Dix exigences aux études d'équivalence thérapeutique

ou Pourquoi absence de preuve d'une différence ne signifie pas la même chose qu'équivalence

Peter Kleist
PFC Pharma Focus AG, Volketswil

Quintessence

- Par rapport aux études traditionnelles de preuve de supériorité d'un traitement contre placebo ou traitement standard, la réalisation d'études d'équivalence ou de non-infériorité impose des exigences méthodologiques nettement plus élevées.
- La «sensibilité d'assay», c.-à-d. si une étude d'équivalence est en mesure ou non de démontrer une différence entre deux traitements, a une importance capitale.
- La valeur d'une étude d'équivalence dépend dans une large mesure d'un design adéquat, du choix du témoin positif, du nombre de cas suffisant pour la question posée, du domaine d'équivalence fixé, d'une analyse adéquate et, non des moindres, de sa réalisation correcte et respectant le protocole.
- La plupart des études effectuées jusqu'ici et concluant à l'équivalence thérapeutique de deux traitements ne remplissent pas ces conditions.

Summary

Ten demands to therapeutic equivalence studies. Or why lack of evidence of difference and equivalence does not mean the same thing

- *Compared with traditional studies to prove the superiority of a treatment vs. placebo or standard treatment, the conduct of equivalence or noninferiority studies imposes markedly higher methodological demands.*
- *Of essential importance is assay-sensitivity, i.e. whether an equivalence study is at all capable of demonstrating an existing difference between two treatments.*
- *The validity of an equivalence study depends decisively on appropriate design, the choice of the positive control, a sufficient sample size for the question under study, the predefined equivalence margins, appropriate evaluation and, not least, its correct per-protocol conduct.*
- *Up to the present the majority of studies which draw conclusions concerning the therapeutic equivalence of two treatments have failed to meet these demands.*

Introduction

La majorité des études randomisées et contrôlées sont effectuées pour prouver la supériorité d'un traitement sur un autre (études dites de supériorité ou «superiority trials»). Le traitement témoin est souvent l'administration d'un placebo. Mais cela peut présenter quelques difficultés:

1. Il y a souvent un traitement efficace, ce qui fait que l'usage d'un placebo est souvent non éthique [1, 2]
2. L'examen de paramètres cliniques solides recherché dans de nombreux domaines thérapeutiques rend plus difficile la preuve de différences entre deux traitements efficaces, car les standards thérapeutiques du moment sont encore très élevés et les paramètres solides – comme accidents cardiovasculaires après infarctus du myocarde – sont toujours plus rares [3, 4]. La découverte d'une petite différence impose des études avec de très grands collectifs.
3. Aucune percée thérapeutique n'est à prévoir dans de nombreuses maladies. Les différences entre un traitement nouveau et un établi se limitent éventuellement à une meilleure tolérance, une plus grande sécurité, une plus grande simplicité d'emploi ou des coûts inférieurs, mais l'efficacité reste comparable [5, 6].

C'est pour de telles situations qu'il existe les études d'équivalence.

Définition et limites

Les études d'équivalence sont censées montrer que cliniquement, deux traitements ne sont pas significativement différents, ce qui veut dire que leurs effets sont dans un domaine dans lequel leur équivalence peut être constatée, mais qu'il n'est pas possible d'attendre un effet absolument identique. Les études d'équivalence pures, dans lesquelles des écarts encore tolérables dans deux directions (c.-à-d. «plus» ou «moins») sont définis pour prouver l'équivalence, sont rares [7]; un exemple serait une étude comparative de biodis-

Vous trouverez les questions à choix multiple concernant cet article à la page 809 ou sur internet sous www.smf-cme.ch.

ponibilité entre un générique et son original: le pic de concentration plasmatique et la surface sous la courbe concentration/temps ne doivent – pour prouver la bioéquivalence – s'écarter que dans des marges fixées officiellement (généralement 80–125%).

Dans les études d'équivalence avec questions thérapeutiques, il s'agit essentiellement d'études dites de non-infériorité («noninferiority trials»; seul le terme le plus usité de non-infériorité sera employé par la suite). Le domaine d'équivalence défini ne va généralement que dans une seule direction, celle de l'infériorité cliniquement non significative [6]. La question est donc asymétrique. Un nouveau traitement peut être recommandé si son effet est au moins comparable, mais pas inférieur à celui du traitement de référence. La démonstration d'une supériorité est à considérer comme «bonus» [7].

Il est difficile d'estimer la proportion de ce type d'étude parmi toutes les études randomisées et contrôlées, car les termes d'équivalence et de non-infériorité sont souvent utilisés à tort pour ce qui est de la question posée, du design de l'étude et des conclusions tirées des résultats obtenus. Deux études très récentes ont démontré de manière impressionnante que seulement 20% environ des études concluant à une équivalence ou à une non-infériorité respectent les critères spécifiques des études d'équivalence [8, 9]. Les deux tiers étaient des études de supériorité, dans lesquelles les tests statistiques utilisés n'ont pu démontrer aucune différence entre les groupes de traitement [8]. L'absence de preuve d'une différence, c.-à-d. que l'hypothèse zéro («Il n'y a *aucune* différence entre les traitements.») ne peut *pas* être rejetée, ne signifie toutefois pas que ces traitements sont équivalents [5]. Ils le sont peut-être, mais une étude axée sur la preuve d'une différence n'est pas faite pour prouver leur équivalence [10]. Ou comme le disent nos collègues anglophones: «Absence of evidence (of a difference) is not evidence of absence (of a difference).» [11, 12]

Au contraire des études de supériorité, les études d'équivalence ou de non-infériorité doivent respecter des exigences méthodologiques particulières pour pouvoir produire des résultats concluants:

1. Une étude de supériorité ne dépend – en plus de sa validité interne – que de l'incertitude statistique, mais pas de facteurs externes, c.-à-d. hors de l'étude [13]. Un collectif important augmente la chance de montrer une différence, mais une étude n'ayant pas une puissance statistique suffisante est sujette à une erreur de type II, c.-à-d. la fausse conclusion d'une différence qui n'existe pas entre deux traitements [14, 15]. Pour que les études d'équivalence ou de non-infériorité soient valides, toute une série de facteurs indépendants de ces études jouent un rôle essentiel.
2. La validité des études d'équivalence ou de non-infériorité est très étroitement liée à la «sensibilité d'assay», c.-à-d. à leur capacité à démontrer une différence entre deux traitements, qui consiste à distinguer un traitement efficace d'un inefficace ou moins efficace [3, 13]. Une étude de supériorité «réussie» a simultanément confirmé la sensibilité d'assay. Une sensibilité d'assay insuffisante pour les études d'équivalence a par contre pour conséquence qu'avec la démonstration d'une équivalence, nous ne savons pas si une mauvaise réalisation de ces études a empêché de démontrer une différence existant bel et bien, ou si les deux traitements sont pareillement efficaces ou inefficaces. Une condition importante est donc que le traitement témoin ait fait la preuve indubitable de son efficacité [5].

C'est de là que découlent les exigences spécifiques posées aux études d'équivalence ou de non-infériorité. L'extension récemment publiée du CONSORT-Statement sur les études d'équivalence souligne l'importance du fait qu'elles soient respectées pour garantir des résultats et conclusions d'études valides.

Dix exigences aux études cliniques visant à examiner l'équivalence thérapeutique

Si vous faites vous-même des études d'équivalence, ou si en tant que lecteur de littérature médicale vous voulez juger de leur validité, les dix points ci-dessous vous y aideront concrètement.

1. Une étude d'équivalence a, à sa base, une hypothèse claire sur l'étude de l'équivalence ou de la non-infériorité. Les hypothèses zéro et alternative sont permutées, en comparaison des études de supériorité. L'hypothèse zéro d'une étude d'équivalence veut dire que les traitements étudiés sont différents; une erreur de type I est d'admettre par erreur une équivalence, et une erreur de type II d'admettre une différence qui n'existe pas dans la réalité [5, 7]. Il n'est pas admissible de tirer la conclusion d'équivalence d'une étude conçue pour démontrer une supériorité dont le résultat est «négatif» (voir point 9). Mais cela est souvent le reflet de la réalité [8]. Par exemple, seules deux études sur 25 ayant déclaré une équivalence pour les antibiotiques utilisés dans la méningite bactérienne infantile, ont véritablement été planifiées pour répondre à cette question [16].
2. Le domaine d'équivalence ou de non-infériorité doit être défini au préalable, c.-à-d. que les limites d'une différence encore acceptable, cliniquement parlant, doivent être fixées avant le début de l'étude [7]. Ce domaine doit

– s'il n'y a pas de consensus général – être fixé par des experts indépendants [14] – pour éviter d'une part tout éventuel biais par une fixation post hoc subjective, et de l'autre un intervalle trop généreux [17]. Dans les études avec la mortalité comme paramètre surtout, il faut veiller à ce que l'écart toléré ne soit pas fixé trop large pour pouvoir admettre une non-infériorité encore acceptable par rapport à un traitement efficace [18].

Une valeur limite de non-infériorité souvent préconisée est de <50% de l'effet préalablement démontré par le traitement de référence contre placebo. Si par ex. l'effet hypotenseur supplémentaire du traitement de référence contre placebo est au moins de 10 mm Hg (limite inférieur de l'intervalle de confiance), un domaine d'équivalence imaginable serait de 6 à 14 mm Hg (soit 10 mm Hg ± 40%). Si la valeur limite était de >50%, l'efficacité du traitement test – après avoir épuisé son domaine – serait plus proche de celle du placebo que de celle du traitement témoin, ce qui est en contradiction avec l'idée même d'équivalence (ou de non-infériorité). Dans les études sur la mortalité, le domaine de non-infériorité doit généralement être défini plus étroit (voir aussi exemple au point 9) [14].

3. Le collectif est-il suffisant? Le nombre de cas d'une étude d'équivalence et de non-infériorité dépend de trois facteurs: l'intervalle de confiance, la puissance statistique et le domaine d'équivalence cité au point 2. Ce dernier est généralement plus étroit que la différence examinée dans une étude de supériorité, ce qui explique pourquoi le nombre de cas dans les études d'équivalence ou de non-infériorité est plus élevé [5]. Si le collectif est trop petit, et donc la puissance trop faible, la sensibilité d'assay diminue, c'est-à-dire la chance de démontrer une différence existant bel et bien. Un exemple extrême est celui d'une étude comparative entre octréotide et sclérothérapie dans les hémorragies sur varices œsophagiennes: la conclusion de l'équivalence thérapeutique a été tirée sur la base de 100 patients admis dans une étude, malgré le fait que la méthode a jugé qu'il fallait 1800 cas [19]. La probabilité de pouvoir découvrir une différence a été réduite à 5% dans cette étude.
4. Pour le choix du témoin positif, donc du traitement actif comparatif, il doit s'agir d'un standard bien établi [5, 10]. Une admission officielle à elle seule (surtout si elle date d'un certain temps) n'est pas suffisante, dans la mesure où elle n'exprime pas toujours un effet significatif par rapport à un placebo. De nombreux médicaments dont l'efficacité a été prouvée, par ex. différents antidépresseurs, analgésiques, antiangoreux ou antihypertenseurs, ne sont parvenus à démontrer aucune supériorité sur le placebo dans un nombre non négligeable d'études [13]. Le problème que cela pose a été illustré par Tramèr et al. à l'exemple d'études avec l'ondansétron dans la prévention du vomissement postopératoire [20]: comment supposer une efficacité de l'ondansétron – comme traitement témoin dans une étude comparative – si ce médicament n'a pas été significativement meilleur que le placebo dans 30% des études précédemment effectuées? En l'absence de preuve suffisante de l'efficacité du traitement témoin, il est impossible de répondre à la question de savoir si une équivalence démontrée exprime l'efficacité comparable ou la non-efficacité comparable des traitements étudiés.
5. Le design d'une étude doit se référer le plus près possible aux études dans lesquelles le traitement témoin a déjà fait la preuve de son efficacité [5, 7]. D'importants éléments à cet égard sont la sélection des patients, la dose, les autres traitements admis ou pas admis, sans oublier la durée du traitement et du suivi [5, 6, 13]. Les traitements témoins utilisés dans quelques grandes études comparatives sur les antihypertenseurs n'ont souvent pas démontré une efficacité suffisante dans les collectifs de patients étudiés [21]. Il est possible de conclure par erreur à une équivalence (ou non-infériorité) thérapeutique même avec des patients ayant une symptomatologie très discrète ou s'écartant très peu des normes, car l'effet thérapeutique exigé est trop faible (domaine suprathérapeutique). Des schémas de titration inadéquats et des doses thérapeutiques cibles insuffisantes peuvent avoir une importance considérable pour le résultat (partiellement douteux) d'études comparatives contre traitement actif, ce qui est par exemple le cas dans le domaine de la psychiatrie.
6. Le paramètre de l'étude est-il adéquat? Il devrait d'une part être le même que celui utilisé préalablement pour la preuve de l'efficacité du traitement actif témoin; et de l'autre, il doit remplir certaines conditions: il ne doit pas être «trop variable», car cela augmente en soi la probabilité d'équivalence au terme de l'étude. La variabilité du paramètre est certainement l'une des causes les plus fréquentes de la variabilité des résultats des substances citées au point 4 [13, 20]. Si le nombre prévu d'incidents est faible (paramètres se produisant relativement rarement), il faut parfois examiner un paramètre combiné – mais attention: les composantes de ce paramètre ne répondant pas de la même manière au traitement, elles augmentent elles aussi la variabilité et donc le manque de sécurité des études d'équivalence [23]. C'est pourquoi par exemple un «time-to-treatment-failure» n'est pas adéquat, car une interruption de traite-

ment pour effets indésirables ou une sortie d'étude volontaire a peu d'importance pour la maladie de fond ou l'efficacité pharmacologique d'un traitement [24].

7. Le standard de la réalisation de l'étude doit être très élevé pour les études d'équivalence ou de non-infériorité tout spécialement [5], car les écarts par rapport au protocole peuvent effacer ou annuler les différences existant entre deux traitements. De nombreux non-respects du protocole, un nombre élevé d'interruptions ou de crossovers thérapeutiques, de même que des mesures insuffisamment standardisées augmentent la probabilité d'équivalence. Une randomisation insuffisante peut fausser le résultat de l'étude dans les deux sens. Le lecteur de publications sur des études doit être particulièrement attentif à ces facteurs et à la qualité de leur réalisation.
8. Il ne faut pas sous-estimer la (non-)compliance des patients, surtout dans la fiabilité de la prise des médicaments d'étude [10]. Au terme de grandes études sur des antihypertenseurs par exemple, les deux tiers des patients en moyenne recevaient encore le médicament auquel ils avaient été randomisés au départ [21]; le niveau de compliance est en outre sujet à caution – car en fin de compte c'est le manque de compliance dans la prise des médicaments qui est la principale raison du fait que la tension artérielle n'est suffisamment contrôlée que chez un quart des hypertendus. Pour ne pas tirer de fausses conclusions sur l'équivalence (car moins la compliance est bonne, plus il est probable qu'il y aura équivalence), il est indispensable de prouver que la compliance est suffisante et comparable dans les deux groupes de traitement; en plus du classique «pill count» («drug account»), il faut en principe envisager de recourir à des distributeurs de médicaments qui enregistrent la date et l'heure de la prise des comprimés (MEMS, ou «medication event monitoring systems»).
9. Une étude d'équivalence doit être jugée sur la base des intervalles de confiance; les tests statistiques conventionnels ne jouent aucun rôle pour ce type d'étude [5, 7]. Même pour les études de non-infériorité, il faut éventuellement poser une double question (intervalle de confiance 95% bilatéral) pour pouvoir faire une analyse (consécutive) en matière de supériorité [18]. Car si dans une étude planifiée pour la supériorité le test d'équivalence est généralement interdit, il est en principe possible dans une étude d'équivalence d'étendre l'analyse vers une éventuelle supériorité [3]. Le plan de l'analyse doit exactement préciser sur quelle base il est procédé à une mutation de l'analyse de non-infériorité vers supériorité.

Un intervalle de confiance 95% dit – en simplifiant – qu'en répétant 100 fois l'étude 95 résultats seraient dans le même domaine. Formulé autrement: il montre avec une probabilité de 95% le domaine dans lequel le résultat effectif se trouverait si un très important collectif de patients était examiné dans de très nombreuses études. L'intervalle de confiance correspond ainsi à un témoin de la fiabilité d'un seul résultat d'une seule étude, ce qui est forcément sujet à caution. Pour tirer la conclusion d'équivalence ou de non-infériorité, l'intervalle de confiance doit se trouver dans le domaine d'équivalence ou de non-infériorité préalablement fixé. La différence de mortalité constatée dans une étude entre les traitements A et B pourrait par exemple être de 10% en faveur du traitement A (c.-à-d. –10% par rapport à B). Un test statistique standard ne montre pas de différence significative ($p > 0{,}05$). Mais avec un intervalle de confiance large (par ex. de –24% à +14%), et un domaine de non-infériorité préalablement fixé (–x% jusqu'à max. +10%), il n'est pas possible de conclure à la non-infériorité du traitement A, car il n'est pas exclu que le traitement B soit cliniquement et statistiquement meilleur que le traitement A. Le p du test statistique ne répond donc pas grand-chose à la question d'équivalence ou de non-infériorité.

Un autre exemple bien réel illustre l'importance des intervalles de confiance pour l'analyse. Dans l'étude COBALT, la double administration en bolus de t-PA a été comparée à la perfusion continue chez plus de 7000 patients victimes d'un infarctus du myocarde [26]. La mortalité sous bolus a été un peu plus élevée, et le haut de l'intervalle de confiance était de 1,49%, ce qui est supérieur au seuil de non-infériorité préalablement fixé à 1,4%, c.-à-d. qu'une augmentation maximale de 0,4% de la mortalité aurait encore été considérée comme équivalente. Cette étude n'a donc pas pu démontrer l'équivalence de ces deux schémas de traitement. D'autres exemples sont présentés dans la figure 1 .

10. Dans une étude de supériorité, l'analyse «intention-to-treat» (ITT) est l'option conservatrice. Dans une étude d'équivalence et de non-infériorité, l'analyse ITT est généralement sujette à un biais dans le sens équivalence, car chaque écart de la répartition de traitement randomisée (par ex. suite à interruption d'étude ou crossover de traitement) diminue potentiellement les différences entre les traitements et augmente ainsi la susceptibilité à une erreur de type I [18]. Il faut donc procéder à une analyse «per-protocol (PP, ou «as treated»), et les résultats obtenus dans une étude d'équivalence sont

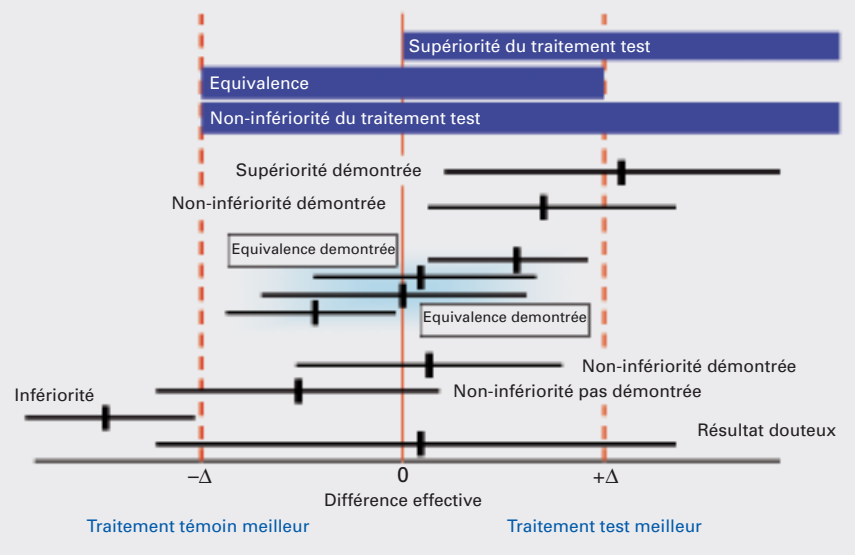

Figure 1
Exemples de résultats possibles d'une étude d'équivalence ou de non-infériorité, en fonction des intervalles de confiance. Les pointillés verticaux correspondent aux marges du domaine d'équivalence préalablement fixées (–Δ à +Δ) (figure adapté d'après Jones et al. [5], Pater [10] et Gomberg-Maitland et al. [14]).

plus dignes de confiance si les analyses ITT et PP ne diffèrent pas de manière notable [5, 7, 27]. L'analyse exclusivement ITT faite dans une étude de deux schémas d'administration d'érythropoïétine par exemple (avec une compliance de 63%) est certainement à considérer comme inadéquate [27, 28].

Trois autres particularités méthodologiques des études d'équivalence

1. Contrairement aux études de supériorité, la mise en aveugle des traitements joue un rôle secondaire dans les études d'équivalence, car cela ne confère pas de protection efficace contre un biais dans le sens d'une équivalence des résultats de traitement, par ex. par des scores ou répartitions par stades les plus semblables possibles [3].
2. Si les analyses intermédiaires confirment l'équivalence ou la non-infériorité, il n'y a aucune raison médicale d'interrompre l'étude (pour autant que la tolérance et la sécurité soient comparables) [7]. Une étude peut se poursuivre pour éventuellement montrer à son terme la supériorité du traitement test [18].
3. Si le traitement de référence n'est qu'insuffisamment établi, il faut envisager d'effectuer une étude d'équivalence à trois bras, contrôlée en plus contre placebo. Ce n'est que si le traitement de référence démontre une supériorité sur le placebo que la question de l'éventuelle équivalence entre les traitements test et de référence pourra être abordée à l'étape suivante [6, 13, 29]. Une telle marche à suivre correspond à l'établissement d'un standard interne pour déterminer la sensibilité d'assay.

Conclusion

La réalisation d'études d'équivalence est une alternative aux études de supériorité traditionnelles, s'il existe déjà un traitement efficace dans l'indication étudiée. Les exigences méthodologiques posées aux études d'équivalence sont cependant un peu plus élevées. De très nombreux facteurs peuvent faire conclure à l'équivalence bien que les traitements soient en réalité différents.

La validité d'une étude d'équivalence ou de non-infériorité dépend dans une large mesure de son design, du choix du témoin positif, du collectif suffisant pour la question étudiée, du domaine d'équivalence fixé, d'une analyse adéquate et surtout de sa réalisation correcte, respectant le protocole.

Pour l'admission de nouveaux médicaments ou de nouvelles indications, les instances de contrôle des médicaments exigent également explicitement une justification pour le domaine d'équivalence choisi [30, 31] et une sensibilité d'assay suffisante [32]. Si les traitements de référence ne se sont pas toujours avérés efficaces contre placebo, il faut une étude à trois bras, contrôlée contre placebo, pour suivre le standard officiel, par ex. pour étudier de nouveaux antidépresseurs [33].

Correspondance:
Dr Peter Kleist
PFC Pharma Focus AG
Chriesbaumstrasse 2
CH-8604 Volketswil
peter.kleist@pfc.ch

Bibliographie recommandée

– Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ. 1996;313:36–9.
– Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW, for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials. JAMA. 2006;295:1152–60.
– Pocock SJ. The pros and cons of noninferiority trials. Fundamental Clin Pharmacol. 2003;17:483–90.
– Temple R, Ellenberg SS. Placebo-controlled trials and active-controlled trials in the evaluation of new treatments. Part I: ethical and scientific issues. Ann Intern Med. 2000;133:455–63.

Vous trouverez la bibliographie complète [1–33] dans la version en ligne de l'article sous www.medicalforum.ch/pdf/pdf_f/2006/2006-37/2006-37-119.pdf.