité globale de l'étude	Kunz 1998[12]	Essais de bonne qualité vs faible qualité	Effet variable de la faible qualité de l'essai	[0,27 à 1,0]	Np
	Moher 1998[13]	id°	Faible qualité surestime l'effet du tt	0,66 [0,59 à 0,71]	0.005
	Verhagen 2002[14]	id°	Pas de corrélation entre qualité et résultat du tt	Np	Np
	Jüni 1999[53]	id°	id°	1,13 [0,70 à 1,82]	P<0,05
Résumé					
Concordance entre le résumé et les résultats	Poccock 1987[15]		Non-concordance surestime l'effet du tt	Np	Np
Méthode					
Recueil du consentement	Marcus 1997[18]	Patients consentants vs non-consentants	Population consentante répond différemment au tt	Np	P=0,02
	Kleijnen 1994[17]	Placebo vs produit actif	Diminue la différence	Np	Np
Critères d'inclusion et d'exclusion	Zimmermann 2002[21]	Critères restrictifs vs non restrictifs	Critères restrictifs: population non représentative	Np	Np
	Licht 1997[22]	id°	id°	Np	Np
Proportion de patient enrôlés	Gross 2002[24]	Patients enrôlés vs non enrôlés	Proportions faible: population non représentative	Np	Np
	Dowd 2000[25]	id°	id°	Np	Np
	Forestier 2001[26]	id°	Patients éligibles répondent mieux au tt	Np	P=0,04
Randomisation	Kunz 1998[12]	Essais randomisés vs non randomisés	Effet variable de la randomisation	[0,76 à 1,6]	Np
	Colditz 1989[27]	id°	Non-randomisation surestime l'effet du tt	1.15	P=0,004
	Schlutz 1995[29]	id°	id°	0,70 [0,62 à 0,79]	P<0,001,
	Kunz 1998[12]	Randomisation en insu vs ouverte	Randomisation ouverte surestime le tt	[30 à 40]	Np
	Moher 1998[13]	id°	id°	0,63 [0,45 à 0,88)	P=0,36
	id°	Randomisation adéquate vs non-adéquate	Méthode de randomisation non adéquate surestime l'effet tt	0,89 [0,67 à 1,20]	P=0,23
	Schlutz 1995[29]	id°	id°	0,95 [0,81 à 1,12]	P=0,58,

Thème de l'article	Étude	Plan expérimental	Conclusion de l'article	Importance de l'effet [IC 95%]	Signification statistique
Taille de l'étude	Moher 1994[33]	calcul à posteriori du nombre de sujet nécessaire dans différents essais médicaux	64 à 84% de conclusions invalides par sous estimations de la différence statistique	Np	Np
	Dimick 2001[34]	id°	Taille suffisante dans 16 à 36% des cas : erreur par sous estimations de la différence statistique	Np	Np
Insu	Colditz 1989[27]	Patients en insu vs sans insu	Absence d'insu surestime l'effet du tt préféré des patients	Np	P=0,02
	Schulz 1995[29]	Double insu vs absence d'insu	Absence d'insu surestime l'effet tt	0,83 [0,71 à 0,96]	P=0,01
	Moher 1998[13]	id°	id°	0,63 [0,45 à 0,88]	P=0,46
Type de comparaison	Kisch 1993[36]	Produit présenté comme actif vs présenté comme inactif	Tt présenté comme inactif moins efficace	Np	Np
	Kisch 1988[39]	id°	id°	Np	Np
	Hobjartsson 2001[38]	Placebo vs absence de tt	Placebo > absence de tt	-0,28 [-0,38 à -0,19]	P=0,001
	Nash 2002[37]	Information de l'existence d'un placebo vs non-information	Informer de l'existence d'un placebo diminue l'effet des deux tt	Np	Np
	Colditz 1989[27]	Ancien vs nouveau tt	Ancien < "nouveau" tt	Np	P=0,04
	Gotzche 1989[40]	id°	id°	Np	Np
	Miller 1989[28]	id°	id°	1.56	Np
Résultats					
Sélection des critères à posteriori	Jüni 2002[43]	Nouvel AINS vs Ancien AINS	Crée une différence artificielle	Np	Np
Méthode d'analyse	Schulz 1995[26]	Analyse de tous les sujets randomisation vs exclusion	Exclusion après randomisation surestime l'effet tt	1,07 [0,94 à 1,21]	
Conclusion					
Conclusion concordante avec les résultats de l'article	Gotzche 1989[26]	comparaison article avec sa conclusion	Conclusion invalide ou douteuse dans 75% des cas	Np	Np

Np = non précisé

Tableau II : Influence supposée de la méthodologie sur les résultats

Thème de recherche	Plan expérimental suggéré	Conséquences supposées
Méthode		
Proportion de patients enrôlés	Recrutement consécutif vs non planifié	Population non représentative
Restriction à une population homogène de patients	Population homogène vs non homogène	Population non homogène: biais méconnus
Comparabilité du pronostic	Pronostic comparable vs non comparable	Pronostic différent: réponse différente au tt
Critères pronostiques validés	Critères validés vs non validés	Populations non comparables
Abandons d'études	Comparaison abandons vs non-abandon	Abandon des non répondeurs surestime l'effet du tt
Perdus de vue	Comparaison perdus de vue vs non perdus de vue	Perte de vue des non répondeurs surestime l'effet du tt
Description précise du tt délivré	Description précise vs imprécise	Variabilité de l'effet traitement
Traitements associés	Tt associés comparables vs non comparables	Effet imputable au tt associé plutôt qu'au tt évalué
Insu du thérapeute	Insu du thérapeute vs absence d'insu	Avantage le tt préféré du thérapeute
Résultats		
Critères de jugement validés	Critères validés vs non validés	Mesure pertinente de l'effet du tt
Courte période de suivi	Courte vs Longue période de suivi	Diminue les perdus de vue et le risque d'apparition de facteur de confusion
Longue période de suivi	id°	Mesure l'effet réel du tt dans une pathologie chronique
Timing différent entre les deux groupes	Timing identique vs différent	Différence attribuable au timing et pas au tt
Description des effets indésirables		Non-description surestime le bénéfice des tt
Insu de l'évaluateur	Insu de l'évaluateur vs absence d'insu	Absence d'insu avantage le tt préféré de l'évaluateur
Tests statistiques adaptés	Tests inadaptés vs adaptés	Tests inadaptés manquent de puissance: sous estime la différence
Insu du statisticien	Insu du statisticien vs absence d'insu	Absence d'insu avantage le tt préféré du statisticien
Discussion		
Situation de l'étude par rapport aux connaissances actuelles	Enquêtes d'opinion chez les prescripteurs	Mauvaise détermination de la place du tt évalué
Conflit d'intérêt		
Intérêt financier	Conflit d'intérêt vs absence de conflit	Avantage le tt qui rapporte le plus à l'auteur de l'article
Intérêt intellectuel	id°	Avantage le traitement préféré de l'auteur

tt = traitement