Trois pièges dans les études de non-infériorité

Peter Kleist

Quintessence

- Les études de non-infériorité cachent toutes sortes de pièges. Les plus importants sont: inefficacité du traitement témoin, erreurs dans la fixation de la limite pour une infériorité encore juste acceptable du traitement d'étude et faiblesses dans le design et la réalisation de l'étude.
- L'efficacité du traitement témoin n'est pas mesurée directement, mais donnée à l'avance. Les bases de l'hypothèse que le traitement témoin est efficace sont souvent insuffisantes.
- Une fixation trop généreuse du seuil de non-infériorité peut amener à la conclusion de non-infériorité malgré que la différence ait effectivement une importance clinique.
- Un mauvais design d'étude et une réalisation déficiente effacent les éventuelles différences entre les traitements.
- Il est souvent impossible de répondre à la question «les traitements sont-ils aussi efficaces ou aussi inefficaces l'un que l'autre?».
- Cet article donne des indications pour pouvoir juger la qualité d'une étude de non-infériorité.

«Nous gagnons en jouant non pas bien mais mieux.»
Savielly Tartakover,
maître d'échecs russe (1887–1956)

Remarques préliminaires provocantes

Une recherche sur Pub-Med donne pour l'expression *noninferiority* plus de 1000 résultats, dont plus de la moitié se réfèrent à des publications des années 2009 à 2011 – donc tendance en forte augmentation. En tant que lecteur de littérature médicale, vous êtes donc forcément confronté à des études de non-infériorité.

Les études de non-infériorité sont établies dans la science médicale. Deux éléments ont notablement contribué à leur utilisation devenue inflationniste: s'il y a déjà un traitement efficace pour une maladie, une étude contrôlée contre placebo pour prouver l'efficacité d'un nouveau médicament est pratiquement injustifiable éthiquement. L'efficacité de nombreuses nouvelles options est cependant équivalente aux actuelles, dans le meilleur des cas; s'intéresser à une éventuelle supériorité du nouveau traitement n'a donc aucun sens. Derrière les études de non-infériorité se trouve en principe la question de savoir si deux traitements se distinguent l'un de l'autre tout au plus dans une mesure non significative.

Les études de non-infériorité ont cependant aussi quelque chose d'obscur. Premièrement, cette expression en tant que telle n'est pas correcte. Il ne s'agit en fait pas de non-infériorité mais de «pas trop d'infériorité». Secundo, elles sont une forme particulière d'études avec contrôle historique. Nous savons depuis longtemps que les comparaisons à des données historiques peuvent être problématiques. Et tertio: quel autre type d'étude peut prétendre obtenir un résultat voulu avec un mauvais design et une mauvaise réalisation?

Ce sont là les trois grands pièges des études de non-infériorité. Il ne s'agira par la suite que des études comparant des traitements prétendument actifs sans contrôle contre placebo.

Peter Kleist

L'auteur n'a pas déclaré des obligations financières ou personnelles en rapport avec l'article soumis.

1er piège: le traitement témoin n'est pas efficace

Une étude de non-infériorité met en relation l'effet d'un traitement d'étude à celui d'un traitement témoin. L'effet du traitement témoin n'est toutefois pas mesuré directement dans l'étude, mais donné par les résultats de précédentes études contrôlées contre placebo. En d'autres termes: la conclusion de la non-infériorité d'un traitement par rapport à un autre ne repose pas sur des faits solides, mais sur une hypothèse – à savoir que les traitements d'étude et témoin se seraient révélés efficaces dans une étude contrôlée contre placebo.

Ce qui ne pose aucun problème si le traitement témoin a nettement et régulièrement montré une supériorité sur le placebo dans de précédentes études. Ce n'est cependant pas le cas de nombreux médicaments courants, y compris ceux admis officiellement. Les données sont inconsistantes surtout pour les traitements symptomatiques, même si les études avaient une puissance de test suffisante et ont été effectuées correctement, notamment avec antidépresseurs, antipsychotiques, analgésiques, antihistaminiques, antihypertenseurs, antiasthmatiques et antiémétiques [1]. Les causes possibles de résultats hétérogènes sont une efficacité modérée seulement du traitement, des proportions de réponse variables, une bonne réponse au placebo, une compliance insuffisante, une interférence par effets *regression-to-the-mean* ou l'évolution spontanée d'une maladie. D'après une analyse de l'autorité sanitaire américaine FDA, aucune différence statistiquement significative par rapport au placebo n'a pu être démontrée dans 46% des études sur 8 antidépresseurs admis et 25% de celles sur 5 neuroleptiques admis [2].

Si une étude de non-infériorité fait une comparaison à un traitement témoin dont il n'est pas certain qu'il sera efficace dans cette étude aussi, elle ne permettra pas de tirer

Tableau 1. Résultats de 6 études dans lesquelles l'antidépresseur nomifensine a été comparé à l'imipramine et au placebo (adapté d'après [3]).

Etude	Valeur au départ (échelle de dépression d'Hamilton)	Valeur ajustée après 4 semaines de traitement (échelle de dépression d'Hamilton)			p (nomifensine contre imipramine)
		Nomifensine	Imipramine	Placebo	
1	23,9	13,4	12,8	14,8	0,78
2	26,0	13,0	13,4	13,9	0,86
3	28,1	19,4	20,3	18,9	0,81
4	29,6	7,3	9,5	23,5	0,63*
5	37,6	21,9	21,9	22,0	1,0
6	26,1	11,2	10,8	10,5	0,85

* p <0,001 pour les comparaisons nomifensine contre placebo et imipramine contre placebo.

Tableau textuel 1. Questions permettant de juger la validité d'un traitement témoin dans les études de non-infériorité.

- Le traitement témoin est-il un standard, au moins largement utilisé ou la plus mauvaise de plusieurs alternatives thérapeutiques?
- Le traitement témoin a-t-il déjà fait la preuve de son efficacité dans plusieurs études randomisées et contrôlées contre placebo aves des collectifs suffisants?
- L'efficacité du traitement témoin est-elle prévisible et régulière ou au contraire très variable?
- Y a-t-il des études bien faites avec le traitement témoin, dans lesquelles aucune efficacité n'a pu être démontrée?
- Y a-t-il des résultats d'une méta-analyse qui a tenu compte des études aussi bien positives que négatives (de préférence avec modèle de *Random Effects* pour juger l'hétérogénéité de ces études et de leurs résultats)?
- Le traitement témoin est-il utilisé dans l'étude de non-infériorité de la même manière qu'il a auparavant fait la preuve de son efficacité (par ex. concernant la dose, la forme galénique et la durée de traitement)?
- La réalisation de l'étude de non-infériorité est-elle comparable à celle des précédentes études avec le traitement témoin (par ex. concernant la séléction des patients, la gravité de la maladie ou le paramètre d'étude)?
- L'efficacité du traitement témoin dans l'étude de non-infériorité est-elle identique à celle de précédentes études?

de conclusion valable. Car la comparabilité des effets peut aussi bien exprimer l'efficacité que l'inefficacité des deux traitements. Un exemple pour l'illustrer (tab. 1, d'après [3]): la nomifensine, antidépresseur finalement non admis en raison de ses effets toxiques, a été comparée dans 6 études à l'imipramine et au placebo. La diminution marquée des chiffres sur l'échelle de dépression d'Hamilton a été cliniquement significative. Surtout: dans aucune de ces 6 études, la différence par rapport à l'imipramine n'a été significative, mais la supériorité de l'imipramine sur le placebo avait précédemment été démontrée dans 6 études. Ce n'est que le contrôle contre placebo effectué dans ces études qui a démontré que ces deux antidépresseurs n'ont effectivement été efficaces que dans une seule étude (tab. 1, étude n° 4).

Le problème des traitements témoins invalides dans les études de non-infériorité surtout est présent dans le domaine psychiatrique. Mais il peut se présenter dans n'importe quelle indication, comme le montre un exemple de l'otorhinolaryngologie. Une étude récemment publiée a montré que chez des patients ayant une surdité brusque l'efficacité des corticostéroïdes intratympaniques n'est pas inférieure à celle d'une corticothérapie orale [4]. Les preuves de l'efficacité des corticostéroïdes oraux – le traitement témoin – sont cependant insuffisantes et reposent essentiellement sur une étude d'il y a 30 ans. La proportion des rémissions dans l'étude actuelle est en plus comparable à celle des rémissions spontanées chez les patients non traités [5]. La question «aussi efficace ou aussi inefficace l'un que l'autre?» reste ouverte.

Le tableau textuel 1 donne quelques questions à se poser pour juger l'adéquation d'un traitement témoin dans une étude de non-infériorité.

2e piège: fausses hypothèses pour la non-infériorité

L'argument le plus évident pour démontrer l'équivalence de deux traitements serait de prouver leur équivalence absolue. En clinique, cela n'est pas possible – sauf si la supériorité du traitement d'étude est prouvée. Mais c'est justement cela qui est également impossible dans le contexte des études de non-infériorité. Ces études examinent donc si le traitement d'étude est tout au plus moins bon que le témoin, mais de manière non significative (raison pour laquelle l'expression *non-infériorité* n'est pas parfaitement adéquate).

Pour cela, il faut fixer le domaine de non-signification *avant* le début de l'étude, concrètement: le seuil dit de non-infériorité doit être défini. Si ce n'est pas le cas, ou s'il est fixé seulement après la fin de l'étude, elle est inutilisable. Car ce seuil a une importance déterminante aussi bien pour l'hypothèse examinée que pour connaître le nombre nécessaire de patients. La fixation subjective post-hoc du seuil en connaissant les données est en outre sujette à un biais et permettrait en plus de manipuler l'étude. Quelque chose qui ne saurait être suffisamment répété: une étude de supériorité qui a échoué en ne montrant aucune différence statistiquement significative entre deux traitements n'est pas adaptée pour faire des affirmations sur la non-infériorité. Car le fait que le traitement d'étude soit éventuellement inférieur au traitement témoin ne peut être exclu en raison du design de l'étude [6]. Dans les études de non-infériorité méritant cette appellation, il faut donc impérativement trouver les indications suivantes: a) une hypothèse claire sur la non-infériorité, b) la fixation d'un seuil correspondant et c) un nombre de cas qui en découle.

Il n'y a pas de règle d'or pour fixer le seuil de non-infériorité. Nous prenons habituellement la valeur inférieure de l'intervalle de confiance du traitement témoin dans de précédentes études contrôlées contre placebo, comme le montre l'exemple hypothétique suivant: si le traitement témoin a abaissé de 2,5% de plus que le placebo (intervalle de confiance 2,0–3,1%) le nombre d'accidents (par ex. infarctus du myocarde), le chiffre inférieur (2,0%) – exprimé de manière simplifiée – reflète l'effet minimal du traitement. En acceptant maintenant un seuil de 2% pour un nouveau traitement dans une étude de non-infériorité (également chiffre inférieur de l'intervalle de confiance, dans ce cas pour la différence entre traitement d'étude et témoin), et profitant de l'intervalle, l'effet complet du traitement témoin serait épuisé et il n'y aurait

plus aucune efficacité par rapport au placebo. C'est pour cette raison que pour fixer le seuil nous ne prenons généralement qu'une fraction du seuil de confiance inférieur pour le traitement témoin contre placebo – par ex. 50%, soit la moitié de l'effet obtenu. Dans notre exemple, cela donnerait un seuil de non-infériorité de 0,5 × 2% = 1%.

La fixation du seuil n'est finalement pas une décision statistique mais doit faire suite à une réflexion sur quelques points de vue cliniques. Deux conditions doivent toujours être remplies: a) le traitement d'étude est toujours encore meilleur que le placebo et l'écart est b) au maximum d'une différence cliniquement non significative par rapport au traitement témoin. Il est parfois difficile de savoir quelle différence a une importance clinique car il n'y a souvent aucune recommandation officielle d'experts. La règle d'or est: plus la maladie est grave plus la fixation du seuil doit être conservatrice. Chaque «compromis» en rapport avec le paramètre mortalité coûte la vie de patients.

Un seuil trop généreux est choisi dans certains cas, et ceci pour deux raisons: a) un traitement supposé moins efficace doit quand même être muni du sceau de la non-infériorité et b) cette manière de faire est censée réduire le nombre de cas pour améliorer la faisabilité de l'étude. Le problème est qu'un seuil inadéquat peut favoriser la conclusion de non-infériorité malgré que le traitement d'étude soit cliniquement significativement plus mauvais que le traitement témoin et que son effet soit identique ou à peine différent de celui du placebo. Un exemple concret de la littérature va l'illustrer.

L'inhibiteur de la thrombine qu'est le ximélagatran a été comparé à la warfarine dans une étude de non-infériorité chez des patients en fibrillation auriculaire. Le seuil de non-infériorité a été fixé à une augmentation annuelle de 2% des accidents vasculaires cérébraux (AVC) et thromboemboliques par rapport à la warfarine [7]. Lors de la planification de l'étude, les résultats de deux méta-analyses avaient déjà été publiés, qui montraient pour la warfarine une diminution annuelle des AVC de 3,1 et 2,8% en moyenne par rapport au placebo [8, 9]. Dans la méta-analyse Random Effects sur la warfarine [9], l'intervalle de confiance a été de 1,4 à 4,2%. En en prenant la valeur inférieure (1,4%), il aurait été d'emblée possible de démontrer une non-infériorité du ximélagatran par rapport à la warfarine sans pouvoir exclure que son effet soit plus mauvais que celui du placebo. En réalité, les accidents annuels à la fin de l'étude ont été de 1,6% sous ximélagatran et 1,2% sous warfarine, avec un intervalle de confiance de 0,13 à 1,03% pour la différence entre ces traitements [10]. Avec le seuil précédemment fixé de 2%, les auteurs de cette étude ont constaté la non-infériorité du ximélagatran par rapport à la warfarine – mais s'ils avaient mieux tenu compte des résultats de la warfarine même un seuil de non-infériorité de 1% (au lieu de 2%) n'aurait pas pu empêcher l'échec de leur étude.

3e piège: le design et la réalisation de l'étude sont insuffisants

Contrairement à une étude sur la preuve d'efficacité, dans une étude de non-infériorité le fait que les traitements soient en aveugle n'offre aucune protection contre un biais. Théoriquement, tous les échantillons prélevés chez les patients pourraient être mélangés ou toujours le même résultat pourrait être inscrit dans les dossiers pour supprimer toutes les différences. Les exemples ci-dessous sont moins extrêmes mais plus proches de la réalité: en sélectionnant des patients ayant une maladie de faible gravité, nous donnons au traitement le moins efficace une bonne chance d'avoir un effet comparable à celui du traitement témoin. Ou alors en insistant sur les effets indésirables potentiels lors de l'information donnée aux patients, ou avec une longue durée d'étude, nous faisons en sorte que seule une minorité des patients suivent les recommandations thérapeutiques; alors le résultat de l'étude est largement le reflet de l'évolution spontanée de la maladie [11].

Intentionnellement ou pas: les faiblesses dans le design de l'étude et dans sa réalisation surtout ont le même effet: les éventuelles différences entre les traitements sont effacées et une non-infériorité est favorisée comme résultat final de l'étude. Raison pour laquelle des mesures d'assurance de qualité et l'examen objectif des paramètres de l'étude ont une grande importance. Le tableau textuel 2 donne une liste de plusieurs facteurs influençant un résultat d'étude dans le sens d'une non-infériorité. Ils sont censés vous aider à poser les bonnes questions et pouvoir apprécier la qualité d'une étude de non-infériorité.

Les points cités dans ce tableau ne sont pas que de la théorie mais la réalité, comme le confirment les exemples concrets de la littérature ci-dessous: une étude chez des schizophrènes a eu comme résultat qu'un traitement par olanzapine n'est pas inférieur à celui par clozapine; mais le protocole n'a pas prévu suffisamment d'étapes de titration ce qui a fait que la majorité des patients traités par olanzapine n'a pas reçu les doses thérapeutiques cibles jusqu'à la fin de l'étude [12]. Une autre étude chez des patients anémiques a constaté l'équivalence de deux schémas de traitement par érythropoïétine, malgré que 63% seulement des patients aient reçu un traitement suivant le protocole [13]. Et au terme de grandes études comparatives sur des antihypertenseurs, les deux tiers seulement des patients prenaient encore le traitement prévu au départ [14].

Remarques conclusives décevantes

Si une étude de non-infériorité est valide ou pas, cela dépend étroitement de sa dite sensibilité de test. Il s'agit de la capacité d'une étude de faire une différenciation entre différents traitements, ou en d'autres termes: entre un traitement moins ou pas du tout efficace. Tous les pièges dont il a été question dans cet article concernent des situations dans lesquelles la sensibilité de test n'est pas (suffisamment) donnée. Avec le choix inadéquat du traitement témoin, nous ne savons pas s'il est vraiment efficace ni s'il peut servir de référence; avec un seuil de non-infériorité trop généreux ou une mauvaise réalisation de l'étude, il est possible que les différences entre le traitement d'étude et le témoin soit gommées. Dans de tels cas finalement, la conclusion

Tableau textuel 2. Facteurs influençant un résultat d'étude dans le sens non-infériorité.

• Schéma de titration inadéquat/posologie inadéquate du traitement témoin
• Population de patients inhomogène
• Maladie de faible gravité → besoin moins important d'un effet marqué
• Durée de traitement ou de suivi inadéquate
• Aucune mesure standardisée
• Grande variablilité paramétrique
• Autre traitement autorisé influençant le paramètre d'étude
• Beaucoup de non-répondeurs
• Beaucoup d'interruptions de traitement ou de l'étude
• Beaucoup d'entorses au protocole
• Faible compliance
• Nombreuses mesures manquantes
• Analyse *intention-to-treat* (plutôt que per-protocole)

de non-infériorité est tirée à tort. En réalité reste donc dans la plupart des cas la question de savoir si deux traitements sont aussi efficaces ou aussi inefficaces l'un que l'autre.

S'agit-il de cas isolés? Certainement pas, car il y a quelques années encore une revue systématique a eu pour résultat que même pas une étude sur cinq suffit aux exigences de qualité requises pour les études de non-infériorité [15]. Vous ne devriez donc ne croire que la moitié de celles qui ont été publiées. Espérons que cet article vous aidera à trouver une réponse à la question «mais quelle moitié?».

Correspondance:
Dr Peter Kleist
GlaxoSmithKline AG
Talstrasse 3–5
CH-3053 Münchenbuchsee
peter.m.kleist@gsk.com

Référence recommandée

– Piaggio G, Elbourne DR, Altman, DG, et al. Reporting of noninferiority and equivalence randomized trials. An extension of the CONSORT statement. JAMA. 2006;295:1152–60.

Vous trouverez la liste complète et numérotée des références dans la version en ligne de cet article sous www.medicalforum.ch.

CME www.smf-cme.ch

1. Le résultat d'une étude confirme la non-infériorité du traitement d'étude par rapport au traitement témoin. Lequel des arguments ci-dessous permet-il de considérer comme *douteuse* la conclusion de non-infériorité?

A Un design groupes parallèles a été utilisé pour l'étude.
B Le pourcentage des patients n'ayant pas pris régulièrement leur traitement a été de 2% sous traitement d'étude et 4% sous traitement témoin.
C Les résultats des analyses *intention-to-treat* et per-protocole n'ont été que très peu différents.
D Au cours des 6 mois de l'étude, les résultats du paramètre ont été stables après 4 mois déjà sous traitement témoin et après 5 mois seulement sous traitement d'étude.
E Dans les deux groupes, plus de la moitié des patients ont interrompu leur traitement avant terme.

2. Le seuil de non-infériorité du traitement d'étude par rapport au traitement standard est fixé à un pourcentage relatif d'accidents 20% plus élevé. Au terme de l'étude, le pourcentage absolu a été exactement de 1% dans les deux groupes. Quel intervalle de confiance pour la différence entre ces deux traitements parle en faveur de la non-infériorité du traitement d'étude?

A 0,5–1,5%.
B 0,5–0,75%.
C 0,25–2%.
D 0,9–1,1%.
E Le pourcentage d'accidents est le même dans les deux groupes, ce qui fait qu'il est impossible de calculer l'intervalle de confiance.

Drei Fallgruben bei Nicht-Unterlegenheitsstudien /

Trois pièges dans les études de non-infériorité

Literatur (Online-Version) / Références (online version)

1 Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Ann Intern Med 2000; 133: 455-63.
2 Laughren TP. The scientific and ethical basis for placebo-controlled trials in depression and schizophrenia: an FDA perspective. Eur Psychiatry 2001; 16: 418-23.
3 Leber P. Hazards of inference: the active control investigation. Epilepsia 1989; 30 (Suppl 1): S57-63.
4 Rauch SD, Halpin CF, Antonelli PJ, et al. Oral vs intratympanic corticosteroid therapy for idiopathic sudden sensorineural hearing loss. A randomized trial. JAMA 2011; 305: 2071-9.
5 Piccirillo JF. Steroids for idiopathic sudden sensorineural hearing loss. Some questions answered, others remain (editorial). JAMA 2011; 305: 2114-5.
6 Anderson P. Absence of evidence is not evidence of absence. BMJ 2004; 328: 476-7.
7 Halperin JL, and the Executive Steering Committee, on behalf of the SPORTIF III and V Study Investigators. Ximelagatran compared with warfarin for prevention of thromboembolism in patients with nonvalvular atrial fibrillation: Rationale, objectives, and design of a pair of clinical studies and baseline characteristics (SPORTIF III and V). Am Heart J 2003; 146:431-8.
8 Hart RG, Benavente O, McBride R, et al. Antithrombotic therapy to prevent stroke in patients with atrial fibrillation: A meta-analysis. Ann Internal Med 1999; 131-492-501.
9 [No authors listed]. Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation. Analysis of pooled data from five randomized controlled trials. Arch Intern Med 1994; 154: 1449-57.
10 SPORTIF Executive Steering Committee for the SPORTIF V Investigators. Ximelagatran vs Warfarin for stroke prevention in patients with nonvalvular atrial fibrillation. A randomized trial. JAMA 2005; 293: 690-8.
11 Lüdtke R. Nicht unterlegen, oder doch? Editorial. Forsch Komplementärmed 2006; 13: 332-3.
12 Tollefson GD, Birkett MA, Kiesler GM, et al. Double-blind comparison of olanzapine versus clozapine in schizophrenic patients clinically eligible for treatment with clozapine. Biol Psychiatry 2001; 49: 52-63.
13 Steensma DP, Molina R, Sloan JA, et al. Phase III study of two different schedules of erythropoietin in anemic patients with cancer. J Clin Oncol 2006; 24: 1079-89.
14 McAlister FA, Sackett DL. Active-control equivalence trials and antihypertensive agents. Am J Med 2001; 111: 553-8.
15 Le Henaff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. JAMA 2006; 295: 1147-51.