Société Française de Pharmacologie et de Thérapeutique

Groupe de Travail Méthodologie

Livre blanc SFPT

De la nécessité de la méthodologie dans l'évaluation des médicaments

Document compagnon

Dossier 2 – Les biais de l'essai de supériorité

14 février 2022

Comité de rédaction et relecture (par ordre alphabétique)

Jean Luc Cracowski

Michel Cucherat

Dominique Deplanque

Behrouz Kassai

Charles Khouri

Silvy Laporte

Clara Locher

Florian Naudet

Edouard Ollier

Matthieu Roustit

Table des matières

1 Introduction 7
2 Apport de la méta-épidémiologie 10
3 Biais prévenus par la randomisation imprévisible 11
4 Biais prévenus par le double insu vis-à-vis de la mesure du critère de jugement 13
4.1 Limitation des biais liés à la mesure dans les essais en ouvert 14
5 Biais prévenus par le double insu vis-à-vis de la réalisation de l'essai 16
6 Biais prévenus par l'analyse en ITT 17
6.1 Les méthodes de remplacement des données manquantes 18
7 Évaluation globale du risque de biais 21

1 Introduction

Un essai de supériorité est biaisé quand un autre facteur que le traitement étudié induit la différence en faveur du nouveau traitement observée au niveau du ou des critères de jugement.

> *Par exemple dans un essai évaluant un antiagrégant plaquettaire versus placebo pour prévenir les AVC dans la FA, si tous les patients du groupe aspirine reçoivent en plus des AVK et aucun dans le groupe placebo, il est évident que l'on aura moins d'AVC dans le groupe traité que l'aspirine prévienne ou pas en réalité l'AVC.*

Un biais est donc une cause de résultat faux positif[1].

Sauf cas exceptionnel, il est impossible de déterminer a posteriori si un résultat est effectivement biaisé ou non (étant donné que l'on ne connait pas le réel effet du traitement). Il n'est donc pas possible de diagnostiquer a posteriori si un résultat est biaisé ou pas.

> *De plus, il s'agirait d'une argumentation a posteriori, basée entièrement sur un raisonnement exploratoire (inductif) consistant à une recherche tous azimuts de « signes de biais » et qui finalement serait très subjective et influencée par l'opinion du lecteur (risque de procès à charge ou de cécité élective vis-à-vis des problèmes).*

Cependant, il a été possible d'identifier toutes les causes de biais qui peuvent survenir dans un essai (cf. Tableau 1) et d'inventer des principes méthodologiques qui empêchent leur survenu (randomisation imprévisible, double insu, analyse en intention de traité avec remplacement des données manquantes).

En appliquant ces principes méthodologiques, il est possible de mettre un essai à l'abri des biais (de le protéger contre les biais).

Ainsi, si ces principes méthodologiques de protection contre les biais ont été correctement mis en œuvre, les résultats « positifs » obtenus ne peuvent pas provenir de biais (mais il est encore possible d'avoir une erreur aléatoire). Il sera donc possible de les considérer comme réels et de conclure à l'intérêt du traitement (après analyse de la signification statistique, cf. dossier compagnon 1 – le risque alpha).

Pour produire des résultats avec un degré de certitude suffisant pour baser un changement de pratique, **un essai doit être à l'abri des biais** (au niveau de sa conception et de sa réalisation).

- Si c'est le cas, un résultat positif ne pourra pas être un faux positif dû à un biais et pourra être accepté comme tel.

[1] Au sens large, un biais peut aller dans les deux sens. Mais dans l'essai thérapeutique de supériorité, on se préoccupe uniquement des biais qui pourraient faire conclure à tort à l'intérêt du traitement. Les biais qui conduisent à faire que l'essai ne peut pas conclure à l'effet du traitement n'ont pas pour conséquences de faire adopter un traitement sans intérêt. Cette problématique concerne surtout le chercheur ou l'industriel qui développe le traitement et non pas le clinicien qui se pose la question de la fiabilité du résultat sur lequel il s'apprête d'adopter le nouveau traitement. L'interprétation des essais non concluants (« négatifs ») est particulière et complètement différente de celle des essais « positifs » que nous développons ici (cf. section **Erreur ! Source du renvoi introuvable.**)

- Si l'étude n'est pas complètement à l'abri des biais (mise en œuvre partielle des principes méthodologiques ou perversion de ces principes lors de la réalisation), **l'étude est à risque de biais**. Les résultats « positifs » produits par une telle étude peuvent être potentiellement dus en totalité aux biais et donc être faussement positifs. Il n'est donc pas possible de considérer ces résultats pour baser un changement de pratique (car il y a un risque de recommander ce changement de pratique à tort).

La validité interne est remise en cause, non pas parce que l'on a la preuve évidente d'un biais, mais parce que l'essai est à risque de biais, car insuffisamment protégé contre les biais.

Il est donc abusif de dire qu'un essai est biaisé, car la seule conclusion objective qui puisse être faite est que l'essai est protégé contre les biais ou non

Tableau 1 – Présentation des 4 biais pouvant affecter un essai thérapeutique de supériorité, classés en fonction du principe méthodologique correspondant

Biais	**Mécanisme du biais prévenu**	**Nom du biais[2]**
Biais prévenu par la randomisation imprévisible	Biais survenant quand le groupe traité est favorisé par la sélection de patients moins graves que ceux inclus dans le groupe contrôle	Biais de sélection (ATTENTION ne correspond pas en totalité au biais de sélection des études épidémiologiques)
Biais prévenu par le double insu au niveau de la mesure du critère de jugement	Biais survenant quand la mesure du critère de jugement favorise le groupe traité	Biais de mesure
Biais prévenu par le double insu au niveau de la réalisation et du suivi de l'essai	Biais survenant quand la prise en charge des patients favorise le groupe traité	Biais de suivi (réalisation)
Biais prévenu par l'analyse en intention de traiter avec remplacement des données manquantes	Biais survenant quand le groupe traité est favorisé par la « sortie de l'analyse » de certains patients	Biais d'attrition

Pour qu'il y ait biais conduisant à un résultat faussement positif dans l'essai de supériorité, il faut donc qu'un facteur conditionnant le critère de jugement soit asymétrique entre les 2 groupes et favorise le groupe traité. Ainsi les facteurs qui n'influencent pas le critère de jugement ne peuvent pas induire de biais ainsi que les facteurs dont la répartition est symétrique[3] (y compris en moyenne) entre les 2 groupes.

Dans le discours courant, le terme biais est souvent utilisé de manière inappropriée pour désigner tout problème perçu avec un essai. En fait les biais ne représentent qu'un type, parfaitement bien défini, des réserves que l'on peut émettre vis-à-vis d'une étude. Il est par exemple totalement inapproprié de parler de biais statistiques pour désigner un problème lié au risque alpha.

Par exemple, le terme biais de sélection est souvent utilisé à tort pour parler d'un défaut de représentativité des patients inclus, par exemple, si un essai qui voulait inclure des patients âgés se retrouve avec très peu de ces patients. Il ne s'agit pas d'un biais étant donné que le problème survient en amont de l'inclusion, mais bien d'un problème de pertinence clinique. C'est un problème de validité externe et non pas de validité

[2] Les noms de biais sont très variables d'un auteur à l'autre avec de nombreux synonymes parfois ambigus entre le monde de l'essai clinique et celui de l'épidémiologie. Il n'est pas très important de mémoriser le nom de biais. L'important est de comprendre les mécanismes des biais et en quoi les principes méthodologiques les évitent.

[3] Cela ne s'applique pas à l'essai de non-infériorité où les biais problématiques sont ceux qui diminuent la différence entre les 2 traitements et font apparaitre un traitement non inférieur au standard un traitement en réalité très inférieur.

interne. Un biais est un facteur qui fait que le résultat que l'on obtient dans l'étude est différent de celui qu'il aurait dû être compte tenu des patients inclus. Le fait que l'étude ne permet pas de répondre à la question posée en termes de représentativité, de contexte de réalisation, de définition de la maladie est un problème de validité externe et fait que le résultat (pourtant intrinsèquement correct) ne peut pas servir à guider la pratique, car il ne reflète pas forcément le réel bénéfice qu'apporterait éventuellement ce traitement chez les patients à traiter dans la vraie vie (qui ne correspondront pas à ceux qui ont été effectivement étudiés dans l'étude).

2 Apport de la méta-épidémiologie

Plusieurs études méta-épidemiologique ont recherchés les facteurs associés à la survenue d'un biais dans les essais thérapeutiques [1, 2, 3, 4, 5]. Le principe consiste à comparer, à partir d'un ensemble d'essais évaluant tous le même traitement chez les mêmes patients, l'influence de variables liées à la méthodologie des études (absence de double aveugle, absence de randomisation, randomisation prévisible, absence d'analyse en ITT, etc.) sur la taille d'effet observé.

Ces études ont ainsi permis d'identifier, de manière empirique et non pas théorique, les types d'essais à risque de biais. Cette recherche a ensuite permis l'élaboration d'outils d'évaluation du risque de biais, validés de manière empirique : le score de Jadad [6], le ROB Cochrane puis le ROB 2.0 [7] actuellement recommandé pour l'évaluation méthodologique des études (en méta-analyse entre autres).

Un autre apport fondamental de ces études a été de montrer que les études biaisées surestiment aussi la taille de l'effet traitement, ce qui est entièrement logique vu que pour conclure à tort à un effet du traitement dans une situation où le traitement a un effet nul, il faut bien qu'un effet non nul soit observé, donc surestimé par rapport à la réalité. Ainsi il s'est avéré que les essais en ouvert surestimé d'environ 15 à 20% la taille de l'effet. Avec un traitement sans effet le risque ratio attendu est de 1. Dans un essai en ouvert le risque ratio observé pourra être entre 0.85 et 0.80 en moyenne.

3 Biais prévenus par la randomisation imprévisible

Contrairement à ce qui est fréquemment présenté, le but de la randomisation n'est pas de créer deux groupes identiques [8]. En effet, rien ne garantit que la randomisation (qui est un processus purement aléatoire[4]) « mette » exactement le même nombre de femmes dans les 2 groupes, ou le même nombre de diabétiques. À l'issue d'une randomisation, il peut y avoir des différences de patients entre les 2 groupes, mais qui sont des différences dues uniquement au hasard, pouvant certes fausser l'estimation de l'effet du traitement, mais qui ne sont pas systématiques. Il s'agit alors d'une erreur aléatoire gérée par le test statistique et non pas d'une cause de biais (car non systématique, la re-randomisation des mêmes patients ne conduira pas aux mêmes différences entre les 2 groupes).

La randomisation assure que la nature du traitement reçu par un patient ne dépend en rien de ses caractéristiques. Elle ne garantit pas que les 2 groupes seront identiques.

On dit souvent que la randomisation donne 2 groupes comparables. Cela ne veut pas dire que les 2 groupes sont identiques (cf. supra), mais qu'ils permettent de faire une comparaison loyale qui ne sera pas systématiquement influencée par autre chose que le traitement étudié (comparable veut dire apte à faire une comparaison non biaisée et non pas que les deux groupes sont identiques).

Il est ainsi sans intérêt de vérifier que les 2 groupes issus de la randomisation (table 1, des caractéristiques à l'inclusion) sont effectivement identiques [8][5]. La comparabilité des groupes est garantie par l'allocation aléatoire des traitements (et par sa non-perversion lors de la réalisation de l'étude, cf. ci-dessous). Il est tout à fait possible que les 2 groupes diffèrent sur certaines caractéristiques, mais cela sera uniquement du fait du hasard. Ces différences n'introduiront donc pas un biais, mais éventuellement une erreur aléatoire, prise en compte naturellement par le test statistique.

> *Par exemple si un test statistique était effectué pour chaque caractéristique des patients, 5% d'entre elles seraient significatives au seuil de 5%, par définition. Cela montre l'inutilité de faire de tels tests qui ne sont pas effectués en pratique pour cette raison.*

En pratique, il suffit que l'allocation des traitements soit vraiment aléatoire et imprévisible pour pouvoir conclure que l'essai est à l'abri des biais à ce niveau.

Pour être efficace, une randomisation doit être imprévisible, c'est-à-dire que les investigateurs ne peuvent pas connaitre la nature du traitement que devrait recevoir dans l'essai un nouveau patient avant de l'avoir effectivement inclus.

---

[4] Toute la justification du test statistique dans un essai repose sur le fait que le hasard peut créer des différences (sur le critère de jugement qui pourraient faire conclure à tort à l'existence d'un effet du traitement). Il est donc impossible que le même hasard, lorsqu'il est utilisé pour allouer les traitements, ne puisse plus créer des différences et veillerait à bien répartir toutes les caractéristiques des patients entre les 2 groupes.

[5] Sauf si le but est de montrer qu'il y a eu une perversion de la randomisation, ce qui remettrait en cause le control des biais de l'étude.

Un exemple de randomisation prévisible est la randomisation par enveloppes dans un essai en ouvert. En principe, après avoir inclus le patient dans l'étude, l'investigateur doit ouvrir la première enveloppe disponible pour connaitre le traitement alloué à ce patient[6]. Mais rien ne l'empêche d'ouvrir l'enveloppe avant de formaliser l'inclusion du patient et de ne le faire que si la nature du traitement lui convient (certains investigateurs préfèrent tel ou tel traitement en fonction des caractéristiques des patients).

Pour éviter cela, il faut que la nature du traitement ne soit communiquée à l'investigateur d'après l'inclusion effective du patient dans l'essai. Cela est obtenu par une procédure centralisée par le Web ou téléphone. Si l'investigateur, après, décide ne pas donner ce traitement au patient, cela n'introduira pas de déséquilibre entre les groupes puisque ce patient sera maintenu dans son groupe de randomisation du fait de l'analyse en intention de traiter.

> "Investigators used an interactive voice- or Web response system to determine treatment assignment"
> "Patients were randomly assigned in a 1:1:1 ratio by means of an interactive voice–Web response system to one of two secukinumab dose groups or a placebo group" [10.1056/NEJMoa1412679]
> "The allocation was performed using a sealed envelope system. The treatment allocations were not masked to the patients and the treating physicians."

Dans un essai en double insu, tout type de randomisation est imprévisible, mais les procédures centralisées sont aussi largement utilisées.

	Biais lié à la randomisation
Essai en ouvert + randomisation prévisible (non centralisée comme les enveloppes)	Risque de biais
Essai en ouvert + randomisation imprévisible (centralisée, téléphone, WEB)	À l'abri des biais
Essai en double insu (quel que soit la méthode de randomisation)	À l'abri des biais

[6] Comme l'essai est en ouvert, les enveloppes révèlent la nature du traitement reçu. Dans un essai en double aveugle, l'enveloppe contient un n° de boite dont il est impossible de savoir la nature du traitement contenu dans cette boite

4 Biais prévenus par le double insu vis-à-vis de la mesure du critère de jugement

Le double insu (*double blind, double masked, blindness*) empêche que la mesure du critère de jugement soit influencée par la nature du traitement reçu et puisse favoriser systématiquement le traitement évalué.

> Un essai en ouvert compare HBMP et héparine non fractionnée (HNF) dans la thromboprophylaxie en chirurgie orthopédique. Le critère de jugement est la thrombose veineuse superficielle (TVP) suspectée cliniquement et confirmée par phlébographie. En réalité les 2 produits sont strictement équivalents et donnent une fréquence de TVP postopératoire de 5%. Cependant les investigateurs sont convaincus que les HBPM sont plus efficaces. Devant une suspicion de phlébite, ils demanderont plus facilement une vérification phlébographie pour les patients du groupe contrôle que pour ceux du groupe traité. Ainsi, une plus grande proportion des TVP existantes sera détectée dans le groupe contrôle que dans le groupe traité. Si avec ce plus grand recours à la phlébographie, 95% des TVP sont détectées dans le groupe contrôle contre seulement 60% dans le groupe traité, la fréquence observée de TVP (le critère de jugement) sera de 5%*60% = 3% contre 5%*95%=4.75%, faisant croire à une supériorité des HBPM par rapport à l'HNF.

Un essai est en double insu quand personne ne connait la nature du traitement reçu par le patient, ni le patient lui-même, ni les médecins ou les autres soignants qui s'occupent de lui : médecins qui appliquent le traitement, qui prennent en charge le patient, qui mesurent le critère de jugement. Le double insu est en réalité un quadruple insu. Si un des éléments de la chaine connait le traitement reçu, la possibilité de biais réapparait.

Le double insu est obtenu grâce à l'utilisation d'un placebo identique en tout point au traitement évalué (*matching placebo*)

> "patients were randomly assigned to receive either memantine (20 mg per day; Merz) or an identical appearing placebo."
> "Enrolled, eligible patients were randomly assigned to receive either ticagrelor or matching placebo, in accordance with the sequestered, fixed-randomization schedule"

En cas de galéniques très différentes (comme avec la comparaison entre une forme orale et une forme intraveineuse) la technique du double placebo (*double-dummy*) est utilisée.

> Rocket a comparé le rivaroxaban à la warfarine dans la FA [10.1056/NEJMoa1009638]
> *"Rocket was a multicenter, randomized, double-blind, double-dummy, event-driven trial. … Patients were randomly assigned to receive fixed dose rivaroxaban or adjusted-dose warfarin (target international normalized ratio [INR], 2.0 to 3.0). Patients in each group also received a placebo tablet in order to maintain blinding."*
> La réalisation de cet essai en double aveugle était un défi, car la warfarine nécessite un ajustement de dose en fonction de l'INR et pas le rivaroxaban. L'utilisation d'un double placebo n'est donc pas suffisante pour assurer que les 2 bras de l'essai soient indistinguables. Il faut en plus que les investigateurs ajustent les doses de la « warfarine » (verum ou placebo) dans les 2 groupes. Dans le groupe rivaroxaban, ils ajusteront le placebo à partir d'INR factice (sham INR).
> *"A point-of-care device was used to generate encrypted values that were sent to an independent study monitor, who provided sites with either real INR values (for patients in the warfarin group in order to adjust the dose) or sham values (for patients in the rivaroxaban group receiving placebo warfarin) during the course of the trial. Sham INR results were generated by means of a validated algorithm reflecting the*

distribution of values in warfarin-treated patients with characteristics similar to those in the study population."

4.1 Limitation des biais liés à la mesure dans les essais en ouvert

Si l'essai ne peut pas être réalisé en double aveugle (chirurgie conservatrice par rapport à une chirurgie d'amputation par exemple), les biais liés à la mesure pourront être évités si le critère est parfaitement objectif (c'est-à-dire non sujette à une quelconque interprétation).

Il n'y a guère que la mortalité totale qui soit un critère parfaitement objectif, n'entrainant pas des erreurs de classement. Même la détermination de la cause du décès (cardiovasculaire, traumatique, etc.) peut être sujette à interprétation dans les cas compliqués (patients avec de nombreuses comorbidités) ou ambiguës (absence d'autopsie).

Au mieux, si le critère est partiellement subjectif, les biais pourront être partiellement limités par le recours à un comité d'adjudication des évènements en aveugle. Cependant cela ne remplace pas le double insu, car 1) il reste les biais liés au suivi et 2) la documentation médicale des cas est transmise à ce comité par les investigateurs eux-mêmes, et la connaissance du traitement reçu peut influencer la quantité et la précision de l'information transmise.

The clinical-events committee of the TIMI Study Group adjudicated all components of the primary outcomes and key components of other safety and efficacy outcomes

Dans certaines situations, la réalisation du double aveugle est impossible. Les essais, réalisés forcément en ouvert, ne sont pas pour autant à l'abri des biais. Dans ces domaines, il y a une impossibilité structurelle à contrôler tous les biais qui fait qu'il sera impossible d'obtenir des preuves totalement fiables.

En absence de double insu, les biais peuvent être important, pouvant faire croire à un bénéfice important du traitement même si celui n'en apporte aucun en réalité. Pour cette raison il est recommandé que les essais se fassent en double insu même si cette réalisation est compliquée. Ainsi, le double insu est devenu aussi le standard de réalisation des essais en chirurgie (et des dispositifs médicaux). Pour assurer cet insu, le groupe contrôle reçoit une intervention chirurgicale factice (sham intervention).

Au début des années 2000, la greffe de cellule souche semblait être un traitement efficace dans la maladie de Parkinson sévère. Des essais randomisés en ouvert, où le groupe contrôle était seulement observé, avait montré une amélioration du score d'évaluation de l'intensité du Parkinson. Mais compte tenu du fait que ces essais pouvaient être biaisés, un nouvel essai a été entrepris en double aveugle cette fois-ci [10.1056/NEJM200103083441002].
L'introduction de l'article justifie le recours au double insu de la façon suivante : « *We and others have reported that transplanted dopamine neurons survive and that patients may have progressive clinical improvement over a period of three to four years. All these studies were unblinded, and the number of patients in each was small… We conducted a double-blind, sham-surgery–controlled trial of the implantation of embryonic dopamine neurons in patients with severe Parkinson's disease*."
La réalisation de cet essai méthodologiquement sans faille fut d'un grand apport, car il n'a pas confirmé le bénéfice de cette thérapie cellulaire. Sans cet essai, la prise en charge des parkinsons sévères se serait fourvoyée dans une voie thérapeutique lourde, non dénuée de risque pour les patients, couteuse et finalement n'apportant pas le bénéfice escompté.

Au premier abord la « sham intervention » semble posé un problème éthique (anesthésie, incision cutanée), mais il faut bien réaliser que l'enjeu éthique d'un essai est surtout ne pas conduire à valider à tort un traitement sans intérêt. Ainsi, il n'y a rien de moins éthique qu'un essai de faible qualité méthodologique, exposant à un risque de mauvaise prise de décision. Bien entendu les patients participant à l'étude doivent être informés et volontaires.

	Biais lié à la mesure
Essai en double insu	À l'abri des biais
Essai en ouvert, mais critère de jugement parfaitement objectif	À l'abri des biais
Essai en ouvert, comité d'adjudication en aveugle	Risque de biais
Essais en ouvert, critère subjectif	Risque de biais

5 Biais prévenus par le double insu vis-à-vis de la réalisation de l'essai

Le double insu empêche aussi que la prise en charge des patients soit différente entre les 2 groupes en termes de soin complémentaire, de traitements concomitants ou de secours (*rescue treatment*), de décision d'abandon des traitements curatifs et de passage en soins palliatifs, etc.

Si dans un essai en double insu, un investigateur décide de donner un traitement actif en plus des traitements de l'étude, il le fera de la même façon dans les deux groupes (qui sont indistinguables). Cela peut conduire à ce que les patients des 2 groupes soient traités de façon identique. L'essai sera « négatif » même si le nouveau traitement est supérieur au contrôle en réalité. Mais cela ne peut pas favoriser le traitement évalué, car il ne sera pas possible de favoriser uniquement ces patients.

	Biais lié à la réalisation
Essai en double insu	À l'abri des biais
Essais en ouvert	Risque de biais

6 Biais prévenus par l'analyse en ITT

L'analyse en intention de traiter consiste à inclure dans l'analyse de l'effet du traitement sur le critère de jugement tous les patients randomisés, dans le groupe où ils ont été randomisés, sans tenir compte des évènements intercurrents qui auraient pu survenir : erreur de traitement, arrêt du traitement (*treatment discontinuation, treatment stopped prematurely*), retrait de l'étude (*withdrawal*), recours au traitement de l'autre groupe, inclusion à tort (*included in error*), patient perdu de vue (*lost to follow-up*), etc.).

> All efficacy and safety analyses were based on the intention-to-treat principle and included all the patients who underwent randomization. [10.1056/NEJMoa1916870]

Cette analyse empêche que l'on puisse conditionner le résultat de l'étude en sortant des patients de l'analyse.

Si des patients sont « perdus de vue » (*lost to follow-up*), c'est-à-dire s'ils ne sont pas venus à la dernière visite de l'étude où les critères de jugement étaient mesurés, il est impossible de les faire contribuer à l'analyse. La valeur de leur critère de jugement est manquante, on parle de données manquantes (*missing value*).

> *Un essai a comme critère de jugement la fraction d'éjection ventriculaire gauche (FEVG) mesurée par échocardiographie cardiaque lors de la dernière visite de suivi. L'effet du traitement sera mesuré par la différence de moyenne entre les 150 patients inclus dans le groupe traité et les 148 du groupe contrôle. Douze patients dans chaque groupe ne sont pas venus à la dernière visite (ils ont été perdus de vue). Même avec la volonté de les faire participer à l'analyse, cela sera impossible, car ils ne peuvent pas être pris en compte dans le calcul des 2 moyennes (les prendre en compte qu'au niveau du dénominateur entraine une sous-estimation de la moyenne). Pour les faire contribuer à la moyenne, il faut remplacer les valeurs manquantes de VEFG pour ces patients par une valeur arbitraire, mais en veillant bien à ce qu'elles ne puissent pas favoriser le groupe traité.*

Les données manquantes représentent un risque de biais, car elles peuvent survenir en fonction du traitement reçu et de l'évolution du patient.

> *Un essai évalue un antidépresseur versus placebo. Le critère de jugement est l'échec du traitement à la 12ème semaine mesuré avec un score de dépression. Les patients qui présentent des effets secondaires vont avoir tendance à abandonner l'étude surtout s'ils ne ressentent pas d'amélioration de leur état. Ainsi les patients, qui auraient été des échecs du traitement s'ils étaient allés jusqu'au bout de l'essai, le quitteront plus fréquemment dans le groupe traité que dans le groupe contrôle (étant donné que les EI sont plus fréquents dans le groupe traité que dans le groupe placebo). Si le traitement n'est pas efficace, le même nombre d'échecs du traitement devrait être observé dans les 2 groupes. Mais en raison des perdus de vue, il y aura moins de patients en échec au terme de l'essai dans le groupe traité que dans le groupe contrôle conduisant à un résultat biaisé faisant conclure à tort à une efficacité du nouveau traitement.*

Avec les analyses de survie (*time to event analysis*) les perdus de vues sont souvent considérés comme des censures. Cette approche ne protège pas contre les biais, car il n'y a pas de remplacement conservateur.

> "Efficacy outcomes were examined in the intention-to-treat population with the use of time-to-event analyses; data on patients who withdrew from the trial or were lost to follow-up were censored at the last available follow-up time."

Les données manquantes sur les critères de jugements doivent être remplacées par une valeur arbitraire choisie de façon que cette imputation ne puisse pas favoriser le groupe traité. La méthode de remplacement doit être conservatrice, c'est-à-dire handicaper l'apparition de la supériorité du traitement. Si après ce remplacement conservateur, la supériorité du traitement est toujours présente, le résultat est robuste, car il vient d'être montré qu'il n'était pas conditionné par les données manquantes sur le critère de jugement.

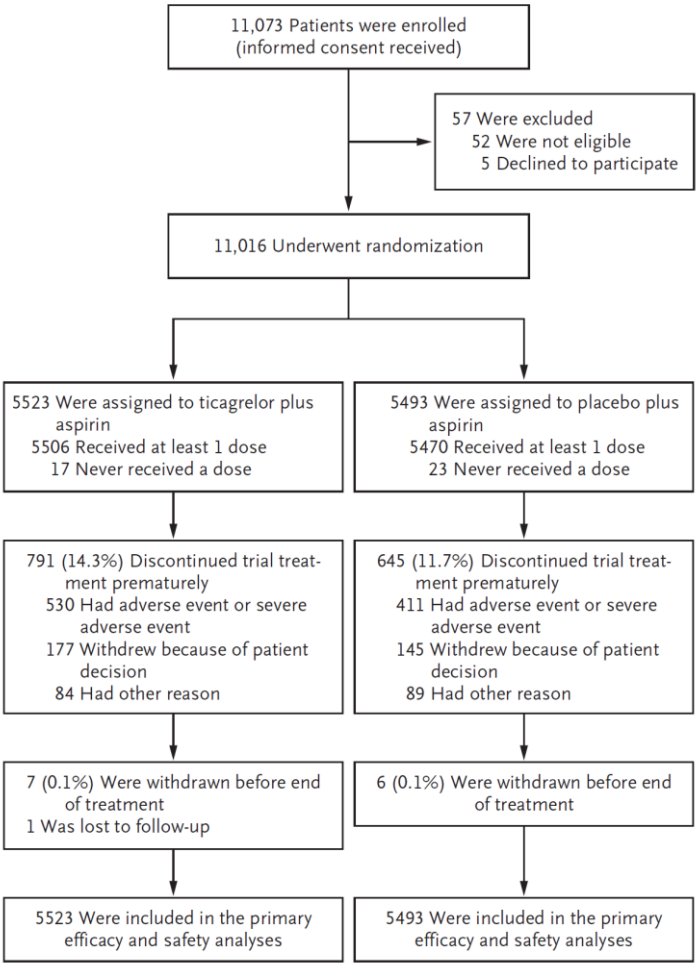

Figure 1 – Flow chart
Le flow chart permet de vérifier que l'analyse est faite en intention de traiter en comparant les effectifs randomisés (5523 et 5493) aux effectifs analysés (inclus dans l'analyse primaire). Un seul perdu de vue a été observé dans cette étude dans le groupe traité. Les patients non traités restent bien dans l'analyse. [10.1056/NEJMoa1916870]

6.1 Les méthodes de remplacement des données manquantes

Avec les critères de jugement continu (avec lesquelles l'effet du traitement est recherché en comparant par exemple les moyennes), la valeur manquante est remplacée par la dernière valeur connue pour ce patient, provenant de la dernière visite à laquelle le patient s'est rendu. Cette méthode est appelée LOCF (*last observation carry forward*). Une autre méthode consiste à utiliser la valeur à la baseline (BOCF : *baseline observation carry forward*).

"Missing observations were imputed by using the most recent previous observation (the last observation carried forward)."
"Missing values were imputed by means of the last-observation-carried-forward method."

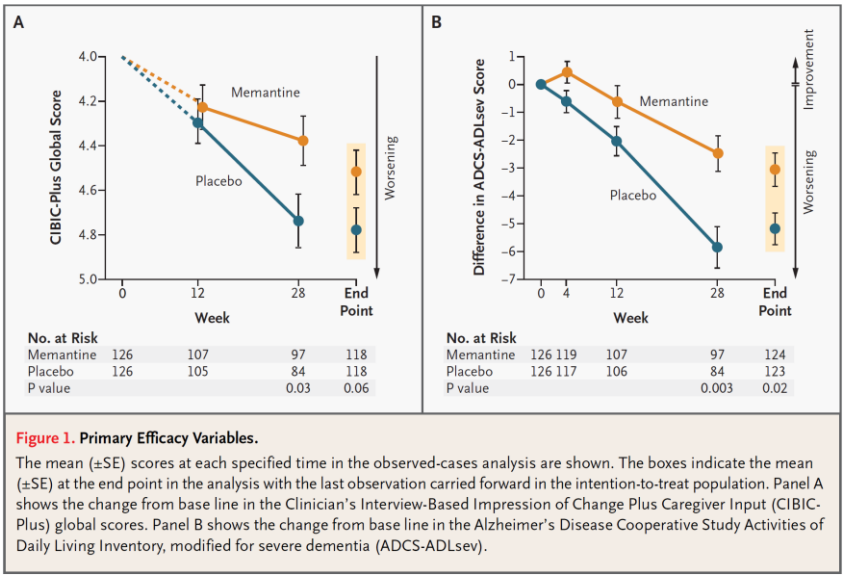

Figure 1. **Primary Efficacy Variables.**
The mean (±SE) scores at each specified time in the observed-cases analysis are shown. The boxes indicate the mean (±SE) at the end point in the analysis with the last observation carried forward in the intention-to-treat population. Panel A shows the change from base line in the Clinician's Interview-Based Impression of Change Plus Caregiver Input (CIBIC-Plus) global scores. Panel B shows the change from base line in the Alzheimer's Disease Cooperative Study Activities of Daily Living Inventory, modified for severe dementia (ADCS-ADLsev).

Figure 2 – Exemple de résultat avec un remplacement des données manquantes par la méthode LOCF.
Le critère de jugement de cette étude était mesuré à 28 semaines. Pour la sous-figure A, seulement 97 et 84 patients ont une mesure effective à cette date (*No at risk*) à comparer avec les effectifs randomisés indiqués en dessous du t0 (126 et 126). Il existe donc de nombreux patients avec la valeur du critère de jugement manquante. Le résultat présenté à droite (appelé Endpoint) est les moyennes obtenues après remplacement des données manquantes par LOCF (on remarque qu'il reste des patients non pris en considération 118 à la place des 126 initiaux, cette analyse n'est donc pas une analyse en intention de traiter. On remarque aussi le côté conservateur de la méthode LOCF, car a différence entre les 2 groupes est plus faible après remplacement qu'avant et la signification est perdu (p=0.06). La conclusion est que le résultat initial « *observed case analysis* » est non robuste vis-à-vis des perdus de vue existant dans cette étude et qu'il n'apporte pas de démonstration de l'effet du traitement sur ce critère de jugement. [N Engl J Med 2003;348:1333-41]

Pour les données binaires (avec lesquelles l'effet du traitement est recherché en comparant la fréquence des évènements), la méthode la plus conservatrice est celle du biais maximum (*worst case scenario*). Les perdus de vue du groupe traité sont considérés comme ayant fait le critère de jugement et pas ceux du groupe contrôle.

Dans un essai avec les AVC comme critère de jugement, il y a eu 25 AVC chez les 200 patients du groupe traité et 30 chez les 200 patients du groupe contrôle. Il y a aussi 5 perdus de vus dans le groupe traité et 6 dans le groupe contrôle. La question qui se pose est : est-ce que les perdus de vu du groupe ont pu produire ce résultat en faveur du traitement. La réponse est oui, car si les 5 perdus de vue sont des patients qui ont en réalité fait un AVC après avoir quitté l'étude, les résultats auraient été 25+5=30 AVC sous traitement comparé à 30 AVC dans le groupe contrôle. Le résultat brut de cet essai n'est donc pas à l'abri d'un biais lié au perdu de vue.

"a maximum-bias hypothesis was also applied, in which thyroid ablation of patients who could not be evaluated or those with persistent disease was considered incomplete in the groups receiving recombinant human thyrotropin or 1.1 GBq and as complete in groups receiving thyroid hormone withdrawal or 3.7 GBq." [N Engl J Med 2012;366:1663-73.]
"Missing assessments were imputed with the use of either the last-observation-carried-forward method or a method that imputed data according to a worst-case scenario"
"A post hoc sensitivity analysis of the worst-case scenario for mortality at 6 months did not alter the results"

Cette imputation des données manquantes peut aussi se faire avec une technique sophistiquée appelée imputation multiple. Juger de la pertinence de cette méthode est au-delà des objectifs de ce document.

Au total pour écarter le risque de biais lié au données manquantes sur le critère de jugement il est nécessaire :

- Qu'il soit bien mentionné que l'analyse se fait en intention de traiter ou porte sur la population d'analyse en intention[7] de traiter (full set analysis)
- Que dans le flow chart, l'effectif des patients analysés soit identique à celui des patients randomisés dans chaque groupe
- Qu'il n'y ait pas de perdu de vue (de patients pour lequel le critère de jugement n'est pas disponible) ou, s'il y en a, que les données manquantes sur le critère de jugement ont été remplacées par une méthode conservatrice (BOCF, biais moyen ou biais maximum)
- Si les données manquantes sur le critère de jugement n'ont pas été remplacées, leur nombre ne remet pas en cause la robustesse du résultat.

	Biais lié aux données manquantes
Analyse en intention de traiter avec remplacement des données manquantes par une méthode conservatrice	À l'abri des biais
Analyse en intention de traiter sans remplacement conservateur des données manquantes, mais nombre de perdus de vues ne remettant pas en cause la robustesse du résultat	À l'abri des biais
Analyse en intention de traiter sans remplacement conservateur des données manquantes, robustesse du résultat non assuré étant donné le nombre de perdu de vu	Risque de biais
Analyse en per protocol ; mauvaise définition de l'ITT	Risque de biais

[7] Le terme population d'analyse désigne le sous-ensemble des patients qui seront pris en considération par l'analyse. La population ITT correspond à la totalité des patients inclus (moins les retraits de consentement à être suivi)

7 Évaluation globale du risque de biais

À l'issue de l'évaluation du risque de biais, les résultats sont classés en 2 catégories :

- Résultats à l'abri des biais, pouvant potentiellement faire changer les pratiques (s'ils permettent de conclure de manière statistiquement significative et s'ils sont cliniquement pertinents)
- Résultats non à l'abri des biais, insuffisamment solide pour faire changer les pratiques

Références

1 Savovic J, Turner RM, Mawdsley D, et al. Association Between Risk-of-Bias Assessments and Results of Randomized Trials in Cochrane Reviews: The ROBES Meta-Epidemiologic Study. *Am J Epidemiol* 2018;187:1113–22 doi:10.1093/aje/kwx344; PMID:29126260;

2 Sterne JAC, Jüni P, Schulz KF, et al. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;21:1513–24 doi:10.1002/sim.1184; PMID:12111917;

3 Balk EM, Bonis PAL, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287:2973–82 doi:10.1001/jama.287.22.2973; PMID:12052127;

4 Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408–12 doi:10.1001/jama.273.5.408; PMID:7823387;

5 Savović J, Jones H, Altman D, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technol Assess* 2012;16:1–82 doi:10.3310/hta16350; PMID:22989478;

6 Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *The Lancet* 1998;352:609–13 doi:10.1016/S0140-6736(98)01085-X;

7 Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898 doi:10.1136/bmj.l4898; PMID:31462531;

8 Senn S. Testing for baseline balance in clinical trials. *Stat Med* 1994;13:1715–26 ; PMID:7997705;