La lecture critique des essais thérapeutiques pour la pratique médicale

Faculté de médecine Lyon Est

2ème cycle

Pr Michel Cucherat

Service Hospitalo-Universitaire de Pharmacologie et de Toxicologie

Version 1.0

26/4/2021

Prérequis

Un certain nombre de notions de base doivent être acquises avant d'entreprendre la lecture de ce cours :

- Principe de base de l'essai contrôlé randomisé
    - Groupes de traitement comparés (bras de l'essai)
    - Critère d'inclusion (et de non inclusion/exclusion)
    - Suivi longitudinal des patients, visites de suivi
    - Notion de critère de jugement
- Définition du double insu (double aveugle), d'un essai en ouvert, analyse en ITT, en per protocol
- Test d'hypothèse, hypothèse nulle, erreur statistique alpha et beta, risque alpha, signification statistique, p value
- Effet en statistique (effet traitement)
- Mesures d'un effet : risque ratio, odds ratio, hazard ratio, différence des risques, NNT, différence des moyennes, différences des changements avant après
- Analyse de survie, censure, date de point, courbe de Kaplan Meier, hazard ratio
- Principes éthiques de la recherche clinique
- Cadre règlementaire de base de la recherche clinique

Avant-propos

Ce cours ne couvre pas de manière exhaustive tout le champ de la lecture critique des essais cliniques. Il vise seulement la formation initiale et a pour but d'apporter au lecteur les compétences de base pour comprendre et commencer à lire sans complexe les essais thérapeutiques publiés dans la littérature médicale de son domaine.

La finalité de cet apprentissage est de pouvoir s'autodéterminer le plus possible par rapport aux nouvelles propositions thérapeutiques et d'éviter ainsi d'être un bouchon balloté au gré des flots de la communication promotionnelle ou des opinions des uns et des autres.

Ce cours peut être ponctuellement réducteur, en particulier sur les points statistiques, pour éviter de noyer le lecteur sous des aspects techniques. La maitrise complète de la lecture critique sera acquise par des compléments abordés en 3ème cycle ou par des formations complémentaires pour ceux qui souhaitent développer une expertise avancée (pour participer à des travaux d'élaboration de recommandations de pratique, pour enseigner, ou pour siéger dans des commissions consacrées à la régulation ou l'évaluation des thérapeutiques par exemple).

Objectifs pédagogiques

Ce cours vise à apporter les compétences de base de la lecture des essais thérapeutiques aux médecins cliniciens (qu'ils aient une activité de recherche ou non).

À l'issue du cours, l'étudiant pourra :

- Déterminer si les résultats produits par l'essai clinique permettent d'utiliser ou non le nouveau traitement en pratique courante (en termes de fiabilité et de pertinence clinique)

Pour cela l'étudiant saura :

- Identifier si un résultat est à l'abri des biais
- Identifier si un résultat permet de conclure à l'intérêt clinique du traitement avec un risque alpha global de conclure à tort contrôlé
- Identifier si au contraire un résultat est seulement exploratoire (et donc insuffisamment démontré pour constituer un argument suffisamment fiable pour justifier l'utilisation du nouveau traitement)
- Identifier si un résultat démontré est suffisamment cliniquement pertinent pour constituer un réel progrès thérapeutique (pertinence du comparateur, du/des critères de jugement, balance bénéfice risque favorable)

Ces aptitudes concerneront les essais cliniques contemporains du type de ceux auxquels les futurs médecins seront majoritairement confrontés : essais de supériorité et de non-infériorité, type phase 3, portant sur des critères cliniques[1].

Ces compétences lui permettront :

- De comprendre la justification des recommandations de pratique (et éventuellement de détecter celles qui ne sont pas entièrement basées sur des preuves) et des décisions réglementaires (HAS)
- D'être armée vis-à-vis de de la communication promotionnelle et de savoir déjouer ses pièges (y compris dans le discours des leaders d'opinion)
- De s'autodéterminer en connaissance de cause lors de l'actualisation de ses choix thérapeutiques (sur la base des données factuelles et non pas uniquement à partir de la communication promotionnelle)

[1] Les essais précoces dont l'objectif n'est pas de guider la pratique, mais de préparer la réalisation des essais de confirmation ne font pas partie des objectifs pour les cliniciens (concernent seulement les futurs chercheurs) ainsi que les designs autres que les essais en bras parallèles et les plans factoriels.

Table des matières

1 Introduction 1

1.1 Pourquoi avoir une lecture critique : les spins de conclusion 1

1.2 Pourquoi faire attention à la méthodologie 3

1.3 Pourquoi faut-il avoir des preuves du bénéfice clinique des traitements 3

2 Finalité de la lecture critique d'un essai thérapeutique 5

3 Grille de lecture 7

3.1 Pour la validité interne 7

3.2 Pour la pertinence clinique 8

4 Évaluation de la solidité statistique des résultats 8

4.1 Risque alpha (*type I error rate*) 8

4.2 Risque alpha global (overall type I error rate) 9

4.3 Multiplicité et inflation du risque alpha global global 11

4.4 Technique de contrôle du risque alpha global gérant la multiplicité 12

4.4.1 Répartition 12

4.4.2 Hiérarchisation (*closed testing*) 13

4.4.3 Combination des deux approches 16

4.5 Nouvelle politique de présentation des p value 17

4.6 Critères de jugement secondaires 18

4.6.1 Essai avec un critère de jugement principal unique 18

4.6.2 Essai gérant la multiplicité par un plan de contrôle du risque global 18

4.7 Les analyses en sous-groupes 19

4.8 L'analyse finale et les analyses intermédiaires 23

5 Évaluation du risque de biais 25

5.1 Biais prévenus par la randomisation imprévisible 27

5.2 Biais prévenus par le double insu vis-à-vis de la mesure du critère de jugement 28

5.3 Biais prévenus par le double insu vis-à-vis de la réalisation de l'essai 30

5.4 Biais prévenus par l'analyse en ITT 31

5.5 Évaluation globale du risque de biais 34

6 Lecture critique et fraude scientifique" 34

7 Évaluation de la pertinence clinique (clinical relevance) 35

7.1 Pertinence du comparateur 36

7.2 Pertinence clinique du critère de jugement 37

7.3 Interprétation du résultat obtenu sur un critère composite ? 38

7.4 La balance bénéfice risque 41

8 Le cas des essais « négatifs » 43

9 Conclusion pratique de la lecture critique 43

10 Références 45

1 Introduction

1.1 Pourquoi avoir une lecture critique : les spins de conclusion

Une première justification de la nécessité pour tout clinicien de savoir interpréter les résultats des essais cliniques est constituée par l'existence des spins de conclusions.

Les spins de conclusion (ou enjolivement de la conclusion) sont des situations où l'essai est non-concluant (n'apportant pas de démonstration de l'intérêt clinique du nouveau traitement), mais où la conclusion de l'article est positive, en faveur de l'intérêt du traitement.

Les spins sont fréquents et on peut estimer que pour la moitié des essais non concluants il existe un spin dans l'abstract, la discussion ou la conclusion de l'article.

Les conséquences des spins sont graves, car ils conduisent à faire croire au lecteur, qui se limiterait à la conclusion de l'abstract, que le nouveau l'intérêt a un intérêt clinique alors qu'en réalité celui-ci n'est pas démontré (voir la section 1.3 pour la nécessité d'avoir une démonstration de l'intérêt d'un traitement avant de l'utiliser).

Une des finalités de l'enseignement de la lecture critique est d'armer les futurs cliniciens à ne pas tomber dans les pièges des spins de conclusions.

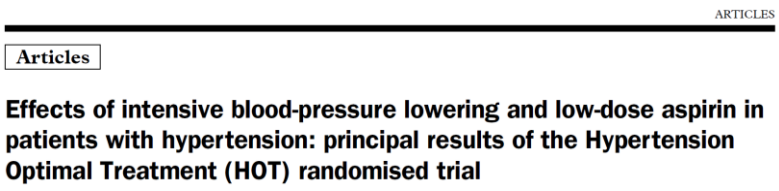
Le but de la lecture critique est de faire sa propre conclusion à partir de son évaluation de la rigueur méthodologique de l'étude et des résultats obtenus, pour ne pas tomber dans le piège d'un spin de conclusion

Les spins sont aussi très fréquents dans les supports de communications promotionnelles ou dans le discours de certains leaders d'opinion. L'éminence d'un locuteur ne doit vous faire perdre votre regard critique et votre volonté de revenir aux faits prouvés.

Exemple : HOT

L'essai HOT [10.1016/S0140-6736(98)04311-6] avait comme objectif de montrer que la prise en charge intensive de l'hypertension permettait de maximiser le bénéfice sur les évènements cardiovasculaires. Pour cela, 3 cibles de PAD étaient comparées : la cible standard de l'époque 90mmHg et deux cibles d'intensification 85 et 80 mmHg. L'essai était randomisé et a inclus 18790 patients. Le critère de jugement principal était les évènements cardiovasculaires mortels et non mortels.

Le caractère randomisé de l'essai est mentionné dans le titre de l'article :


ARTICLES

Articles

Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial

Et décrit dans la section méthode de l'abstract :

« Methods 18 790 patients, from 26 countries, aged 50–80 years (mean 61·5 years) with hypertension and diastolic blood pressure between 100 mm Hg and 115 mm Hg (mean 105 mm Hg) were randomly assigned a target diastolic blood pressure. 6264 patients were allocated to the target pressure 90 mm Hg,

6264 to 85 mm Hg, and 6262 to 80 mm Hg. Felodipine was given as baseline therapy with the addition of other agents, according to a five-step regimen. ».

Le traitement utilisé était un nouvel inhibiteur calcique, la félodipine, et cet essai appartenait au plan de développement de ce produit, même si cette étude n'évaluait pas le bénéfice de cette molécule. Il était sponsorisé par le fabricant du produit comme les autres phases 3 du plan de développement.

Comme il s'agissait d'un essai randomisé, son niveau de preuve était élevé et à même de faire changer les pratiques.

La conclusion de l'abstract est la suivante :

"Intensive lowering of blood pressure in patients with hypertension was associated with a low rate of cardiovascular events. The HOT Study shows the benefits of lowering the diastolic blood pressure down to 82·6 mm Hg."

Cette conclusion incite donc à changer les pratiques et à abandonner la cible standard de 90mmHg pour chercher à atteindre 82.6mmHg. Cependant, cette conclusion apparait d'emblée surprenante du fait de la valeur de la cible mise en avant : 82.6 mmHg. Cet essai ne peut pas conduire à ce résultat. En effet, les seules conclusions possibles sont 85 ou 80 mmHg en fonction du groupe dans lequel il y aurait le moins d'évènements cardiovasculaires. Il est donc impossible de conclure à 82.6mmHg. De plus les résultats présentés sont les suivants :

Event	Number of events	Events/ 1000 patient-years	p for trend	Comparison	Relative risk (95% CI)
Major cardiovascular events					
≤90 mm Hg	232	9·9		90 *vs* 85	0·99 (0·83–1·19)
≤85 mm Hg	234	10·0		85 *vs* 80	1·08 (0·89–1·29)
≤80 mm Hg	217	9·3	0·50	90 *vs* 80	1·07 (0·89–1·28)

Il apparait qu'aucune différence de fréquence des évènements cardiovasculaire n'a été obtenue entre les 3 cibles (9.9, 10.0 et 9.3 /1000 patients années respectivement pour les cibles de 90, 85 et 80mmHg) avec un p = 0.5. L'interprétation de ces résultats est : cette étude a échoué à montrer le bénéfice d'intensifier la baisse de PAD. Il n'y a pas lieu de changer les pratiques.

À la place de cette conclusion, le papier, comme nous l'avons vu, conclut positivement et propose une nouvelle cible. La valeur de 82.6 mmHg qui ne peut pas être validée par le plan d'expérience qui avait été utilisé provient en réalité d'une analyse de la relation entre la PAD obtenue après l'instauration du traitement, quel que soit le groupe alloué par la randomisation et la fréquence des évènements cardiovasculaires. Une régression curvilinéaire a été utilisée pour modéliser cette relation. Le minimum de la courbe de régression est 82.6mmHg. Cette analyse est donc une recherche d'association, purement observationnelle, et n'a pas du tout le niveau de preuve d'un essai randomisé. C'est d'ailleurs la nième observation d'une courbe en U (ou en J) obtenue en épidémiologie de l'HTA.

Cet article est trompeur sur plusieurs aspects. Non seulement il conclut positivement à partir de résultats négatifs, mais il met aussi en avant un résultat obtenu par une toute méthodologie que celle qui est décrite. L'article aurait donc dû décrire une étude observationnelle et non pas un essai randomisé. Le lecteur rapide peut ainsi ne pas percevoir ce problème et considérer que la conclusion est parfaitement bien établie, car elle provient d'un essai randomisé et a donc un niveau de preuve maximal.

Bien entendu les reviewers ont certainement signalé ces problèmes, mais l'éditeur en chef a quand même pris la décision de publier cet article sous cette forme sans demander aux auteurs de reformater le manuscrit. L'enjeu pour le Lancet de publier cet essai était certainement important en termes de citations et de ventes de tirés à part compte tenu du buzz qu'aller faire ce résultat [1].

1.2 Pourquoi faire attention à la méthodologie

Un essai peut donner un résultat en faveur de l'intérêt du nouveau traitement alors que ce dernier n'apporte aucun bénéfice en réalité et cela en raison d'un biais ou de l'erreur statistique alpha (ou d'une découverte fortuite). On parle alors de résultat faussement positif (faux positif) dans le sens où l'essai est « positif » (il a apparemment atteint son objectif, montrer que le traitement à un intérêt) mais à tort, le résultat est donc faux par rapport à la réalité.

Un résultat faussement positif a de graves conséquences, car il conduit à recommander/utiliser un traitement qui n'apportera pas le bénéfice escompté aux patients.

Comme il est impossible de savoir si un résultat est faussement positif (car on ne connait pas la réalité de l'effet du traitement), le seul moyen d'éviter l'utilisation indue d'un traitement à la suite d'un résultat d'essai faussement positif est d'empêcher leurs survenus. C'est le but de la méthodologie et de l'analyse statistique :

- La méthodologie a pour but d'éviter les biais.
- L'analyse statistique réduit (contrôle) à un niveau faible (2.5%) le risque de conclure à tort à l'intérêt du traitement du fait d'une erreur statistique alpha (due uniquement au hasard, consécutive aux fluctuations d'échantillonnages).

Le vintafolide a été développé dans le cancer de l'ovaire. Une première étude [10.1200/JCO.2013.49.7685] randomisée, multicentrique, en ouvert, comparant vintafolide plus doxorubicine à la doxorubicine seule s'avère « positive » en montrant une amélioration de la PFS (survie dans progression du cancer). Devant l'absence de traitement efficace dans cette situation, il aurait été tentant d'utiliser ce nouveau traitement sur la base de ce résultat d'essai randomisé. Cependant, il fallait attendre les résultats de la phase 3 PROCEED dont le design était très comparable à l'étude précédente : randomisé, multicentrique, en double aveugle, comparant vintafolide plus doxorubicine à la doxorubicine seule + placebo. Cette phase 3 a été arrêtée prématurément lors d'une analyse intermédiaire pour futilité. Le produit a ensuite été abandonné, car s'avérant être sans utilité dans cette situation. Rétrospectivement il s'avère donc que le résultat de la première étude était faussement positif, faux positif qu'il est possible d'imputer à l'absence d'aveugle (étude en ouvert) alors que la phase 3 était protégée contre les biais grâce à l'utilisation du double aveugle (placebo). Cet exemple illustre bien l'importance de l'utilisation d'une méthodologie rigoureuse pour éviter la survenue des résultats faussement positifs.

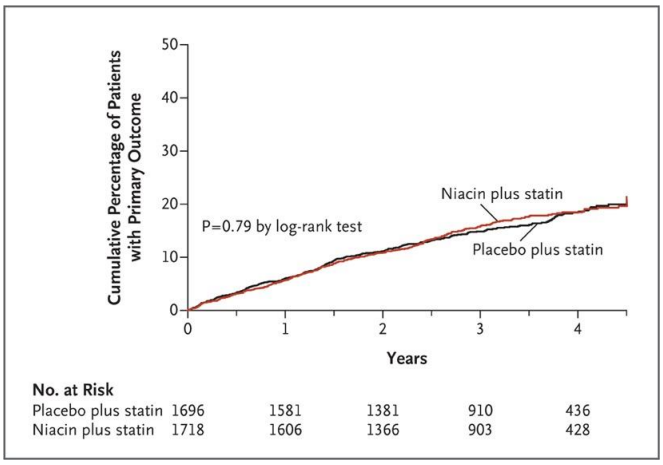
1.3 Pourquoi faut-il avoir des preuves du bénéfice clinique des traitements

La vérification par des faits prouvés (*evidence*) que les nouveaux traitements apportent bien le bénéfice clinique escompté s'est avérée indispensable au cours du temps pour plusieurs raisons.

Les mécanismes ne prédisent pas avec certitudes le bénéfice. Un effet pharmacologique ne produit pas forcément le bénéfice clinique attendu, car nos connaissances sur les mécanismes d'action et la physiopathologie des maladies sont encore parcellaires.

Avec les statines, la baisse de LDL cholestérol s'est traduire en une réduction de la fréquence des évènements cardiovasculaires parfaitement bien démontrés dans de nombreux essais.

L'acide nicotinique (niacin) est un produit qui entraine une baisse substantielle du LDL cholestérol. Cependant dans HPS2-THRIVE, un essai de morbi-mortalité de grande taille avec 25,673 patients et un suivi médian de 3.9 ans, n'a pas été en mesure de confirmer un bénéfice clinique en termes d'évènements cardiovasculaires [10.1056/NEJMoa1300955].



Ainsi même quand un mécanisme d'action (baisse provoquée des LDL) a déjà montré qu'il entrainait un bénéfice clinique avec une certaine classe de médicaments, il n'est pas garanti qu'une autre classe ayant le même effet produise le même bénéfice.

Même après des études de phase 2 concluantes, les échecs des phase 3 (dont le but est de démontrer le bénéfice) sont fréquents, aux alentours de 50 [2]. Ce résultat montre qu'il n'est pas possible de se passer de la vérification des théories et des hypothèses thérapeutiques par les essais cliniques conçus pour mettre en évidence le bénéfice clinique.

L'utilisation sans preuve de traitements expose à un risque de perte de chance pour les patients si le traitement est en réalité sans intérêt clinique. De plus il n'est pas possible d'exclure un effet délétère.

Au début des années 1980, il avait été pris comme habitude de donner des antiarythmiques de classe 1c aux patients qui présentent de nombreuses extrasystoles ventriculaires (ESV) après un infarctus du myocarde, dans le but de prévenir la mort subite. Le raisonnement partait de la constatation que la fréquence des ESV était corrélée avec le risque de mort subite et que les antiarythmiques de classe 1c réduisent drastiquement les ESV. Cette pratique a été établie sans preuve de la prévention de la mort subite. Ce n'est qu'au bout de 10 ans que l'essai de mortalité a été entrepris (alors qu'il aurait dû l'être avant l'établissement de cette pratique thérapeutique). A la place de confirmer la prévention des morts subites, cet essai [10.1056/NEJM199103213241201] à montrer une multiplication par 3 de la mortalité.

Plus récemment, l'hydroxychloroquine a été proposée comme traitement dans la COVID-19 sans preuve clinique, sur la base d'étude in vitro d'activité de ce produit sur le virus et d'études cliniques non randomisées, présentant de nombreuses limites méthodologiques. Il s'agissait d'une utilisation compassionnelle motivée par la gravité de la maladie et absence de ressources thérapeutiques curatrices. Malgré tout, en s'affranchissant de ce contexte émotif, des essais randomisés ont été entrepris, comme RECOVERY [10.1056/NEJMoa2022926]. Ces essais n'ont pas permis de mettre en évidence de bénéfice et leur méta-analyse montre une augmentation de la mortalité [10.1038/s41467-021-22446-z]. Cet exemple montre que, même dans un contexte d'urgence grave, il n'est pas possible d'envisager un usage compassionnel sans preuve, compte tenu des limites des données préliminaires.

Une des raisons pouvant expliquer pourquoi des effets pharmacologiques prometteurs ne débouchent pas toujours sur un bénéfice clinique réside dans les limites des études in-vitro et des études cliniques préliminaires. Il est connu que les modèles animaux ou cellulaires peuvent ne pas être prédictifs de ce qui se passe chez l'homme. Les études précliniques et cliniques précoces souffrent souvent de lacunes méthodologiques majeures. De plus un biais de publication ou de « selective reporting » des résultats peut distordre complètement la perception de la réalité en ayant conduit à la publication exclusive des résultats faussement positifs. Les raisonnements qui extrapolent le bénéfice à partir de l'effet sont

aussi parfois trop simplistes ou optimistes et néglige les problématiques de pharmacocinétique (est-il possible d'atteindre la concentration efficace à l'effecteur ?) ou de toxicité.

Pour l'hydroxychloroquine, l'idée de son potentiel intérêt provenait d'une étude in vitro montrant une activité de cette molécule sur le virus dans un modèle cellulaire Vero [10.1038/s41422-020-0282-0]. Ce résultat a ensuite été contredit par une autre étude, réalisée sur des lignées cellulaires de primates plus prédictives de l'action chez l'homme [10.1038/s41586-020-2575-3]. De plus il avait été négligé que, compte tenu de la pharmacocinétique très particulière de la molécule, il était impossible d'obtenir la concentration nécessaire suivant la 1er étude avec les doses maximales tolérées.

Ils s'avèrent aussi que l'essai clinique randomisé ne peut pas être remplacé par d'autres types d'études comme les études observationnelles utilisant les données de vraie vie en raison des limites méthodologiques de ces approches difficilement surmontables [3]. Régulièrement des essais randomisés échouent à confirmer des bénéfices suggérés par des études observationnelles montrant ainsi le manque de fiabilité de ces études.

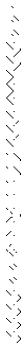
Des études observationnelles suggéraient qu'une supplémentation en vitamine D et calcium pourrait prévenir les adénomes coliques. L'essai randomisé entrepris pour vérifier cette hypothèse à échouer à mettre en évidence un bénéfice [10.1056/NEJMoa1500409]

2 Finalité de la lecture critique d'un essai thérapeutique

La finalité, pour la pratique médicale, de la lecture critique d'un article d'essai thérapeutique est de déterminer si cet essai apporte la **preuve** (*evidence*) de **l'intérêt** (*effectiveness, utility*) du nouveau traitement.

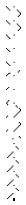
- Si c'est le cas, le nouveau traitement pourra être introduit dans la stratégie thérapeutique de la pathologie.
- Si ce n'est pas le cas, le traitement ne peut pas être utilisé et éventuellement d'autres essais doivent être entrepris pour apporter la preuve de son intérêt

L'intérêt clinique du traitement est démontré par un résultat fiable, statistiquement significatif en termes de contrôle du risque alpha global[2] et cliniquement pertinent (avec entre autres une balance bénéfice risque favorable). La simple mise en évidence d'un « effet » du traitement n'est pas suffisante si cet effet est mesuré sur un critère de jugement non cliniquement pertinent. L'intérêt clinique du traitement n'est pas non plus démontré, même en cas de mise en évidence d'un bénéfice cliniquement pertinent, si la balance bénéfice risque est défavorable (effet indésirable contrebalançant en totalité le bénéfice obtenu).



Le bamlanivimab est un anticorps monoclonal neutralisant du SARS-Cov-2. Il a été développé comme traitement ambulatoire précoce des cas de Covid ne nécessitant pas d'oxygénothérapie. Une publication relate les résultats préliminaires obtenus dans un essai de phase 2 BLASEZ 1 partie 1 [10.1056/NEJMoa2029849]. Bien qu'utilisé par certains pour inciter à l'utilisation rapide de ce produit, ces résultats présentent de nombreuses limites méthodologiques : critère de jugement principal réduction de la charge virale non statistiquement significatif ; réduction de la fréquence des hospitalisations non statistiquement significative ; le seul résultat statistiquement significatif provenait d'une analyse en sous-groupe post hoc. Ils ne démontrent donc pas l'intérêt clinique du bamlanivimab dans cette indication.

[2] Et pas seulement nominalement significatif



De ce fait la société de pathologie infectieuse de langue française a conclu dans ses recommandations : « Le groupe recommandations de la SPILF considère que l'utilisation du Bamlanivimab ne doit pas être recommandée en monothérapie, en raison de l'absence d'intérêt clinique démontré dans les essais. Seule une utilisation dans des essais cliniques est pour l'instant concevable. » (https://www.infectiologie.com/fr/actualites/place-du-bamlanivimab_-n.html)

Le but de la lecture critique est de juger si un résultat est suffisamment fiable et cliniquement pertinent pour constituer une preuve de l'intérêt du traitement suffisante pour le faire adopter en pratique médicale

---

La lecture critique ne consiste pas à lire de manière linéaire l'article. Elle a pour but de répondre à la question : cet essai (rapporté dans cet article) apporte-t-il la preuve de l'intérêt du nouveau traitement ? Pour cela, l'article et son supplément électronique ainsi que le protocole, vont être utilisés comme source documentaire pour vérifier, les uns après les autres, les différents points qui permettront de répondre à cette question (cf. section **Erreur ! Source du renvoi introuvable.**). La lecture critique (des RCT) est une démarche structurée et standardisée et non pas une lecture passive guidée uniquement par le texte lui-même.

En évitant de lire discussion et conclusion des auteurs, on évite de se faire influencer et d'éventuellement accepter un résultat discutable. Le but de la lecture critique est de se faire sa propre discussion (compte tenu des patients que l'on est amené à traiter) et sa propre conclusion.

> *En effet, ces articles sont écrits par des professionnels de la rédaction médicale et un des objectifs de l'optimisation de la rédaction et de suggérer que les éventuels problèmes d'un essai n'en pas. La discussion doit comporter une section sur les limites de l'étude, mais il a été montré que cette partie minorée souvent les limites éventuelles du papier.*

En premier il convient d'identifier la question clinique à laquelle l'essai cherche à apporter une réponse (par un PICO par exemple).

Ensuite, il convient d'identifier le ou les résultats qui pourraient éventuellement servir à justifier l'intérêt du traitement. La lecture critique va alors vérifier s'ils sont parfaitement bien démontrés et s'ils apportent bien la preuve de l'intérêt clinique du traitement.

Pour cela il faut se poser, pour chacun d'entre eux, la question de la fiabilité du résultat (validité interne), c'est-à-dire celle de la protection de l'essai contre les biais, celle de la signification statistique des résultats (sont-ils « suffisamment » statistiquement significatifs pour permettre de conclure à l'intérêt du traitement avec un risque alpha global bien contrôlé ?).

Puis il convient d'évaluer la pertinence clinique : représentent-ils une plus-value médicale certaine ?

Pour chacune de ces questions, la réponse est recherchée dans la publication (et ses compléments). Les articles sont rédigés en suivant des guides de rédaction (CONSORT) qui garantit que la réponse aux questions de lecture critique figure bien dans la publication (à la place attendue).

3 Grille de lecture

3.1 Pour la validité interne

S'assurer que le résultat correspond bien à un objectif de l'essai prédéfini ?

> *Si ce n'est pas le cas, le résultat est dit post hoc et il peut être une découverte fortuite. Ne correspondant pas à une hypothèse faite préalablement, ce résultat est insuffisamment solide pour entrainer l'utilisation du traitement, il est purement exploratoire. Tout au plus il permet de générer une nouvelle hypothèse à tester dans un nouvel essai.*

Vérifier que le résultat était dans le plan de contrôle du risque alpha global et qu'il est statistiquement significatif en termes de risque alpha global ?

- Déterminer le type de contrôle du risque alpha global utilisé : hiérarchisation, répartition du risque alpha (co-primary endpoints) ou critère de jugement principal unique (rare actuellement) ? Vérifier si des analyses intermédiaires ont été réalisées ?
- Déduire la condition nécessaire (seuil de signification, algorithme) pour que le résultat puisse permettre de conclure à l'intérêt du traitement avec un risque alpha global contrôlé (= signification statistique en termes de risque alpha global de l'essai) :
    - Répartition du risque alpha global entre plusieurs critères : déterminer le risque alpha attribué au critère analysé (seuil ajusté de la signification) (cf. section 4.4.1).
    - Hiérarchisation : déterminer la place dans la hiérarchie du critère analysé, déterminer le seuil de la signification à utiliser (0.05 bilatéral ou moins si des analyses intermédiaires étaient prévues) et déterminer si tous les tests au-dessus de ce critère dans la hiérarchie étaient bien significatifs (cf. section 4.4.2).
    - Réalisation d'analyses intermédiaires : déterminer le seuil de la signification à franchir à l'analyse. Ce surajoute à la répartition et à la hiérarchisation (cf. section 4.8).
    - Critère de jugement principal unique : déterminer qu'il est bien unique et prédéfini a priori, déterminer le seuil de la signification à utiliser (0.05 bilatéral ou moins si des analyses intermédiaires étaient prévues) (cf. section 4.6.1).
    - Si le résultat est un résultat de sous-groupe, vérifiez que le sous-groupe était prévu dans le plan de contrôle du risque alpha global. Si ce n'est pas le cas, le résultat est purement exploratoire (cf. section 4.7).

Vérifier si le résultat est correctement mis à l'abri des biais par la méthodologie employée ? (Cf. section 5)

- S'assurer qu'une randomisation imprévisible a été employée et qu'elle n'a pas été pervertie ;
- S'assurer de la réalisation en double insu pour écarter la possibilité de biais durant le suivi et la prise en charge des patients ;
- S'assurer de la mesure en double insu du critère de jugement pour écarter la possibilité de biais au niveau de la mesure du critère ;
- S'assurer que l'analyse a été effectuée en intention de traiter et que les éventuelles valeurs manquantes pour le critère de jugement ont été remplacées par une méthode conservative (essai de supériorité).

3.2 Pour la pertinence clinique

Vérifier si le critère de jugement est clinique (et non pas intermédiaire) (cf. section 7.2)

Vérifier si le comparateur est approprié et loyal (cf. section 7.1).

S'assurer que la taille de l'effet est suffisante pour être cliniquement pertinente.

Si le critère est composite, s'assurer de l'homogénéité des effets sur les composantes (cf. section 7.3).

Vérifier si la balance bénéfice risque est favorable (cf. section 7.4) :

- Qualitativement : il n'existe pas d'événement indésirable de gravité disproportionnée par rapport au bénéfice apporté.
- Quantitativement : les événements indésirables de même gravité que les événements évités ne contrebalancent pas numériquement le bénéfice.

4 Évaluation de la solidité statistique des résultats

4.1 Risque alpha (*type I error rate*)

La détermination qu'un traitement à un effet sur un critère de jugement s'effectue en comparant la valeur du critère de jugement entre les 2 bras de l'essai pour chercher s'il y a une différence en faveur d'un effet bénéfique du traitement (moins de décès dans le groupe traité que dans le groupe contrôlé par exemple).

Or une telle différence peut survenir du fait des fluctuations aléatoires d'échantillonnage (liées purement au hasard) même si le traitement n'a aucun effet en réalité sur le critère de jugement.

Ainsi, si l'on ne prenait pas ce risque en compte, on pourrait conclure à tort à l'existence d'une différence dans les cas où le traitement n'a pas d'effet. C'est l'erreur statistique alpha (aussi appelé de première espèce ou de type 1, *type I error*) : conclure à tort à une différence qui n'existe pas en réalité.

Cette erreur statistique a de lourdes conséquences dans le cadre d'un essai clinique, car elle conduit à conclure à tort que le nouveau traitement apporte un bénéfice aux patients et donc conduit à le recommander et à l'utiliser indument en pratique.

Pour limiter au maximum cette possibilité catastrophique on utilise un test statistique qui va limiter le risque de commettre une erreur alpha à une valeur faible (on parle de contrôle du risque alpha), 5% en bilatéral en général.

Ainsi on ne conclura que si la différence est statistiquement significative. Dans le cas où le traitement n'a pas d'effet sur le critère considéré, on ne conclura à une différence que dans 5% des cas.

N.B. : En absence de prise en considération du risque d'erreur alpha, on serait amené à utiliser tous les traitements qui n'apportent pas de bénéfice. En travaillant avec un seuil de la signification statistique à 5%, on n'accepte plus que 5% des traitements sans effet (en réalité 2.5%, car seule la moitié des différences dues au hasard sont en faveur du nouveau traitement et conduisent à l'utiliser).

L'utilisation de la signification statistique permet de réduire le risque de conclure à tort du fait du hasard, mais ne le réduit pas à zéro. Avec un seuil de signification à 5% bilatéral, on admet encore 2.5% des traitements sans effet. Un résultat significatif ne signifie pas qu'il est démontré avec certitude.

4.2 Risque alpha global (overall type I error rate)

Il existe en fait deux niveaux de risque alpha.

Risque alpha nominal	Risque que l'on prend de conclure à tort à l'existence d'un effet du traitement au niveau d'un test particulier, dans le cas où le traitement n'a pas d'effet au niveau de ce test particulier.
Risque alpha global de l'essai	Risque que l'on prend de conclure à tort à un quelconque intérêt du traitement à l'issue de l'essai. C'est l'unique **risque alpha d'intérêt** dans l'essai thérapeutique, qu'il convient de garder strictement inférieur à 5% en bilatéral.

Le risque alpha au niveau d'un test (sur un critère de jugement par exemple) est le risque de conclure à tort à une différence au niveau de ce test particulier. Ce niveau correspond à la présentation classique du risque alpha dans les cours de statistique.

L'autre niveau est celui du risque alpha global de l'essai. Celui-ci est au centre de la problématique statistique de l'essai. Un essai est entrepris pour faire la conclusion de l'intérêt du traitement. Or cette conclusion va reposer sur un test statistique. Elle pourra donc être prise à tort du fait du hasard. C'est le risque alpha global de l'essai qui doit être parfaitement bien contrôlé à moins de 5% en bilatéral (2.5% en unilatéral).

> "To control the overall type I error, …"
>
> "to preserve the overall type I error rate at 0.05 (two-sided) after accounting for one interim analysis."

Si la conclusion de l'essai ne peut être faite qu'à partir d'un seul et unique test statistique, le risque de conclure à tort au niveau de l'essai est celui de conclure à tort au niveau du test.

Dans les essais modernes, on souhaite aller au-delà de la contrainte de la conclusion unique et pouvoir conclure à l'intérêt du traitement à partir de plusieurs tests (par exemple à partir de plusieurs critères de jugement, ou en effectuant plusieurs analyses et éventuellement en considérant des sous -groupes de patients). Il est donc nécessaire d'autoriser une multiplicité des comparaisons sans que cela entraine une inflation du risque alpha global.

> Statistical testing in the COMPASS study [10.1056/NEJMoa1709118] involves multiple testing in 3 main areas:
>
> 1. Multiple intervention comparisons: Rivaroxaban 2.5 mg bid + aspirin 100 mg od (rivaroxaban plus aspirin) compared to active control aspirin 100 mg od (aspirin); Rivaroxaban 5.0 mg bid (rivaroxaban) compared to active control aspirin 100 mg od (aspirin)
> 2. Multiple outcomes: One primary efficacy outcome and 3 key secondary efficacy outcomes.
> 3. Multiple decision points: A first interim analysis was to be conducted after approximately 50% of the target number of subjects had experienced an unrefuted primary efficacy outcome, a second interim analysis was to be conducted after approximately 75% of the target number of outcomes, and a final analysis was to be conducted after the target number of 2,200 unrefuted primary efficacy outcomes.

Testing multiple hypotheses may increase the Type I error rate and we used a variety of statistical procedures to control the overall Type I error.

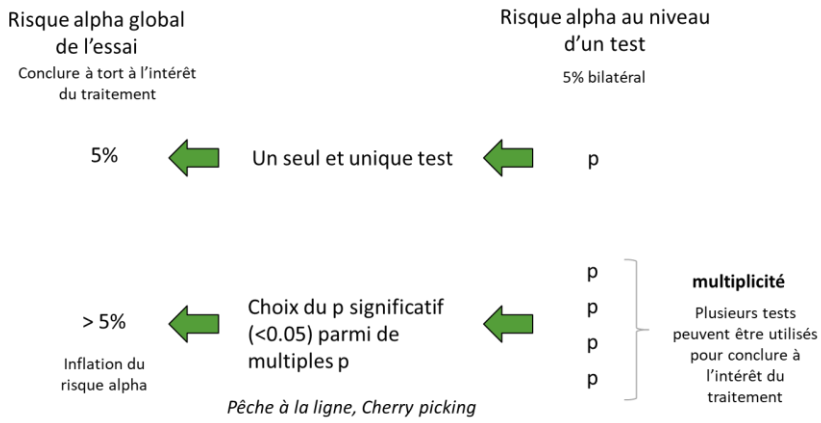
Cependant cette multiplicité des tests statistiques, chacun pouvant conduire à la conclusion de l'intérêt du traitement, augmente le risque alpha global (les tests unitaires ont toujours le même risque alpha nominal de 0.05). On parle d'inflation du risque alpha global (voir section suivante).

> *En multipliant les tests, on augmente le risque de trouver au moins un test avec un p<0.05 même si le traitement n'a aucun effet au niveau d'aucun des tests. 5% c'est 1/20. Sous l'hypothèse nulle, on s'attend à avoir en moyenne un p<0.05 (faux positif dû au hasard) tous les 20 tests réalisés.*

Dans un essai thérapeutique, un résultat statiquement significatif signifie qu'il permet de conclure à l'intérêt du traitement avec un risque alpha global parfaitement bien contrôlé[3]

Figure 1 – Les deux niveaux de risque alpha et les conséquences de la multiplication des tests pour conclure à l'intérêt du traitement

Le risque alpha global au niveau de l'essai (colonne de gauche) peut conduire à conclure à tort à l'intérêt du traitement et à recommander l'utilisation d'un traitement en réalité sans intérêt. C'est le risque alpha qui doit être parfaitement contrôlé dans l'essai. Quand on dit qu'un résultat est statiquement significatif dans un essai randomisé, cela signifie qu'il permet de conclure à l'intérêt du traitement avec ce risque alpha global parfaitement bien contrôlé. Cependant ce risque alpha global peut augmenter abusivement si cette conclusion peut effectuer à partir de multiple test (partie inférieure du schéma). Il faut donc mettre en œuvre des techniques particulières pour gérer la multiplicité et empêcher cette inflation du risque alpha global (cf. section 4.4).

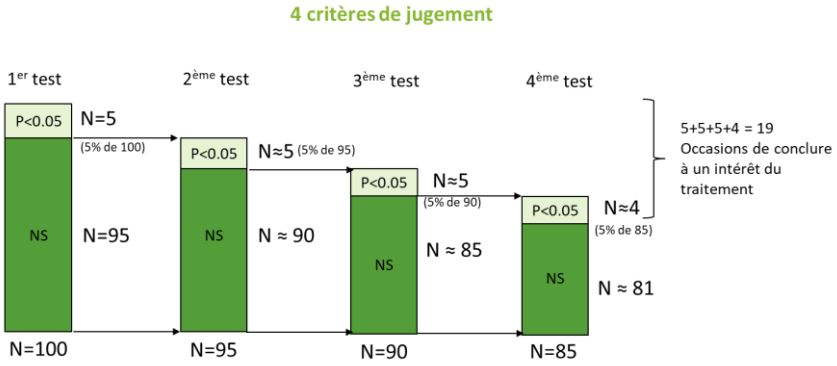


Signification statistique nominale	Contrôle le risque alpha de conclure à tort sur le test considéré	H0 : absence d'effet sur ce test particulier
Signification statistique en termes de risque alpha global de l'essai	Contrôle le risque alpha de conclure à tort à l'intérêt du traitement à l'issue de l'essai	H0 : absence d'intérêt du traitement, c'est-à-dire absence d'effet sur tous les tests qui sont réalisés

[3] Ainsi des p<0.05 pourront ne pas être statistiquement significatifs, car il ne contrôle pas le risque alpha global (on peut dire à la rigueur qu'ils sont nominalement significatifs).

4.3 Multiplicité et inflation du risque alpha global global

L'inflation du risque alpha gobal peut être illustrée de manière assez simple sans recours à aucune formule mathématique. Considérons un traitement sans aucun effet et imaginons que 100 essais randomisés versus placebo ont été réalisés. Sur ces 100 essais, on accepte donc de conclure à tort à l'intérêt du traitement dans cinq d'entre eux, mais pas plus ! (Risque alpha à 5%, on ne rentre pas dans la problématique du test bilatéral pour simplifier[4]). Imaginons aussi que 4 critères de jugement complètement différents sont analysés, et qu'il est possible de trouver un intérêt au traitement à partir du moment où un quelconque de ces 4 critères montre un effet du traitement statistiquement significatif à son niveau.



Parmi ces 100 essais réalisés, 5 d'entre eux auront un p<0.05 sur le 1er critère de jugement et permettront de conclure à l'intérêt du traitement. L'examen du 2ème critère aura lieu pour 95 essais (100-5). Parmi ces 95 essais, 5 auront un p<0.05 (5% de 95 ≈5) et permettront de conclure à l'intérêt du traitement. Cela laisse ≈90 essais qui sont négatifs sur le 1er et le 2ème critère. Parmi ces 90, ≈5 auront un p<0.05 sur le 3ème critère et permettront de conclure à l'intérêt du traitement et finalement il restera ≈85 essais pour lesquels le 4ème critère sera examiné et qui donneront ≈4 nouvelles occasions (5% de 85) de conclure à l'intérêt du traitement. Au total, globalement, il y aura eu 5+5+5+4=19 occasions de conclure à tort à l'intérêt d'un traitement qui en est dépourvu en réalité. Le risque alpha global est donc de 19/100 = 19% ce qui montre l'importance du processus d'inflation du risque alpha global lorsqu'une multiplicité (*multiplicity*) des comparaisons statistiques est présente.

Maintenant, comment pourrait-on toujours continuer à envisager 4 critères pour déterminer l'intérêt du traitement sans que cela n'induise d'inflation du risque alpha global. Une solution simple est de ne retenir que les tests où le p est inférieur à 5%/4 = 1.25%. Cela conduira à donner lors du 1er test que 1.25 occasion de conclure à l'intérêt du traitement, puis 1.25% * (100-1.25) ≈1.25 nouvelles occasions lors de l'examen du 2ème test, puis encore 1.25 et 1.25 pour le 3ème et 4ème test. Au total il y aura 1.25+1.25+1.25+1.25 = 5 occasions de conclure à l'intérêt du traitement à tort, soit un risque alpha global de 5%.

[4] Dans les démonstrations par l'exemple données dans ce document, des approximations simplificatrices sont susceptibles d'être faites dans un but pédagogique. Toute la complexité mathématique du problème sous-jacent n'est pas abordée pour éviter de noyer le lecteur dans des détails inutiles pour la compréhension générale et l'appropriation des concepts (comme, entre autres, l'indépendance en probabilité des tests multiples, le risque alpha de l'hypothèse de supériorité qui n'est que de 2.5% et non pas 5%, etc.). Le lecteur ayant une expertise en statistique comprendra le pourquoi de ces simplifications compte tenu du public visé.

Diviser le risque alpha global par le nombre de tests induit par la multiplicité (méthode de Bonferroni) permet ainsi d'éviter l'inflation tout en autorisant la multiplicité. Cependant pour cela la règle de décision change ($p<0.0125$) et la signification statistique n'est plus $p<0.05$.

Une autre façon d'éviter l'inflation est de limiter à 1 le nombre de test permettant de faire la conclusion recherchée (le 1er dans notre exemple ci-dessus). C'est le principe du critère de jugement principal unique, mais dont l'usage a été progressivement abandonné depuis 2010.

4.4 Technique de contrôle du risque alpha global gérant la multiplicité

4.4.1 Répartition

Dans la méthode par répartition, le risque alpha global est réparti entre les différents critères de jugement. Il peut s'agir d'une équirépartition ou non, cela n'a pas de conséquence.

Cette approche est parfois appelée co primary enpoints pour bien insister sur le fait que les 2 critères auront le statut de critère de jugement principal, c'est-à-dire permettant de faire des démonstrations. La répartition peut se faire entre 2 ou plusieurs critères de jugement

Pour pouvoir conclure sur un critère de manière statistiquement significative, il faudra que la p-value soit inférieure au risque alpha attribué à ce critère.

Ainsi il apparait que la signification statistique en termes de risque alpha global n'est plus du tout synonyme de p inférieur à 0.05.

L'avantage de la méthode par répartition et de pouvoir conclure sur l'un **ou** l'autre des critères de jugement en fonction de la valeur de p obtenue

Dans cet essai d'oncologie, 3 critères de jugements (OS, PFS et ORR[5]) étaient envisagés pour chercher un quelconque intérêt au traitement évalué (association nivolumab et ipilumab). Le risque alpha global de trouver un intérêt à tort au nivolumab plus ipilumab a été réparti entre ces 3 critères :

« *The coprimary end points were overall survival (alpha level, 0.04), objective response rate (alpha level, 0.001), and progression-free survival (alpha level, 0.009) among patients with intermediate or poor prognostic risk*" [10.1056/NEJMoa1712126]

Pour être significatifs (c'est-à-dire suffisamment fort pour permettre de conclure à l'intérêt du nivolumab plus ipilumab), les p obtenus pour chaque critère (p nominal) doivent être inférieur au risque alpha attribué au critère.

Les résultats obtenus sont les suivants : « *overall survival rate was 75% with nivolumab plus ipilimumab and 60% with sunitinib (hazard ratio for death, 0.63; P<0.001). The objective response rate was 42% versus 27% (P<0.001. The median progression-free survival was 11.6 months and 8.4 months, respectively (hazard ratio for disease progression or death, 0.82; P=0.03, not significant per the prespecified 0.009 threshold).*"

Pour interpréter ce type d'analyse statistique, le plus simple est de faire un tableau du type :

Critère	Risque alpha attribué (seuil de signification)	P nominal	Verdict
OS	0.04	P<0.001	Significatif, permet de conclure à l'intérêt du nivolumab plus ipilumab en raison d'un bénéfice statistiquement démontré sur la survie
PFS	0.009	P=0.03	Non significatif

[5] OS : survie globale, PFS : survie sans progression, ORR : réponse objective (réduction de taille de la masse tumorale

ORR	0.001	P<0.001	Significatif permet de conclure à un effet démontré, mais ce critère n'a que très peu de pertinence clinique. Si la survie n'avait pas été significative, ce résultat démontré n'aurait pas été suffisant pour justifier l'utilisation du traitement.

Dans les essais modernes, la signification statistique n'est plus du tout synonyme de p<0.05

4.4.2 Hiérarchisation (*closed testing*)

L'approche par hiérarchisation consiste à hiérarchiser dans le protocole les critères de jugement (le 1er, le 2ème, le 3ème, etc.).

> We used a closed testing procedure, with prespecified hierarchical testing of the primary and secondary outcomes.

> a hierarchical sequential testing approach of outcomes was used to control for the type 1 error rate, with testing of outcomes as described in the order listed in the Outcomes section, beginning with the CDR-SB score [10.1056/NEJMoa1812840]

> Analyses followed a predefined hierarchical hypothesis-testing strategy to adjust for multiplicity to maintain a familywise type I error of 5%. According to this strategy, the statistical significance of each secondary end point could be investigated only if the previous end point was significant (P<0.05 for pooled analyses). The statistical-hierarchy testing order was as follows: ACR20 response, PASI 75, PASI 90, DAS28-CRP, physical component summary of SF-36, HAQ-DI, ACR50, mTSS, dactylitis and enthesitis, and mTSS. Adapté à partir de [10.1056/NEJMoa1412679]

Une fois les résultats obtenus, ils sont analysés dans l'ordre de la hiérarchie. Tous les premiers critères de la hiérarchie où p<0.05 sont alors significatifs et permettent de conclure au bénéfice du traitement sur ces critères. Dès l'obtention d'un p>=0.05, l'analyse s'interrompt et tous les autres critères situés en dessous dans la hiérarchie deviennent non concluants (quelle que soit la valeur du p, y compris si p<0.05).

> For the primary and key secondary outcomes only, the type I error was controlled by a hierarchical gatekeeping procedure, wherein each successive outcome was tested only if the preceding comparison was significant at a two-sided P value of 0.05. [10.1056/NEJMoa1714631]

> L'essai DAPA-HF [10.1056/NEJMoa1911303] a utilisé une hiérarchisation pour contrôler le risque alpha global sur plusieurs critères :
>
> « We used a closed testing procedure, with prespecified hierarchical testing of the primary and secondary outcomes. The type I error was controlled at a two-sided alpha level of 0.0499 for multiple comparisons across primary and secondary outcomes, with one interim efficacy analysis taken into account. »
>
> NB : Le seuil de signification dans la hiérarchie n'est pas 0.05, mais 0.0499, car une partie du risque alpha global a été attribué à une analyse intermédiaire (cf. section 4.8).

Pour interpréter les résultats, il convient en premier d'identifier les critères inclus dans la hiérarchie et leur position respective :

"The primary outcome was 1 a composite of worsening heart failure or death from cardiovascular causes. …. A key secondary outcome was 2 a composite of hospitalization for heart failure or cardiovascular death. The additional secondary outcomes were the total number of 3 hospitalizations for heart failure and cardiovascular deaths; 4 the change from baseline to 8 months in the total symptom score on the Kansas City Cardiomyopathy Questionnaire; 5 a composite of worsening renal function; and 6 death from any cause"

Le tableau des résultats se lit alors dans l'ordre de cette hiérarchie. Le p pour les critères 1 à 5 est inférieur au seuil de 0.0499 et permet donc de conclure au bénéfice du traitement sur ces critères. Dans ce tableau (cf. note de bas de tableau) le sigle NA est utilisé pour les tests qui ne permettent pas de conclure. Il s'avère donc que le p du critère n° 5 ne permettait pas de conclure (il n'est pas rapporté, mais c'est le premier NA de la hiérarchie). De ce fait le p du critère n° 6 n'est pas rapporté (cf. section 4.5).

Table 2. Primary and Secondary Cardiovascular Outcomes and Adverse Events of Special Interest.*

	Variable	Dapagliflozin (N=2373)	events/100 patient-yr	Placebo (N=2371)	events/100 patient-yr	Hazard or Rate Ratio or Difference (95% CI)	P Value
	Efficacy outcomes						
1	Primary composite outcome — no. (%)†	386 (16.3)	11.6	502 (21.2)	15.6	0.74 (0.65 to 0.85)	<0.001
	Hospitalization or an urgent visit for heart failure	237 (10.0)	7.1	326 (13.7)	10.1	0.70 (0.59 to 0.83)	NA
	Hospitalization for heart failure	231 (9.7)	6.9	318 (13.4)	9.8	0.70 (0.59 to 0.83)	NA
	Urgent heart-failure visit	10 (0.4)	0.3	23 (1.0)	0.7	0.43 (0.20 to 0.90)	NA
	Cardiovascular death	227 (9.6)	6.5	273 (11.5)	7.9	0.82 (0.69 to 0.98)	NA
	Secondary outcomes						
2	Cardiovascular death or heart-failure hospitalization — no. (%)	382 (16.1)	11.4	495 (20.9)	15.3	0.75 (0.65 to 0.85)	<0.001
3	Total no. of hospitalizations for heart failure and cardiovascular deaths‡	567	—	742	—	0.75 (0.65 to 0.88)	<0.001
4	Change in KCCQ total symptom score at 8 mo§	6.1±18.6	—	3.3±19.2	—	1.18 (1.11 to 1.26)	<0.001
5	Worsening renal function — no. (%)¶	28 (1.2)	0.8	39 (1.6)	1.2	0.71 (0.44 to 1.16)	NA STOP
6	Death from any cause — no. (%)	276 (11.6)	7.9	329 (13.9)	9.5	0.83 (0.71 to 0.97)	NA

* Plus–minus values are means ±SD. NA denotes not applicable because P values for efficacy outcomes are reported only for outcomes that were included in the hierarchical-testing strategy.

Dans cet exemple, il ne faut surtout pas tomber dans le piège de conclure à un résultat significatif pour les décès de toute cause (critère n° 6) en se basant sur l'intervalle de confiance. En effet il s'agirait d'une signification nominale qui n'a rien à voir avec la signification en termes de risque alpha global. Sur ce critère aucune conclusion ne peut être portée, car le test de la hiérarchie s'arrête au-dessus (au niveau du critère n° 5).

Il ne faut pas déduire la signification statistique de l'intervalle de confiance quand le p n'est pas rapporté

Les p<0.05 pour des critères situés dans la hiérarchie en dessous du premier « non significatif » ne doivent pas être considérés et ne permettent pas de conclure au bénéfice du traitement.

Les résultats de l'essai Odyssey Outcome ont d'abord été présentés à un congrès de cardiologie avec la diapositive suivante :

ACC.18

Main Secondary Efficacy Endpoints:
Hierarchical Testing

Endpoint, n (%)	Alirocumab (N=9462)	Placebo (N=9462)	HR (95% CI)	Log-rank P-value
CHD event	**1199 (12.7)**	**1349 (14.3)**	**0.88 (0.81, 0.95)**	**0.001**
Major CHD event	**793 (8.4)**	**899 (9.5)**	**0.88 (0.80, 0.96)**	**0.006**
CV event	**1301 (13.7)**	**1474 (15.6)**	**0.87 (0.81, 0.94)**	**0.0003**
Death, MI, ischemic stroke	**973 (10.3)**	**1126 (11.9)**	**0.86 (0.79, 0.93)**	**0.0003**
CHD death	**205 (2.2)**	**222 (2.3)**	0.92 (0.76, 1.11)	0.38
CV death	**240 (2.5)**	**271 (2.9)**	0.88 (0.74, 1.05)	0.15
All-cause death	**334 (3.5)**	**392 (4.1)**	**0.85 (0.73, 0.98)**	**0.026***

*Nominal P-value

ODYSSEY OUTCOMES 33

http://clinicaltrialresults.org/Slides/ACC2018/ODYSSEY_Steg.pdf

Un bénéfice de l'alirocumab a été montré sur les 4 premiers critères de jugement secondaires. La valeur du p sur le premier critère de mortalité (CHD, coronary heart disease death) interrompt la hiérarchie et il est donc impossible de conclure sur les 3 derniers critères, y compris sur les décès de toute cause, même si son p nominal est inférieur à 0.05. Cette subtilité statistique n'a cependant pas été perçue par tout le monde et ce résultat de mortalité a ensuite été largement repris dans des sources secondaires[6] et en communication promotionnelles pour mettre en avant une réduction de la mortalité de toute cause comme le montre les titres suivants :

http://www.news.sanofi.us/2018-11-11-ODYSSEY-OUTCOMES-investigators-highlight-at-AHA-that-Praluent-R-alirocumab-Injection-was-associated-with-fewer-deaths-from-any-cause

[6] Les sources secondaires sont des revues journalistiques ou des revues professionnelles, très nombreuses et souvent distribuées gratuitement aux médecins. Elles sont souvent des revues promotionnelles (publirédaction-nel).

Dans la publication dans le NEJM [10.1056/NEJMoa1801174], aucune p value n'est rapportée pour les décès de toute cause conformément aux pratiques de ce journal (cf. section 4.5) afin de prévenir ce genre de surinterprétation des résultats et les spins de conclusions qui pourraient être engendrés.

Table 2. Composite Primary End Point and Secondary End Points (Intention-to-Treat Population).

End Point	Alirocumab (N=9462)	Placebo (N=9462)	Hazard Ratio (95% CI)	P Value
	number of patients (percent)			
Primary end point: composite of death from coronary heart disease, nonfatal myocardial infarction, fatal or nonfatal ischemic stroke, or unstable angina requiring hospitalization	903 (9.5)	1052 (11.1)	0.85 (0.78–0.93)	<0.001
Major secondary end points, in order of hierarchical testing				
Any coronary heart disease event*	1199 (12.7)	1349 (14.3)	0.88 (0.81–0.95)	0.001
Major coronary heart disease event†	793 (8.4)	899 (9.5)	0.88 (0.80–0.96)	0.006
Any cardiovascular event‡	1301 (13.7)	1474 (15.6)	0.87 (0.81–0.94)	<0.001
Composite of death from any cause, nonfatal myocardial infarction, or nonfatal ischemic stroke§	973 (10.3)	1126 (11.9)	0.86 (0.79–0.93)	<0.001
Death from coronary heart disease	205 (2.2)	222 (2.3)	0.92 (0.76–1.11)	0.38¶
Death from cardiovascular causes	240 (2.5)	271 (2.9)	0.88 (0.74–1.05)	
Death from any cause	334 (3.5)	392 (4.1)	0.85 (0.73–0.98)	

Cet exemple illustre bien les limites des sources secondaires et de la communication promotionnelle et montre l'intérêt de pouvoir interpréter par soi-même les résultats des essais pour se forger sa propre opinion sur le réel intérêt clinique d'un nouveau traitement, en toute indépendance.

Dans les méthodes hiérarchiques, le seuil de la signification n'est pas toujours 0.05. Il peut être plus petit en raison, par exemple, de la réalisation d'analyses intermédiaires ou d'une répartition en amont pour gérer plusieurs doses de traitement (cf. section 4.4.3).

We used a closed testing procedure, with prespecified hierarchical testing of the primary and secondary outcomes. The type I error was controlled at a two-sided alpha level of 0.0499 for multiple comparisons across primary and secondary outcomes, with one interim efficacy analysis taken into account.

4.4.3 Combination des deux approches

Les deux approches de la hiérarchisation et de la répartition sont assez fréquemment combinées. Par exemple, lorsque deux doses du nouveau traitement sont évaluées dans un essai (essai à 3 bras) le risque alpha global est d'abord réparti entre les 2 doses (2.5% pour chaque dose par exemple) et ensuite pour chaque dose une série de critères de jugement sont testés de manière séquentielle.

A Bonferroni approach (splitting the overall α between the two dose levels of verubecestat) in conjunction with a hierarchical sequential testing approach of outcomes was used to control for the type 1 error rate, with testing of outcomes as described in the order listed in the Outcomes section.

Separately for each dose level, if significant superiority was not shown, all subsequent outcomes were assumed not to have differed significantly between the groups [10.1056/NEJMoa1812840]

Dans d'autres cas, les 2 doses peuvent être testées de manière complètement hiérarchique

Hypotheses were tested in the following order: the 20-mg cannabidiol group, followed by the 10-mg cannabidiol group, was compared with the placebo group with respect to the primary outcome; the 20-mg cannabidiol group was then compared with the placebo group with respect to each key secondary

outcome in the order listed above, and then the 10-mg cannabidiol group was compared with the placebo group with respect to each key secondary outcome in the same order [10.1056/NEJMoa1714631]

4.5 Nouvelle politique de présentation des p value

Depuis 2018, des revues comme le NEJM ne présentent plus dans les articles d'essais cliniques les p des tests qui ne permettent pas d'éventuellement conclure à l'intérêt du traitement[7], afin d'éviter à leur lecteur de commettre des surinterprétations abusives de p values nominales inférieures à 0.05, mais qui ne sont pas pour autant statistiquement significatives en termes de risque alpha global.

Figure 2 – Exemple de nouvelle présentation des p values
Pour les tests qui ne peuvent pas conduire à conclure à l'intérêt du traitement les p values ne sont plus rapportées. Il s'agit des tests non pris en compte dans le plan de contrôle du risque alpha global ou des résultats non significatifs en termes de risque alpha global. [10.1056/NEJMoa1811090]

Table 2. Primary and Secondary End Points in the Modified Intention-to-Treat Population.*

End Point	Rimegepant (N = 537)	Placebo (N = 535)	Absolute Difference	P Value
	no./total no. (%)†		*percentage points (95% CI)*	
Primary end points				
Freedom from pain 2 hours after the dose	105 (19.6)	64 (12.0)	7.6 (3.3 to 11.9)	<0.001
Freedom from the most bothersome symptom 2 hours after the dose	202 (37.6)	135 (25.2)	12.4 (6.9 to 17.9)	<0.001
Secondary end points				
Freedom from photophobia 2 hours after the dose	183/489 (37.4)	106/477 (22.3)	15.1 (9.4 to 20.8)	<0.001
Freedom from phonophobia 2 hours after the dose	133/362 (36.7)	100/374 (26.8)	9.9 (3.2 to 16.6)	0.004
Pain relief 2 hours after the dose	312 (58.1)	229 (42.8)	15.3 (9.4 to 21.2)	<0.001
Freedom from nausea 2 hours after the dose	171/355 (48.1)	145/336 (43.3)	4.8 (–2.7 to 12.2)	
Use of rescue medication within 24 hr after the dose	113 (21.0)	198 (37.0)	–16.0 (–21.3 to –10.6)	
Sustained freedom from pain 2 to 24 hr after the dose	66 (12.3)	38 (7.1)	5.2 (1.7 to 8.7)	
Sustained pain relief 2 to 24 hr after the dose	229 (42.6)	142 (26.5)	16.1 (10.5 to 21.7)	
Sustained freedom from pain 2 to 48 hr after the dose	53 (9.9)	32 (6.0)	3.9 (0.7 to 7.1)	
Sustained pain relief 2 to 48 hr after the dose	195 (36.3)	121 (22.6)	13.7 (8.3 to 19.1)	
Pain relapse 2 to 48 hr after the dose	52/105 (49.6)	32/64 (50.0)	–0.4 (–15.8 to 15.1)	
Ability to function normally 2 hr after the dose	175 (32.6)	125 (23.4)	9.2 (3.9 to 14.6)	

* The modified intention-to-treat population included patients who underwent randomization, had a migraine attack with pain of moderate or severe intensity, took a dose of rimegepant or placebo, and had at least one efficacy assessment after administration of the dose. To maintain the type I statistical error rate at 0.05, a prespecified hierarchical testing procedure was applied; end points are presented in the sequence in which they were evaluated. Because the incidence of freedom from nausea did not differ significantly between the groups, all statistical tests below this end point in the hierarchy are reported without P values, and no inferences can be made from those results. Percentages,

Dans ces papiers, l'absence de p value **ne doit absolument pas être compensé** en cherchant la signification statistique à partir de l'intervalle de confiance à 95%. Cela conduirait à déterminer une signification nominale, mais qui ne permet pas de conclure à la signification du résultat vis-à-vis du risque alpha global. Si le p n'a été rapporté, c'est que le test ne peut pas être utilisé pour déterminer l'intérêt du traitement (il est en est en dehors du plan de contrôle du risque alpha global ou non significatif).

[7] Ces tests sont parfois appelés non inférentiels, car ils ne permettent pas d'inférer si le traitement à un intérêt ou pas, compte tenu du plan de contrôle du risque alpha de l'essai.

4.6 Critères de jugement secondaires

4.6.1 Essai avec un critère de jugement principal unique

Les 2 terminologies « critère de jugement principal » (*primary endpoint ou outcome*) et « critères de jugement secondaires » (*secondary endpoint/outcome*) n'ont vraiment de sens qu'avec les essais utilisant un critère de jugement principal unique.

Dans ce cas, seul ce critère principal unique peut permettre de conclure à l'intérêt du traitement. C'est le seul qui peut être statistiquement significatif en termes de risque alpha global de l'essai et qui peut apporter une démonstration de l'intérêt du traitement.

Dans ce cadre, les critères secondaires ne permettent pas de démontrer et ne peuvent pas être statistiquement significatifs en termes de risque alpha global de l'essai, quelle que soit la valeur nominale de leur p (dans le NEJM ces p ne sont plus rapportés pour cette raison, cf. section 4.5).

Ces critères ne peuvent pas permettre de conclure à l'intérêt du traitement. Tout au plus, ils peuvent faire générer de nouvelles hypothèses à tester dans un nouvel essai.

4.6.2 Essai gérant la multiplicité par un plan de contrôle du risque global

Dans les essais cliniques modernes les termes critère de jugement principal et critères de jugement secondaires n'ont plus beaucoup d'intérêt et, surtout, l'interprétation ne doit s'arrêter aux termes utilisés, mais doit rentrer dans le détail du plan de contrôle du risque alpha global.

Par exemple, en cas d'utilisation d'une hiérarchisation, le terme critère principal désigne le premier de la hiérarchie. Les suivants sont souvent encore appelés critères de jugement secondaires, mais en précisant hiérarchisés ou clés (*key secondary endpoints, main secondary endpoints*, etc.).

Ainsi dans ce cas, des « critères de jugement secondaires » peuvent permettre de conclure à l'intérêt du traitement, mais parce qu'ils sont hiérarchisés (entre eux et avec le critère principal). On voit ainsi toute l'ambigüité actuelle du terme critère secondaire et la nécessité de ne pas s'arrêter aux termes et de bien disséquer la méthode.

De plus dans ces études avec hiérarchisation, il peut aussi exister des critères de jugement secondaires qui ne sont pas dans le plan de contrôle du risque alpha et qui n'ont donc qu'une valeur exploratoire.

	Essai avec un critère de jugement principal unique (espèce en voie d'extinction !)	**Essai gérant la multiplicité avec un plan de contrôle du risque alpha global**
Critères **pouvant permettre de conclure** à l'intérêt du traitement **Démonstration possible**	• Critère de jugement principal	• Critère de jugement principal • Critères de jugement secondaires hiérarchisés • Co primary endpoints (répartition)
Critères **ne pouvant pas permettant de conclure** à l'intérêt du traitement = critère exploratoire **Pas de démonstration possible**	• Critères de jugement secondaires • Critères tertiaires • Critères exploratoires	• Critères de jugement secondaires non hiérarchisés (non inclus dans le plan de contrôle du risque alpha global) • Critères tertiaires • Critères exploratoires

Figure 3 – Exemple d'un critère de jugement secondaire avec contrôles du risque alpha global et un critère de jugement secondaire ordinaire [10.1056/NEJMoa1714631]

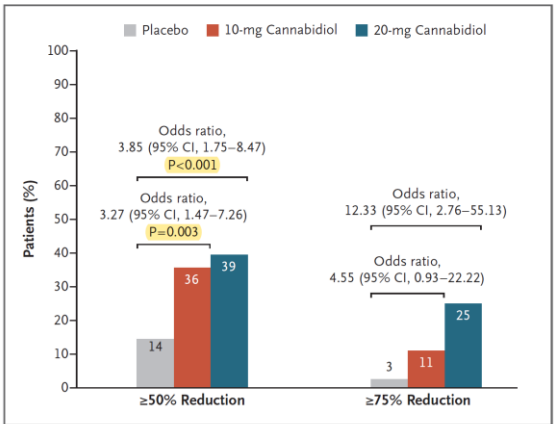


Figure 3. Reductions of at Least 50% and 75% from Baseline in Drop-Seizure Frequency during the Treatment Period.

The odds ratios are for the comparisons between each cannabidiol group and the placebo group. The P values were calculated with the use of a Cochran–Mantel–Haenszel test, with stratification according to age group (2 to 5 years, 6 to 11 years, 12 to 17 years, and 18 to 55 years). P values for reductions of at least 75% from baseline in drop-seizure frequency are not shown because this was not a key secondary outcome, and type 1 error was not controlled.

4.7 Les analyses en sous-groupes

Les analyses en sous-groupes mesurent l'effet du traitement pour des sous-catégories de patients (et non plus pour la totalité des patients inclus dans l'étude).

Les analyses en sous-groupes ont des limites statistiques importantes qui empêchent de les utiliser pour conclure à l'effet du traitement ou à son absence au niveau des sous-types de patients.

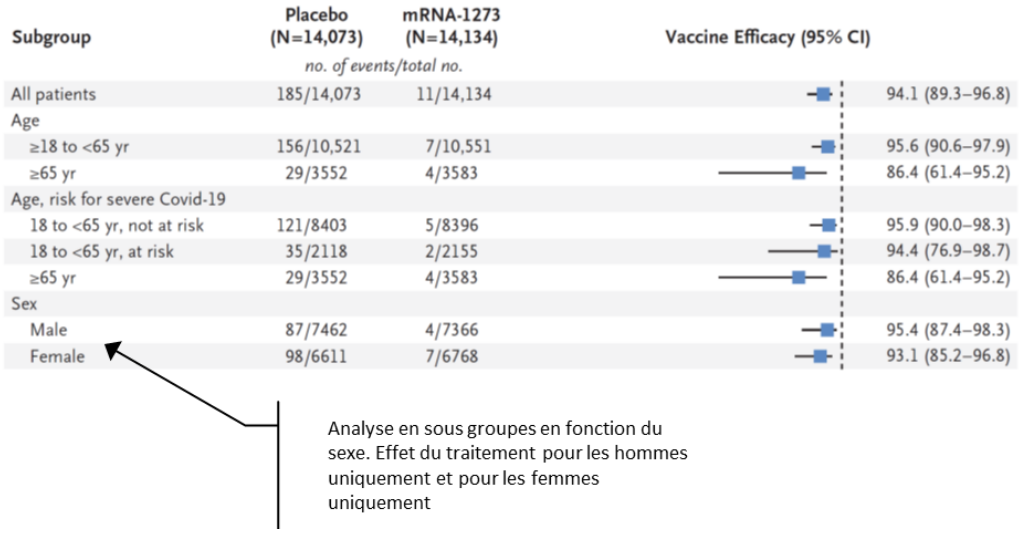
Sauf méthodes particulières (cf. ci-dessous ???), les analyses en sous-groupes n'ont qu'une valeur exploratoire et leurs résultats ne permettent pas de faire des conclusions fiables, à même de faire changer les pratiques.

Au mieux, elles permettent de générer de nouvelles hypothèses, à vérifier dans de nouveaux essais entrepris spécialement.

Cela étant dit, il faut noter qu'en pratique, hors du champ d'une interprétation rigoureuse, les analyses en sous-groupe sont largement surinterprétées, exposant à des prises de décisions potentiellement erronées. La littérature regorge d'exemples de résultats de sous-groupes qui ont été invalidés par les études de vérification ultérieures.

Par exemple, dans un essai négatif, les sous-groupes sous souvent entrepris de manière erronée pour chercher si le traitement ne serait pas utile que pour un sous-groupe particulier de patients, où l'on trouverait un résultat « statistiquement significatif ».

Figure 4 – Exemple d'analyses en sous-groupes réalisées dans un essai de vaccins de la COVID19 [10.1056/NEJMoa2035389].
Le résultat de l'essai (all patient) est rappelé en haut du graphique. Plusieurs analyses en sous-groupes sont représentées. La première est l'analyse en fonction de l'âge avec 2 modalités : entre 18 et 65 et supérieure à 65. L'efficacité du vaccin est estimée spécifiquement pour chacune de ces 2 modalités : 95.6% pour les 18-65 ans et 86.4% pour les 65 et plus. L'efficacité vaccinale est la réduction relative du risque (= 1-risque relatif*100%).



Subgroup	Placebo (N=14,073)	mRNA-1273 (N=14,134)	Vaccine Efficacy (95% CI)
	no. of events/total no.		
All patients	185/14,073	11/14,134	94.1 (89.3–96.8)
Age			
≥18 to <65 yr	156/10,521	7/10,551	95.6 (90.6–97.9)
≥65 yr	29/3552	4/3583	86.4 (61.4–95.2)
Age, risk for severe Covid-19			
18 to <65 yr, not at risk	121/8403	5/8396	95.9 (90.0–98.3)
18 to <65 yr, at risk	35/2118	2/2155	94.4 (76.9–98.7)
≥65 yr	29/3552	4/3583	86.4 (61.4–95.2)
Sex			
Male	87/7462	4/7366	95.4 (87.4–98.3)
Female	98/6611	7/6768	93.1 (85.2–96.8)

Dans un essai « négatif » (non concluant), les sous-groupes ne permettent pas de conclure à l'effet du traitement pour un sous-type de patients particulier, car il existe une inflation du risque alpha global lié à la multiplicité des comparaisons induites par les sous-groupes (souvent plusieurs dizaines, voire centaines).

La préspécification des sous-groupes au protocole ne solutionne pas la problématique (inflation du risque alpha liée à la multiplicité).

Pour permettre de conclure à l'intérêt particulier du traitement pour un ou des sous-groupes de patients, ce ou ces sous-groupes doivent être inclus dans le plan de contrôle du risque alpha global (par hiérarchisation ou répartition du risque alpha).

L'essai Plato [10.1056/NEJMoa0904327] évaluait le ticagrelor lors d'un syndrome coronarien aigu (SCA). Dans cette pathologie les patients peuvent bénéficier d'une procédure invasive de revascularisation (PCI, stent). Ces patients représentent une population bien particulière par rapport à ceux traités exclusivement de manière médicale. Il convient d'avoir la preuve formelle que le ticagrelor apporte bien un bénéfice chez eux. Il y a donc nécessité de pouvoir conclure sur un sous-groupe de patients (ceux traités invasivement). Ce sous-groupe a pour cela été inclus dans la hiérarchie en 2ème position :

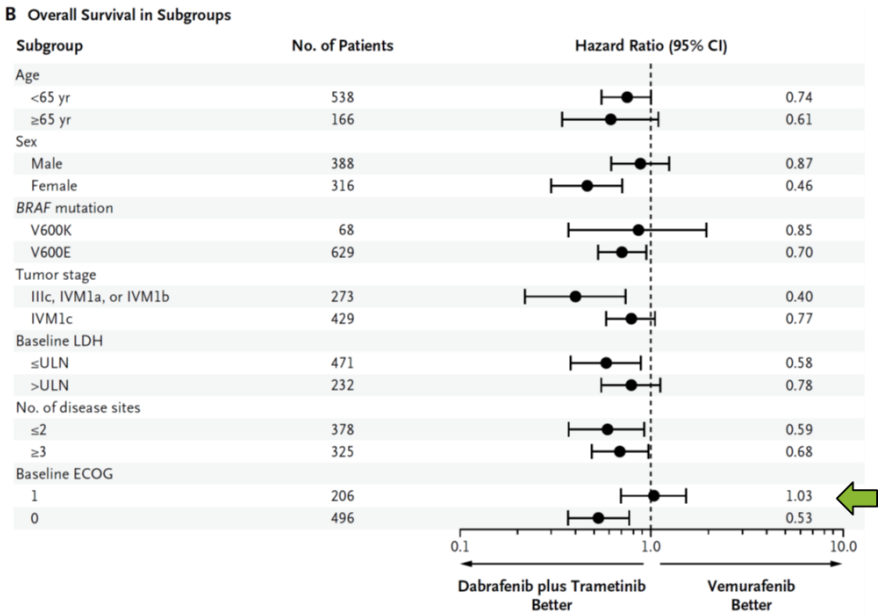
« The primary efficacy variable was the time to the first occurrence of composite of death from vascular causes, myocardial infarction, or stroke. … The principal secondary efficacy end point was the primary efficacy variable studied in the subgroup of patients for whom invasive management was planned at randomization. Additional secondary end points (analyzed for the entire study population) were …".

Sans cette inclusion dans le plan de contrôle du risque alpha global, ce résultat aurait été uniquement exploratoire et n'aurait pas permis de décider si le ticagrelor avait sa place dans la prise en charge des SCA traités invasivement.

Dans un essai concluant (montrant l'intérêt du traitement au niveau), les analyses en sous-groupe ne permettent pas non plus de conclure à l'absence d'effet pour certains sous-types de patients pour plusieurs raisons :

- Inflation du risque beta (de ne pas conclure à tort à l'effet du traitement) liée à la multiplicité.
- Réduction d'effectif, entrainant une réduction de la précision des estimations (largeur des intervalles de confiance) et de la puissance statistique.
- Conclusion à l'absence d'effet à partir d'une différence non significative impossible (cf. section 8)

Figure 5 – Exemple d'un sous-groupe « non significatif » dans un essai concluant.
Une réduction de mortalité a été parfaitement bien démontrée par cet essai chez des patients ayant un mélanome métastatique. Dans le sous-groupe des patients avec un ECOG de 1 (flèche en bas), le hazard ratio est de 1.03 avec un IC incluant de-facto 1. Compte tenu de ses limitations statistiques, ce résultat ne doit pas faire conclure à l'absence de bénéfice du traitement chez ces patients et les faire exclure de l'indication [10.1056/NEJMoa1412690]



Il n'est donc pas possible de conclure que certains patients ne sont pas répondeurs au traitement à partir d'analyse en sous-groupes ordinaires. Pour cela il faut l'utilisation d'une méthode statistique adaptée (hors sujet ici).

Il est extrêmement risqué de conclure à l'efficacité du traitement ou à son absence pour certains sous-types de patients à partir des analyses en sous-groupes ordinaires

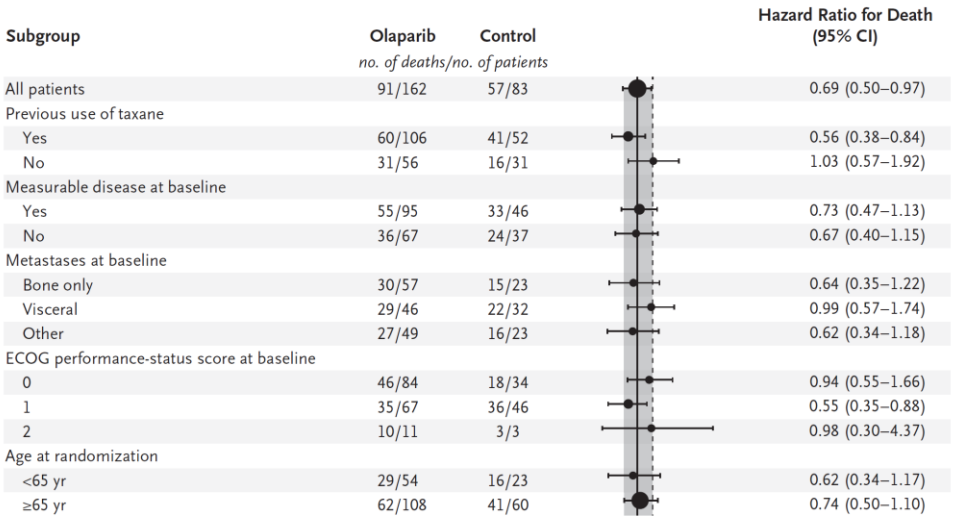
---

❖ *Vérification de la généralisabilité du résultat*

Les analyses en sous-groupes ont pour seul but de s'assurer de la généralisabilité du résultat global de l'essai à l'ensemble des sous-types de patients inclus dans l'essai. C'est-à-dire s'assurer que l'hypothèse qui a été faite lors de l'élaboration des critères d'éligibilité du protocole, que tous les patients

inclus bénéficieraient de la même façon du traitement n'est pas remise en cause par les résultats des analyses en sous-groupes.

Figure 6 – Utilisation des analyses en sous-groupes pour s'assurer de la généralisabilité du résultat de l'essai à tous les sous-types de patients inclus dans l'essai.
La bande grise verticale représente la projection de l'intervalle de confiance du résultat de l'essai (All patients) sur les résultats des analyses en sous-groupes. Les intervalles de confiances des sous-groupes ont tous des parties communes avec l'intervalle de confiance du résultat de l'essai. Il n'existe pas de cas où le résultat obtenu par un sous-groupe serait discordant avec le résultat global compte tenu de l'incertitude des estimations. Le résultat de l'essai est donc généralisable à l'ensemble des catégories de patients inclus.
[10.1056/NEJMoa2022485]



Subgroup	Olaparib	Control	Hazard Ratio for Death (95% CI)
	no. of deaths/no. of patients		
All patients	91/162	57/83	0.69 (0.50–0.97)
Previous use of taxane			
Yes	60/106	41/52	0.56 (0.38–0.84)
No	31/56	16/31	1.03 (0.57–1.92)
Measurable disease at baseline			
Yes	55/95	33/46	0.73 (0.47–1.13)
No	36/67	24/37	0.67 (0.40–1.15)
Metastases at baseline			
Bone only	30/57	15/23	0.64 (0.35–1.22)
Visceral	29/46	22/32	0.99 (0.57–1.74)
Other	27/49	16/23	0.62 (0.34–1.18)
ECOG performance-status score at baseline			
0	46/84	18/34	0.94 (0.55–1.66)
1	35/67	36/46	0.55 (0.35–0.88)
2	10/11	3/3	0.98 (0.30–4.37)
Age at randomization			
<65 yr	29/54	16/23	0.62 (0.34–1.17)
≥65 yr	62/108	41/60	0.74 (0.50–1.10)

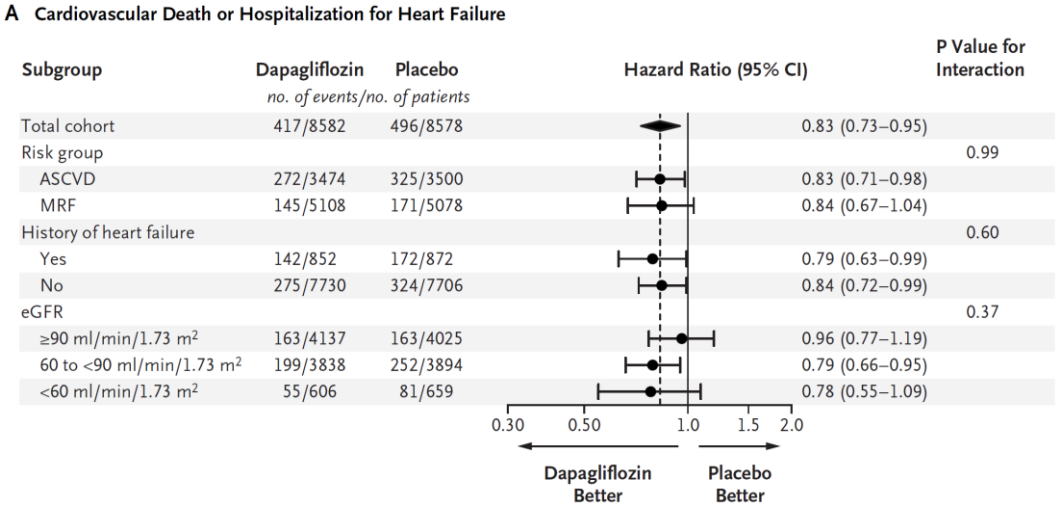
❖ *Interaction*

Les analyses en sous-groupes testent fréquemment l'interaction, c'est-à-dire recherche si la taille de l'effet du traitement (le risque relatif, le hazard ratio, l'odds ratio ou la différence de moyenne) varie quantitativement entre les sous-groupes. Par exemple, un test d'interaction fait sur le sexe recherche si la taille de la taille de l'effet du traitement est différente entre les hommes et les femmes.

Interaction et hétérogénéité sont synonymes dans ce contexte.

Dans un essai de nombreux tests d'interaction sont réalisés, mais comme ils ne sont pas utilisés pour conclure à l'intérêt du traitement, cette multiplicité n'entraine pas d'inflation du risque alpha global de l'essai. Cette multiplicité conduit cependant à découvrir des interactions à tort qui n'existe pas en réalité. Il existe aussi un risque non négligeable de conclure tort à l'absence d'interaction (les tests d'interactions sont en général peu puissants).

Figure 7 – Exemple d'analyse en sous-groupes avec test d'interaction.
Pour l'analyse en sous-groupe en fonction des antécédents d'insuffisance cardiaque, le p du test d'interaction est de 0.60, ne permettant pas de conclure qu'il existe une différence statistiquement significative entre l'effet du traitement chez les patients ayant un antécédent (HR=0.79) par rapport à l'effet chez les patients sans antécédents (HR=0.84). Compte tenu de l'incertitude entourant ces 2 estimations, il n'est pas possible de conclure que ces 2 hazard ratio (0.79 et 0.84) sont différents. [10.1056/NEJMoa1812389]



Utilisation **erronée** des sous-groupes pour conclure vis-à-vis de l'effet du traitement	Analyse de la signification statistique (nominale) pour chaque sous-groupe, par exemple chez les hommes, chez les femmes	Correspond au même objectif que l'essai, entraine donc une inflation du risque alpha global de conclure à tort à un quelconque intérêt du traitement
Utilisation **appropriée** des sous-groupes pour rechercher si un facteur modifie la taille de l'effet du traitement (interaction)	Comparer la taille de l'effet entre les sous-groupes (par exemple entre les hommes et les femmes pour explorer si le sexe est un facteur modifiant l'effet du traitement)	Ne cherche pas à conclure à l'intérêt du traitement Résultat purement exploratoire, cognitif. Ne permettant pas de faire des conclusions sur l'intérêt du traitement, cette analyse n'entraine pas d'inflation du risque alpha global de l'essai

4.8 L'analyse finale et les analyses intermédiaires

Le moment de l'analyse d'un essai doit être parfaitement bien défini a priori pour éviter que l'essai soit poursuivi ou arrêté en fonction des résultats du moment.

> *Si le moment de l'analyse n'est pas préfixé, l'essai sera alors analysé a un moment arbitraire qui peut dépendre des résultats (des analyses sont répétées régulièrement jusqu'à ce que le résultat s'avère satisfaisant si cela arrive).*

Cette analyse survient, après l'inclusion de tous les patients nécessaires, soit à une date de point prédéfini, soit lorsque la durée de suivi voulue (mortalité à 1an par exemple) a été atteinte pour tous les patients, soit, le plus souvent actuellement, lorsque le nombre d'évènements nécessaires a été atteint (tous groupes confondus).

« We estimated that 288 events would be required to detect a hazard ratio for death of 0.675 with an alpha level of 0.05" [10.1056/NEJMoa1412690]

Lors de cette analyse finale de l'essai, le bénéfice du traitement est recherché en comparant le critère de jugement entre les 2 groupes et la signification statistique de la différence observée est appréciée en calculant le p.

Parfois d'autres analyses sont réalisées avant cette analyse finale. Il s'agit des analyses intermédiaires (AI) qui sont en général au nombre d'une ou deux. Ces analyses ont aussi pour but de mettre en évidence le bénéfice du traitement (analyse d'efficacité) si les résultats le permettent et reposent donc sur une comparaison statistique des 2 groupes.

Si le bénéfice du traitement est démontré à une analyse intermédiaire, l'objectif de l'essai est atteint et il n'ait plus nécessaire de le poursuivre (pour cet objectif, mais parfois l'étude se poursuit pour répondre à un autre objectif, sur un autre co primary endpoint par exemple). On dit que l'essai a été arrêté prématurément pour démonstration anticipée de l'efficacité.

Cependant si une analyse intermédiaire ne permet pas de conclure au bénéfice du traitement, l'essai se poursuit jusqu'à la prochaine analyse intermédiaire ou jusqu'à l'analyse finale.

La réalisation des AI entraine une répétition potentielle des comparaisons statistiques cherchant à conclure au bénéfice du traitement. Il y a donc potentiellement une inflation du risque alpha global de l'essai.

Les AI sont donc réalisées à l'aide de méthode statistique adaptée (peu importe leur nom) qui ajuste le seuil de la signification statistique.

Pour pouvoir conclure à une analyse intermédiaire, il faut que le p (nominal) soit inférieur au seuil ajusté calculé par la méthode statistique (on dit alors que la frontière de la signification a été franchie). Le seuil ajusté est en général assez faible (0.0025 par exemple) et il est calculé en fonction du nombre d'évènements observé au moment de l'analyse. Il est rapporté dans la publication.

> "At the data-cutoff date of April 17, 2014, the interim analysis was performed after 222 events had occurred. For the overall survival analysis, 100 patients (28%) in the combination-therapy group and 122 (35%) in the vemurafenib group had died (hazard ratio for death in the combination-therapy group, 0.69; 95% confidence interval [CI], 0.53 to 0.89; P=0.005) (Figure 1A). The prespecified stopping boundary (P<0.0214) was crossed, and the study was stopped for efficacy on July 14, 2014" [10.1056/NEJMoa1412690]

Si l'essai n'est pas arrêté lors des analyses intermédiaires et arrive à l'analyse finale, le seuil de la signification est aussi ajusté à la baisse pour prendre en compte le risque alpha « consommé » lors des analyses intermédiaires (répartition du risque alpha global entre les différentes analyses).

> During the course of the trial, two interim analyses were conducted after 50% and 75%, respectively, of the target number of 1,400 participants had experienced a primary cardiovascular endpoint. To conserve alpha for the final analysis and to limit the possibility of a chance positive interim finding, each interim analysis followed the same closed testing procedure, with a one-sided significance level of 0.01% allotted to the first efficacy interim analysis, and a one sided significance level of 0.04% allotted to the second efficacy interim analysis, and thus a one-sided significance level of 2.45% retained for the final analysis. [10.1056/NEJMoa1707914 supplement]

Des spins de conclusion sont fréquemment observés quand l'analyse intermédiaire ne permet pas de conclure formellement, car le p nominal n'est pas inférieur au seuil ajusté, mais qu'il est cependant inférieur à 0.05.

> "Although the difference in overall survival did not cross the prespecified superiority boundary (P<0.0096), continuous lenalidomide–dexamethasone reduced the risk of death, as compared with MPT (hazard ratio, 0.78; 95% CI, 0.64 to 0.96; P=0.02)" [10.1056/NEJMoa1402551]

5 Évaluation du risque de biais

Un essai de supériorité est biaisé quand il existe un autre facteur que le traitement étudié qui induit une différence en faveur du nouveau traitement 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[8].

Sauf cas exceptionnel, il est impossible de déterminer 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 encore d'erreur aléatoire). Il sera donc possible de les considérer comme réels et de conclure à l'intérêt du traitement (après analyse de la robustesse statistique, cf. section 4).

Au niveau des biais, la lecture critique consiste donc à vérifier si l'étude est complètement à l'abri des biais (correctement conçue et réalisée).

- 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.

[8] 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 posent la question de la fiabilité du résultat sur lequel il s'apprête d'adopter le nouveau traitement. La lecture critique des essais « négatifs » est particulière et complètement différente de celle des essais « positifs » que nous développons ici (cf. section 8)

- 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[9]**
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[10] (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,

[9] 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. Pour l'essai thérapeutique, 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.

[10] 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.

mais bien d'un problème de pertinence clinique. C'est un problème de validité externe et non pas de validité 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).

5.1 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. En effet, rien ne garantit que la randomisation (qui est un processus purement aléatoire) « 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, mais qu'ils permettent de faire une comparaison loyale qui ne sera pas 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 seront identiques).

La conséquence de cela pour la lecture critique est qu'il est sans intérêt de vérifier que les 2 groupes issus de la randomisation (table 1, des caractéristiques à la baseline) sont effectivement identiques. 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 convient seulement de vérifier que l'allocation des traitements était 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.

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[11]. 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.

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

5.2 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 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 phlébo, 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

[11] comme l'essai est un 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

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."*

❖ *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. 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).

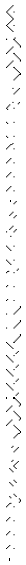
Au mieux, si le critère est partiellement subjectif, les biais pourront être partiellement limités par le recours à un comité d'adjudication des évènements en aveugle. Cependant cela ne remplace pas le double insu, car 1) il reste les liais liés au suivi et 2) la documentation médicale des cas est transmise 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 mauvaises prises décision. Bien entendu les patients participant à l'étude doivent être informés et volontaires.

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

5.3 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.

Essai en double insu	À l'abri des biais

5.4 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.*

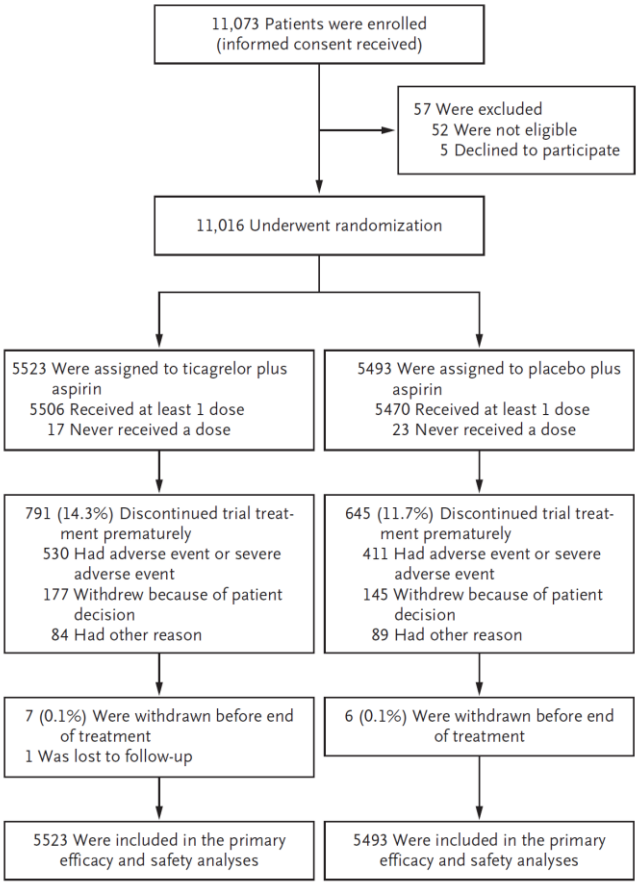
Avec les analyses de survie (*time to event*) les perdus de vues sont souvent considérés comme des censures. Cette approche ne protège pas contre les biais, car il n'y a pas de remplacement conservateur.

> "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 8 – 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]



❖ *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 9 – 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]

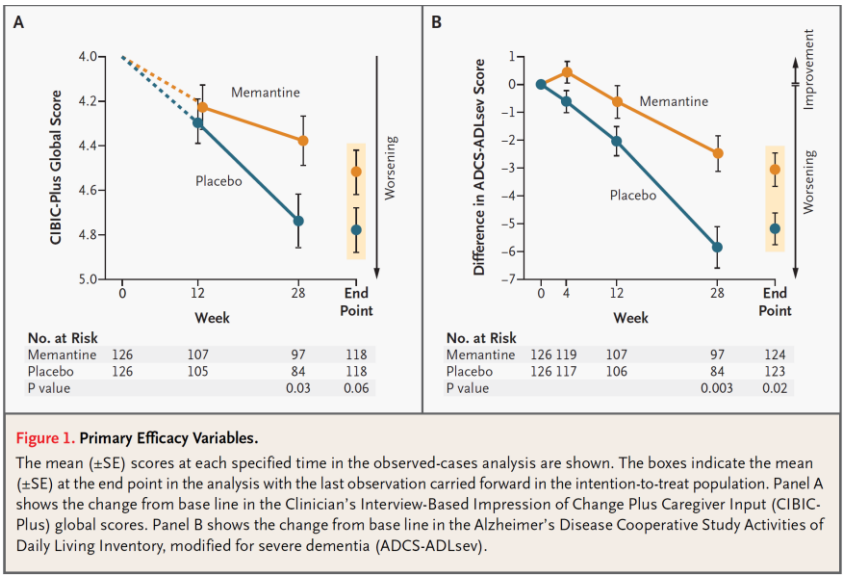


Figure 1. **Primary Efficacy Variables.**
The mean (±SE) scores at each specified time in the observed-cases analysis are shown. The boxes indicate the mean (±SE) at the end point in the analysis with the last observation carried forward in the intention-to-treat population. Panel A shows the change 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).

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 cours.

La lecture critique s’assure :

- Qu’il soit bien mentionné que l’analyse se fait en intention de traiter ou porte sur la population d’analyse en intention[12] 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.

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

5.5 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

6 Lecture critique et fraude scientifique"

Les essais sont réalisés par des promoteurs/sponsors (industriels ou académiques) qui ont un intérêt direct à ce que l’essai donne un résultat « positif ». Il pourrait y avoir une tentation de faire que cela arrive coute que coute, en manipulant les données, par exemple, en escamotant certains patients traités dont l’évolution n’a pas été favorable.

[12] 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)

Depuis toujours des moyens sont mis en œuvre pour empêcher la possibilité de telles pratiques sous la forme d'un système d'assurance qualité comprenant :

- Un protocole et un manuel de procédures qui définissent parfaitement comment doit être réalisé l'essai sur le terrain ;
- Une formation spécifique des investigateurs ;
- Un processus de contrôle des données avec des audits sur site, comparant les données recueillies pour l'essai par rapport aux dossiers des patients ;
- Une traçabilité de toutes les modifications des données, rendant visible par exemple la suppression d'un patient ;
- Accès impossible pour le promoteur/sponsor à la liste de randomisation (qui est la seule à donner la nature du traitement reçu par les patients dans un essai en double aveugle) ;
- Des procédures qui font que les données et les résultats des analyses intermédiaires ne sont pas connus du promoteur/sponsor (mais seulement du comité de surveillance, DSMB, indépendant et externe à l'étude) ;
- Tous les acteurs de l'essai, y compris les investigateurs, prennent des engagements contractuels (convention ou contrat de travail) et engagent leur responsabilité ;
- Une éventuelle réanalyse des données par la FDA par exemple

"data accuracy and integrity were ensured by the contract research organization and checked extensively before database freeze and statistical analysis."

"All the authors wrote or contributed to the writing of the manuscript, … and vouch for the accuracy and completeness of the data and analysis."

Toutes ces mesures font que l'on peut avoir confiance dans les résultats des études. Bien sûr, il existe de temps en temps, des histoires de fraudes avérées, mais elles sont exceptionnelles et sont souvent les conséquences des malversations d'un seul individu. Ce risque est très faible dans les essais industriels où les moyens sont disponibles pour mettre en place ces mesures, et un peu plus présents dans les essais académiques qui peuvent être encore réalisés sans système d'assurance qualité.

La méthodologie a aussi une action dans la prévention de la fraude en rendant hasardeuses d'éventuelles malversations. Par exemple, dans un essai en double aveugle, un promoteur ne s'aventura pas à exclure des patients d'évolution défavorable, car il ne saura pas ce qu'il fait exactement : favoriser ou défavoriser le traitement évalué.

Au total, la méthodologie et le système d'assurance qualité assurent que l'étude donnera toujours les mêmes résultats, quels que soient les liens d'intérêts des acteurs. En revanche, ces liens d'intérêts vont influencer la discussion des résultats, l'interprétation et la communication faites autour de l'étude.

La lecture critique n'a pas pour but de rechercher la fraude.

7 Évaluation de la pertinence clinique (clinical relevance)

Après s'être assuré que le résultat était réel (validité interne), la lecture critique s'assure que ce résultat représente bien un réel progrès thérapeutique (une amélioration du service médical rendu si l'on reprend les termes de la commission de la transparence).

Il s'agit d'évaluer l'utilité médicale démontrée du traitement, en quoi il amènera un changement notable, cliniquement pertinent, du devenir des patients.

En effet, un résultat peut montrer avec un degré de certitude (niveau de preuve) élevé que le traitement à un effet, mais, par exemple, sur un critère de jugement qui ne mesure pas directement le bénéfice attendu par les médecins ou les patients. Dans un cancer au stade métastatique, un résultat sur la réponse tumorale, sans effet démontré sur la survie (ou la qualité de vie) ne documente en rien si le traitement répondra aux attentes des patients en termes de survie. Dans ce cas, ce résultat n'apporte pas la preuve que le traitement est un progrès thérapeutique notable. Dans d'autres cas il pourra s'agir d'une taille d'effet trop petite, d'un comparateur non loyal ou d'une balance bénéfice risque non favorable.

Note. L'évaluation de la pertinence clinique demande de bien connaitre le domaine médical de l'essai (critères de jugement pertinents, traitements déjà validés, etc.). Il s'agit d'une problématique qui sera davantage détaillée au niveau du 3eme cycle, mais pour le 2eme cycle il convient de connaitre les principes et de savoir identifier les situations caricaturales.

7.1 Pertinence du comparateur

Le traitement comparateur (traitement du groupe contrôle) peut être un placebo ou un traitement actif.

Avec le placebo, 2 cas de figure sont possibles :

- Les patients reçoivent dans les 2 groupes un traitement de base identique. On dit que l'évaluation du nouveau traitement se fait « on top » la stratégie de base ou « en add-on ». Le but est de montrer que l'adjonction du nouveau traitement à la stratégie thérapeutique actuelle apporte un bénéfice supplémentaire aux patients.

  > L'essai Pegasus compare le ticagrelor à l'aspirine en prévention cardiovasculaire secondaire à distance de l'évènement initial [10.1056/NEJMoa1500857]. Tous les patients reçoivent aussi de l'aspirine dont le bénéfice a été démontré dans un essai précédent (aspirine seule versus placebo seul). La comparaison effectuée dans Pegasus est donc celle d'une bithérapie ticagrelor + aspirine versus aspirine seule (+ placebo qui est là pour assurer le double aveugle).
  >
  > We randomly assigned, in a double-blind 1:1:1 fashion, 21,162 patients who had had a myocardial infarction 1 to 3 years earlier to ticagrelor at a dose of 90 mg twice daily, ticagrelor at a dose of 60 mg twice daily, or placebo. All the patients were to receive low-dose aspirin.

- Les patients du groupe contrôle ne reçoivent que le placebo. L'évaluation ne se fait pas pardessus d'autres traitements, car aucun traitement n'a démontré son intérêt dans cette situation. Le placebo est alors un comparateur pertinent, car il n'existe pas de traitement de référence. En revanche, si un traitement a déjà montré son intérêt, faire un essai versus placebo n'est pas pertinent, car 1) cela représente une perte de chance pour les patients du groupe contrôle (sauf si l'abstention thérapeutique est aussi une option) et 2) ne répond pas à la question clinique qui se pose dans cette situation : peut-on avoir un traitement plus efficace à celui déjà disponible ?

  > Several studies have tested the administration of systemic therapy with curative intent after patients had disease control with chemoradiotherapy. However, to date, these therapies have proved ineffective. PACIFIC was a randomized, placebo-controlled, phase 3 trial evaluating the immune checkpoint inhibitor durvalumab [10.1056/NEJMoa1809697]

Lorsque le comparateur est un traitement actif, l'essai répond à la question : le remplacement du traitement de référence par le nouveau traitement apporte-t-il un bénéfice supplémentaire aux patients.

Warfarin and other vitamin K antagonists are highly effective treatments, reducing the risk of stroke by about two thirds, but their use is limited by a narrow therapeutic range, drug and food interactions, required monitoring, and risk of bleeding. In the ARISTOTLE trial, we compared apixaban with warfarin for the prevention of stroke or systemic embolism in patients with atrial fibrillation and at least one additional risk factor for stroke. [10.1056/NEJMoa1107039]

Le traitement comparateur actif doit être le meilleur traitement disponible, utilisé de manière optimale. Autrement la comparaison est déloyale, car ne s'effectuant pas par rapport à la meilleure référence possible.

Note : Cette évaluation demande des connaissances sur le domaine médical, et au niveau du 2ème cycle elle ne sera abordée qu'avec la mise à disposition d'une documentation permettant de juger de la situation sans connaissance préalable.

7.2 Pertinence clinique du critère de jugement

La démonstration de l'intérêt clinique d'un traitement nécessite l'utilisation d'un critère clinique.

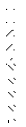
La mise en évidence d'un effet sur un critère intermédiaire est nettement insuffisante pour justifier l'utilisation d'un traitement en pratique

Les critères cliniques correspondent directement à la problématique de la pathologie à traiter. Il s'agit souvent d'évènements cliniques (décès, AVC), mais aussi de signes fonctionnels (douleur, succès thérapeutique sur une échelle de dépression, etc.).

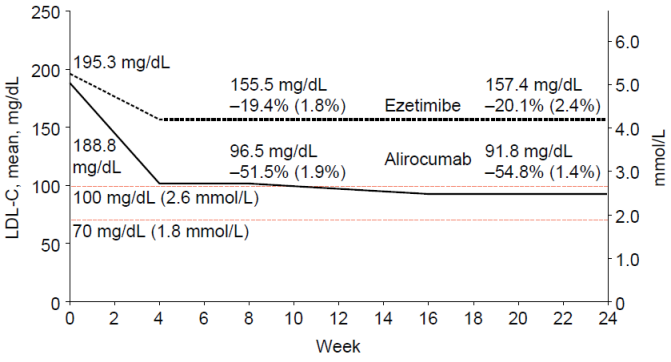
Les critères intermédiaires sont proches du mécanisme d'action des traitements et correspondent le plus souvent à des paramètres biologiques (ou d'imagerie). Mais un effet du traitement sur un critère intermédiaire ne s'accompagne pas toujours d'un bénéfice sur le critère clinique (cf. section 1.3). Pour cette raison, les critères intermédiaires sont insuffisants pour évaluer l'intérêt clinique des traitements.

Tableau 2 – Exemples de critères cliniques et intermédiaires

Pathologie	Critères cliniques	Critères intermédiaires
Ostéoporose	Fractures vertébrales Fractures os long	Densité osseuse
Hypertension	Événements cardiovasculaires (infarctus, AVC, décès de cause cardiovasculaires)	Pression artérielle
COVID-19	Mortalité, aggravation clinique, décès ou ventilation mécanique	Charge virale
Arthrose	Échelle de douleur (EVA), Indice algofonctionnel de Lequesne	Pincement articulaire



L'alirocumab est un hypolipémiant de dernière génération. Les premiers essais réalisés avec cette molécule avaient comme critère de jugement le LDL-cholesterol, comme, par exemple dans les essais Odyssey ALTERNATIVE [10.1016/j.jacl.2015.08.006] ou ODYSSEY COMBO II [10.1111/dom.12909].



No. Pts.	Baseline	Week 12	Week 24
Alirocumab	123	109	90
Ezetimibe	118	95	78

Mais ces essais sur critères intermédiaires n'étaient pas suffisamment cliniquement pertinents pour apporter la preuve de l'intérêt du traitement. Un grand essai de morbi-mortalité, ODYSSEY OUTCOMES [10.1056/NEJMoa1801174], incluant 18,924 avec un suivi médian de 2.8 ans, a donc été entrepris. L'introduction de la publication justifie ainsi la réalisation de cet essai :

"Studies have shown that mutations conveying gain or loss of function of PCSK9 result in a higher or lower level of LDL cholesterol, respectively, which in turn is associated with a corresponding higher or lower risk of incident coronary heart disease. These findings have led to the development of monoclonal antibodies to PCSK9 that produce substantial reductions in LDL cholesterol when administered alone or with a statin.

To date, the potential for a PCSK9 antibody to reduce cardiovascular risk after an acute coronary syndrome remains undetermined. In the ODYSSEY OUTCOMES trial, we tested the hypothesis that treatment with alirocumab, a fully human monoclonal antibody to PCSK9, would result in a lower risk of recurrent ischemic cardiovascular events than placebo … ".

7.3 Interprétation du résultat obtenu sur un critère composite ?

Un critère composite est un critère qui regroupe plusieurs entités cliniques (plusieurs types d'évènements cliniques) comme les évènements cardiovasculaires dont les composantes sont les infarctus du myocarde non mortels, les AVC non mortels et les décès de causes cardiovasculaires.

Un patient sera comptabilisé comme « ayant fait » le critère composite à partir du moment où il présente un des évènements composants. Ainsi, dans le tableau de résultat de l'essai, le nombre de critères composites est le nombre de patients ayant présentés au moins un des évènements composants.

Figure 10 – Exemple de tableau de résultat rapportant un critère composite
La première ligne présente les résultats du critère évènement cardiovasculaires. Dans le groupe ticagrelor 90mg, 493 patients ont présenté au moins un évènement composant ce critère : un décès d'ordre cardiovasculaire, un AVC non mortel ou un infarctus non mortel. La présence de cooccurrences (patients ayant présenté plusieurs de ces évènements, par exemple un premier infarctus puis 6 mois après un second mortel) est attestée par le fait de la somme des évènements composants ne donne pas le nombre du composite (100 AVC + 275 infarctus + 182 décès CV =557 pour 493 patients pour le composite) [10.1056/NEJMoa1500857]

End Point	Ticagrelor, 90 mg (N=7050)	Ticagrelor, 60 mg (N=7045)	Placebo (N=7067)
	number (percent)		
Cardiovascular death, myocardial infarction, or stroke	493 (7.85)	487 (7.77)	578 (9.04)
Death from coronary heart disease, myocardial infarction, or stroke	438 (6.99)	445 (7.09)	535 (8.33)
Cardiovascular death or myocardial infarction	424 (6.79)	422 (6.77)	497 (7.81)
Death from coronary heart disease or myocardial infarction	350 (5.59)	360 (5.75)	429 (6.68)
Cardiovascular death	182 (2.94)	174 (2.86)	210 (3.39)
Death from coronary heart disease	97 (1.53)	106 (1.72)	132 (2.08)
Myocardial infarction	275 (4.40)	285 (4.53)	338 (5.25)
Stroke			
Any	100 (1.61)	91 (1.47)	122 (1.94)
Ischemic	88 (1.41)	78 (1.28)	103 (1.65)

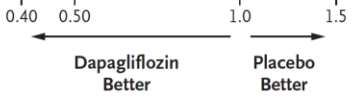
Une limite des critères composites est qu'ils peuvent regrouper des évènements dont la signification clinique est très inégale (en termes de gravité par exemple). Ces composantes les moins cliniquement signifiantes peuvent être majoritaires, ce qui fait que l'effet observé au niveau du critère composite représente principalement l'effet du traitement sur l'évènement le moins cliniquement signifiant.

Les critères composites regroupent fréquemment des évènements mortels et non mortels. Souvent ce qui fait la valeur d'un nouveau traitement dans une aire thérapeutique déjà pourvue est une réduction de la mortalité. D'où l'enjeu de pouvoir conclure d'une façon ou d'une autre à cette réduction. Quand un critère de morbi-mortalité est réduit par le traitement, la conclusion que le traitement à réduit la fréquence du critère « décès ou évènement non mortels » peut être facilement compris comme une réduction « des décès » et « des évènements mortels ». Cependant, dans certains cas, cette réduction du composite n'a été obtenue que par un effet sur les évènements non mortels. L'impression que donne alors la conclusion est excessive par rapport à la réalité des résultats.

❖ *Exemple d'un résultat de critère composite problématique*

Dans un essai [10.1056/NEJMoa1812389] chez le patient diabétique de type 2, le critère de jugement principal était défini par *« The primary efficacy outcomes were MACE and a composite of cardiovascular death or hospitalization for heart failure*. ». Les résultats obtenus sont les suivants :

Outcome	Dapagliflozin (N=8582) no. (%)	Dapagliflozin (N=8582) rate/1000 patient-yr	Placebo (N=8578) no. (%)	Placebo (N=8578) rate/1000 patient-yr	Hazard Ratio (95% CI)	P Value
Cardiovascular death or hospitalization for heart failure	417 (4.9)	12.2	496 (5.8)	14.7	0.83 (0.73–0.95)	0.005
MACE	756 (8.8)	22.6	803 (9.4)	24.2	0.93 (0.84–1.03)	0.17
≥40% decrease in eGFR to <60 ml/min/1.73 m², ESRD, or death from renal or cardiovascular cause	370 (4.3)	10.8	480 (5.6)	14.1	0.76 (0.67–0.87)	
Death from any cause	529 (6.2)	15.1	570 (6.6)	16.4	0.93 (0.82–1.04)	
Hospitalization for heart failure	212 (2.5)	6.2	286 (3.3)	8.5	0.73 (0.61–0.88)	
Myocardial infarction	393 (4.6)	11.7	441 (5.1)	13.2	0.89 (0.77–1.01)	
Ischemic stroke	235 (2.7)	6.9	231 (2.7)	6.8	1.01 (0.84–1.21)	
Death from cardiovascular cause	245 (2.9)	7.0	249 (2.9)	7.1	0.98 (0.82–1.17)	
Death from noncardiovascular cause	211 (2.5)	6.0	238 (2.8)	6.8	0.88 (0.73–1.06)	
≥40% decrease in eGFR to <60 ml/min/1.73 m², ESRD, or death from renal cause	127 (1.5)	3.7	238 (2.8)	7.0	0.53 (0.43–0.66)	



La réduction (HR=0.83) du critère composite décès cardiovasculaire ou hospitalisation a été obtenu uniquement par une réduction des hospitalisations (HR=0.73). Aucune tendance n'est observée sur les décès (HR=0.98). Finalement le résultat sur le composite ne représente qu'un effet sur les hospitalisations et non pas un effet homogène sur les 2 composantes (cf. ci-dessous). Conclure à une réduction des décès ou des hospitalisations devient excessif contrairement à une situation où les 2 composantes auraient été réduites de la même manière.

Pour éviter cette confusion, l'article conclut de la façon suivante : « *In patients with type 2 diabetes who had or were at risk for atherosclerotic cardiovascular disease, treatment with dapagliflozin … did result in a lower rate of cardiovascular death or hospitalization for heart failure, a finding that reflects a lower rate of hospitalization for heart failure* ».

❖ *Exemple d'une situation ne posant pas de problème d'interprétation*

Dans un essai chez des patients insuffisants cardiaques [10.1056/NEJMoa1911303], le critère de jugement principal était : « The primary outcome was a composite of worsening heart failure (hospitalization or an urgent visit resulting in intravenous therapy for heart failure) or cardiovascular death. ».

Une réduction du critère composite a été obtenue identique en taille (HR=0.74) avec la réduction des hospitalisations (HR=0.70) et celle des décès (HR=0.82). La conclusion que le traitement réduit le critère décès ou hospitalisation n'est donc pas excessif. Les résultats sont en concordance avec le sens trivial de la notion de composite.

Table 2. Primary and Secondary Cardiovascular Outcomes and Adverse Events of Special Interest.*

Variable	Dapagliflozin (N=2373) values	Dapagliflozin (N=2373) events/100 patient-yr	Placebo (N=2371) values	Placebo (N=2371) events/100 patient-yr	Hazard or Rate Ratio or Difference (95% CI)	P Value
Efficacy outcomes						
Primary composite outcome — no. (%)†	386 (16.3)	11.6	502 (21.2)	15.6	0.74 (0.65 to 0.85)	<0.001
Hospitalization or an urgent visit for heart failure	237 (10.0)	7.1	326 (13.7)	10.1	0.70 (0.59 to 0.83)	NA
Hospitalization for heart failure	231 (9.7)	6.9	318 (13.4)	9.8	0.70 (0.59 to 0.83)	NA
Urgent heart-failure visit	10 (0.4)	0.3	23 (1.0)	0.7	0.43 (0.20 to 0.90)	NA
Cardiovascular death	227 (9.6)	6.5	273 (11.5)	7.9	0.82 (0.69 to 0.98)	NA

La conclusion de l'article est "Among patients with heart failure and a reduced ejection fraction, the risk of worsening heart failure or death from cardiovascular causes was lower among those who received dapagliflozin than among those who received placebo".

La démonstration d'un bénéfice sur un critère composite ne nécessite pas que les composantes soient statistiquement significatives individuellement. Cela montre bien qu'une réduction sur un critère « de morbi ou de mortalité » n'a pas valeur d'une démonstration sur la morbidité et d'une démonstration sur la mortalité. Si dans un domaine particulier, il est nécessaire d'avoir une démonstration spécifique d'un bénéfice en termes de mortalité par exemple (car d'autres produits ont déjà montré leur aptitude à réduire la mortalité), une démonstration ne portant que sur le composite ne sera pas suffisante pour apporter la preuve de l'intérêt du traitement. Il faudra, dans ce cas, une démonstration spécifique de la réduction de la mortalité. Les décès devront alors être inclus dans le plan de contrôle du risque alpha global (hiérarchisation le plus souvent).

Si le critère composite a pour objectif d'évaluer le bénéfice net (cf. section 7.4) en regroupant des évènements que le traitement prévient et des effets indésirables qu'il induit, l'homogénéité des effets n'est plus une condition d'acceptabilité du résultat, vu que par construction des effets opposés sont attendus.

Dans un essai d'un anticoagulant dans la FA [10.1056/NEJMoa1107039], un critère de bénéfice net a été utilisé regroupant les AVC et embolies systémiques que le traitement cherche à prévenir avec les hémorragies majeures. Une réduction de ce critère est obtenue montrant que les hémorragies ne contrebalancent pas le bénéfice obtenu. La balance bénéfice risque du traitement est favorable.

Outcome	Apixaban Group (N=9088)		Warfarin Group (N=9052)		Hazard Ratio (95% CI)	P Value
	Patients with Event	Event Rate	Patients with Event	Event Rate		
	no.	*%/yr*	*no.*	*%/yr*		
Net clinical outcomes						
Stroke, systemic embolism, or major bleeding	521	3.17	666	4.11	0.77 (0.69–0.86)	<0.001
Stroke, systemic embolism, major bleeding, or death from any cause	1009	6.13	1168	7.20	0.85 (0.78–0.92)	<0.001

7.4 La balance bénéfice risque

Dans de très nombreux domaines, les traitements présentent des risques avérés parfois très sérieux (les anticoagulants augmentent la fréquence des hémorragies sévères, les fibrinolytiques à la phase aigüe de l'infarctus augmentent les hémorragies intracrâniennes, des traitements anticancers induisent des décès toxiques), mais apportent aussi des bénéfices notables aux patients, en augmentant leur survie par exemple. Dans ces situations, le principe « *primum non nocere* » ne s'applique plus, mais il faut s'assurer que la balance bénéfice risque reste favorable malgré ces effets indésirables sérieux (*serious adverse event*), c'est-à-dire que les effets délétères du traitement ne contrebalanceront pas en total le bénéfice produit.

Dans l'essai Pegasus [10.1056/NEJMoa1500857], l'adjonction du ticagrelor à l'aspirine en prévention cardiovasculaire secondaire à distance de l'évènement initial a été évaluée.

Un bénéfice notable a été obtenu pour la dose de 60mg avec une réduction de la fréquence des évènements ischémiques (décès cardiovasculaire, infarctus du myocarde et AVC) de 16% en relatif.

Table 2. Efficacy End Points as 3-Year Kaplan–Meier Estimates.

End Point	Ticagrelor, 90 mg (N=7050)	Ticagrelor, 60 mg (N=7045)	Placebo (N=7067)	Ticagrelor, 90 mg vs. Placebo		Ticagrelor, 60 mg vs. Placebo	
				Hazard Ratio (95% CI)	P Value	Hazard Ratio (95% CI)	P Value
	number (percent)						
Cardiovascular death, myocardial infarction, or stroke	493 (7.85)	487 (7.77)	578 (9.04)	0.85 (0.75–0.96)	0.008	0.84 (0.74–0.95)	0.004

À l'opposé sur le critère de jugement principal de sécurité une augmentation de 132% en relatif des hémorragies majeures est observée, nominalement significative avec un p<0.001.

Table 3. Safety End Points as 3-Year Kaplan–Meier Estimates.*

End Point	Ticagrelor, 90 mg (N=6988)	Ticagrelor, 60 mg (N=6958)	Placebo (N=6996)	Ticagrelor, 90 mg vs. Placebo		Ticagrelor, 60 mg vs. Placebo	
				Hazard Ratio (95% CI)	P Value	Hazard Ratio (95% CI)	P Value
	number (percent)						
Bleeding							
TIMI major bleeding	127 (2.60)	115 (2.30)	54 (1.06)	2.69 (1.96–3.70)	<0.001	2.32 (1.68–3.21)	<0.001

Pour déterminer la balance bénéfice risque on ne peut pas comparer les réductions relatives et les augmentations relatives, car la fréquence de base des évènements ischémique est 9-fois plus élevée (9.04%) que celle des hémorragies (1.06%).

Il faut déterminer le nombre d'évènements ischémiques qui seront évités avec le ticagrelor 60mg pour le comparer avec le nombre d'hémorragies sévères induites. Par définition le critère principal de sécurité représente les évènements indésirables de même gravité clinique que les évènements que l'on a évités, ces 2 nombres pourront être comparés. Ils permettront de déduire le bénéfice clinique net (le bénéfice, net des effets indésirables).

Ce calcul se fait pour 1000 patients traités par exemple en utilisant les différences des risques. Pour le bénéfice, la différence des risques est 7.77% - 9.04% = -1.27%, ce qui signifie que le traitement de 1000 patients sur la même durée que celle de l'étude permet d'éviter 12.7 évènements supplémentaires par rapport à l'aspirine seule. Pour les hémorragies, la différence des risques est 2.30% - 1.06% = +1.24%. Ainsi, durant la même période, le traitement de ces 1000 patients aura induit 12.4 hémorragies sévères supplémentaires par rapport à l'aspirine seule. La balance bénéfice risque n'est donc pas du tout favorable avec autant d'hémorragies induites que d'évènements cardiovasculaires évités.

Ces résultats ont conduit la commission de la transparence à conclure que « En l'absence de bénéfice clinique net mis en évidence, le bénéfice modeste observé en termes d'efficacité est susceptible d'être contrebalancé en totalité par des effets indésirables de même importance clinique»[13].

[13] https://www.has-sante.fr/upload/docs/evamed/CT-15256_BRILIQUE_PIC_INS_Avis3_CT15256.pdf

8 Le cas des essais « négatifs »

Les résultats d'essais randomisés « négatifs », non concluants, car non statistiquement significatifs, ne permettent pas de conclure, évidemment, à l'intérêt du traitement et ne peuvent pas justifier un changement de pratique.

Ils ne permettent pas, non plus, de conclure à l'absence formelle d'intérêt, car un résultat non statistiquement significatif ne permet pas de conclure à l'absence d'effet (« l'absence de preuve n'est pas la preuve de l'absence »).

La formulation de la conclusion doit être du type : l'essai à échouer à mettre en évidence le bénéfice du traitement ou l'essai n'a pas permis de mettre en évidence le bénéfice. À la rigueur : aucun bénéfice n'a été montré dans cet essai ou l'étude ne montre pas de bénéfice du traitement.

Dans de rares cas, il est cependant possible de conclure à l'absence d'intérêt du traitement, quand l'intervalle de confiance est très étroit autour de l'absence d'effet (RR=0.99, IC 95% entre 0.96 et 1.02 par exemple) ou que le résultat montre une tendance non significative à un effet délétère (RR=1.9, IC 95% entre 0.98 et 2.82 par exemple).

La non-signification statistique peut provenir d'un manque de puissance qui ferait que l'essai est faussement négatif (risque beta). Cependant, même s'il est possible d'expliquer la négativité d'un résultat par un manque de puissance, cela ne permet pas néanmoins de conclure à l'intérêt du traitement. Cela peut seulement laisser penser qu'un nouvel essai, suffisamment puissant cette fois-ci, pourrait être concluant. Mais en attendant ce résultat, aucune preuve n'est disponible.

> *La lecture critique des essais négatifs n'intéresse pas le médecin, car ces résultats ne l'amènent pas à se poser la question de s'il doit utiliser ou non le nouveau traitement. Pour ce type de résultats, le raisonnement de lecture critique est complètement inversé par rapport aux résultats positifs. La question qui se pose avec les résultats négatifs est celle d'un résultat faussement négatif, pour savoir s'il faut faire un autre essai ou abandonner définitivement la molécule. Les biais recherchés ne sont donc plus du tout les mêmes. Ce ne sont plus les circonstances qui peuvent faire apparaitre une différence à tort qui entrainent un biais, mais bien celles qui auraient pu faire disparaitre la différence existant entre les 2 produits. De même, au niveau statistique ce n'est plus une question d'erreur alpha, mais d'erreur beta.*

9 Conclusion pratique de la lecture critique

À l'issue de l'analyse critique de l'essai, il convient de décider ou pas :

1. Si l'essai démontre avec un haut degré de certitude[14] que le traitement apporte un bénéfice,
2. Et si ce résultat démontré constitue bien un réel progrès thérapeutique, apportant une réelle plus-value médicale par rapport à l'existant

Si ces deux critères sont vérifiés, le nouveau traitement peut être considéré pour changer la pratique et s'inscrire dans la stratégie thérapeutique de la situation pathologique considérée.

[14] Le terme « degré de certitude » remplace actuellement la notion de niveau de preuve. Le niveau de preuve était la capacité intrinsèque d'un type d'étude à apporter des preuves solides et ne prenait pas en compte les résultats obtenus. Le degré de certitude (utilisé dans les recommandations GRADE) connote directement la fiabilité de chaque résultat produit par un essai et permet de décider s'il est suffisamment sûr pour justifier un changement de pratique.

L'essai Aristotle [10.1056/NEJMoa1107039] a comparé l'apixaban à la warfarine dans la FA. L'essai était randomisé, en double aveugle, avec 35 et 34 patients perdus de vue pour une différence sur le critère de jugement principal de 53 évènements. Une hiérarchisation avait été utilisée pour contrôler de risque alpha global sur 3 critères de jugement en supériorité : les AVC et embolies systémiques, les saignements majeurs et les décès de toutes causes. Un résultat statistiquement significatif a été obtenu sur ces 3 critères.

La conclusion est « In patients with atrial fibrillation, apixaban was superior to warfarin in preventing stroke or systemic embolism, caused less bleeding, and resulted in lower mortality".

Ces résultats ont permis d'introduire l'apixaban dans la stratégie thérapeutique de la FA.

Si ce n'est pas le cas, le traitement n'a pas (encore) fait ses preuves pour une utilisation en pratique et ne peut qu'être utilisé dans le cadre d'autres essais thérapeutiques. Éventuellement, si un bénéfice un quel que soit peut cliniquement pertinent est démontré, le traitement peut représenter une alternative dans la stratégie thérapeutique.

« Les données actuelles sur l'utilisation de l'ivermectine pour traiter les patients atteints de COVID-19 ne sont pas probantes. En attendant que davantage de données soient disponibles, l'OMS recommande de n'administrer ce médicament que dans le cadre d'essais cliniques. » https://www.who.int/fr/news-room/feature-stories/detail/who-advises-that-ivermectin-only-be-used-to-treat-covid-19-within-clinical-trials

Attention : un résultat insuffisamment cliniquement pertinent, mais parfaitement bien démontré peut conduire à une AMM. Le but de l'AMM est de statuer si le médicament est actif avec une balance bénéfice risque acceptable et non pas de juger si le traitement a sa place dans la stratégie thérapeutique. Cette dernière évaluation est, en France, du ressort de la commission de transparence et des recommandations de pratique.

10 Références

1 Lundh A, Barbateskovic M, Hróbjartsson A, et al. Conflicts of interest at medical journals: the influence of industry-supported randomised trials on journal impact factors and revenue - cohort study. *PLOS Medicine* 2010;7(10):e1000354.

2 Hwang TJ, Carpenter D, Lauffenburger JC, et al. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Intern Med* 2016;176(12):1826–33. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2565686.

3 Gerstein HC, McMurray J, Holman RR. Real-world studies no substitute for RCTs in establishing efficacy. *The Lancet* 2019;393(10168):210–11.