Société Française de Pharmacologie et de Thérapeutique

Groupe de Travail Méthodologie

Livre blanc SFPT

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

14 février 2022

Préface

La rédaction d'un livre blanc et sa diffusion constituent des actions de communication qui confinent parfois au marketing. Ici, rien de cela. Rédigé par un groupe d'expert dans une période où les éléments les plus simples et les plus évidents de la méthodologie des essais cliniques sont parfois remis en cause ou à tout le moins oubliés, ce livre blanc a été construit comme un guide pratique pour le lecteur préoccupé de la juste évaluation des médicaments. Dans ce contexte, ce document apporte une vue d'ensemble sur les éléments qui justifient les conditions de réalisation des essais cliniques contribuant ainsi à affiner le regard critique sur les conditions de mise à disposition de nouveaux médicaments. Bien qu'assez technique sur plusieurs aspects, ce document reste néanmoins accessible à un large lectorat, les plus aguerris ou les plus curieux trouveront dans les documents compagnons des éléments supplémentaires pour les aider dans la compréhension de tel ou tel point particulier. Enfin, ce livre blanc a aussi pour but de désacraliser la méthodologie des essais cliniques qui ne doit pas rester l'apanage de quelques initiés mais bien devenir ou redevenir un outil largement diffusé, dont les préceptes demeurent indispensables à l'évaluation des médicaments et des thérapeutiques non médicamenteuses. Pour faciliter la diffusion de ce livre blanc, nous avons fait le choix de le rendre disponible à tous sans restriction. Il me reste alors à vous en souhaiter une bonne lecture et à vous demander de bien vouloir le partager avec toute personne préoccupée de près ou de loin par l'évaluation clinique des médicaments.

Dominique DEPLANQUE

Président de la SFPT

Table des matières

Avant-propos

Ce document est principalement destiné aux utilisateurs des résultats de la recherche clinique sur l'efficacité et la sécurité des médicaments. Il présente et justifie les exigences méthodologiques nécessaires pour décider d'utiliser, de recommander ou d'enregistrer un nouveau médicament ou une molécule déjà utilisée dans une autre indication.

Il se veut être une aide destinée à tous ceux qui sont confrontés dans leur exercice professionnel aux résultats de la recherche clinique sur le médicament : cliniciens prescripteurs, pharmaciens, autres professionnels de santé, professionnels travaillant dans les agences d'enregistrement ou régulation, décideurs de santé publique, médico-économistes, professionnels participant à l'élaboration de recommandations de pratiques (agences, sociétés savantes), enseignants, patients, etc.

En étant continuellement actualisé pour suivre au plus près les évolutions méthodologiques rencontrées dans les dernières publications d'essais cliniques, ce document devrait aussi être utile à ceux qui souhaitent parfaire leur expertise méthodologique pour participer, par exemple, à des groupes collaboratifs internationaux (conception d'essais multicentriques ou élaboration de recommandations).

Ce n'est pas un manuel de méthodologie destiné aux professionnels de la recherche. Il est destiné à remplir simplement un maillon manquant entre les ouvrages techniques de méthodologie pour les « producteurs de résultats » qui conçoivent et réalisent les études ou les analyses statistiques et les « utilisateurs des résultats ».

Ce n'est pas directement un ouvrage sur la lecture critique des essais thérapeutique mais il fournit néanmoins les éléments de compréhension nécessaires pour évaluer une publication en lecture critique. Les concepts abordés dans ce document devraient permettre au lecteur d'appréhender correctement la finalité de la lecture critique et d'acquérir l'expertise nécessaire pour la réalisation de cette évaluation.

Chapitre 1. Pourquoi ce livre blanc ?

L'élaboration de ce livre blanc a été motivée par le constat d'une tendance récente vers l'abandon des fondamentaux de la recherche de la preuve de l'intérêt clinique des nouveaux médicaments avant de les utiliser en pratique médicale quotidienne [1, 2, 3, 4, 5, 6, 7, 8, 9].

Même si les principes méthodologiques sont généralement connus, la récente crise de la Covid a révélé une certaine propension à facilement les abandonner [7, 9, 10, 11, 12, 13], révélatrice du fait que ces principes sont plus acceptés par convention que par conviction de leur intérêt fondamental. Sous la pression induite par la gravité de la situation sont apparues de nombreuses revendications d'usage compassionnel de traitements, rejetant la nécessité d'une évaluation a priori de ces candidats médicaments. La majorité des études entreprises [14, 15, 16] ne respectaient pas les principes méthodologiques[1], revenant à une approche intuitive et naïve de l'appréciation de l'intérêt clinique d'un traitement.

En effet, pour la totalité des molécules qui ont finalement échoué à montrer un quelconque intérêt dans des essais bien conduits il existe de nombreuses études faisant miroiter des résultats positifs [1]. Cette production de masse de mauvaises études a induit une cacophonie tapageuse tant au niveau du débat professionnel qu'auprès du grand public, conduisant à donner une impression d'amateurisme et de décisions arbitraires dans l'évaluation des traitements. La production d'études aux résultats discordants a fait que chacun pouvait brandir une étude pour soutenir sa position, fût-elle scientifiquement infondée.

Il serait trompeur de penser que cette situation était uniquement circonstancielle, liée au stress induit par cette crise sanitaire majeure. Elle s'est développée dans un contexte d'évolution des méthodes d'évaluation du médicament ces dernières années, au détriment de l'obtention d'un niveau de preuve satisfaisant. En témoigne, par exemple, l'augmentation des enregistrements accélérés (conditionnels) en oncologie, des revendications d'intérêt clinique de traitements reposant de plus en plus sur des études non comparatives [17] et/ou des critères intermédiaires sans réelle valeur de critère de substitution (« *surrogate* ») [18], voire même sur des résultats non concluants. Le récent enregistrement par la FDA de l'aducanumab dans la maladie d'Alzheimer sur la base d'un unique résultat post hoc d'une étude négative arrêtée initialement pour futilité, et contre l'avis de son comité consultatif [19, 20], est un exemple de cette rupture de paradigme appelée de leurs vœux par certains [3, 21, 22].

> *L'aducanumab est un médicament anti-plaque amyloïde développé dans la maladie d'Alzheimer [19, 20] dans un contexte où tous les autres candidats visant la même cible ont échoué. Le plan de développement de l'aducanumab comprenait 2 essais de phase 3. Le premier a été non concluant et le second a été arrêté prématurément pour futilité[2]. Mais, alors que la molécule avait été enterrée comme les autres de la classe, le sponsor a revendiqué ultérieurement que l'essai était en fait positif en récupérant de nouvelles données obtenues après l'arrêt de l'étude. Avec ces données qui n'ont aucune valeur méthodologique de confirmation, l'industriel a pu obtenir l'enregistrement par la FDA alors que le comité consultatif avait émis*

[1] Expérience empirique individuelle du médecin, absence de groupe contrôle ou de raisonnement contrefactuel, comparaisons naïves de groupes traités et non traités, etc.
[2] Impossibilité quasi certaine de montrer un bénéfice

un avis négatif, entrainant la démission de 3 membres de ce comité[3]. Le coût annuel est de 56 k$[4]. Plusieurs institutions et payeurs ont clairement exclu ce traitement du cadre de leurs prestations[5]. Une demande d'enregistrement a été faite à l'EMA qui a été naturellement rejetée il y a quelques jours [6].

La méthodologie actuelle d'évaluation des nouveaux traitements par des essais randomisés de confirmation n'est pas basée sur des principes gratuits, arbitraires. Ces principes sont nécessaires pour garantir la fiabilité des résultats et permettent ainsi de répondre à la l'exigence d'avoir des preuves « au-delà de tout doute raisonnable » pour enregistrer, utiliser, recommander et rembourser les nouveaux traitements. La mise en œuvre de ces principes est tout à fait possible et ces essais sont réalisés en grand nombre (plus de 20 000 publications d'essais randomisés dans la base Pubmed en 2020). Cette approche a contribué à la découverte et mise en place de très nombreux progrès thérapeutiques majeurs avec des essais marquants qui ont radicalement transformé la pratique médicale (les IEC dans l'insuffisance cardiaque avec l'essai CONSENSUS [23], les immunothérapies dans le mélanome [24], etc.). Ils révèlent aussi (avec apparemment la même fréquence [25]) que de nombreux traitements prometteurs, du fait de leur mécanisme d'action et des résultats de leurs études cliniques précoces, échouent à montrer qu'ils apportent un réel bénéfice clinique pour le patient.

La justification des principes méthodologiques est rarement expliquée et les acteurs les acceptent sans véritablement se les approprier [26]. Un des buts de ce livre banc est de revenir sur la justification de ces principes et de montrer qu'ils ont été inventés pour solutionner des problématiques scientifiques pratiques [27] et qu'ils ne sont pas de simples conventions d'usage. Faire autrement ne permet pas d'apporter le degré de certitude nécessaire à l'adoption d'un nouveau traitement compte tenu des enjeux scientifiques, éthiques, déontologiques, sociétaux et de santé publique sous-jacents. Cela conduirait à accepter par principe l'utilisation de médicaments sans un potentiel intérêt clinique[7].

La SFPT est une société savante embrassant tous les aspects relatifs au médicament et en particulier son bon usage. Elle regroupe ainsi en son sein une forte expertise méthodologique. En rédigeant ce livre blanc, la SFPT a voulu partager cette expertise afin d'aider tous les médecins, professionnels de santé et autres acteurs du médicament à mieux comprendre l'importance de la méthodologie et à mieux appréhender et exploiter les résultats des essais thérapeutiques modernes.

À travers ce livre blanc, la SFPT souhaite réaffirmer que :

- La recherche des preuves est fondamentale pour proposer aux patients des traitements utiles médicalement et qui répondent à leur attente (comme cela est mentionné dans le code de déontologie).
- Les principes méthodologiques, même s'ils paraissent par moment contraignants et obscurs, sont les garants de l'obtention du degré de certitude souhaitable (cf. section 2.2).
- La méthodologie est l'élément central de l'intégrité scientifique concernant la fiabilité de l'information scientifique [28].
- Il est utile que tous les acteurs du médicament aient une connaissance de base et opérationnelle de ces principes et qu'ils puissent se faire par eux même leur opinion quant à l'opportunité d'adopter un nouveau traitement (cf. section 3.3 et 4.5).

[3] https://www.nytimes.com/2021/07/07/health/aducanumab-aduhelm-alzheimers-elderly.html
[4] https://www.statnews.com/pharmalot/2021/06/07/biogen-fda-alzheimers-medicare-cms/
[5] https://www.formularywatch.com/view/some-blues-plans-won-t-cover-new-alzheimer-s-therapy-aduhelm
[6] https://www.reuters.com/business/healthcare-pharmaceuticals/biogens-alzheimers-drug-gets-negative-vote-ema-panel-2021-11-17/
[7] Ne pas attendre de tout traitement qu'il apporte la preuve formelle de son intérêt revient à considérer que l'utilisation de traitement sans intérêt n'est finalement pas un problème.

- Le mécanisme d'action n'est pas un élément suffisant pour justifier qu'une molécule puisse-t-être utilisée dans la pratique clinique comme médicament. La connaissance des mécanismes d'action est indispensable pour la recherche de nouveaux médicaments mais est insuffisante pour justifier leur utilisation. Seule la démonstration d'un bénéfice clinique justifie son utilisation en pratique.

Ce livre blanc a aussi pour objectif d'aider les médecins et les autres acteurs du médicament à interpréter correctement les résultats des essais, à identifier les limites méthodologiques des études et à distinguer correctement les résultats probants des résultats insuffisamment établis pour induire un changement de pratique. En effet, il devient indispensable pour tous les acteurs du médicament de pouvoir évaluer par eux-mêmes la fiabilité et la solidité des arguments mis en avant pour justifier l'utilisation d'un nouveau traitement (cf. section 4.5). Sans expertise suffisante, le lecteur risque de se méprendre sur le réel intérêt d'un nouveau traitement en tombant dans un des nombreux pièges qui lui sont tendus dans la littérature et la communication des résultats des essais cliniques : spins de conclusions [29, 30, 31] (cf. section 3.3), production de masse d'études de mauvaise qualité méthodologique [32] (cf. section 3.3), communication promotionnelle excessive [33, 34, 35] (section4.4), manque de pertinence des traitements des groupes contrôles [36, 37, 38] ou des critères de jugement non cliniquement pertinents [39], etc.

Cet ouvrage se veut aussi être un manuel de méthodologie de l'évaluation des médicaments destiné à tout « consommateur » (et non pas producteur) de résultats d'études d'évaluation des médicaments. Il introduit les grandes problématiques auxquelles est confrontée l'évaluation de l'intérêt clinique des médicaments et liste les éléments de méthode qui permettent de les résoudre. Il s'y ajoute des documents compagnons qui présentent de façon plus approfondie la justification et l'explicitation des principes méthodologiques nécessaires à l'évaluation des traitements sous la forme de dossiers thématiques.

Son but est d'expliciter toutes les connaissances nécessaires à ceux qui souhaiteraient pouvoir s'autodéterminer 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. Même si la décision finale concernant le nouveau traitement viendra des agences de régulation et des recommandations des sociétés savantes, il est utile que chaque acteur du médicament puisse comprendre, anticiper et prendre part à ces décisions afin d'éviter que son activité se résume à appliquer des recommandations sans en maitriser la justification. La crise de la COVID-19 a bien mis en évidence que tout ne pouvait pas reposer sur la régulation et les recommandations car, dans ce domaine comme dans d'autres, les prescripteurs ont souvent reçu les informations et les résultats bruts avant que ces derniers ne soient intégrés dans les recommandations.

Chapitre 2. Principes fondamentaux

La distinction entre un argument solide en faveur de l'intérêt clinique d'un médicament et ce qui ne peut pas l'être s'effectue suivant le principe, jusque-là couramment admis, de l'impérative nécessité d'obtenir des preuves « au-delà de tout doute raisonnable[8] ». La méthodologie a ainsi été établie en identifiant les conditions nécessaires à l'obtention de résultats robustes « au-delà de tout doute raisonnable » avant de recommander l'utilisation d'un nouveau médicament (ou d'une nouvelle indication).

> *La nécessité d'obtenir des preuves formelles de l'intérêt clinique des nouveaux médicaments est parfois remise en cause, en argumentant que cette recherche est trop exigeante et trop lourde à effectuer, rendant impossible l'évaluation de certains traitements qui pourraient cependant remplir une place vacante dans la stratégie thérapeutique[9] (cf. section 4.2). Cette approche revendique que des éléments de présomptions d'efficacité, comme le mécanisme d'action ou des avis d'expert, sont suffisants pour adopter un nouveau traitement afin de ne pas en priver les patients. Il faut cependant rappeler que la régulation des médicaments a été mise en place dans les années 60[10] sur le constat qu'une telle approche conduisait à la promotion et à l'utilisation de traitements sans utilité en grand nombre [40, 41]. De nombreux exemples ont également montré qu'une telle démarche pouvait s'avérer dangereuse.*

Souvent la méthodologie est réduite au périmètre du design de l'étude (essais contrôlés randomisés en double aveugle) mais les problématiques à régler sont bien plus larges. Certaines sont des éléments de la méthode scientifique de base, communes à tous les domaines des sciences. D'autres sont de l'ordre des pratiques de recherche, de publications ou d'exploitations promotionnelles des résultats et sont plus spécifiques au champ de l'évaluation des médicaments (bien qu'aussi rencontrées dans d'autres domaines).

[8] Ce principe qui régit entre autres les exigences règlementaires d'enregistrement des médicaments a été emprunté au droit britannique (« *beyond a reasonable doubt* ») qui définit de cette manière le degré de certitude nécessaire pour condamner.

[9] À ce niveau on confond souvent la nécessité d'avoir un traitement (pour une condition clinique qui en est dépourvue par exemple) avec l'utilité de celui-ci (le candidat traitement est utile quand il est confirmé qu'il apporte une réelle solution au besoin non couvert). Une nouvelle proposition de traitement est souvent présentée comme « utile » parce que ce traitement aurait tout de suite une utilisation en raison d'un besoin non couvert. Mais solutionne-t-il effectivement le problème correspondant ? Il y a ainsi confusion du besoin avec l'utilité démontrée.

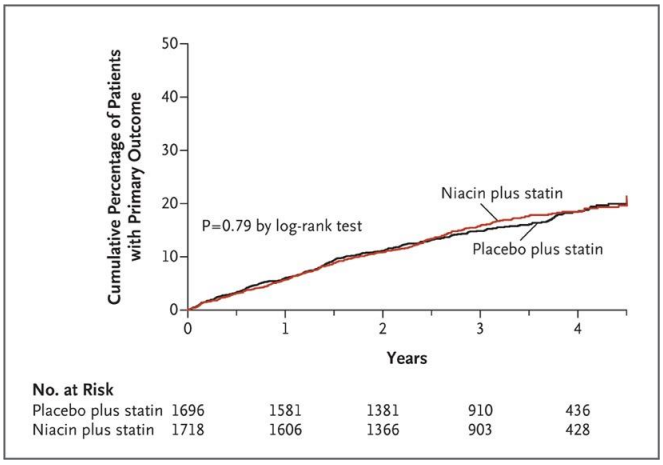
[10] 1962 avec le Kefauver-Harris Amendment to the FD&C Act qui impose que les nouveaux médicaments apportent des preuves substantielles ("substantial evidences") de leur efficacité.

La méthodologie, au sens large, a été construite à partir du constat qu'il est presque toujours possible de produire à partir des études de traitements sans efficacité des résultats apparemment en faveur de leur intérêt. Ainsi, ont été déterminées les conditions de robustesse que doit avoir un résultat pour qu'il puisse conduire à utiliser un nouveau médicament afin d'éviter d'utiliser à tort des traitements en pratique clinique

2.1 Pourquoi ne peut-on pas se contenter d'un mécanisme d'action bien établi ?

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

Les mécanismes d'action ne prédisent pas avec certitude 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 régulièrement parcellaires.

> Avec les statines, la baisse de LDL cholestérol s'est traduite par une réduction de la fréquence des évènements cardiovasculaires bien démontrée dans de nombreux essais.
> L'acide nicotinique (niacine) est un produit qui entraine une baisse du LDL cholestérol. Cependant l'essai HPS2-THRIVE, 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 [42].
>
> 
>
>
> 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 phases 3 (dont le but est de démontrer le bénéfice) sont fréquents, aux alentours de 50% [25]. Ce résultat montre qu'il n'est donc 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 (cf. section 4.2). De plus, il n'est pas possible d'exclure un effet délétère.

> Au début des années 1980, il avait été pris comme habitude de prescrire 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 réduisaient drastiquement les ESV. Cette pratique a été établie sans preuve de la prévention réelle de la mort subite. Ce n'est qu'au bout de 10 ans qu'un essai de mortalité a été entrepris (alors qu'il aurait dû l'être avant l'établissement de cette pratique thérapeutique). Au lieu de confirmer la prévention des morts subites, cet essai [43] a montré une multiplication par 3 de la mortalité chez les patients traités.

Une des raisons pouvant expliquer pourquoi des effets pharmacologiques prometteurs ne débouchent pas toujours sur un bénéfice clinique réside aussi 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 aussi fréquemment de lacunes méthodologiques majeures. De plus, un biais de publication ou de « *selective reporting* » des résultats peut distordre complètement la perception de la réalité de l'effet du traitement en raison de la publication sélective et exclusive des résultats faussement positifs [44, 45] (cf. section 3.2). Les raisonnements qui extrapolent le bénéfice à partir de l'effet sont aussi parfois trop simplistes ou optimistes et négligent les problématiques de pharmacocinétique (est-il possible d'atteindre la concentration efficace au niveau du tissu cible ?), d'observance ou de toxicité.

> Dès le début de l'année 2020, l'hydroxychloroquine a été proposée comme traitement dans la COVID-19. L'idée de son potentiel intérêt provenait d'une étude *in vitro* montrant une activité de cette molécule sur le SARS-CoV2 dans un modèle cellulaire Vero [46]. 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 [47]. 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 dans les conditions de la première étude avec les doses maximales habituellement tolérées. Toutefois, des études cliniques non randomisées, présentant de nombreuses limites méthodologiques, ont été menées, et ont conduit à un usage massif de l'hydroxychloroquine à travers le monde, motivée par la gravité de la maladie et l'absence de thérapeutiques curatrices dans un contexte d'anxiété de la population et de relais médiatiques soutenus. Cependant, les essais randomisés, comme l'étude RECOVERY [48], n'ont pas permis de mettre en évidence de bénéfice et leur méta-analyse montre même une augmentation de la mortalité [49]. Cet exemple montre que, même dans un contexte d'urgence grave, il n'est pas possible d'envisager un usage compassionnel à large échelle sans preuve, compte tenu des limites des données préliminaires.

Ainsi, ce qui fait qu'un candidat médicament peut être reconnu comme traitement utilisable en pratique clinique, ce n'est par son mécanisme d'action ou son effet pharmacologique, mais bien la démonstration dans un essai clinique fiable, que cet effet se traduise en un bénéfice clinique avéré. L'essai clinique est donc l'élément fondamental qui établit le statut d'un médicament, mais il est parfois perçu comme une simple démarche administrative pour obtenir une AMM (Autorisation de Mise sur le Marché), un élément accessoire et indépendant du raisonnement médical qui conduit à utiliser le médicament. Cette perception est aussi renforcée par la façon dont le médicament est enseigné durant les études médicales ou pharmaceutiques.

L'intérêt du mécanisme d'action se situe en amont de l'utilisation du traitement, au niveau de la recherche de nouvelles pistes thérapeutiques. La pharmacologie du mécanisme d'action permet d'élaborer les hypothèses pour proposer, de manière rationnelle, des candidats médicaments. Ces hypothèses seront ensuite évaluées par la pharmacologie clinique dans l'essai randomisé.

2.2 Pourquoi un résultat d'étude « positif » ne reflète-t-il pas toujours une efficacité réelle ? Les résultats faussement positifs.

Un essai clinique contrôlé peut produire 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.

Les conséquences de ces résultats faussement positifs sont très dommageables car ils conduisent à l'enregistrement et à l'utilisation de traitements sans intérêt (ou qui ne sont pas supérieurs au traitement standard en cas de comparaison à ce dernier). S'assurer de sa fiabilité est un aspect essentiel de l'intégrité scientifique.

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 survenues. C'est le but de la méthodologie et de l'analyse statistique :

- La méthodologie a pour but d'éviter les biais, c'est-à-dire qu'il existe un vice de construction ou de réalisation qui fait que l'étude produirait toujours un résultat en faveur du nouveau traitement même si celui-ci n'apporte aucun bénéfice (cf. section 2.2.1).
- 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) (cf. section 2.2.2).

> Le vintafolide a été développé dans le cancer de l'ovaire. Une première étude [50] randomisée, multicentrique, en ouvert, comparant vintafolide plus doxorubicine à la doxorubicine seule s'avère « positive » en montrant une amélioration de la PFS (Progression-Free Survival, survie sans 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, mais 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, s'avérant être sans utilité dans cette situation. Rétrospectivement il se trouve donc que le résultat de la première étude était faussement positif, faux positif qu'il est notamment 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.
> Il apparait dans cet exemple que l'étude de phase 2 avait donné un résultat clairement « positif » avec un traitement qui n'apporte pas de bénéfice en réalité ; illustrant ainsi qu'une étude peut donner un résultat **faussement positif**.

Cette observation n'est pas unique. Dans une étude méta-épidémiologique [25] il apparait qu'environ la moitié des produits échouent en phase 3 (pour manque d'efficacité ou pour effet indésirable). Or, avant d'entrer en phase 3, tous ces produits avaient naturellement obtenu des résultats positifs en phase 2, ce qui souligne le caractère indispensable de l'essai confirmatoire de phase 3 dans le développement du médicament (cf. section 4.3).

Ce point est fondamental et encore insuffisamment perçu. En effet, rien ne garantit, par principe, que les résultats des études soient le reflet exact de la réalité. **Les études peuvent produire des résultats**

faux. Il convient donc d'être d'abord suspicieux devant un résultat positif en faveur de l'intérêt du traitement et de ne l'accepter qu'après s'être assuré de sa fiabilité. C'est fondamentalement la finalité de la lecture critique : détecter les résultats faussement positifs pour éviter de prendre des décisions d'utilisation de nouveaux traitements inappropriés.

En raison des biais et des erreurs aléatoires, un essai peut produire des résultats divergents par rapport au réel effet du traitement

Mais comment détecter ces résultats faussement positifs ? « Ce n'est pas écrit dessus ! ». Rien ne distingue, a priori, un résultat faux positif d'un résultat positif à raison. Comment séparer alors le bon grain de l'ivraie ?

2.2.1 Risque de biais et faux positifs : l'apport de la méthodologie

Historiquement, cette possibilité de faux positif a été identifiée très tôt dans l'évaluation des traitements. Deux phénomènes peuvent contribuer à rendre positifs les résultats d'une étude alors que le traitement n'a pas d'efficacité : les biais et les erreurs aléatoires (erreur et risque alpha). Comme il est impossible de distinguer les faux positifs des vrais positifs, des moyens ont été inventés pour éviter les biais et limiter les erreurs aléatoires. Ces moyens sont les principes méthodologiques et le contrôle du risque alpha par l'inférence statistique (les tests statistiques).

Le but de la méthodologie est d'éviter qu'une étude puisse donner un résultat faux

Si ces principes sont mis en œuvre correctement. L'étude sera à l'abri des biais affectant les résultats[11]. Si l'étude obtient un résultat positif celui-ci est donc le reflet de la réalité, sous réserve que l'erreur aléatoire ait aussi été contrôlée.

Pour écarter la possibilité qu'un résultat positif, en faveur de l'effet du traitement, soit faussement positif, il convient d'analyser la méthodologie et de valider qu'elle était bien en mesure d'éviter les biais. Si c'est le cas, le résultat est fiable, ne reste plus alors que la question du risque d'erreur statistique. C'est le but de la lecture critique d'éviter d'accepter un résultat faux.

Dans le cas du vintafolide, l'étude de phase 2, bien que contrôlée et randomisée, n'était pas en double insu, contrairement à la phase 3. Elle n'était donc pas à l'abri d'un biais de mesure ou de réalisation. Comme le critère PFS est une mesure subjective, un biais de mesure peut tout à fait expliquer la « positivité » observée. Des données montrent que lorsque, dans le même essai, le critère PFS est mesuré par les investigateurs eux-mêmes ou par un comité de validation centralisé et indépendant, des différences jusqu'à 10% peuvent apparaitre entre les 2 mesures, comme dans le groupe placebo de l'exemple suivant (N Engl J Med 2011;364:514-23, 10.1056/NEJMoa1009290) :

[11] Il s'y ajoute les biais survenant dans l'environnement du résultat : biais de publication, biais cognitifs, etc…

Table 2. Progression-free Survival.

Variable	Everolimus (N = 207)	Placebo (N = 203)
Assessment by local investigator		
Progression-free survival events — no. (%)*	109 (53)	165 (81)
Censored data — no. (%)	98 (47)	38 (19)
Median progression-free survival — mo	11.0	4.6
Review by central adjudication committee		
Progression-free survival events — no. (%)*	95 (46)	142 (70)
Censored data — no. (%)	112 (54)	61 (30)
Median progression-free survival — mo	11.4	5.4

Cette incertitude découle des difficultés d'interprétation des examens équivoques. Une incertitude de l'ordre de 10% sur la mesure du critère PFS est bien supérieure à la plupart des effets traitements observés et peut donc expliquer la totalité de la différence obtenue dans un essai en ouvert, si cette incertitude est influencée par la connaissance du traitement reçu par les patients. Il est facile d'imaginer qu'un investigateur aura une certaine réticence à déclarer comme ayant progressé un patient qui reçoit le nouveau traitement prometteur, car cela oblige à l'arrêt du traitement. Pour un patient du groupe contrôle cette décision à moins de conséquences et est donc plus simple à prendre. De plus, elle donne peut-être la possibilité de passer à un traitement de ligne n+1 dont l'efficacité est connue. Un essai en double insu garantit que l'incertitude de mesure sera identique dans les 2 groupes (il y aura toujours une erreur de mesure, mais qui ne sera pas systématiquement différente entre les 2 groupes et ne pourra donc pas causer de différence entre les 2 groupes).

2.2.2 Le risque alpha et son contrôle

La première cause des résultats faussement positifs est le risque alpha[12], risque qu'une différence apparaisse entre les 2 groupes comparés uniquement du fait du hasard. Ce risque alpha provient des fluctuations aléatoires d'échantillonnage et ne peut pas être totalement évité.

Il n'est pas raisonnable de conclure à l'effet du traitement simplement en regardant s'il existe une différence en faveur du traitement étudié entre les 2 groupes au niveau du critère de jugement. Pour éviter de conclure à tort et à travers devant toutes les différences en réalité dues au hasard, il est nécessaire de s'appuyer sur les tests statistiques d'hypothèses. Ainsi, il ne sera possible de conclure à la réalité statistique d'une différence, et donc à l'effet du traitement, que si cette différence est statistiquement significative. À ce moment-là, le risque encouru de conclure à tort est considéré comme suffisamment petit pour être acceptable (moins de 2.5% pour un test classique bilatéral à 5%). La significativité statistique est obtenue quand le paramètre p (la *p value*), calculé à partir des résultats de l'étude, est inférieur au seuil de la significativité statistique dont la valeur correspond au risque alpha consenti (en général 0.05 bilatéral).

[12] En toute rigueur, il s'agirait plutôt de l'erreur statistique alpha, le risque alpha étant la probabilité de survenue d'une erreur alpha. Dans le langage courant, les 2 termes (risque et erreur) sont utilisés sans trop de distinction, car une erreur alpha est un « risque » qui menace les conclusions (le même terme désigne le risque, l'erreur alpha, et la quantification de ce risque).

La significativité statistique est donc classiquement présentée comme p<0.05 dans les ouvrages et les cours de statistique générale. Mais cette présentation ne se préoccupe que du risque de conclure à tort au niveau d'un test statistique donné et uniquement à son niveau.

Dans l'essai thérapeutique, un résultat est statistiquement significatif lorsqu'il permet de conclure au niveau de l'essai à l'intérêt du traitement avec un risque alpha global contrôlé et suffisamment faible (<2.5%)

Dans l'essai thérapeutique, la situation est plus complexe, car l'objectif général est de conclure à l'intérêt ou non du traitement évalué, globalement, à l'issu de l'interprétation de tous les résultats de l'essai. Cette conclusion générale peut donc se faire à partir de multiples comparaisons (multiples critères de jugement par exemple, ou multiples temps de mesure, multiples sous-groupes, multiples analyses intermédiaires, etc.). Cette **multiplicité** fait que les risques de conclure à tort, que l'on consent à prendre au niveau de chacune de ces multiples comparaisons (chaque comparaison donnant lieu à un test), vont s'accumuler, conduisant à un risque général (global) de conclure à tort à un quelconque intérêt du traitement fortement augmenté. Cette **inflation du risque alpha global de l'essai** liée à la multiplicité des comparaisons augmente donc les risques de tirer à tort une conclusion positive de l'essai (cf. dossier n° 1 du document compagnon).

Si cette inflation n'était pas prise en compte, il serait presque toujours possible de trouver une comparaison avec un p<0.05 parmi les multiples *p values* disponibles (car avec un traitement sans intérêt, 5%, soit 1/20° des comparaisons[13] donne un résultat significatif au niveau du test). Presque tous les essais seraient positifs même en l'absence d'efficacité réelle. L'essai thérapeutique n'aurait donc plus aucun sens.

Pour éviter cette situation, la signification statistique dans l'essai thérapeutique est envisagée au niveau global, en termes de **risque alpha global** de conclure à tort à un (quelconque) intérêt du traitement à l'issu de l'essai, et non plus au niveau d'un critère de jugement particulier (niveau du **risque alpha nominal** de conclure à tort à un effet du traitement sur ce critère particulier).

Il existe plusieurs risques alpha.
Celui qui est pertinent dans l'essai thérapeutique est le risque alpha global, risque de conclure à tort à l'intérêt du traitement, globalement, à l'issue de l'essai

Des techniques de gestion de la multiplicité sont alors utilisées et font que la *p value* obtenue au niveau d'un test particulier (*p value* nominale) ne donne plus la signification statistique recherchée. La signification statistique en termes de risque alpha global sera déduite de cette *p value* nominale en fonction de la méthode de gestion de la multiplicité qui aura été fixée dans le protocole : répartition du risque alpha entre des co-critères principaux (*co-primary endpoints*), hiérarchisation des comparaisons qui seront effectuées selon une séquence prédéfinie, combinaison des deux méthodes avec ou sans réallocation (« recyclage ») du risque alpha. Ces méthodes sont détaillées dans le dossier compagnon n° 1.

[13] Des comparaisons indépendantes entre elles

Ces méthodes font souvent que la *p value* à atteindre pour conclure à la significativité (seuil ajusté de la signification) est différente pour chaque critère de jugement et pour les analyses intermédiaires par exemple. Ces seuils ajustés sont très souvent inférieurs à 0.05 en bilatéral et fréquemment des valeurs de p inférieurs à 0.05, mais supérieurs au seuil ajusté, ne sont donc pas statistiquement significatifs en termes de risque alpha global.

p<0.05 n'est plus synonyme de signification statistique avec les méthodes mises en œuvre dans les essais modernes pour contrôler le risque alpha global

Seuls les résultats statistiquement significatifs en termes de contrôle du risque alpha global peuvent être considéré comme une démonstration du bénéfice du traitement. Les résultats qui sont seulement nominalement significatifs ne permettent pas de conclure. Les *p values* correspondantes sont d'ailleurs de moins en moins rapportées dans les publications (pour éviter une mauvaise interprétation).

Seuls les résultats significatifs en termes de contrôle du risque alpha global permettent de conclure à l'intérêt du traitement évalué

2.3 Raisonnement inductif ou raisonnement déductif, *« HARKing »*

Schématiquement deux types de raisonnement peuvent être distingués. Le raisonnement inductif et le raisonnement déductif. Ils n'ont cependant pas la même aptitude à produire des preuves (cf. dossier compagnon n° 4).

Le raisonnement inductif consiste à édicter une règle générale à partir de l'observation. Par exemple, conclure que l'aspirine à un effet hémorragipare à partir de l'observation d'accidents hémorragiques sous aspirine, ou conclure à un surcroit inattendu de cancer avec un sartan à partir d'une fréquence plus élevée d'occurrence de cancer dans le groupe traité par rapport au groupe contrôle dans un essai thérapeutique d'insuffisance cardiaque. La limite des résultats exploratoires provient du fait qu'ils sont issus par essence d'un raisonnement inductif.

Le raisonnement inductif produit des résultats exploratoires, qui doivent être confirmés par une approche hypothético-déductive avant d'être considérés

Le raisonnement déductif (hypothético-déductif) consiste à construire une expérience *ad hoc* à partir d'une hypothèse émanant de la théorie afin de vérifier si les résultats de l'expérience réfutent ou non cette hypothèse. La réfutation implique que la théorie est fausse et qu'elle doit être révisée. Si l'expérience ne réfute pas l'hypothèse, la théorie (l'hypothèse) en sort renforcée (jusqu'à la prochaine réfutation peut être). Le raisonnement hypothetico-déductif a une logique interne (*modus ponens*) et permet l'établissement d'énoncés solides et généralisables[14].

[14] Dans une certaine mesure seulement car, en termes de logique formelle, il est impossible de démontrer la véracité de la théorie : si l'expérience corrobore l'hypothèse, celle-ci ne peut pas être qualifiée de « vraie » mais sa crédibilité ou sa crédence est renforcée.

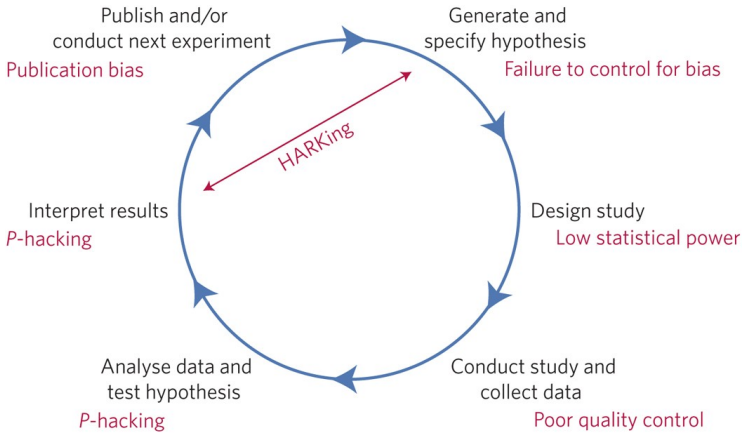


Figure 1 – Illustration de la démarche hypothético-déductive (adaptée d'après la réf. [51])

Dans l'évaluation des médicaments cette distinction prend la forme de l'opposition entre les études/résultats exploratoires et les études de confirmation (cf. ICH E10).

Les essais de confirmation, qui sont entrepris spécialement pour vérifier une hypothèse préalable, respectent le raisonnement hypothético-déductif : cette approche hypothético-déductive permet de « confirmer » les hypothèses et d'obtenir des faits prouvés, afin de démontrer l'intérêt clinique d'un traitement

Beaucoup d'études observationnelles sont de nature purement exploratoire, entreprises sans réel objectif. Les résultats produits exposent donc au risque de découverte fortuite et sont, au mieux, générateurs d'hypothèse à confirmer dans des études ad hoc[15].

> An observational post-licensure (Phase IV) study was conducted at Kaiser Permanente Northern California (KPNC), a US managed care organization, to assess the safety of zoster vaccine in people 60 years of age or older, vaccinated in routine medical care.
> Methods: We performed a cohort study, comparing rates of clinical events resulting in hospitalizations or emergency department visits in a 42-day risk time period immediately following vaccination with rates in the same cohort in a subsequent comparison time period. The study data were reviewed and interpreted by an external safety review committee of 3 independent experts [52].

2.3.1 Le « HARKing »

Les résultats des études purement exploratoires sont assez souvent présentés comme s'ils étaient obtenus par une approche hypothético-déductive. Cette démarche est parfois qualifié de « HARKing » (*Hypothesizing After the Results are Known*) [53]. Il s'agit donc d'une perversion de la démarche hypothético-déductive qui fait passer un raisonnement purement inductif (par exemple issu d'une découverte fortuite) en argument solide issue d'une démarche de confirmation déductive. Seule une lecture attentive de l'étude et une bonne connaissance du domaine permettent de détecter alors la réelle nature du résultat mis en avant.

[15] Cependant en raison de l'infodémie et de la production de masse de ce type de recherche, la vraisemblance de ces résultats devient de plus en plus faible et il convient de se poser la question du bienfondé de leur vérification pour éviter un gaspillage des ressources de recherche.

La publication en 2009 de l'essai ABCSG-12[54] conclu à un effet « anti-cancer de l'acide zolédronique » dans le cancer du sein pré-ménopausique en traitement adjuvant. Bien que surprenant au premier abord, ce résultat semble être en lien avec l'objectif de l'étude concernant ce produit (l'étude est un plan factoriel comparant aussi goserelin et tamoxifen versus goserelin et anastrozole). En effet, des références fondamentales laissant envisager un effet « anti-cancer » sont citées. Cependant, les dates de ces références (publiées entre 2007 et 2008) posent problème, car le premier patient a été inclus dans cet essai en 1999. La justification de l'hypothèse concernant l'acide zolédronique est donc postérieure au début de l'essai. De plus, dans le protocole et le plan d'analyse statistique, il apparait que l'objectif initial concernant l'acide zolédronique était la prévention de la perte osseuse induite par les autres traitements (goserelin et tamoxifen versus goserelin et anastrozole).
Ainsi il devient probable que lors d'une analyse non prévue de l'effet de l'acide zolédronique sur le critère « Disease-Free Survival » (DFS), un résultat significatif est découvert fortuitement. Ce résultat est ensuite justifié en formulant a posteriori l'hypothèse correspondante et en la justifiant avec des études fondamentales postérieures au début de l'étude. L'aspect exploratoire de ce premier résultat a cependant été perçu par une autre équipe qui met alors en place l'essai AZURE [55] qui ne confirmera pas l'hypothèse générée par la première étude.

L'ampleur de cette pratique du « HARKing » en science est bien illustrée par un article d'un blog où l'auteur, très naïvement, décrit avec précision des pratiques de p-hacking et les érige en bonne pratique permettant aux étudiants en thèse d'atteindre l'excellence [16].

2.3.2 Fragilité des explications a posteriori

Souvent la possibilité d'expliquer a posteriori un résultat exploratoire (donc de découverte fortuite) est utilisée comme validation de ce résultat. Il est alors argüé que cela revient au même qu'un raisonnement déductif : les mécanismes ne sont simplement pas invoqués a priori pour construire une hypothèse, mais a posteriori pour l'expliquer. Il existe cependant une différence majeure. Dans la démarche hypothético-déductive l'hypothèse issue des connaissances et des mécanismes est unique et choisie parmi de nombreuses hypothèses physiopathologiques et pharmacologiques, mais surtout indépendamment des futurs résultats qui seront obtenus. Dans la démarche inductive, en revanche, il est possible de trouver a posteriori une infinité d'explications compte-tenu de la complexité des mécanismes physiopathologiques et pharmacologiques. De plus les explications a posteriori sont soigneusement sélectionnées parmi tout ce qu'il est possible de développer comme théorie pour justifier le résultat obtenu. Dans le raisonnement déductif, une seule hypothèse est choisie parmi toutes celles qui pourraient être faites en fonction de cette complexité d'une manière complètement indépendante des futurs résultats qui seront obtenus d'où l'intérêt de la validation prospective des hypothèses.

Les associations a posteriori des rapports anecdotiques (études de cas) relèvent également d'une démarche inductive ce qui explique leur très faible niveau de preuve. Tout au plus ces approches peuvent conduire à générer de nouvelles hypothèses qui seront alors à vérifier selon une démarche hypothético-déductive *ad hoc* ultérieure.

Exemple de la fragilité des explications basées sur les mécanismes

L'essai CAST [56] des anti-arythmiques de classe 1c en post infarctus dans le but de prévenir la mort subite a montré que ces médicaments induisaient en réalité une surmortalité 3 fois plus élevée. L'hypothèse initiale de l'étude était basée sur des mécanismes physiopathologiques et pharmacologiques concernant

[16] https://web.archive.org/web/20170312041524/http:/www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no

> l'électrophysiologie qui permettait de spéculer sur l'obtention d'un bénéfice clinique en termes de prévention des morts rythmiques. Aussitôt après la publication des résultats de CAST, ces mêmes données physiopathologiques et pharmacologiques ont alors été interprétées différemment pour expliquer qu'il était logique que ces produits induisent une surmortalité [57].

Comme nous l'avons vu précédemment (cf. section 2.1), le raisonnement mécaniciste est indispensable au stade précoce de la recherche de nouveaux médicaments mais ne permet pas de recommander ou d'utiliser un traitement. Il permet une approche raisonnée de la sélection des candidats à développer. Certes son efficience, lorsqu'elle a été évaluée, semble modérée (50% des phases 3 sont des échecs [25]) mais il est le garant d'une démarche rationnelle dans la recherche thérapeutique. De plus ces raisonnements, malgré leurs limites, permettent de donner un rationnel aux essais, rendant ainsi acceptables leur coût financier et la prise de risque par les patients. Toutefois l'exemple ci-dessus montre à quel point ils peuvent êtres dévoyés, de par la flexibilité des raisonnements basés sur les mécanismes à s'adapter pour conclure à ce qui est recherché et à se modifier ensuite pour expliquer, a posteriori, n'importe quel résultat observé, même si ces résultats contredisent l'explication initiale.

2.4 L'erreur de raisonnement par « *Post hoc ergo propter hoc* »

« *Post hoc ergo propter hoc* » (après cela donc à cause de cela)[17] [58] est une erreur de raisonnement qui se rencontre fréquemment dans les argumentations en faveur du bénéfice ou d'un effet indésirable de traitement. L'erreur consiste à prendre comme cause ce qui n'est qu'un antécédent, sans relation avec le phénomène observé. Cela revient à établir comme fait une coïncidence.

Cette problématique est très connue en pharmacovigilance. Ce n'est pas parce qu'un évènement indésirable survient pendant la prise d'un médicament (donc après la prise du médicament) que ce médicament est la cause de l'évènement, et cela d'autant plus que l'évènement est banal.

En logique formelle la cause est forcément antécédente, mais le caractère antécédent n'établit pas à lui seul la causalité. Le sens commun (le bon sens) se fait fréquemment abuser par cette erreur de raisonnement.

En médecine, ce type de raisonnement est à la base des observations anecdotiques (*anecdotal reports* ou études des cas) [59]. Ce n'est pas parce qu'il est possible de noter chez un patient (ou plusieurs) une amélioration de leur état après l'administration d'un traitement que cette amélioration est due au traitement. Cela peut provenir d'une simple coïncidence, qui peut se répéter s'il existe un facteur de confusion liée à l'administration du médicament comme le temps avec la guérison spontanée, la régression à la moyenne, ou encore la co-administration d'autres traitements. Bien entendu, si le traitement est efficace il provoquera bien l'amélioration des patients après son administration, mais la réciproque n'est pas vraie (amélioration après traitement n'implique pas forcément que le traitement soit efficace)[18]. Ainsi un « anecdotal report » ne peut pas constituer un argument en faveur d'un changement dans la stratégie thérapeutique[19][20].

L'appellation courante « résultat *post hoc* » désigne un résultat ne correspondant pas à un objectif initial de l'étude, mais qui est mis en avant du fait de sa nature. Par exemple dans un essai d'un

[17] https://en.wikipedia.org/wiki/Post_hoc_ergo_propter_hoc
[18] C'est la différence entre $A \Rightarrow Y$ et $Y \Rightarrow A$.
[19] https://www.evidentlycochrane.net/personal-experiences-unreliable-evidence/
[20] https://s4be.cochrane.org/blog/2017/06/16/1-2-anecdotes-are-unreliable-evidence/

traitement de prévention cardiovasculaire, une moindre fréquence du cancer de la prostate avec la molécule testée est notée par rapport au placebo, amenant à conclure que cette molécule prévient le cancer de la prostate. Ce type de résultat est très fragile, car la conclusion est faite uniquement d'après le résultat observé, donc « post hoc propter ergo hoc » ! Le raisonnement est entièrement tautologique (paralogisme, dans le pire des cas sophisme), car on a l'idée de l'association uniquement d'après le résultat observé : le résultat obtenu est utilisé à la fois pour générer l'hypothèse et montrer que cette hypothèse est vraie (la vérification est donc obligatoire). C'est le raisonnement à la base du « HARKing » (cf. section 2.3).

Les autres limites des résultats post hoc proviennent de leur caractère purement exploratoire (ils n'étaient pas un objectif de l'étude) exposant à un risque important de découverte fortuite du fait de la multiplicité des comparaisons sous-jacente (nombre de cancers spécifiques envisagé dans notre exemple du cancer de la prostate).

Cette erreur de raisonnement (*fallacy*) donne aussi une autre façon d'illustrer les limites des raisonnements purement inductifs (cf. section 2.3 et dossier compagnon n° 4)

2.5 Raisonnement contrefactuel, groupe contrôle

Le but de la recherche thérapeutique est de mettre en évidence les effets des médicaments : ce que cause l'administration du traitement aux patients.

Le terme « effet » est trompeur, car il possède de nombreux sens. Avec les systèmes déterministes (comme en physique classique) l'effet est observé directement par le changement d'état du système étudié, une même cause produisant toujours le même effet. Certains phénomènes étudiés par la pharmacologie fondamentale peuvent être de ce champ. C'est également le célèbre exemple du parachute : nul besoin de faire un essai démontrant le bénéfice du parachute sur le risque de traumatisme grave lorsque l'on saute d'un avion en vol [60].

Dans le domaine de l'évaluation des traitements, cette situation ne se rencontre quasiment jamais[21] en raison de la variabilité du vivant. En médecine, il est facile d'observer que des patients similaires, porteurs de la même forme de la maladie, au même stade pourront avoir des évolutions très différentes. De même une variabilité de réponse aux traitements entre patients similaires est facilement observable.

De ce fait, il est impossible de déterminer l'effet d'un traitement par l'observation de ce qui se passe après son administration. Ce n'est pas parce qu'un patient hypercholestérolémique à risque n'a pas fait d'infarctus 1 an après l'initiation d'une statine que cela est dû à la statine car la majorité (de l'ordre de 95%) des patients hypercholestérolémiques à risque non traités ne font pas ce type d'évènement à 1 an. On pourrait d'ailleurs tenir de la même façon le raisonnement selon lequel les statines ne servent à rien si un patient hypercholestérolémique à risque n'a pas fait d'infarctus pendant 1 an alors qu'il ne recevait pas de statine. De même, si le patient fait un infarctus sous statine cela ne veut pas dire que le traitement n'a pas eu d'effet. Il a peut-être retardé le moment de survenue de l'infarctus. Sauf cas exceptionnel, l'observation de ce qui se passe chez des patients avec le traitement ne permet pas de savoir si celui-ci à un effet (et apporte un bénéfice).

L'objectivation d'un effet nécessite ainsi obligatoirement un raisonnement contrefactuel et une approche probabiliste. C'est-à-dire de comparer ce qui s'est passé avec le traitement étudié à ce qui

[21] Sauf pour les forts toxiques et quelques molécules bien particulières (produits anesthésiants par exemple)

se serait passé sans ce traitement : le contrefait. Bien entendu cette comparaison ne peut pas se faire chez les mêmes patients selon un mode avant/après en raison de l'évolution naturelle des maladies ou lorsque les évènements cliniques d'intérêts sont définitifs (comme les décès) ou qu'ils changent l'état des patients (événements cardiovasculaires par exemple, qui même s'ils peuvent se reproduire chez les mêmes patients n'ont plus alors le même pronostic) [22].

> Durant la crise de la COVID beaucoup d'arguments avancés pour justifier de l'efficacité de certains traitements négligeaient complètement la contre-factualité : « Depuis quelques semaines, nous avons prescrit ce traitement à tous nos patients atteints du coronavirus. Pour ma part, cela représente plus de 200 patients. J'ai eu seulement deux cas graves nécessitant une hospitalisation et qui sont sortis depuis. » Bien évidement cette observation ne permet pas d'en déduire l'efficacité du traitement car il n'y a aucune possibilité de la relativiser par rapport au contrefait : le nombre de ces mêmes patients qui auraient été hospitalisés s'ils n'avaient pas été traités.

Le groupe contrôle est indispensable pour apporter le contrefait et évaluer l'effet du médicament, mais toutes les façons de former ce groupe contrôle ne se valent pas.

En recherche clinique, c'est le but du groupe contrôle d'apporter le contrefait, mais ce n'est qu'avec la randomisation que l'on obtient un contrefait (probabiliste) qui permettra de démontrer la causalité entre les interventions testées et les résultats obtenus. Dans les études observationnelles, le groupe contrôle est indispensable pour assurer un raisonnement contrefactuel, mais la preuve de la causalité est moindre car les traitements n'auront pas été alloués au hasard aux patients [61]. Avec ces études, le contrefait permettra de mettre en évidence des associations.

> *Une nouvelle approche, appelée inférence causale, se développe actuellement pour proposer un cadre d'analyse qui permettrait d'obtenir la causalité avec des études non randomisées [62]. Cette approche basée sur une mathématisation de la causalité permet de construire des stratégies et des modèles d'analyses des données permettant de conclure à la causalité*

La problématique de l'absence de raisonnement contrefactuel redevient d'actualité avec les études mono-bras (cf. dossier compagnon n° 12). Ces études ne permettent pas de conclure en elles-mêmes à l'effet du traitement et débouchent sur des comparaisons par rapport à des références externes plus ou moins fiables [63].

2.6 Le jugement « clinique » peut-il dispenser d'une validation statistique des résultats ?

Parfois il est plaidé qu'un résultat non statistiquement significatif, mais correspondant à une différence importante sur un critère cliniquement important et compatible avec le mécanisme d'action, a un « sens clinique » et doit être pris en considération[23].

[22] Le contrefait ne peut être obtenu chez les mêmes patients que lorsqu'une comparaison avant/après fiable est possible (c'est-à-dire que l'évolution est constante au cours du temps).

[23] Soit comme argument « primaire », soit comme argument secondaire pour renforcer un premier résultat cliniquement peu pertinent par exemple.

Exemples de surinterprétation des résultats non concluants

« Bien que non statistiquement significatif en raison du faible nombre d'évènements[24], un plus faible taux d'hospitalisation est également observé chez les sujets traités par la bithérapie par rapport au placebo. »

«Bien que non statistiquement significatif, ce résultat est cliniquement important. »

« Bien que non statistiquement significatif, le traitement a montré une réduction cliniquement significative dans la progression de l'invalidité confirmée à 12 semaines, telle que mesurée par l'Échelle élargie de l'état d'invalidité (EDSS). »

C'est par exemple un Hazard Ratio de 0.4 sur la mortalité dans un essai de petite taille. Il est argumenté alors qu'il n'y a pas besoin de signification statistique pour considérer que ce résultat est réel, car il s'agit des décès (critère objectif et cliniquement important), que la réduction est importante et que le résultat est attendu compte tenu du mécanisme d'action. En fait, aucun de ces arguments ne permet de garantir que ce résultat est réel.

Ce n'est pas parce que le critère de jugement est cliniquement important qu'une comparaison de 2 groupes de patients n'est plus sujette aux fluctuations aléatoires d'échantillonnage.

Les fluctuations aléatoires d'échantillonnage n'affectent pas les évènements, mais bien les statistiques et en particulier ce qui est observé dans un échantillon (un groupe) de patients. Quelle que soit la nature du critère de jugement, la valeur statistique d'un échantillon est sujette à des fluctuations (des sur- ou sous-estimations) dues uniquement au hasard.

En l'absence de signification statistique, il est risqué de conclure à l'existence de l'effet. De plus, dans un essai de faible puissance (cf. dossier compagnon n° 8), un résultat significatif est peu en faveur de l'effet (faible valeur prédictive positive). Dans une approche d'inférence bayésienne les résultats de ce type (grande différence non significative dans un essai de faible puissance) sont associés à une probabilité a posteriori que le traitement soit efficace faible.

Ce point est illustré de manière empirique par une étude de méta-épidémiologie, portant sur des traitements pour lesquels un effet de grande taille « *very large effect* » avait été observé dans une étude préliminaire (phase 2 le plus souvent) et pour lesquels une étude de confirmation avait aussi été réalisée. L'objectif était d'évaluer si les résultats préliminaires de grande taille sont prédictifs à coup sûr d'un véritable effet du traitement [64]. Ce travail ne trouve pas de corrélation entre les premiers résultats de grande taille et la taille du résultat de l'essai de confirmation. Seulement 43% des essais de confirmation obtiennent un résultat statistiquement significatif. Un effet de grande taille dans un essai préliminaire ne permet donc pas de conclure qu'un essai de confirmation n'est pas nécessaire.

[24] Le raisonnement est ici complètement fallacieux. Le résultat non significatif est expliqué par un manque de puissance en partant du principe implicite que le traitement est efficace. Ce n'est pas le résultat qui induit la conclusion mais une justification tautologique : si le traitement était efficace on aurait quand même pu obtenir un résultat négatif par manque de puissance, donc ce résultat non-concluant montre que le traitement est efficace ! Le traitement est considéré comme étant efficace car une différence est observée. Mais cette conclusion purement intuitive néglige complètement la justification des tests statistiques qui sont nécessaires par la possibilité de fausses différences dues au hasard. Elle considère implicitement que ce qui est observé ne peut pas être faux et qu'il n'y a donc pas de fluctuations aléatoires d'échantillonnages.

De la même façon, la plausibilité biologique n’est pas non plus un argument très probant compte tenu de sa faible valeur prédictive d’un bénéfice clinique (cf. section 2.3.2) [25].

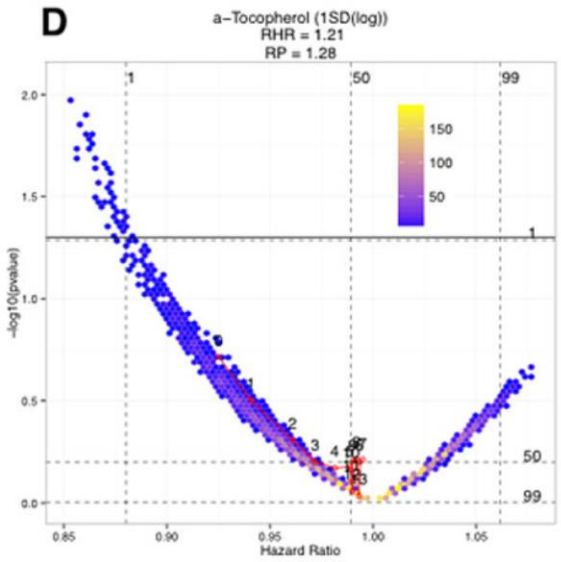
Chapitre 3. Distorsion de la connaissance scientifique

3.1 *P hacking, data dredging*

Les termes « *p hacking* » ou « *data dredging* » désignent l'adaptation de l'analyse statistique en cours de réalisation, en fonction des résultats qu'elle produit. Ces adaptations peuvent concerner aussi bien la méthode statistique (choix de la méthode, transformation de variables, choix des covariables d'ajustement, etc.) que le jeu de données (exclusion de patients, gestion des évènements intercurrents, restriction de l'analyse à une sous population, etc.). Ces adaptations sont d'autant plus faciles à effectuer que l'étude nécessite une analyse statistique complexe, comme avec les études observationnelles par exemple. Sur le plan de l'intégrité scientifique, ces pratiques questionnables de recherche (mentionnées dans le rapport Corvol 2016) contribuent à produire des résultats faux positifs.

Avec cette pratique, il est ainsi possible d'orienter les résultats dans la direction souhaitée, tout du moins en termes de signification statistique (d'où le nom de *p hacking*) [65, 66]. Il a ainsi été montré qu'avec un même jeu de données, confié à des équipes scientifiques différentes ayant des conceptions théoriques antithétiques, il était possible d'obtenir des résultats très différents et même opposés [67, 68]. L'étude perd ainsi sa valeur scientifique (assurée par le fait que la réponse à la question posée est fournie uniquement par les données) pour devenir une simple opération pour produire les résultats escomptés. Il ne s'agit plus d'un test loyal d'une hypothèse thérapeutique où seule la réalité pourra la réfuter ou la confirmer, mais d'une démarche de recherche active de la façon d'analyser des données afin d'obtenir un résultat le plus proche de la réponse voulue ! Un *p-hacking reverse* a aussi été mis en évidence où l'analyse statistique est construite pour ne pas donner de différence significative [69].

> *Cette potentialité peut être illustrée par le concept de vibration des effets [66]. Il s'agit de visualiser l'ampleur suivant laquelle « vibrent » les différents résultats (taille d'effet et p value) obtenus par toutes les possibilités d'analyse d'une même recherche d'association. Ces vibrations peuvent déboucher dans certains cas sur des effets Janus où des résultats opposés sont obtenus à partir du même jeu de données.*



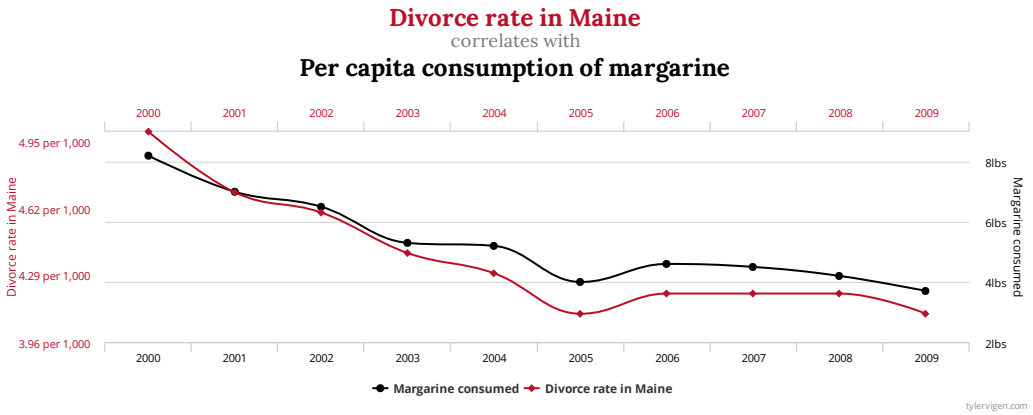
Effet Janus produits par la Vibration des résultats de la recherche d'une association entre le tocopherol et la mortalité totale à partir des données de la National Health and Nutrition Examination Survey [66].

L'abscisse représente le Hazard Ratio et l'axe des ordonnées la p value (échelle logarithmique, 0.05 correspond à 1.3, trait plein horizontal)

Dans la littérature ces aspects sont souvent introduits par l'aphorisme dû à Ronald Coase : « if you torture the data long enough, it will confess to anything »[25]. On parle aussi de *data-dredging* ou partie de pêche [70, 71].

Cette problématique est assez pernicieuse, car l'analyse statistique est parfois présentée comme une démarche dont la finalité est de rechercher ce qu'un jeu de données est à même de révéler. Cette conception est inappropriée aux questionnements d'évaluation des médicaments, domaine où les données sont éminemment bruitées (la variabilité du vivant), et où les effets à détecter sont petits par rapport au bruit. Ainsi le rapport signal bruit est faible et propice à donner de nombreux artéfacts. Une analyse purement exploratoire va vouloir proposer comme fait généralisable un artéfact de découverte fortuite (qui n'apparait que dans les données concernées, avec la méthode d'analyse utilisée). Cette problématique représente aussi la limite des approches de fouilles de données (« *data mining* ») et elle est connexe à l'opposition entre raisonnement inductif et raisonnement déductif (cf. dossier n° 4).

Le site WEB « spurious correlations » (https://www.tylervigen.com/spurious-correlations) réalise à but pédagogique une recherche intensive de corrélations entre toutes les données rendues publiques dans le cadre de la politique d'open data. Les corrélations très importantes avec un coefficient de corrélation supérieur à 0.9 sont présentées. Il est ainsi possible de s'apercevoir de l'existence de très nombreuses relations insoupçonnées que seul le « *data mining* » permet d'identifier, par exemple une parfaite corrélation entre le niveau de consommation de margarine et le taux de divorce dans le Maine entre 2000 et 2009.



L'explication a posteriori d'une découverte plus ou moins fortuite est souvent considérée comme une validation. Ici, il est bien évident que si, dans un couple, l'un impose la margarine au petit déjeuner à la place du beurre cela ne peut que mal finir, d'où cette corrélation. CQFD !

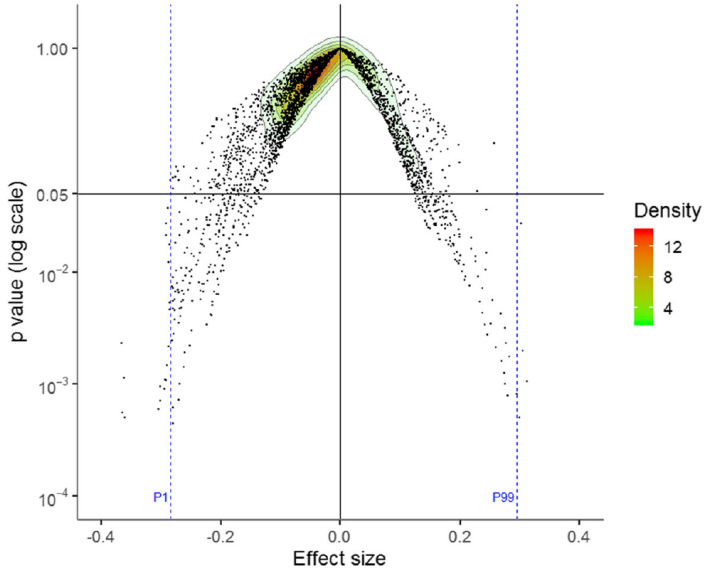
La solution réside dans la conception a priori de l'analyse statistique, complètement indépendante des données et des résultats produits. Cela est obtenu par l'élaboration d'un plan d'analyse statistique (« *statistical analysis plan* », SAP) en amont de la disponibilité des données elles-mêmes rendues disponible pour la reproduction des analyses. Ainsi aucune adaptation de la stratégie d'analyse ne peut

[25] https://en.wikiquote.org/wiki/Ronald_Coase

s'effectuer au moment de sa réalisation (sans que cela soit détectable en comparant le plan d'analyse statistique et l'analyse effectivement réalisée).

Pour les études observationnelles, particulièrement exposées au « *p hacking* » en raison de la complexité des analyses réalisées, la solution réside aussi dans l'approche du SAP. Cependant, pour les études sur données historiques (rétrospectives), le SAP sera par définition élaboré alors que les données sont déjà disponibles. Pour donner la garantie de l'absence de tout « *p hacking* » ou autre opération de « *data dredging* », il est impératif que soit explicitement mentionné dans le protocole et le rapport de l'étude que l'analyse a été conçue indépendamment des données et des résultats produits [72]. On pourrait aussi concevoir un SAP écrit par un comité indépendant dès lors que l'analyse va concerner des données déjà recueillies. Il est possible que des changements soient nécessaires dans le SAP au moment de l'analyse. Le SAP fera alors l'objet d'un amendement et l'analyse initialement prévue sera elle aussi rendue disponible pour assurer la transparence de ces changements.

> *La méta-analyse (cf. dossier compagnon n° 9) est le plus souvent une démarche purement rétrospective entreprise, par définition, après la disponibilité des données (ici les résultats des études) et souvent après la prise de connaissance des résultats de ces études. Le »* p hacking » *et* le « data dredging » *sont des points sensibles de ces études [73]. La potentialité de pouvoir obtenir le résultat souhaité en faisant « les bons choix » dans le protocole, en particulier au niveau des critères d'éligibilité des études, est visualisable par le concept de vibration des effets. Il s'agit de réaliser toutes les méta-analyses possibles en fonction des différentes combinaisons des critères d'éligibilité des études réalisables. La représentation graphique de tous ces résultats visualise alors dans quelle mesure il est possible de produire des résultats différents, voir opposés, dans « la même » méta-analyse. Dans une comparaison indirecte [55] comparant nalmefene et naltrexone pour la réduction de la consommation d'alcool, 9217 méta-analyses en réseau différentes sont possibles à partir des 9 essais du nalmefene et 51 du naltrexone. Les résultats produits par toutes ces méta-analyses (cf. ci-dessous) permettraient de faire toutes les conclusions possibles en termes de supériorité d'un produit ou de l'autre, de manière significative ou non significative.*



3.2 Biais de publication et « *selective reporting* »

3.2.1 Biais de publication

Le biais de publication provient de la publication ou non des études (les essais thérapeutiques entre autres) en fonction de la positivité ou non de leur résultat [74, 75]. Les études négatives (ou, plus largement, celles qui produisent des résultats ne confirmant pas l'hypothèse testée) sont moins fréquemment publiées que les études positives (donnant les résultats qui étaient attendus). Il s'en suit une distorsion de la réalité au niveau des résultats publiés [76, 77]. Au-delà de cette forme binaire publication/non-publication, ce biais prend aussi une forme plus continue : les études négatives finissent par être publiées, mais avec beaucoup de retard, tandis que les études positives le sont immédiatement. À un temps donné, les deux formes de ce biais sont équivalentes.

> Dans la recherche de marqueurs pharmacogénétiques de réponse aux sels de platine dans le cancer du poumon non à petites cellules, plusieurs études de pronostic sous traitement ont rapporté une association des polymorphismes de ERRC1 avec les critères PFS et « Overall Survival » (OS). Une méta-analyse confirme ce résultat [78] en regroupant 18 études du même type. Pour établir définitivement la valeur prédictive de ce marqueur, un essai d'interaction (le design approprié pour l'évaluation des marqueurs prédictifs, les études de pronostic sous traitement pouvant être confondues par une valeur pronostique ordinaire du marqueur [79]) est réalisée [80]. Cet essai ne trouve pas de valeur prédictive de ce marqueur et dans le groupe traité avec le traitement à base de platine ne retrouve pas non plus la valeur pronostique sous traitement. Ce résultat fait suspecter un biais de publication au niveau des études précédentes, qui sont des études faciles à réaliser (rétrospectives, entre 50 et 200 patients) et réalisables en grand nombre compte-tenu de la multitude de séries de patients avec du matériel biologique conservé qui existent à travers le monde. Ainsi, après un premier résultat communiqué suggérant l'association, il est possible de penser que de nombreuses équipes ont cherché à le répliquer. Si le résultat initial était un faux positif dû au hasard, la majorité des réplications seront négatives (à raison), mais sur le nombre, plusieurs résultats statistiquement significatifs (à tort) seront quand même obtenus (le risque alpha, au moins de 2.5% mais en général plus important du fait du « *p hacking* »). Il est probable que les résultats ne retrouvant pas l'association ne sont pas publiés et que seuls les faux positifs le sont (ne serait-ce que sous la forme d'un poster ou d'une communication à un congrès). Ainsi apparait le biais de publication, suivant un scénario logique compte-tenu des enjeux et des pratiques de recherche actuelles dans les domaines exploratoires.

Actuellement, dans le champ des essais thérapeutiques, en raison des mesures mise en place pour le prévenir (principalement l'obligation d'enregistrement de l'essai avant l'inclusion des patients), il semble être moins présent qu'auparavant [81, 82] même s'il n'a pas complètement disparu et s'il persiste sous la forme du « *selective reporting bias* » au niveau des critères de jugement exploratoires (cf. ci-dessous).

> *Les études de méta-épidémiologie récentes révèlent que le nombre d'essais non publiés reste encore élevé malgré les mesures mises en place (enregistrement) [83, 84]. Des initiatives ont été mises en place pour tenter de faire changer les pratiques comme https://www.alltrials.net/ qui maintient un tracker de non-publication des essais enregistrés TrialsTracker (https://trialstracker.ebmdatalab.net/#/).*

Le biais de publication, en rendant publique préférentiellement les faux positifs dans une série de réplication de la recherche de la même association est en mesure de sacraliser (canoniser) ces faux positifs en fausses découvertes [85].

Le biais de publication a été découvert et formalisé dans les années 80 à propos des essais thérapeutiques. Mais il est présent quel que soit le type d'étude et affecte tout autant les études observationnelles [45]. Plusieurs éléments font penser que le biais de publication est très fréquent

avec les études observationnelles rétrospectives qui sont moins lourdes à réaliser (en particulier en raison de leur caractère rétrospectif) et qui sont réalisées en grand nombre [76, 86, 87, 88].

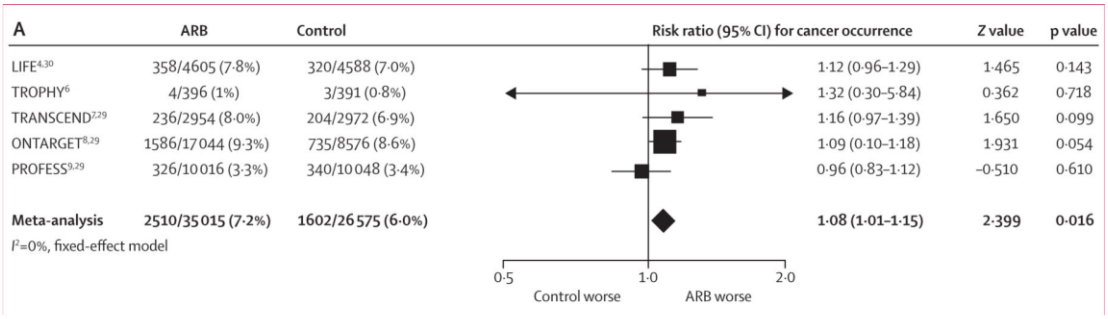
> La survenue de pneumopathies infectieuses est classiquement considérée comme étant un effet indésirable des inhibiteurs de la pompe à proton (IPP). De nombreuses études observationnelles retrouvent un lien entre utilisation d'un IPP et la survenue d'une pneumopathie infectieuse. L'explication mécanistique fait appel à la baisse d'acidité du contenu gastrique qui pourrait favoriser la pullulation microbienne et la contamination pulmonaire à l'occasion de reflux par exemple. La méta-analyse confirme cette association, mais met aussi en évidence un biais de publication important [89]. L'existence du sur-risque de pneumopathies infectieuses est aussi fortement remis en cause par les résultats obtenus par l'essai randomisé en double aveugle COMPASS comparant le pantoprazole au placebo chez 17 598 patients traités et suivi durant 3.01 ans en médiane [90]. L'absence de sur-risque est aussi retrouvée par des études observationnelles en particulier avec des designs autocontrôlés et dont les résultats suggèrent l'existence d'un fort biais d'indication dans les études plus standards (lié au fait que la prescription d'IPP est plus fréquente en cas de terrain propice à la survenue des pneumopathies infectieuses, comme le reflux).
>
> Au total, les données actuellement disponibles ne sont pas en faveur de l'existence de cet effet indésirable qui a cependant été sacralisé en fait établi et universellement mentionné dans les ouvrages et les cours. Depuis la mise en évidence du biais de publication et les résultats de l'essai COMPASS, cette allégation d'effet indésirable (EI) n'a pas encore été récusée et perdure. De plus, les biais cognitifs rentrent aussi en action, rendant encore plus difficile le rétablissement des faits, en raison de la notoriété passée de cet EI. Le fait de l'avoir appris, de le rencontrer dans tous les textes et de l'entendre sur toutes les bouches, le rend réel même s'il est possible de récuser maintenant des résultats l'ayant suggéré initialement. En quelque sorte son existence s'est autonomisée des arguments qui étaient à son origine. Montrer le côté erroné de ces arguments n'a plus de conséquence sur l'acceptation de l'existence de cet EI (le lien entre résultat conduisant à évoquer cet EI et mention de l'EI est rompu au niveau de la connaissance collective). Cet EI est devenu une évidence (au sens français du terme), partagée par tous, et qui n'a plus besoin d'origine. Sa réalité n'a plus besoin d'être justifiée, elle est devenue immanente. En plus, ces biais cognitifs vont être renforcés par l'expérience sensible qui notera la survenue de pneumopathies chez des patients sous IPP. Cette observation ne sera pas rare compte-tenu de la fréquence des prescriptions d'IPP et de la pneumopathie infectieuse, mais cette coïncidence renforcera le biais cognitif par un raisonnement de type « *post hoc proper ergo hoc* » (cf. section 2.4).
>
> Cependant toujours considérer l'existence de ce risque, aujourd'hui sans fondement, peut conduire à priver à tort certains patients de ces médicaments qui leur seraient pourtant utiles (le bénéfice de ces molécules a été démontré pour plusieurs conditions cliniques [91]).

Le biais de publication est évoqué systématiquement dans le cadre des méta-analyses mais ce n'est pas un biais spécifique de la méta-analyse. Il opère en amont, sur la disponibilité des études (et des résultats dans les études, cf. ci-dessous, biais de « *selective reporting* ») et affecte donc la perception de l'efficacité ou de la sécurité d'un traitement intrinsèquement, qu'une méta-analyse ou non soit entreprise. Même l'inventaire discursif des études facilement disponibles (publiées) est alors abusé par le biais de publication.

3.2.2 Selective reporting bias

Une variante du biais de publication affectant les études (les études sont publiées ou non en bloc en fonction de leur résultat) est le « *selective reporting bias* » qui affecte la façon dont sont rapportés les résultats au sein d'une même étude, en fonction de leur positivité [92]. Ainsi les critères de jugement exploratoires (critères de « safety » par exemple) dont la présence n'est pas « obligatoire » peuvent être présents ou non dans la publication en fonction de leur résultat.

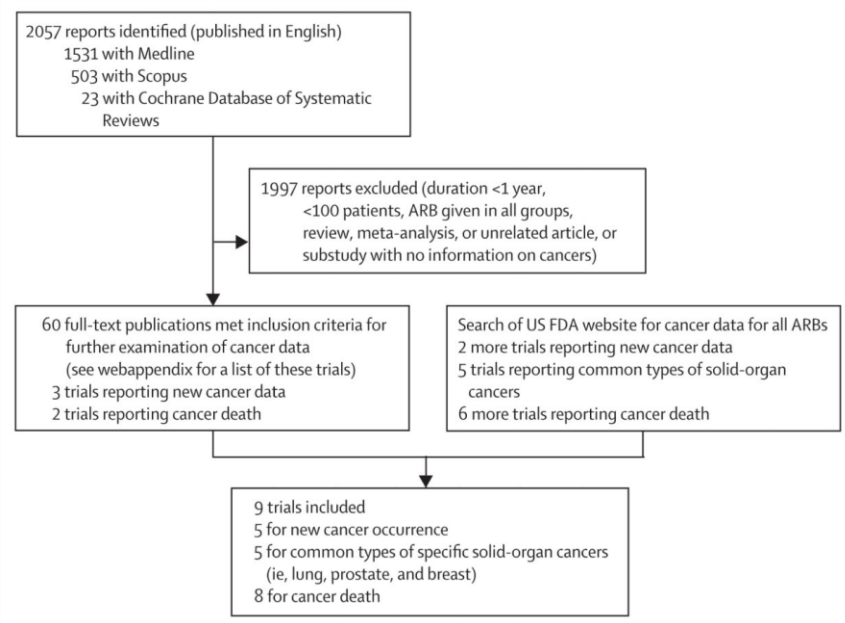
Une méta-analyse publiée dans le Lancet Oncology en 2010 [93] conclut à un sur-risque de cancer avec les sartans (antagonistes des récepteurs de l'angiotensine II). Même si le sur-risque est faible (risque ratio de 1.08), le fait que l'effet indésirable puisse être la survenue d'un cancer sème un certain émoi avec des propositions par certains de retrait de ces traitements dans l'hypertension.
Ce résultat est obtenu à partir de la méta-analyse de 5 essais qui rapportaient les cancers dans leur tableau d'effets indésirables.



A	ARB	Control	Risk ratio (95% CI) for cancer occurrence	Z value	p value
LIFE[4,30]	358/4605 (7·8%)	320/4588 (7·0%)	1·12 (0·96–1·29)	1·465	0·143
TROPHY[6]	4/396 (1%)	3/391 (0·8%)	1·32 (0·30–5·84)	0·362	0·718
TRANSCEND[7,29]	236/2954 (8·0%)	204/2972 (6·9%)	1·16 (0·97–1·39)	1·650	0·099
ONTARGET[8,29]	1586/17044 (9·3%)	735/8576 (8·6%)	1·09 (0·10–1·18)	1·931	0·054
PROFESS[9,29]	326/10016 (3·3%)	340/10048 (3·4%)	0·96 (0·83–1·12)	−0·510	0·610
Meta-analysis	**2510/35015 (7·2%)**	**1602/26575 (6·0%)**	**1·08 (1·01–1·15)**	**2·399**	**0·016**

I^2=0%, fixed-effect model

Cependant le flow chart révèle que 60 essais randomisés (dans l'insuffisance cardiaque ou l'hypertension) étaient éligibles pour cette méta-analyse. Mais seulement quelques essais rapportent les données liées à la survenue d'un cancer (ou des décès par cancer) ce qui explique que le résultat de la méta-analyse ne porte que sur 5 essais.



Il est ainsi possible de craindre que la présence des cancers dans le tableau des effets indésirables d'essais en cardiologie dépende de ce qui est observé. Ces critères ne sont certainement rapportés que s'il existe quelque chose d'intéressant à discuter, comme une différence numérique entre les 2 groupes en défaveur du produit expérimental (il s'agit d'un tableau de « *safety* »). Dans les autres cas, les données sont dans les rapports des essais, mais ne sont pas mis en avant dans la publication, car rien de remarquable n'est à rapporter à leur sujet. Toutes les conditions d'un « *selective reporting bias* » sont alors rassemblées et le résultat portant sur 5 essais sur 60 doit être pris avec circonspection.
Une méta-analyse ultérieure a cherché à récupérer les données sur les cancers de tous les essais, principalement à partir des rapports des études déposées auprès de la FDA [94, 95] et ne retrouve pas d'association, étayant ainsi le rôle d'un « *selective reporting bias* » dans la première méta-analyse.

Le « *selective reporting bias »* montre que la problématique du biais de publication dépasse le cadre de la publication ou non d'une étude et qu'il est susceptible d'affecter les résultats au sein d'une étude (au niveau des critères de jugements [96, 97], des objectifs exploratoires, des sous-groupes [98], des analyses, des ajustements pour les études observationnelles [99], etc.). Ce biais est maintenant pris en considération dans les grilles d'évaluation du risque de biais des essais comme le ROB 2.0 [100].

> *Le «* selective reporting bias » *ne concerne pas seulement les résultats accessoires ou exploratoires des essais, mais affecte aussi les critères majeurs (principaux ou secondaires). La non-publication de ces résultats passe alors fréquemment par un switch de critère entre le protocole et la publication [97, 97, 101, 102, 102]. Une initiative avait été mise en place en 2011 pour surveiller ces switches https://www.compare-trials.org/ et réagir chaque fois qu'une situation de ce type était détectée en envoyant une lettre à l'éditeur, mais elle ne semble plus active en 2021.*

Ce biais explique pourquoi il convient d'être prudent avec les résultats inattendus obtenus en dehors des objectifs initialement fixés dans une étude (en plus des problématiques liées à l'absence de démarche hypothético-déductive, à la multiplicité, etc.), en particulier vis-à-vis de tous les résultats exploratoires de « *safety* » visant à établir un EI à partir d'un raisonnement purement inductif et non testé par la suite par une étude randomisée [103].

3.3 Pourquoi est-il utile de juger par soi-même de la solidité des études ? Les spins de conclusion.

Les essais cliniques étant publiés majoritairement dans des revues à comité de lecture, il pourrait être perçu comme superflu de rejuger par soi-même de la méthodologie et du bien-fondé des conclusions de l'article. Dans un contexte professionnel chargé, où le temps est précieux, on pourrait se dire qu'il suffit de prendre connaissance de la conclusion, l'évaluation de la solidité de celle-ci ayant été réalisée par les reviewers qui possèdent en général des expertises méthodologiques et statistiques bien plus importantes que le simple lecteur.

Il s'avère cependant que cette attitude n'est pas complètement possible en raison notamment de l'existence des spins, dans tous les parties des articles, mais surtout les spins de conclusions .

Les spins de conclusion (ou enjolivement de la conclusion) [104] sont des situations où la conclusion de l'article rapportant un essai clinique est positive, en faveur de l'intérêt du traitement, alors même que l'essai est non-concluant (n'apportant pas de démonstration de l'intérêt clinique du nouveau 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 [105, 106, 107].

La présence de spins peut s'expliquer par les conflits d'intérêts des revues elles-mêmes [108]. Les revues biomédicales, sauf cas exceptionnels, appartiennent à des éditeurs privés et reposent sur un modèle économique de profitabilité. Ces revues sont alors fortement intéressées par les articles rapportant des études industrielles (essais de phase 3 pivotales des nouveaux médicaments par exemple) car ces articles sont susceptibles de leur assurer des retours financiers substantiels par la vente de tirés à part et autres droits de traduction au sponsor de l'étude [109, 110, 111]. Ce type d'article assure aussi un maintien du niveau d'impact factor de la revue du fait du nombre de fois où ces articles, qui font souvent le buzz, sont cités. Ainsi pour ne pas risquer de rater la publication d'un manuscrit à fort potentiel de citations, les éditeurs en chefs sont parfois amenés à relativiser les commentaires des reviewers et à accepter des articles tels qu'ils sont proposés, y compris avec les

éventuels spins présents. Cet état de fait est particulièrement bien décrit dans deux ouvrages écrits par d'anciens éditeurs en chefs d'une revue de tout premier plan [33, 34].

Les conséquences des spins sont graves, car ils conduisent à faire croire au lecteur, qui se limiterait à la lecture de la conclusion de l'abstract, que le nouveau traitement a un intérêt clinique alors qu'en réalité celui-ci n'est pas démontré [112].

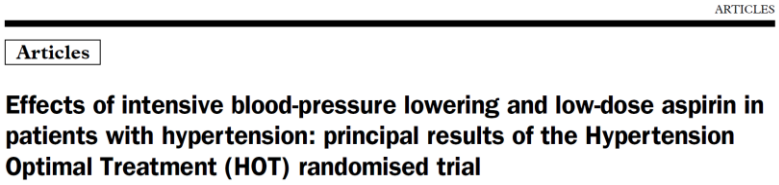
Une des finalités de l'enseignement de la lecture critique durant les études médicales est d'armer les futurs cliniciens à ne pas tomber dans les pièges des spins de conclusions.

Ainsi, il incombe au lecteur de ces publications la charge de faire une évaluation de la rigueur méthodologique de l'étude et des résultats obtenus avant de dériver sa propre conclusion pour ne pas tomber dans le piège d'un spin de conclusion

Les spins sont aussi très fréquents dans les supports de communication promotionnelle ou dans le discours de certains leadeurs d'opinion. L'éminence d'un locuteur ne doit pas faire perdre le regard critique et la volonté de revenir aux faits prouvés.

Exemple : HOT

L'essai HOT avait comme objectif de montrer que la prise en charge intensive de l'hypertension artérielle permettait de maximiser le bénéfice sur les évènements cardiovasculaires [113]. Pour cela, 3 cibles de pression artérielle diastolique (PAD) étaient comparées : la cible standard de l'époque (90 mm Hg) et deux cibles d'intensification (85 et 80 mm Hg). 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 à l'origine 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 90 mm Hg pour chercher à atteindre 82.6 mm Hg. Cependant, cette conclusion apparait d'emblée surprenante du fait de la valeur de la cible mise en avant : 82.6 mm Hg. Cet essai ne peut pas conduire à ce résultat. En effet, les seules conclusions possibles sont 85 ou 80 mm Hg 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 80 mm Hg) avec un p = 0.5. L'interprétation adéquate 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 mm Hg, 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 82.6 mm Hg. 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 autre 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.

On peut imaginer que les reviewers ont signalé ces problèmes, mais l'éditeur en chef aura 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'allait produire ce résultat [108].

3.4 « Infodémie » et course à la publication

Le domaine de l'évaluation des médicaments connait, comme tous les domaines de la science [114], une activité de publication sans précédent [115]. On assiste à une véritable « infodémie », induite par une profusion de publications sur chaque sujet, qui complique le travail de synthèse des résultats nécessaire à la construction des stratégies thérapeutiques.

3.4.1 Production de masse de publications

Cette profusion de publications provient certes d'une activité sans précédent de la recherche thérapeutique, mais aussi d'une production de masse de publications de faible intérêt [116] motivée principalement par la course aux publications à laquelle sont obligés de se livrer les chercheurs (le fameux « *publish or perish* »). Cette dérive trouve ses origines, entre autres, dans l'évaluation des chercheurs et l'attribution des financements d'une partie de l'activité de recherche sur la base d'indices bibliométriques purement quantitatifs. Du fait de ces incitations purement quantitatives, l'objectif de l'activité de « recherche » s'est adapté en devenant progressivement « publier de plus en plus pour abonder les scores bibliométriques » à la place de produire des connaissances validées, utiles pour les patients et opérables [116]. En France, il est ainsi possible de mettre en évidence un effet délétère du système basé sur l'attribution de points SIGAPS [117].

Cette pression à publier mise sur tous les chercheurs a aussi débouché sur l'émergence de pratiques déviantes, encore minoritaires, à l'encontre des principes de l'intégrité scientifique comme les « *ghost*

authors » [118], les auteurs hyper-prolifiques [119], la vente de place dans « l'authorship » [120, 121], l'autocitation excessive et non justifiée [122], les revues de complaisance payantes ou népotiste d'autopromotion [123], les revues prédatrices [124, 125, 126], et à l'extrême la fraude scientifique [127].

> *Cette pression de la publication pour la carrière des chercheurs a aussi conduit progressivement à un dévoiement insidieux des principes de la publication scientifique[26]. Les éditeurs des journaux biomédicaux et les reviewers, connaissant les difficultés des chercheurs, ont progressivement adaptés, peut-être même sans s'en apercevoir, les niveaux d'exigences scientifiques, pour pouvoir publier en nombre ces articles sans intérêts [128]. Originellement, le but de la publication était de communiquer les découvertes scientifiques, maintenant, dans bien des cas, son but est de publier pour publier ! Le monde de l'édition scientifique est devenu un processus autosuffisant. Cette dérive a conduit par exemple à la création de revues de second ordre (prédatrices ou non) destinées uniquement à offrir des opportunités pour publier plus facilement, plus rapidement, et d'offrir des postes d'éditeurs à foison. On a ainsi assisté à l'émergence de revue où une proportion importante des articles publiés provenait des mêmes institutions que celles des responsables éditoriaux.*

Cette production de masse de mauvaise science (« *mass production of garbage science* ») a des conséquences très importantes. Les essais de mauvaise qualité méthodologique, qui n'apporteront pas de réponse fiable, représentent au mieux un gaspillage des ressources de recherche [129] mais exposent surtout au risque de faire adopter à tort des traitements n'apportant pas le bénéfice prétendu. Ces essais peuvent être ainsi vu comme dangereux et non éthiques [130, 131, 132]. De plus, ils donnent le mauvais exemple au sein des équipes de recherche en suggérant une image de normalité aux chercheurs en formation.

Hormis le gaspillage des ressources de recherche [133, 134, 135] que représentent ces mauvaises pratiques (« *misconduct* »), l'infodémie et la production de masse de mauvaises études représentent un nouveau chalenge inédit pour les médecins et autres acteurs du médicament qui doivent exploiter la littérature pour l'actualisation de leurs connaissances, la construction des stratégies thérapeutiques ou l'élaboration des recommandations. Les résultats utiles issus d'études fiables et pertinentes qu'il ne faut pas rater, car ils représentent de véritables progrès thérapeutiques, sont noyés parmi de multiples résultats sans intérêt.

> Par exemple, les études méta-épidemiologique dans la COVID ont fait le constat d'une production d'études de faible qualité méthodologique [14, 15], dispersées et largement redondantes [16]. Ce phénomène a conduit l'OMS à formuler une mise en garde et des propositions de solutions [136, 137].

Il est donc important de savoir identifier le bon grain de l'ivraie et de ne pas se retrouver submergé par une littérature sans intérêts[27].

3.4.2 Les prépublications (*preprints*)

L'avènement des *preprints*, c'est-à-dire de publications sans « reviewing »[28], représente une évolution intéressante de l'écosystème de la publication scientifique. Un des avantages majeurs pour l'évaluation des médicaments est l'accélération de la diffusion des résultats, qui permet, lorsqu'il s'agit

[26] https://www.redactionmedicale.fr/integrite-scientifique

[27] Cette problématique n'est pas propre à la médecine et touche tous les domaines de la science (https://www.nature.com/articles/d41586-019-00381-w)

[28] Certaine plateforme propose un « *folks reviewing* » qui est un modèle de reprise en main par la communauté scientifique du processus d'édition assez séduisant

de réelles innovations thérapeutiques, de ne pas retarder leur mise à disposition pour les patients. Le recours au « *preprints* » a explosé pendant la pandémie de Covid-19, dans un contexte d'urgence et de recherche d'information rapide sur de potentiels traitements.

Cependant l'apparition des « *preprints* » a amplifié l'infodémie et offre une audience potentielle aux études ineptes ou complètement irrecevables méthodologiquement. Pour les publications classiques, le « *peer reviewing* » a certes des limites (sans parler de détournement du « *peer reviewing* », le « *fake peer reviewing* » [138]) et laisse passer des publications ou des études problématiques (tel que l'objectivent les spins de conclusion, les rétractations post publication [139] et les études méta-épidémiologiques montrant les limites des études publiées [15, 140]), mais il effectue néanmoins un certain filtre et évite que de telles études soient trop facilement publiées.

Ces barrières n'existant plus avec les « *preprints* », il faut donc s'attendre à voir surgir des revendications d'efficacité et de place dans la stratégie thérapeutique à partir de publications de cet acabit. La « Publication », qui n'a jamais été d'ailleurs un garant absolu de la validité de l'étude et des résultats présentés [141], ne donne avec ces « *preprints* » plus aucune garantie de sérieux[29]. Il sera donc du ressort du lecteur de juger par lui-même de la fiabilité et de la pertinence des résultats de l'étude rapportée.

Une solution à ce problème serait de systématiquement rejeter les « *preprints* », par principe. Mais cela reviendrait aussi à rejeter d'emblée les apports positifs de cette nouvelle forme de dissémination des résultats de la recherche.

> Le « *preprint* » de l'essai RECOVERY sur la dexaméthasone publié en plein crise COVID le 22 juin 2020[30] illustre parfaitement bien cet atout. Avec ce « *preprint* », la dexaméthasone devenait le premier traitement à avoir démontré une réduction de la mortalité chez les formes sévères. Le « *preprint* » permettait d'appréhender parfaitement bien l'intérêt majeur de ces résultats. La publication « traditionnelle », avalisée par le « *peer reviewing* », n'a été disponible que le 17 juillet 2020 sur NEJM.org en version préliminaire (et le 25 février 2021 ! en version paginée « papier »[31]) et n'apportait rien de plus pouvant faire changer la perception de l'intérêt de ce traitement. La rapidité de la dissémination de ces résultats était d'autant plus justifiée que la dexaméthasone étant disponible, elle permettait une instauration immédiate du traitement chez les patients en soins intensifs dès que la preuve de leur intérêt avait été acquise. En effet, une transparence totale avait été aménagée avec la publication in extenso du protocole et du plan d'analyse statistique. Tout était disponible pour évaluer correctement par soi-même le degré de certitude des résultats et le « *peer reviewing* » n'était pas nécessaire pour servir de tiers de confiance garantissant la solidité méthodologique de l'étude. Dès le communiqué de presse d'ailleurs, il était possible de se faire une idée correcte de l'intérêt clinique de la dexaméthasone.
> Des voix se sont pourtant élevées, y compris en provenance de professionnels médicaux avancés, pour temporiser l'adoption du traitement en arguant qu'il n'y avait pas eu de « *peer reviewing* »,. Cette anecdote reflète le manque d'expertise méthodologique de certains acteurs et les difficultés à décider par eux même directement à partir des résultats « bruts » que peuvent avoir des professionnels pourtant émérites dans leur domaine.

[29] Devant l'explosion de ce phénomène avec la COVID-19, les serveurs de « *preprints* » se sont sentis obligés de rajouter un warning, comme celui de medRxiv : « *Caution: preprints are preliminary reports of work that have not been certified by peer review*. *They should not be relied on to guide clinical practice or health-related behavior and should not be reported in news media as established information.* »
[30] https://www.medrxiv.org/content/10.1101/2020.06.22.20137273v1
[31] https://www.nejm.org/doi/10.1056/NEJMoa2021436

L'exploitation de cette potentialité des « *preprints* » nécessite cependant que les décideurs et prescripteurs puissent juger par eux même de la fiabilité et de la pertinence des résultats « bruts » tels que rapportés dans une publication informative (CONSORT).

> Il est souvent objecté à des résultats uniquement publiés à l'instant t dans un « *preprin »t*, qu'ils sont issus d'un « *preprint* » !
>
> > « L'article cité dans ce post est encore au statut de « *preprint* », aussi faut-il considérer les résultats présentés ici au conditionnel dans l'attente d'une publication définitive »
>
> Le fait d'être publié sous forme de « *preprint »* n'est pas une limitation méthodologique et ne peut pas être un motif de rejet. Lorsque l'étude sera publiée dans une revue traditionnelle, la méthode de l'étude n'aura pas changé, le résultat sera identique. Avoir été revue par un comité de relecture ne change en rien la valeur intrinsèque des résultats d'une étude[32], et il n'est d'ailleurs pas possible d'exclure, que dans certains cas, le reviewing puisse détériorer le papier (l'auteur souhaitant à tout prix être publié peut accepter de suivre des commentaires erronés de reviewers ne disposant pas de toute l'expertise nécessaire). Les « *preprints* » sont une publication anticipée du manuscrit soumis à la revue. Il contient donc, pour les bons papiers, toute l'information nécessaire pour juger de la robustesse des résultats et de leur intérêt pour la pratique. Il est donc tout à fait possible de juger une étude avec le « *preprint* » de la même façon qu'avec une publication d'une revue traditionnelle. À la rigueur on pourrait comprendre que des personnes se sentent mal à l'aise pour suivre une conclusion d'un « *preprint* » car il n'y a pas eu de reviewing et ils ne s'estiment pas en mesure de juger par eux même de la méthode. Mais ce n'est pas une limite du support, mais plutôt du lecteur. Une étude doit être jugée sur le fond et non pas sur la forme (revue ou pas revue par un comité de relecture), sans oublier que le processus de « *peer reviewing »* a aussi des limites (cf. ci-dessus).

Le fait que tout un chacun sera à l'avenir confronté à ce type de publications est un argument supplémentaire pour une meilleure diffusion de la culture du médicament à tous les professionnels de santé et au-delà (grand public, société civile, journalistes, décideurs politiques) ; une certaine expertise sur ces sujets étant déjà nécessaire dans le cadre de la diffusion habituelle des résultats de la recherche thérapeutique comme nous l'avons vu (spin, communication promotionnelle, etc.).

Durant la crise de la COVID, les observateurs ont noté que ce déficit généralisé en culture sur le médicament avait été un facteur important de la confusion et de la désinformation autour de l'efficacité des traitements et des vaccins. [33]

3.4.3 La fraude et les essais zombies

Un autre phénomène a pris récemment de l'ampleur, la fraude scientifique avec la création de publications de A et Z, sans aucune donnée réelle à leur origine, y compris pour les essais thérapeutiques. Il y a ainsi création d'essais que l'on peut qualifier de « zombies » [142]. La fraude scientifique a toujours existé, mais elle était marginale et l'on ne se posait jamais la question de l'invention complète des données devant les papiers. L'ancien éditeur en chef du BMJ, Richard Smith, pense qu'il est temps (juillet 2021) de changer de point de vue devant l'augmentation de la fraude scientifique [143] et d'être plutôt systématiquement suspicieux jusqu'à preuve du contraire.

> *Cette fraude scientifique concernant la création de faux travaux de recherche est à distinguer de la fraude dans les essais cliniques multicentriques qui consiste à l'inclusion de faux patients dans l'essai par certains investigateurs indélicats. Cette fraude n'est pas le fait de l'investigateur principal mais conduit aussi à*

[32] Il est possible que le reviewing constate des défauts (comme le lecteur du preprints) et fasse procéder à des améliorations de méthode, mais dans ce cas le print devient une nouvelle publication (qui n'a pas fait l'objet d'un préprint).

[33] https://sfpt-fr.org/images/covid19/communique-pharmacovid-SFPT.pdf

remettre en cause la réalité des résultats. Ces aspects sont pris en considération de manière standard dans les grands essais industriels à l'aide de différents moyens : bonnes pratiques cliniques, audits sur site de vérification des données saisies par rapport aux dossiers sources des patients, et, in fine, procédure de détection de suspicion de fraude lors de l'analyse [144]. Régulièrement des cas de fraude de ce type sont détectés et solutionnés en retirant le centre dans lequel la fraude a eu lieu. L'exclusion de ces patients ne pose pas de problème méthodologique, car il s'effectue non pas en raison des résultats, mais à la suite d'évènements externes (la détection de la fraude). La randomisation est en plus stratifiée presque toujours sur le centre. Ces exclusions ne remettent pas en cause la puissance statistique lorsque ces patients retirés auront été compensés par des nouveaux inclus dans les autres centres.

> Dans l'essais ESPS-2 [145], une suspicion de fraude fut détectée dans un centre. Tous les patients inclus dans ce centre furent retirés de l'analyse sans compromettre le résultat de l'essai [146]. Les opérations de ce type doivent cependant donner la garantie que l'exclusion de centres ne s'effectue pas en fonction des résultats.

Les premiers à être confrontés à ces essais « zombies », complètement inventés, sont les éditeurs des journaux scientifiques. Récemment la revue *Anesthesia* a quantifié l'importance de ce phénomène dans les soumissions de manuscrits qu'elle reçoit [142] et estime que les essais zombie représentent 8% des soumissions et que 14% contiennent probablement des fausses données.

> Dans le domaine de la COVID, 3 situations de ce type ont défrayé la chronique : les études observationnelles publiées dans le Lancet [147] et le NEJM [148] et un essai randomisé de l'ivermectine [149, 150]. La rétractation de ce dernier essai a entrainé en cascade la rétractation d'au moins une méta-analyse qui l'avait inclus [151][152], illustrant ainsi une nouvelle limite des méta-analyses statiques par rapport aux méta-analyses dynamiques en temps réel (« *living meta-analysis* »).

L'impact est également important sur les méta-analyses. Il devient aussi important de procéder à une veille bibliographique pour intégrer les nouveaux essais que de surveiller les rétractations pour retirer des synthèses et autres méta-analyses les études rétractées qui auraient été retenues au moment de sa réalisation, et de modifier en conséquence les recommandations et les stratégies thérapeutiques correspondantes [153] [154][155].

Chapitre 4. Démarche décisionnelle

Compte tenu des problématiques sous-jacentes à l'évaluation du bénéfice clinique d'un nouveau médicament et de leurs solutions méthodologiques, l'adoption d'un nouveau traitement et la recommandation de son utilisation suivent le processus décisionnel suivant.

4.1 La méthodologie conditionne le niveau de preuve

Les principes méthodologiques des essais cliniques ont été développés pour mettre l'étude à l'abri des biais. Par conséquent, les études qui, comme les études observationnelles, ne relèvent pas de ces principes ne sont pas protégées contre les biais. Ainsi, méthodologiquement, toutes les études ne se valent pas et n'ont pas le même niveau de preuve pour l'évaluation des médicaments. Le niveau de preuve est ainsi une hiérarchisation des types d'études en fonction de leur aptitude intrinsèque à garantir la fiabilité de leurs résultats.

Il s'avère donc 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 [156, 156] et du fait qu'il faudrait d'abord utiliser le nouveau traitement en pratique courante avant de pouvoir évaluer son intérêt (ce qui sous-entend une AMM accordée sur des présuppositions et non sur des faits prouvés[34]).

Régulièrement des essais randomisés échouent à confirmer des bénéfices suggérés par des études observationnelles montrant ainsi le manque de fiabilité de ces études pour cet usage [157, 158, 159, 160, 161, 162, 163].

> 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 a échoué à mettre en évidence un bénéfice [164].

Il existe actuellement de nombreux travaux qui essayent de compenser, par l'analyse statistique, l'absence de contrôle des biais par design des études observationnelles afin que les études observationnelles (de données de « vraie vie ») puissent produire des résultats aussi fiables que les essais randomisés pour l'évaluation de l'efficacité des médicaments [32–38]. À l'avenir, des études observationnelles bâties suivant de nouveaux principes et mettant en œuvre ces techniques modernes d'analyses de données pourront peut-être, avec la disponibilité des données de bonne qualité, produire des résultats de fiabilité suffisante pour guider la construction de la stratégie thérapeutique.

4.2 Les risques d'un accès sans preuve

L'accès sans résultat d'évaluation à des nouvelles propositions thérapeutiques est régulièrement demandé en arguant qu'il n'est pas possible de laisser des patients sans ce traitement dans le groupe contrôle d'un essai. Il est alors avancé qu'à la fin de l'essai il s'avéra que les patients du groupe contrôle

[34] Comme l'évolution du cadre règlementaire vers l'accès précoce

auront été défavorisés par rapport à ceux du groupe traité si le candidat traitement s'avère efficace. Ce raisonnement est bien entendu fallacieux à plusieurs niveaux.

> Par exemple cette demande a été très forte au début de la crise du COVID de la part des patients et des médecins. L'identification rapide *in vitro* de molécules déjà disponibles pouvant avoir une action sur le virus a fait émerger une demande d'utilisation de ces molécules sans attendre la démonstration de leur intérêt clinique dans des essais randomisés en arguant la gravité de la situation, l'urgence et l'absence de traitement spécifique. Un vif débat porté sur la place publique et dans la société civile a d'ailleurs eu lieu. Malgré des appels à la raison [10] et des recommandations qui suivaient scrupuleusement les faits prouvés, un mésusage s'est rapidement instauré avec plusieurs molécules comme l'hydroxychloroquine, l'azithromycine, la vitamine D, etc. [165, 166, 167]. Malgré tout, des groupes collaboratifs ont su garder le cap et réaliser les essais randomisés nécessaires. Ces études ont permis d'identifier des traitements majeurs apportant un réel bénéfice (corticoïdes par exemple) mais ils ont aussi été négatifs pour beaucoup d'autres traitements, illustrant à nouveau les limites des inférences hâtives à partir des mécanismes (dus dans ce cas aux limites des modèles *in vitro* et précliniques, et à des simplifications dans le raisonnement négligeant par exemple les aspects pharmacocinétiques ou les effets indésirables). Malheureusement, cette histoire fournit aussi un exemple d'un des risques majeurs de l'usage compassionnel, celui d'utiliser un traitement avec un effet délétère. En effet la méta-analyse des essais randomisés permet de mettre en évidence une surmortalité avec l'hydroxychloroquine (pas forcément en relation avec une toxicité cardiaque d'ailleurs mais peut être avec une aggravation de la maladie comme le suggère une augmentation de la fréquence de la nécessité d'une ventilation mécanique).
> Si les revendications d'usage compassionnel avaient été suivies, aucun de ces essais randomisés n'aurait donc eu lieu. Il n'aurait pas été possible d'identifier les traitements futiles voir délétères qui seraient toujours utilisés sur la base des arguments précliniques. Les corticoïdes n'auraient pas pu s'imposer universellement non plus, car il existait a priori des réticences à leur utilisation basées sur les résultats de leur évaluation dans d'autres situations d'infection pulmonaire virale (grippe, MERS [168, 169, 170]).

Effectivement, avec un traitement apportant un bénéfice, il s'avèrera a postériori que les patients du groupe contrôle ont été défavorisés durant la réalisation de l'essai par rapport à ceux du groupe traité, contrairement à l'usage compassionnel où aucun patient n'aura été privé du candidat médicament.

Mais ce raisonnement ne tient que si le traitement apporte effectivement un bénéfice. Or, on sait que cela est loin d'être le cas pour toutes les nouvelles propositions thérapeutiques. Il a été montré qu'environ la moitié des nouveaux traitements ne confirme pas le bénéfice escompté lors de leurs phases 3 [25]. En cas d'utilisation compassionnelle de ces traitements, la situation devient identique à ceux des patients du groupe contrôle des essais de traitements apportant un bénéfice, voire pire si ces traitements ont une balance bénéfice-risque défavorable (présence d'effets indésirables sans bénéfice) !

Ainsi, dans les deux approches, il existe la possibilité que des patients soient défavorisés en termes de traitement. Cependant, il existe une différence fondamentale. Dans l'essai, l'utilisation pour certains patients de traitements les défavorisant par rapport aux autres ne dure qu'un temps (celui de la réalisation de l'essai) et sert à quelque chose : à apporter la preuve de l'intérêt ou non du nouveau traitement. Après l'obtention de la démonstration du bénéfice, plus aucun patient ne recevra le traitement le moins favorable. Dans l'usage compassionnel, les situations défavorisant les patients durent *ad vitam aeternam* (car par principe l'intérêt de ces traitements ne sera jamais remis en cause) et ne servent à rien, car ne débouchent pas sur la validation ou la réfutation du bénéfice du traitement (aucun essai ne sera réalisé par principe).

Ainsi l'approche compassionnelle est non opérante, et donc inacceptable. Elle expose les patients au risque d'être défavorisé par le traitement reçu tout autant que l'essai comparatif. Ce risque peut dans

certaines situations paraître minime, notamment lorsque l'espérance de vie est très courte. Ce raisonnement motive d'ailleurs l'usage de plus en plus répandu des essais mono-bras (cf. dossier compagnon n°12) en oncologie. Toutefois, contrairement à l'essai contrôlé, aucune information sur le réel bénéfice du traitement ne sera produite. L'usage compassionnel ne serait plus performant que l'essai qu'à la condition que toutes les nouvelles propositions thérapeutiques apportent un bénéfice net, c'est-à-dire que le préclinique est capable de prédire correctement le bénéfice clinique, et que ce dernier soit supérieur au risque. Ce qui est loin d'être le cas. De plus, l'essai randomisé donne une chance à tous les patients de recevoir le meilleur traitement tandis qu'avec l'usage compassionnel c'est tout-ou-rien en fonction de la chance que le traitement retenu soit réellement bénéfique sans être délétère.

> Toujours dans l'exemple de la COVID, il est apparu que la majorité des traitements envisagés pour un usage compassionnel se sont avérés être sans intérêt ou même délétères lorsque les résultats des essais cliniques ont été disponibles (par exemple hydroxychloroquine, azithromycine, lopinavir/ritonavir, plasma de patients convalescents, anticoagulants à dose curative dans les formes sévères, etc.). L'inscription stricte dans ce principe de l'accès compassionnel aurait fait que ces essais n'auraient jamais eu lieu et ainsi, de nombreux traitements continueraient à être utilisés à tort.

Il faut aussi noter que lorsque l'on parle de patients « défavorisés » dans l'essai, c'est par rapport à ceux recevant l'autre traitement (par exemple les patients ayant reçu le traitement standard par rapport à ceux ayant reçu le nouveau traitement si celui s'avère efficace). Dans un essai, les patients participants ne sont jamais défavorisés par rapport à ceux ne participant pas à l'essai, car tous les patients inclus reçoivent le traitement standard (y compris dans le groupe placebo s'il y a un groupe placebo). Cela signifie plutôt qu'il s'avèrera, a postériori, à la fin de l'étude que ces patients n'ont pas reçu le meilleur des 2 traitements comparés. Mais cela reste un enseignement a postériori et, a priori, il y a équiprobabilité pour tous les patients d'être inclus dans le groupe le plus favorable : le principe de l'équipoise est donc respecté, aucune des deux options n'a a priori d'avantage sur l'autre. C'est d'ailleurs cette situation qui rend éthique la réalisation des essais cliniques. Divers résultats suggèrent, de plus, que les patients participant aux essais sont mieux pris en charge que ceux ne participant pas [171, 172, 173].

Pour finir, l'usage compassionnel revient donc à inclure des patients dans un essai implicite (sans les avertir et sans recueillir leur consentement) qui les expose, tout autant qu'un essai randomisé, à ne pas recevoir le meilleur traitement, sans que cela ne serve les futurs patients ou qu'il puisse être possible de faire marche arrière.

4.3 Une étude préliminaire (de phase 2) peut-elle produire une preuve suffisante ?

Des études de phase 2[35] de bonne facture méthodologique (contrôlées, randomisées, en ouvert ou en double aveugle) sont parfois mises en avant pour justifier de l'intérêt clinique du traitement.

[35] L'appellation des phases est parfois trompeuse. Les phases 2 sont théoriquement des études préliminaires destinées à préparer les études dont la finalité est la démonstration du bénéfice : les phases 3. De ce fait la méthodologie des phases 2 est moins rigoureuse que celle des phase 3. Par collision des concepts, d'éventuelles études de confirmation (donc qui sont des phases 3 par définition) mais qui utilisent la méthodologie habituelle des phase 2, sont appelées à tort phase 2 (cf. page 51). Dans cette section il s'agit des véritables phases 2, à savoir des études préliminaires.

Nonobstant la rigueur de la méthodologie, ce type de résultats ne permet pas de valider l'adoption du nouveau médicament pour de nombreuses raisons.

> L'essai CABOSUN [174] du cabozantinib dans le carcinome rénal métastatique à haut risque ou risque intermédiaire est une étude de phase 2 qui a été finalement utilisée pour justifier l'utilisation de cette molécule dans cette condition clinique[36]. L'essai était randomisé, en ouvert versus sunitinib, incluant 157 patients suivis 21.4 mois (médiane). L'essai n'ayant pas été conçu comme un essai pivot de confirmation, le risque alpha consenti était de 12% unilatéral (24% bilatéral). L'analyse présentée dans la publication utilise des règles de censure simples non conformes avec les recommandations réglementaires.

Comme à l'origine ces études, préparatoires à la phase 3, ne sont pas destinées à apporter la preuve formelle de l'intérêt du traitement, le risque alpha fixé au protocole est souvent supérieur à 5% bilatéral (classiquement 20%). En effet, les faux positifs dus au hasard dans ces études préliminaires n'exposent qu'au risque de réaliser et financer à tort l'étude de phase 3. Et dans ce contexte un tel risque est donc parfois consenti par les industriels pour la décision de go/no go vers la phase 3. En revanche, il est évident qu'un tel niveau de risque est inacceptable dans un contexte de recommandation et d'utilisation d'un nouveau médicament.

Baser la décision sur une phase 2 à la place d'une phase 3 pivot signifie que cette décision est prise sur la base d'une seule étude concluante à la place de deux (ou plus, par exemple dans les domaines où deux phases 3 sont exigées). Il y a donc un manque de vérification qui n'est pas superflu, car d'un point de vue méta-épidémiologique il s'avère qu'une étude de phase 3 sur deux est un échec [25]. Accepter un produit à partir d'une simple étude de phase 2 constitue donc une prise de risque importante.

Ce faible taux de confirmation des résultats de phase 2 par les phases 3 peut s'expliquer par de nombreux facteurs : limites de critères utilisés en phase 2, risque alpha, un risque de biais souvent important avec des études non randomisées ou en ouvert, analysées en per protocole, etc.

Les études de phase 2 ne peuvent pas non plus être considérées comme des preuves au-delà de tout doute raisonnable pour des raisons d'assurance qualité. N'ayant pas comme finalité d'apporter ce type de preuve, ces études sont réalisées avec des standards d'assurance qualité souvent plus légers que celui des études de phase 3 en termes de monitorage, de contrôles des données, etc. Au final, même si les données recueillies sont de qualité, il n'est pas possible de le garantir.

Le nombre de patients et la durée de suivi sont généralement faibles, si bien qu'à l'issue de l'étude de phase 2, l'appréciation de la sécurité reste très limitée, ainsi que celle de l'évolution de l'efficacité au cours du temps. La représentativité des patients est aussi fortement handicapée par ce point. Dans ces conditions (faible puissance pour le critère considéré, absence d'hypothèse spécifique), un résultat $p<0.05$ a une faible valeur prédictive de l'effet du traitement (cf. chapitre sur l'inférence bayésienne).

L'utilisation d'une étude de phase 2 pour revendiquer l'intérêt clinique du médicament est souvent décidée devant un effet traitement très important, faisant penser qu'une confirmation par l'étude de phase 3 est inutile et que retarder l'accès des patients à ce traitement induit une perte de chance. Ce point est contredit par la méta-épidémiologie qui montre une faible valeur prédictive des « *very large treatment effect* » observée en phase précoce [64] (cf. section 2.6, page 17). Comme les études de phase 2 sont par nature exploratoires, ce type de résultat est presque toujours obtenu sur des critères de jugement annexes, non inclus dans un plan de contrôle du risque alpha global et en dehors de toute démarche hypothético-déductive. Les résultats méta-épidemiologiques déjà cités montrent que la

[36] Comme en témoigne la soumission d'un dossier à la commission de transparence https://www.has-sante.fr/upload/docs/evamed/CT-17224_CABOMETYX_PIC_EI_Avis3_CT17224.pdf

taille de l'effet (et/ou l'importance clinique du critère, cf. section 2.6) n'affranchit pas ces résultats de leurs limites méthodologiques.

La décision d'utiliser une étude de phase 2 dépend la plupart du temps de ses résultats. Il s'agit d'une démarche entièrement guidée par les résultats qui présentent ainsi les limites méthodologiques des choix *post hoc* et de l'absence de démarche confirmatoire (cf. section 2.3 et Chapitre 3). La démarche est purement opportuniste en saisissant l'occasion d'exploiter un résultat non initialement prévu pour cela et ce en fonction de ce qu'il permet (« *cherry picking* »).

La mise à disposition du médicament va aussi fortement handicaper la réalisation des études de confirmation (même si celles-ci sont exigées en cas d'enregistrement précoce) [175]. Pire, même en cas de résultats négatifs les produits peuvent rester enregistrés et recommandés, par exemple dans 30% des cas en oncologie [175].

4.4 Mélanges des genres

Un autre élément trivial renforce l'importance de la méthodologie dans l'évaluation des médicaments. En effet, il ne faut pas oublier que, sauf cas exceptionnel, le médicament est un objet commercial. Il est donc tout naturel que les fabricants fassent tout leur possible pour faire acheter leurs produits en s'appuyant sur des arguments commerciaux. La méthodologie permet de délimiter le champ des arguments possibles et d'éviter que ce produit particulier fasse l'objet d'une communication promotionnelle non fondée, voire même trompeuse. Ainsi, dans beaucoup de pays, la communication sur le médicament est strictement encadrée et doit s'appuyer sur des faits prouvés. Cependant, tout est fait pour optimiser la communication à l'intérieur de ces contraintes, compte-tenu des intérêts sous-jacents. Sans enfreindre les règles, une partie de la communication promotionnelle va alors s'appuyer sur des arguments litigieux du point de vue méthodologique. Un regard critique est de ce fait nécessaire.

La réalisation des essais par les industriels eux même (ou des sociétés sous contrat) est quelquefois vue comme un élément rédhibitoire pouvant compromettre la fiabilité des résultats produits. La méthodologie permet d'exclure cette éventualité dans le cadre des essais d'enregistrement. Si les principes méthodologiques sont correctement appliqués et réalisés, les résultats de l'étude ne peuvent être influencés sur le plan de leur validité. Quels que soient les intérêts des personnes impliquées, la méthodologie garantit que les résultats resteront inchangés. Ces considérations ont d'ailleurs prévalu lors de l'élaboration, au cours du temps, des principes de l'essai thérapeutique moderne. Les principes méthodologiques (associés au système d'assurance qualité et au monitorage qui vérifie qu'ils sont correctement appliqués) ont donc aussi cette vertu de rendre recevables les développements industriels.

> Une partie de l'exigence d'un système assurance qualité dans les essais provient de l'histoire de l'essai UGDP [176, 177, 178]. L'essai UGDP (*University Group Diabetes Program*) avait pour objectif de montrer que la baisse provoquée de la glycémie par un quelconque moyen médicamenteux dans le diabète de type 2 prévenait les évènements cardiovasculaires. Les résultats ont été publiés en décembre 1970 et ne montraient pas de réduction de la fréquence des évènements cardiovasculaires et même un effet délétère important du tolbutamide. Ces résultats ont alors conduit à une bataille scientifique, médiatique et judiciaire épique autour de la qualité des données et de la validation de la cause des décès.

L'influence des conflits d'intérêts n'est donc pas tant à craindre aux niveaux de la production des résultats que de la conception de l'essai en terme de pertinence du comparateur, du critère de jugement, de la définition de la population ciblée, du choix de faire un essai de non-infériorité à la

place d'un essai de supériorité, de pertinence médicale de l'objectif, etc. [179], toutes choses évaluables ensuite lors de l'interprétation de la publication. Par exemple, certains essais (appelé « seeding trials »), sous des apparences scientifiques, ont en réalité des objectifs purement marketing [180, 181].

4.5 Pourquoi tous les acteurs du médicament doivent avoir des compétences méthodologiques minimales ?

Les médecins et les autres acteurs du médicament sont régulièrement confrontés aux résultats des essais cliniques et à leurs aspects méthodologiques, en particulier lors de la présentation d'un nouveau traitement, lors de la communication promotionnelle, dans le cadre de la formation continue, etc.

Il devient indispensable pour tous ces acteurs du médicament de pouvoir évaluer par eux-mêmes la solidité des arguments mis en avant pour inciter à un changement de stratégie thérapeutique ou pour justifier l'adoption d'un nouveau traitement. Sans l'expertise nécessaire, le lecteur risque de se méprendre sur le réel intérêt d'un nouveau traitement en tombant dans un des nombreux pièges qui lui sont tendus dans la littérature et la communication des résultats des essais cliniques : spins de conclusions en particulier, production de masse d'études de faible qualité méthodologique, résultats d'études de faible niveau de preuve, arguments reposant sur des raisonnements insuffisamment solides, etc.

Dans ce cadre, la finalité de la lecture de ces articles (ou de résultats extraits de ces articles) est très pratique : déterminer ce qu'apporte le nouveau traitement évalué, pour savoir s'il doit être utilisé ou non en pratique courante.

Même si la décision finale concernant un nouveau traitement viendra des agences de régulation et des sociétés savantes, il est utile que chaque médecin (qui est formé à un niveau doctorat) puisse comprendre et anticiper ces décisions afin d'éviter que son activité ne se résume à appliquer des recommandations dont il ne comprendrait pas la justification. La crise sanitaire de la COVID a bien mis en évidence que tout ne pouvait pas reposer sur la régulation et les recommandations.

4.6 Une présomption de bénéfice est insuffisante

L'obtention de la preuve du bénéfice est présentée de plus en plus comme étant superfétatoire, difficile et couteuse à obtenir, et inutile, privant les patients de traitements innovants qui ne peuvent pas, soi-disant, produire la preuve jusqu'ici demandée. Un débat s'est ouvert où il est avancé qu'une simple présomption de bénéfice serait suffisante et permettrait alors aux patients d'avoir accès à ces traitements sans retard.

La présomption de bénéfice n'est définie nulle part mais correspond à une situation d'absence de preuve formelle. Cette présomption est alors fondée sur différents types d'arguments ou de résultats :

- Rapports anecdotiques
- Mécanisme d'action prometteur
- Résultats montrant un effet pharmacologique, résultats d'essais sur un critère intermédiaire
- Argumentation non contrefactuelle, basée sur une série de patients, une étude mono-bras
- Résultats d'une phase 2 (non pivotale)

- Résultats *post hoc*, de découverte fortuite (sous-groupe, critères de jugement secondaires ou exploratoires)
- Études observationnelles sans inférence causale et ne faisant pas la preuve de sa fiabilité ; études rétrospectives
- Résultats exploratoires issus d'une méta-analyse sans essai concluant par lui-même
- Résultat nominalement significatif mais non significatif en termes de risque alpha global
- Résultats non nominalement statistiquement significatif d'un essai randomisé, mais dont la tendance a un « sens clinique »
- Résultats d'une analyse intermédiaire non concluante

Chacun des points de cette liste correspond à des situations qui ont été identifiées dans l'histoire du développement de la méthodologie comme étant problématiques, pouvant être faussement positives (en faveur du bénéfice du traitement) et conduire à une décision erronée.

Bien entendu lorsque le traitement apporte un bénéfice clinique des arguments sont vérifiés, mais la question est bien leur aptitude à ne pas faillir et à donner suffisamment de garanties sur l'existence du bénéfice clinque *in fine*.

Les principes méthodologiques ne sont pas idéologiques. Ils représentent les solutions aux problématiques que soulève l'évaluation du bénéfice clinique d'un nouveau médicament.

Tableau 1 – Limites et insuffisance des différents arguments avancés comme présomption de bénéfice clinique

Type d'arguments insuffisamment solides	Limite(s) potentielle(s)	Solution
Avis d'expert (non factuel)	Subjectivisme ; spéculatif	Vérification des hypothèses par une étude *ad hoc* (raisonnement déductif)
Résultat anecdotique, étude de cas	Raisonnement du type « *post hoc proper ergo hoc* » ; absence de raisonnement contrefactuel ; pas d'évaluation de la balance bénéfice-risque	Vérification des hypothèses par une étude *ad hoc* (raisonnement déductif)
Résultats *post hoc* (quel que soit le type d'étude), découverte fortuite	Raisonnement inductif ; résultat *post hoc* ; multiplicité ; « harking »; potentiellement renforcés par le biais de publication et le « *selective reporting* »	Vérification des hypothèses par une étude *ad hoc* (raisonnement déductif)
Fouille de donnée (« *data mining* » sur « *big data* » par exemple)	Raisonnement inductif ; Identification des artefacts ; multiplicité ; « harking » ; « p-hacking », « data dredging » et « cherry picking »	Vérification des hypothèses par une étude *ad hoc* (raisonnement déductif)
Mécanisme d'action, effet pharmacologique	Pas une finalité clinique en soi ; manque de prédictibilité sur le bénéfice clinique ; pas d'évaluation des risques et de la balance bénéfice-risque	Vérification par les faits que le mécanisme ou l'effet pharmacologique se traduit bien chez les patients en bénéfice clinique
Critère intermédiaire	Pas de pertinence clinique en soi. Plus proche de l'effet pharmacologique que de l'effet clinique. Les effets pharmacologiques ne prédisent pas systématiquement les bénéfices cliniques ; pas d'évaluation des risques et de la balance bénéfice-risque	Démonstration de la valeur de substitution (« *surrogacy* ») du critère ou essais sur critères cliniques (morbi-mortalité)
Étude rétrospective (quel que soit le type d'étude : analytique, descriptive avec une comparaison externe, méta-analyse, etc.)	Fouille possible, « harking » ; multiplicité ; « p-hacking » ; biais de publication, « *selective reporting* »	Garantir qu'il s'agit d'une réelle validation prospective d'une nouvelle hypothèse sur des données historiques (rétrospectives) ; garantir que l'hypothèse n'a pas été formulée à partir des mêmes données (protocole enregistré, SAP, enregistrement a priori des critères cliniques de succès, etc.)
Argument issu d'une démarche rétrospective (sur données historiques)	Démonstration de l'absence de biais de confusion résiduelle gênant ; ajustement raisonné (DAG) ; contrôle du biais de sélection (DAG) ; design d'émulation d'essai cible ; étude multi cohorte ; inférence causale	À confirmer par une démarche prospective ou par une vraie démarche prospective-rétrospective [182]
Résultat non comparatif (absence de groupe contrôle : étude mono-bras)	Absence de raisonnement contrefactuel direct	Étude comparative à contrôle externe a minima, mais ne remplace par un RCT
Étude observationnelle d'efficacité (sans preuve d'absence de biais de confusion résiduelle)	Biais de confusion ; autres biais (sélection, mesure, classification, etc.)	Études observationnelle reposant sur une approche d'inférence causale, émulant un essai cible et faisant la preuve d'absence de biais de confusion résiduelle

Type d'arguments insuffisamment solides	Limite(s) potentielle(s)	Solution
Phase 2 non pivotale	Souvent méthodologie et qualité de réalisation dégradées (risque alpha global non contrôlé, règles de censure inappropriées, absence de monitorage, standard d'assurance qualité inférieure aux études pivotales) ; Absence de réelle étude de confirmation (phase pivotale) ; évaluation de la sécurité limitée	Réalisation d'une étude pivotale dans un contexte rendant possible sa réalisation.
Étude isolée ; revue de la littérature ou méta-analyse non exhaustive (n'ayant pas cherché les études non publiées)	Biais de publication (sauf si l'étude est unique mais seule une revue systématique ou un plan de développement permet de l'attester) ou sélection d'études d'après leurs résultats (et non pas leur qualité méthodologique). Possibilité de sacralisation d'un faux positif	Décision issue d'un processus « d'*evidence synthesis »* (revue systématique et éventuellement méta-analyse)
Décision après analyses intermédiaires d'efficacité non prévues et non formalisées	Multiplicité, inflation du risque alpha non contrôlé et surtout « p-hacking » (décision d'arrêter l'étude et de décider à partir des résultats produits entièrement guidée par les résultats eux-mêmes, sophisme, tautologie)	Analyse des résultats d'efficacité avant le terme prévu de l'étude uniquement dans un cadre prévu à priori et formalisé avec une méthode de contrôle du risque alpha global (ou si un événement externe justifie l'arrêt de l'étude)
Décision après analyses intermédiaires d'efficacité non concluantes (mais nominalement significatives)	Inflation du risque alpha, avec un risque alpha global non contrôlé ;	La signification nominale ne signifie rien avec les plans modernes de contrôle du risque alpha global. Conclusion au bénéfice que lorsque l'analyse intermédiaire est significative suivant le seuil ajusté ; absence de communication des résultats intermédiaires non concluants
Résultat nominalement « statistiquement significatif » mais non significatif en termes de risque alpha global	Le risque de conclure à tort à l'intérêt du traitement n'est plus contrôlé. Considérer ce résultat revient à ne plus prendre en considération la problématique des différences dues au hasard et des erreurs statistiques.	Ne pas rapporter les p values des résultats non significatifs en termes de risque alpha global. Réplication à entreprendre de manière *ad hoc*
Résultat non significatif mais médicalement « important »	Le résultat est non significatif, le risque de conclure à tort à l'intérêt du traitement n'est plus contrôlé. La pertinence ou l'importance médicale du résultat ne lève pas cette problématique.	Aucune. Les résultats non significatifs ne permettent pas de conclure compte-tenu du choix anticipé du risque maximal consenti de conclure à tort à l'intérêt du traitement (cf. principes fondamentaux de l'inférence statistique)

DAG : Directed acyclic graph

Une autre limite des éléments de présomption est qu'étant spéculatifs par nature et ayant chacun des limites intrinsèques, ils offrent la possibilité de débat sans fin entre experts. Le décideur est alors soumis à une grande variabilité du discours entre experts et d'un jour à l'autre. Aucun de ces arguments, n'ayant valeur de preuve, n'est donc en mesure de statuer définitivement et mettre un terme aux interprétations et spéculations. L'intérêt des preuves de haut niveau est aussi leur aspect opposable, qui s'impose à nous indépendamment d'un schéma de pensée ou d'opinion. L'étude répond directement à la question qui se pose (si l'étude est bien conçue et réalisée) indépendamment de tout raisonnement interprétatif ou spéculatif.

Les éléments de présomption de bénéfice sont presque toujours focalisés sur l'efficacité et n'abordent pas le risque. Ils induisent donc la recherche de démonstration d'une efficacité plus que d'un bénéfice clinique net (qui intègre le risque). Les raisonnements mécanicistes sous-estiment systématiquement le risque des molécules. Les résultats des études intermédiaires ou non comparatives ne sont pas en mesure de documenter correctement le risque.

4.7 Schéma décisionnel à partir des essais méthodologiquement solides

La finalité des essais thérapeutiques (pivots, de « phase 3 ») est d'apporter la **preuve** (« *evidence* ») de **l'intérêt clinique** (« *effectiveness, utility* ») du nouveau traitement au-delà de tout doute raisonnable.

- Si c'est le cas, le nouveau traitement pourra être introduit dans la stratégie thérapeutique de la pathologie.

> L'essai Aristotle [183] 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 le 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 que dans le cadre d'autres essais thérapeutiques. Éventuellement, si un bénéfice 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

Initialement, le but des essais de phase 2 est de préparer la réalisation de ces essais de phase 3 (recherche de la dose, exploration des effets, etc.). Dans ces essais la méthodologie peut être moins rigoureuse car le risque est de réaliser une phase 3 à tort et non pas de recommander à tort l'utilisation d'un nouveau traitement. On rencontre ainsi des seuils de signification statistique à 20% et le groupe contrôle, ou la randomisation, sont fréquemment absents. Ainsi, par abus de langage, phase 2 est presque devenue synonyme d'essai non comparatif.

L'ambiguïté terminologique survient du fait que, parfois, certains plans de développement prévoient d'utiliser des essais non comparatifs comme preuve définitive de l'intérêt du traitement. Aucun essai contrôlé randomisé n'est alors prévu. Dans ce type de plan, la phase 3 est ainsi une étude non contrôlée mais elle est presque toujours dénommée phase 2 eu égard à sa méthodologie qui jusqu'à présent ne se rencontrait que dans les phases préliminaires de phase 2[37] (cf. dossier compagnon n° 14). On note cependant une tendance récente à la correction de cette dérive avec des développements qui présentent clairement une étude mono-bras comme phase 3 (cf. par exemple NCT03312634). Ainsi le numéro de phase correspond à la fonction et non pas à la méthodologie. Il existe par exemple des études de phase 2/3 qui répondent par la même étude (même méthodologie) aux deux besoins (cf. dossier compagnon n°14). Parallèlement à cela on assiste aussi avec l'enregistrement accéléré et l'accès précoce à l'utilisation de vraies phases 2 (une phase 3 est prévue dans le plan de développement) comme justification de l'intérêt clinique du traitement.

L'intérêt clinique du traitement est démontré quand le résultat est fiable, statistiquement significatif en termes de contrôle du risque alpha global[38] **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 par exemple ou par rapport à un traitement contrôle non optimal. 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 BLAZE 1 partie 1 [184] (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 de 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.
> 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)

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.

[37] Le risque d'un résultat faux positif en phase 2 est de conduire à la réalisation à tort d'une phase 3 qui s'avérera négative par la suite et non pas de recommander à tort le nouveau traitement. Il s'agit d'un risque couru par le développeur (l'industriel) et non pas par le patient. Cependant, on assiste depuis quelques années à un renforcement de la méthodologie des phases 2 pour limiter le risque financier de réaliser pour rien une phase 3 mais aussi pour éviter d'accaparer pour rien des patients (devenu rare compte tenu du nombre d'essai en cours dans chaque pathologie) et d'handicaper le recrutement d'autres essais.

[38] Et pas seulement nominalement significatif

Références

1 Prasad VK. Malignant: How bad policy and bad evidence harm people with cancer. Baltimore: Johns Hopkins University Press 2020 ISBN:9781421437637;

2 Kozauer N, Katz R. Regulatory innovation and drug development for early-stage Alzheimer's disease. *N Engl J Med* 2013;368:1169–71 doi:10.1056/NEJMp1302513; PMID:23484795;

3 (2021). F.D.A. to Ease Alzheimer's Drug Approval Rules - The New York Times. *New York Times,* 15 November 2021. Available at: https://www.nytimes.com/2013/03/14/health/fda-to-ease-alzheimers-drug-approval-rules.html?pagewanted=all Accessed November 15, 2021.

4 Hwang TJ, Ross JS, Vokinger KN, et al. Association between FDA and EMA expedited approval programs and therapeutic value of new medicines: retrospective cohort study. *BMJ* 2020;371:m3434 doi:10.1136/bmj.m3434; PMID:33028575;

5 Gyawali B, Hey SP, Kesselheim AS. Assessment of the Clinical Benefit of Cancer Drugs Receiving Accelerated Approval. *JAMA Internal Medicine* 2019;179:906–13 doi:10.1001/jamainternmed.2019.0462; PMID:31135808;

6 Naci H, Davis C, Savović J, et al. Design characteristics, risk of bias, and reporting of randomised controlled trials supporting approvals of cancer drugs by European Medicines Agency, 2014-16: cross sectional analysis. *BMJ* 2019;366:l5221 doi:10.1136/bmj.l5221; PMID:31533922;

7 Schnog J-JB, Samson MJ, Gans ROB, et al. An urgent call to raise the bar in oncology. *Br J Cancer* 2021 doi:10.1038/s41416-021-01495-7; PMID:34400802;

8 Del Paggio JC, Tannock IF. The fragility of phase 3 trials supporting FDA-approved anticancer medicines: a retrospective analysis. *The Lancet Oncology* 2019;20:1065–69 doi:10.1016/S1470-2045(19)30338-9;

9 Tannock IF, Amir E, Booth CM, et al. Relevance of randomised controlled trials in oncology. *The Lancet Oncology* 2016;17:e560-e567 doi:10.1016/S1470-2045(16)30572-1; PMID:27924754;

10 Zagury-Orly I, Schwartzstein RM. Covid-19 - A Reminder to Reason. *N Engl J Med* 2020 doi:10.1056/NEJMp2009405; PMID:32343505;

11 Ladanie A, Schmitt AM, Speich B, et al. Clinical Trial Evidence Supporting US Food and Drug Administration Approval of Novel Cancer Therapies Between 2000 and 2016. *JAMA Netw Open* 2020;3:e2024406 doi:10.1001/jamanetworkopen.2020.24406; PMID:33170262;

12 Salas-Vega S, Iliopoulos O, Mossialos E. Assessment of Overall Survival, Quality of Life, and Safety Benefits Associated With New Cancer Medicines. *JAMA Oncol* 2017;3:382–90 doi:10.1001/jamaoncol.2016.4166; PMID:28033447;

13 Tannock IF, Templeton AJ. Flawed trials for cancer. *Annals of Oncology* 2020;31:331–33 doi:10.1016/j.annonc.2019.11.017; PMID:32067676;

14 Pundi K, Perino AC, Harrington RA, et al. Characteristics and Strength of Evidence of COVID-19 Studies Registered on ClinicalTrials.gov. *JAMA Intern Med* 2020 doi:10.1001/jamainternmed.2020.2904; PMID:32730617;

15 Raynaud M, Zhang H, Louis K, et al. COVID-19-related medical research: a meta-research and critical appraisal. *BMC Med Res Methodol* 2021;21:1 doi:10.1186/s12874-020-01190-w; PMID:33397292;

16 van Nguyen T, Rivière P, Ripoll P, et al. Research response to coronavirus disease 2019 needed better coordination and collaboration: a living mapping of registered trials. *Journal of Clinical Epidemiology* 2020;130:107–16 doi:10.1016/j.jclinepi.2020.10.010; PMID:33096223;

17 Tibau A, Molto C, Borrell M, et al. Magnitude of Clinical Benefit of Cancer Drugs Approved by the US Food and Drug Administration Based on Single-Arm Trials. *JAMA Oncol* 2018;4:1610–11 doi:10.1001/jamaoncol.2018.4300; PMID:30267037;

18 Chen EY, Raghunathan V, Prasad V. An Overview of Cancer Drugs Approved by the US Food and Drug

Administration Based on the Surrogate End Point of Response Rate. *JAMA Internal Medicine* 2019;179:915–21 doi:10.1001/jamainternmed.2019.0583; PMID:31135822;

19 Walsh S, Merrick R, Milne R, et al. Aducanumab for Alzheimer's disease? *BMJ* 2021;374:n1682 doi:10.1136/bmj.n1682; PMID:34226181;

20 Alexander GC, Emerson S, Kesselheim AS. Evaluation of Aducanumab for Alzheimer Disease: Scientific Evidence and Regulatory Review Involving Efficacy, Safety, and Futility. *JAMA* 2021;325:1717–18 doi:10.1001/jama.2021.3854; PMID:33783469;

21 CardioBrief: FDA's Gottlieb Preparing To Lower The Bar To Approval 2017. Available at: https://www.medpagetoday.com/cardiology/cardiobrief/68224 Accessed November 15, 2021.

22 CardioBrief. Cardiac Devices Could Become a Big Problem For Califf And The FDA 2016. Available at: http://www.cardiobrief.org/2016/11/07/cardiac-devices-could-become-a-big-problem-for-califf-and-the-fda/ Accessed November 15, 2021.

23 Effects of enalapril on mortality in severe congestive heart failure. Results of the Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS). *New Engl J Med* 1987;316:1429–35 doi:10.1056/NEJM198706043162301; PMID:2883575;

24 Robert C, Long GV, Brady B, et al. Nivolumab in previously untreated melanoma without BRAF mutation. *N Engl J Med* 2015;372:320–30 doi:10.1056/NEJMoa1412082; PMID:25399552;

25 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;176:1826–33 doi:10.1001/jamainternmed.2016.6008; PMID:27723879;

26 Gigerenzer G. We need statistical thinking, not statistical rituals. *Behav Brain Sci* 1998;21:199–200 doi:10.1017/S0140525X98281167;

27 Marks H. La médecine des preuves: Histoire et anthropologie des essais cliniques / 1900-1990. Le Plessis-Robinson: Institut Synthélabo pour le progrès de la connaissance 1999 ISBN:2-84324-044-1;

28 FR_ALLEA_Code_de_conduite_europeen_pour_lintegrite_en_recherche.

29 Boutron I, Dutton S, Ravaud P, et al. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303:2058–64 doi:10.1001/jama.2010.651; PMID:20501928;

30 Khan MS, Lateef N, Siddiqi TJ, et al. Level and Prevalence of Spin in Published Cardiovascular Randomized Clinical Trial Reports With Statistically Nonsignificant Primary Outcomes: A Systematic Review. *JAMA Netw Open* 2019;2:e192622 doi:10.1001/jamanetworkopen.2019.2622; PMID:31050775;

31 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:e1000354 doi:10.1371/journal.pmed.1000354; PMID:21048986;

32 Ioannidis JPA. Hundreds of thousands of zombie randomised trials circulate among us. *Anaesthesia* 2020 doi:10.1111/anae.15297; PMID:33124075;

33 Angell M. The truth about the drug companies: How they deceive us and what to do about it. New York: Random House 2005 ISBN:9780375760945;

34 Kassirer JP. On the take: How medicine's complicity with big business can endanger your health. Oxford, New York: Oxford University Press 2005 ISBN:9780195300048;

35 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:e1000354 doi:10.1371/journal.pmed.1000354; PMID:21048986;

36 Hilal T, Gonzalez-Velez M, Prasad V. Limitations in Clinical Trials Leading to Anticancer Drug Approvals by the US Food and Drug Administration. *JAMA Intern Med* 2020;180:1108–15 doi:10.1001/jamainternmed.2020.2250; PMID:32539071;

37 Hilal T, Sonbol MB, Prasad V. Analysis of Control Arm Quality in Randomized Clinical Trials Leading to Anticancer Drug Approval by the US Food and Drug Administration. *JAMA Oncol* 2019;5:887–92 doi:10.1001/jamaoncol.2019.0167; PMID:31046071;

38 Mohyuddin GR, Koehn K, Sborov D, et al. Quality of control groups in randomised trials of multiple myeloma enrolling in the USA: a systematic review. *The Lancet Haematology* 2021;8:e299-e304 doi:10.1016/S2352-3026(21)00024-7;

39 Prasad V, Kim C, Burotto M, et al. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. *JAMA Internal Medicine*

2015;175:1389–98 doi:10.1001/jamainternmed.2015.2829; PMID:26098871;

40 Carpenter DP. Reputation and power: Organizational image and pharmaceutical regulation at the FDA. Princeton (N.J.), Oxford: Princeton University Press 2010 ISBN:9780691141794;

41 Temple R. Policy developments in regulatory approval. *Stat Med* 2002;21:2939–48 doi:10.1002/sim.1298; PMID:12325110;

42 Landray MJ, Haynes R, Hopewell JC, et al. Effects of extended-release niacin with laropiprant in high-risk patients. *N Engl J Med* 2014;371:203–12 doi:10.1056/NEJMoa1300955; PMID:25014686;

43 Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *New Engl J Med* 1991;324:781–88 doi:10.1056/NEJM199103213241201; PMID:1900101;

44 Nissen SB, Magidson T, Gross K, et al. Publication bias and the canonization of false facts. *eLife* 2016;5 doi:10.7554/eLife.21451; PMID:27995896;

45 Song F, Parekh-Bhurke S, Hooper L, et al. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC Med Res Methodol* 2009;9:79 doi:10.1186/1471-2288-9-79; PMID:19941636;

46 Wang M, Cao R, Zhang L, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res* 2020;30:269–71 doi:10.1038/s41422-020-0282-0; PMID:32020029;

47 Hoffmann M, Mösbauer K, Hofmann-Winkler H, et al. Chloroquine does not inhibit infection of human lung cells with SARS-CoV-2. *Nature* 2020;585:588–90 doi:10.1038/s41586-020-2575-3; PMID:32698190;

48 Horby P, Mafham M, Linsell L, et al. Effect of Hydroxychloroquine in Hospitalized Patients with Covid-19. *N Engl J Med* 2020;383:2030–40 doi:10.1056/NEJMoa2022926; PMID:33031652;

49 Axfors C, Schmitt AM, Janiaud P, et al. Mortality outcomes with hydroxychloroquine and chloroquine in COVID-19 from an international collaborative meta-analysis of randomized trials. *Nat Commun* 2021;12:2349 doi:10.1038/s41467-021-22446-z; PMID:33859192;

50 Naumann RW, Coleman RL, Burger RA, et al. PRECEDENT: a randomized phase II trial comparing vintafolide (EC145) and pegylated liposomal doxorubicin (PLD) in combination versus PLD alone in patients with platinum-resistant ovarian cancer. *J Clin Oncol* 2013;31:4400–06 doi:10.1200/JCO.2013.49.7685; PMID:24127448;

51 Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nat Hum Behav* 2017;1:21 doi:10.1038/s41562-016-0021; PMID:33954258;

52 Baxter R, Tran TN, Hansen J, et al. Safety of Zostavax™--a cohort study in a managed care organization. *Vaccine* 2012;30:6636–41 doi:10.1016/j.vaccine.2012.08.070; PMID:22963800;

53 Kerr NL. HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev* 1998;2:196–217 doi:10.1207/s15327957pspr0203_4; PMID:15647155;

54 Gnant M, Mlineritsch B, Schippinger W, et al. Endocrine therapy plus zoledronic acid in premenopausal breast cancer. *N Engl J Med* 2009;360:679–91 doi:10.1056/NEJMoa0806285; PMID:19213681;

55 Coleman RE, Marshall H, Cameron D, et al. Breast-cancer adjuvant therapy with zoledronic acid. *N Engl J Med* 2011;365:1396–405 doi:10.1056/NEJMoa1105195; PMID:21995387;

56 Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New Engl J Med* 1989;321:406–12 doi:10.1056/NEJM198908103210629; PMID:2473403;

57 Anderson JL, Platia EV, Hallstrom A, et al. Interaction of baseline characteristics with the hazard of encainide, flecainide, and moricizine therapy in patients with myocardial infarction. A possible explanation for increased mortality in the Cardiac Arrhythmia Suppression Trial (CAST). *Circulation* 1994;90:2843–52 doi:10.1161/01.CIR.90.6.2843; PMID:7994829;

58 Grouse L. Post hoc ergo propter hoc. *J. Thorac. Dis.* 2016;8:E511-2 doi:10.21037/jtd.2016.04.49; PMID:27499984;

59 Bertholf RL. Scientific Evidence, Medical Practice, and the Insidious Danger of Anecdotal Reports. *Laboratory Medicine* 2020;51:555–56 doi:10.1093/labmed/lmaa093; PMID:33095864;

60 Smith GCS, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ* 2003;327:1459–61 doi:10.1136/bmj.327.7429.1459; PMID:14684649;

61 Höfler M. Causal inference based on counterfactuals. *BMC Med Res Methodol*

2005;5:28 doi:10.1186/1471-2288-5-28; PMID:16159397;

62 Pearl J. An introduction to causal inference. *The International Journal of Biostatistics* 2010;6:Article 7 doi:10.2202/1557-4679.1203; PMID:20305706;

63 Cucherat M, Laporte S, Delaitre O, et al. From single-arm studies to externally controlled studies. Methodological considerations and guidelines. *Therapie* 2020;75:21–27 doi:10.1016/j.therap.2019.11.007; PMID:32063399;

64 Nagendran M, Pereira TV, Kiew G, et al. Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. *BMJ* 2016;355:i5432 doi:10.1136/bmj.i5432; PMID:27789483;

65 Bruns SB, Ioannidis JPA. p-Curve and p-Hacking in Observational Research. *PLoS ONE* 2016;11:e0149144 doi:10.1371/journal.pone.0149144; PMID:26886098;

66 Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology* 2015;68:1046–58 doi:10.1016/j.jclinepi.2015.05.029; PMID:26279400;

67 Head ML, Holman L, Lanfear R, et al. The extent and consequences of p-hacking in science. *PLoS Biology* 2015;13:e1002106 doi:10.1371/journal.pbio.1002106; PMID:25768323;

68 Silberzahn R, Uhlmann EL, Martin DP, et al. Many analysts, one dataset: Making transparent how variations in analytical choices affect results 2017.

69 Chuard PJC, Vrtílek M, Head ML, et al. Evidence that nonsignificant results are sometimes preferred: Reverse P-hacking or selective reporting? *PLoS Biol* 2019;17:e3000127 doi:10.1371/journal.pbio.3000127; PMID:30682013;

70 Michels KB, Rosner BA. Data trawling: to fish or not to fish. *The Lancet* 1996;348:1152–53 doi:10.1016/S0140-6736(96)05418-9;

71 Data dredging - Wikipedia 2021. Available at: https://en.wikipedia.org/wiki/Data_dredging Accessed August 30, 2021.

72 Berger ML, Sox H, Willke RJ, et al. Good Practices for Real-World Data Studies of Treatment and/or Comparative Effectiveness: Recommendations from the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. *Value Health* 2017;20:1003–08 doi:10.1016/j.jval.2017.08.3019; PMID:28964430;

73 Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *The Milbank Quarterly* 2016;94:485–514 doi:10.1111/1468-0009.12210; PMID:27620683;

74 Egger M, Davey Smith G, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629–34 doi:10.1136/bmj.315.7109.629; PMID:9310563;

75 Easterbrook P, Gopalan R, Berlin J, et al. Publication bias in clinical research. *The Lancet* 1991;337:867–72 doi:10.1016/0140-6736(91)90201-Y;

76 Song F, Parekh-Bhurke S, Hooper L, et al. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC Med Res Methodol* 2009;9:79 doi:10.1186/1471-2288-9-79; PMID:19941636;

77 Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358:252–60 doi:10.1056/NEJMsa065779; PMID:18199864;

78 Han Y, Liu J, Sun M, et al. A Significant Statistical Advancement on the Predictive Values of ERCC1 Polymorphisms for Clinical Outcomes of Platinum-Based Chemotherapy in Non-Small Cell Lung Cancer: An Updated Meta-Analysis. *Disease Markers* 2016;2016:7643981 doi:10.1155/2016/7643981; PMID:27057082;

79 Smits KM, Schouten JS, Smits LJ, et al. A review on the design and reporting of studies on drug-gene interaction. *J Clin Epidemiol* 2005;58:651–54 doi:10.1016/j.jclinepi.2005.01.001;

80 Lee SM, Falzon M, Blackhall F, et al. Randomized Prospective Biomarker Trial of ERCC1 for Comparing Platinum and Nonplatinum Therapy in Advanced Non-Small-Cell Lung Cancer: ERCC1 Trial (ET). *JCO* 2017;35:402–11 doi:10.1200/JCO.2016.68.1841; PMID:27893326;

81 Phillips AT, Desai NR, Krumholz HM, et al. Association of the FDA Amendment Act with trial registration, publication, and outcome reporting. *Trials* 2017;18:333 doi:10.1186/s13063-017-2068-3; PMID:28720112;

82 Ivanov A, Kaczkowska BA, Khan SA, et al. Review and Analysis of Publication Trends over Three Decades in Three High Impact Medicine Journals. *PLoS ONE* 2017;12:e0170056 doi:10.1371/journal.pone.0170056; PMID:28107475;

83 Psotka MA, Latta F, Cani D, et al. Publication Rates of Heart Failure Clinical Trials Remain Low. *J Am Coll Cardiol* 2020;75:3151–61 doi:10.1016/j.jacc.2020.04.068; PMID:32586589;

84 The BMJ. Paul Glasziou and Iain Chalmers: Can it really be true that 50% of research is unpublished? - The BMJ 2017. Available at: https://blogs.bmj.com/bmj/2017/06/05/paul-glasziou-and-iain-chalmers-can-it-really-be-true-that-50-of-research-is-unpublished/ Accessed August 22, 2021.

85 Nissen SB, Magidson T, Gross K, et al. Publication bias and the canonization of false facts. *eLife* 2016;5 doi:10.7554/eLife.21451; PMID:27995896;

86 Gerber AS, Malhotra N. Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results? *Sociological methods & research* 2008;37:3–30 doi:10.1177/0049124108318973;

87 Schuemie MJ, Ryan PB, DuMouchel W, et al. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med* 2014;33:209–18 doi:10.1002/sim.5925; PMID:23900808;

88 Albarqouni LN, López-López JA, Higgins JPT. Indirect evidence of reporting biases was found in a survey of medical research studies. *J Clin Epidemiol* 2017;83:57–64 doi:10.1016/j.jclinepi.2016.11.013; PMID:28088596;

89 Wang C-H, Li C-H, Hsieh R, et al. Proton pump inhibitors therapy and the risk of pneumonia: a systematic review and meta-analysis of randomized controlled trials and observational studies. *Expert Opinion on Drug Safety* 2019;18:163–72 doi:10.1080/14740338.2019.1577820; PMID:30704306;

90 Moayyedi P, Eikelboom JW, Bosch J, et al. Safety of Proton Pump Inhibitors Based on a Large, Multi-Year, Randomized Trial of Patients Receiving Rivaroxaban or Aspirin. *Gastroenterology* 2019;157:682-691.e2 doi:10.1053/j.gastro.2019.05.056; PMID:31152740;

91 Haute Autorité de Santé. Bon usage des inhibiteurs de la pompe à protons – Note de cadrage 2021. Available at: https://www.has-sante.fr/jcms/p_3221957/fr/bon-usage-des-inhibiteurs-de-la-pompe-a-protons-note-de-cadrage Accessed August 22, 2021.

92 Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2000;49:359–70 doi:10.1111/1467-9876.00197;

93 Sipahi I, Debanne SM, Rowland DY, et al. Angiotensin-receptor blockade and risk of cancer: meta-analysis of randomised controlled trials. *The Lancet Oncology* 2010;11:627–36 doi:10.1016/S1470-2045(10)70106-6;

94 U.S. Food and Drug Administration. Angiotensin Receptor Blockers (ARBs) NaN. Available at: https://www.fda.gov/drugs/drug-safety-and-availability/fda-drug-safety-communication-no-increase-risk-cancer-certain-blood-pressure-drugs-angiotensin Accessed August 22, 2021.

95 Effects of telmisartan, irbesartan, valsartan, candesartan, and losartan on cancers in 15 trials enrolling 138,769 individuals. *Journal of Hypertension* 2011;29:623–35 doi:10.1097/HJH.0b013e328344a7de; PMID:21358417;

96 Wayant C, Scheckel C, Hicks C, et al. Evidence of selective reporting bias in hematology journals: A systematic review. *PLoS ONE* 2017;12:e0178379 doi:10.1371/journal.pone.0178379; PMID:28570573;

97 Chan A-W, Hróbjartsson A, Haahr MT, et al. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457–65 doi:10.1001/jama.291.20.2457; PMID:15161896;

98 Hahn S, Williamson PR, Hutton JL, et al. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Stat Med* 2000;19:3325–36 ;

99 Peters J, Mengersen K. Selective reporting of adjusted estimates in observational epidemiology studies: reasons and implications for meta-analyses. *Eval Health Prof* 2008;31:370–89 doi:10.1177/0163278708324438; PMID:19000980;

100 Ioannidis JPA, Trikalinos TA. An exploratory test for an excess of significant findings. *Clinical Trials* 2007;4:245–53 doi:10.1177/1740774507079441; PMID:17715249;

101 Krsticevic M, Saric D, Saric F, et al. Selective reporting bias due to discrepancies between registered and published outcomes in osteoarthritis trials. *J Comp Eff Res* 2019;8:1265–73 doi:10.2217/cer-2019-0068; PMID:31739691;

102 Zhang S, Liang F, Li W. Comparison between publicly accessible publications, registries, and protocols of phase III trials indicated persistence of selective outcome reporting. *Journal of Clinical Epidemiology* 2017;91:87–94

doi:10.1016/j.jclinepi.2017.07.010; PMID:28757260;

103 Saini P, Loke YK, Gamble C, et al. Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews. *BMJ* 2014;349:g6501 doi:10.1136/bmj.g6501; PMID:25416499;

104 Boutron I, Dutton S, Ravaud P, et al. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303:2058–64 doi:10.1001/jama.2010.651; PMID:20501928;

105 Khan MS, Lateef N, Siddiqi TJ, et al. Level and Prevalence of Spin in Published Cardiovascular Randomized Clinical Trial Reports With Statistically Nonsignificant Primary Outcomes: A Systematic Review. *JAMA Netw Open* 2019;2:e192622 doi:10.1001/jamanetworkopen.2019.2622; PMID:31050775;

106 Gyawali B, Prasad V. Negative trials in ovarian cancer: is there such a thing as too much optimism? *Ecancermedicalscience* 2016;10:ed58 doi:10.3332/ecancer.2016.ed58; PMID:27594913;

107 Reynolds-Vaughn V, Riddle J, Brown J, et al. Evaluation of Spin in the Abstracts of Emergency Medicine Randomized Controlled Trials. *Annals of emergency medicine* 2019;75:423–31 doi:10.1016/j.annemergmed.2019.03.011; PMID:31101371;

108 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:e1000354 doi:10.1371/journal.pmed.1000354; PMID:21048986;

109 Callaway E. Questions raised over medical journals' financial ties to industry. *Nature* 2010 doi:10.1038/news.2010.564;

110 Smith R. Lapses at the new England journal of medicine. *JOURNAL OF THE ROYAL SOCIETY OF MEDICINE* 2006;99:380–82 doi:10.1258/jrsm.99.8.380; PMID:16893926;

111 Richard Smith's non-medical blogs. The New England Journal of Medicine, open access, Plan S, and undeclared conflicts of interest   2019. Available at: https://richardswsmith.wordpress.com/the-new-england-journal-of-medicine-open-access-plan-s-and-undeclared-conflicts-of-interest/ Accessed August 19, 2021.

112 Boutron I, Altman DG, Hopewell S, et al. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *JCO* 2014;32:4120–26 doi:10.1200/JCO.2014.56.7503; PMID:25403215;

113 Hansson L, Zanchetti A, Carruthers SG, et al. Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. *The Lancet* 1998;351:1755–62 doi:10.1016/S0140-6736(98)04311-6;

114 The paper mountain. *Nature* 2014 doi:10.1038/d41586-019-00381-w;

115 Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLOS Medicine* 2010;7:e1000326 doi:10.1371/journal.pmed.1000326; PMID:20877712;

116 van Calster B, Wynants L, Riley RD, et al. Methodology over metrics: Current scientific standards are a disservice to patients and society. *Journal of Clinical Epidemiology* 2021 doi:10.1016/j.jclinepi.2021.05.018; PMID:34077797;

117 CIRST. L'effet SIGAPS : La recherche médicale française sous l'emprise de l'évaluation comptable 2021. Available at: https://www.cirst.uqam.ca/publications/leffet-sigaps-la-recherche-medicale-francaise-sous-lemprise-de-levaluation-comptable-2/ Accessed August 01, 2021.

118 Wislar JS, Flanagin A, Fontanarosa PB, et al. Honorary and ghost authorship in high impact biomedical journals: a cross sectional survey. *BMJ* 2011;343:d6128 doi:10.1136/bmj.d6128; PMID:22028479;

119 Ioannidis JPA, Klavans R, Boyack KW. Thousands of scientists publish a paper every five days. *Nature* 2018;561:167–69 doi:10.1038/d41586-018-06185-8; PMID:30209384;

120 Peer Review Congress. Authorship for Sale: A Survey of Predatory Publishers and Journals - Peer Review Congress 2021 Accessed August 01, 2021.

121 Seife C. For Sale: "Your Name Here" in a Prestigious Science Journal 2021. Available at: https://www.scientificamerican.com/article/for-sale-your-name-here-in-a-prestigious-science-journal/ Accessed August 01, 2021.

122 Baccini A, Nicolao G de, Petrovich E. Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLoS ONE* 2019;14:e0221212 doi:10.1371/journal.pone.0221212; PMID:31509555;

123 Locher C, Moher D, Cristea IA, et al. Publication by association: how the COVID-19 pandemic has shown relationships between authors and editorial board members in the field of infectious diseases. *BMJ Evid Based Med* 2021 doi:10.1136/bmjebm-2021-111670; PMID:33785512;

124 Bagues M, Sylos-Labini M, Zinovyeva N. A walk on the wild side: 'Predatory' journals and information asymmetries in scientific evaluations. *Research Policy* 2019;48:462–77 doi:10.1016/j.respol.2018.04.013;

125 Beall J. Predatory publishers are corrupting open access. *Nature* 2012;489:179 doi:10.1038/489179a; PMID:22972258;

126 Butler D. Investigating journals: The dark side of publishing. *Nature* 2013;495:433–35 doi:10.1038/495433a; PMID:23538810;

127 Yeo-Teh NSL, Tang BL. An alarming retraction rate for scientific publications on Coronavirus Disease 2019 (COVID-19). *Account Res* 2021;28:47–53 doi:10.1080/08989621.2020.1782203; PMID:32573274;

128 Smaldino PE, McElreath R. The natural selection of bad science. *R. Soc. open sci.* 2016;3:160384 doi:10.1098/rsos.160384; PMID:27703703;

129 Yordanov Y, Dechartres A, Porcher R, et al. Avoidable waste of research related to inadequate methods in clinical trials. *BMJ* 2015;350:h809 doi:10.1136/bmj.h809; PMID:25804210;

130 Zarin DA, Goodman SN, Kimmelman J. Harms From Uninformative Clinical Trials. *JAMA* 2019;322:813–14 doi:10.1001/jama.2019.9892; PMID:31343666;

131 Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002;288:358–62 doi:10.1001/jama.288.3.358; PMID:12117401;

132 Cohen PJ. Failure to conduct a placebo-controlled trial may be unethical. *Am J Bioeth* 2002;2:24 doi:10.1162/152651602317533604; PMID:12189067;

133 Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *The Lancet* 2014;383:156–65 doi:10.1016/S0140-6736(13)62229-1;

134 Glasziou P, Chalmers I. Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ* 2018:k4645 doi:10.1136/bmj.k4645;

135 Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *The Lancet* 2009;374:86–89 doi:10.1016/S0140-6736(09)60329-9;

136 Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation 2021. Available at: https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation Accessed August 01, 2021.

137 Infodemic 2021. Available at: https://www.who.int/health-topics/infodemic#tab=tab_1 Accessed August 01, 2021.

138 Washington Post. Major publisher retracts 64 scientific papers in fake peer review outbreak 2015. Available at: https://www.washingtonpost.com/news/morning-mix/wp/2015/08/18/outbreak-of-fake-peer-reviews-widens-as-major-publisher-retracts-64-scientific-papers/ Accessed August 23, 2021.

139 Gaudino M, Robinson NB, Audisio K, et al. Trends and Characteristics of Retracted Articles in the Biomedical Literature, 1971 to 2020. *JAMA Intern Med* 2021;181:1118–21 doi:10.1001/jamainternmed.2021.1807; PMID:33970185;

140 Jung RG, Di Santo P, Clifford C, et al. Methodological quality of COVID-19 clinical research. *Nat Commun* 2021;12:943 doi:10.1038/s41467-021-21220-5; PMID:33574258;

141 Belluz J, Hoffman S. Let's stop pretending peer review works 2015. Available at: https://www.vox.com/2015/12/7/9865086/peer-review-science-problems Accessed August 23, 2021.

142 Carlisle JB. False individual patient data and zombie randomised controlled trials submitted to Anaesthesia. *Anaesthesia* 2021;76:472–79 doi:10.1111/anae.15263; PMID:33040331;

143 The BMJ. Time to assume that health research is fraudulent until proven otherwise? - The BMJ 2021. Available at: https://blogs.bmj.com/bmj/2021/07/05/time-to-assume-that-health-research-is-fraudulent-until-proved-otherwise/ Accessed August 31, 2021.

144 Al-Marzouki S, Evans S, Marshall T, et al. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ* 2005;331:267–70 doi:10.1136/bmj.331.7511.267; PMID:16052019;

145 Diener HC, Cunha L, Forbes C, et al. European Stroke Prevention Study. 2. Dipyridamole and acetylsalicylic acid in the secondary prevention of stroke. *J Neurol Sci* 1996;143:1–13

doi:10.1016/s0022-510x(96)00308-5; PMID:8981292;

146 Enserink M. Fraud and ethics charges hit stroke drug trial. *Science* 1996;274:2004–05 doi:10.1126/science.274.5295.2004; PMID:8984655;

147 Mehra MR, Ruschitzka F, Patel AN. Retraction—Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *The Lancet* 2020;395:1820 doi:10.1016/S0140-6736(20)31324-6;

148 Mehra MR, Desai SS, Kuy S, et al. Retraction: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. DOI: 10.1056/NEJMoa2007621. *N Engl J Med* 2020;382:2582 doi:10.1056/NEJMc2021225; PMID:32501665;

149 the Guardian. Huge study supporting ivermectin as Covid treatment withdrawn over ethical concerns 2021. Available at: https://www.theguardian.com/science/2021/jul/16/huge-study-supporting-ivermectin-as-covid-treatment-withdrawn-over-ethical-concerns Accessed August 20, 2021.

150 Reardon S. Flawed ivermectin preprint highlights challenges of COVID drug studies. *Nature* 2021;596:173–74 doi:10.1038/d41586-021-02081-w; PMID:34341573;

151 Hill A, Garratt A, Levi J, et al. Erratum: Expression of Concern: "Meta-analysis of Randomized Trials of Ivermectin to Treat SARS-CoV-2 Infection". *Open Forum Infectious Diseases* 2021;8:ofab394 doi:10.1093/ofid/ofab394; PMID:34410284;

152 Retraction Watch. Ivermectin meta-analysis to be retracted, revised, say authors 2021. Available at: https://retractionwatch.com/2021/08/10/ivermectin-meta-analysis-to-be-retracted-revised-say-authors/ Accessed August 20, 2021.

153 Authors of meta-analysis on heart disease retract it when they realize a NEJM reference had been retracted – Retraction Watch 2021. Available at: https://retractionwatch.com/2020/12/09/authors-of-meta-analysis-on-heart-disease-retract-it-when-they-realize-a-nejm-reference-had-been-retracted/ Accessed August 01, 2021.

154 Avenell A, Stewart F, Grey A, et al. An investigation into the impact and implications of published papers from retracted research: systematic search of affected literature. *BMJ open* 2019;9:e031909 doi:10.1136/bmjopen-2019-031909; PMID:31666272;

155 Faggion CM. More detailed guidance on the inclusion/exclusion of retracted articles in systematic reviews is needed. *Journal of Clinical Epidemiology* 2019;116:133–34 doi:10.1016/j.jclinepi.2019.07.006; PMID:31306745;

156 Gerstein HC, McMurray J, Holman RR. Real-world studies no substitute for RCTs in establishing efficacy. *The Lancet* 2019;393:210–11 doi:10.1016/s0140-6736(18)32840-x;

157 Banerjee R, Prasad V. Are Observational, Real-World Studies Suitable to Make Cancer Treatment Recommendations? *JAMA Netw Open* 2020;3:e2012119 doi:10.1001/jamanetworkopen.2020.12119; PMID:32729916;

158 Kumar A, Guss ZD, Courtney PT, et al. Evaluation of the Use of Cancer Registry Data for Comparative Effectiveness Research. *JAMA Netw Open* 2020;3:e2011985 doi:10.1001/jamanetworkopen.2020.11985; PMID:32729921;

159 Concato J. Observational versus experimental studies: what's the evidence for a hierarchy? *NeuroRx the journal of the American Society for Experimental NeuroTherapeutics* 2004;1:341–47 doi:10.1602/neurorx.1.3.341;

160 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England journal of medicine* 2000;342:1887–92 doi:10.1056/nejm200006223422507;

161 Criner GJ, Connett JE, Aaron S d., et al. Simvastatin for the prevention of exacerbations in moderate-to-severe COPD. *The New England journal of medicine* 2014;370:2201–10 doi:10.1056/NEJMoa1403086; PMID:24836125;

162 Fletcher AE. Controversy over "contradiction": Should randomized trials always trump observational studies? *American journal of ophthalmology* 2009;147:384–86 doi:10.1016/j.ajo.2008.04.024;

163 Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286:821–30 doi:10.1001/jama.286.7.821; PMID:11497536;

164 Baron JA, Barry EL, Mott LA, et al. A Trial of Calcium and Vitamin D for the Prevention of Colorectal Adenomas. *N Engl J Med* 2015;373:1519–30 doi:10.1056/NEJMoa1500409; PMID:26465985;

165 Bull-Otterson L, Gray EB, Budnitz DS, et al. Hydroxychloroquine and Chloroquine Prescribing Patterns by Provider Specialty Following Initial Reports of Potential Benefit for COVID-19 Treatment - United States, January-June 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:1210–15

doi:10.15585/mmwr.mm6935a4; PMID:32881845;

166 Watts Up With That? Hydroxychloroquine-based COVID-19 Treatment, A Systematic Review of Clinical Evidence and Expert Opinion from Physicians' Surveys 2020. Available at: https://wattsupwiththat.com/2020/07/07/hydroxychloroquine-based-covid-19-treatment-a-systematic-review-of-clinical-evidence-and-expert-opinion-from-physicians-surveys/ Accessed December 11, 2021.

167 Jackson, Coker. 65 Percent of Physicians in New Survey Would Give Anti-Malaria Drugs to Their Own Family to Treat COVID-19 2020. Available at: https://www.prnewswire.com/news-releases/65-percent-of-physicians-in-new-survey-would-give-anti-malaria-drugs-to-their-own-family-to-treat-covid-19-301037543.html Accessed December 11, 2021.

168 Stockman LJ, Bellamy R, Garner P. SARS: systematic review of treatment effects. *PLOS Medicine* 2006;3:e343 doi:10.1371/journal.pmed.0030343; PMID:16968120;

169 Arabi YM, Mandourah Y, Al-Hameed F, et al. Corticosteroid Therapy for Critically Ill Patients with Middle East Respiratory Syndrome. *Am J Respir Crit Care Med* 2018;197:757–67 doi:10.1164/rccm.201706-1172OC; PMID:29161116;

170 Lansbury LE, Rodrigo C, Leonardi-Bee J, et al. Corticosteroids as Adjunctive Therapy in the Treatment of Influenza: An Updated Cochrane Systematic Review and Meta-analysis. *Crit Care Med* 2020;48:e98-e106 doi:10.1097/CCM.0000000000004093; PMID:31939808;

171 Simpson SH, Eurich DT, Majumdar SR, et al. A meta-analysis of the association between adherence to drug therapy and mortality. *BMJ* 2006;333:15 doi:10.1136/bmj.38875.675486.55; PMID:16790458;

172 Vist GE, Hagen KB, Devereaux PJ, et al. Outcomes of patients who participate in randomised controlled trials compared to similar patients receiving similar interventions who do not participate. *Cochrane Database Syst Rev* 2007:MR000009 doi:10.1002/14651858.MR000009.pub3; PMID:17443630;

173 Davis S, Wright PW, Schulman SF, et al. Participants in prospective, randomized clinical trials for resected non-small cell lung cancer have improved survival compared with nonparticipants in such trials. *Cancer* 1985;56:1710–18 doi:10.1002/1097-0142(19851001)56:7<1710:aid-cncr2820560741>3.0.co;2-t; PMID:3896460;

174 Choueiri TK, Halabi S, Sanford BL, et al. Cabozantinib Versus Sunitinib As Initial Targeted Therapy for Patients With Metastatic Renal Cell Carcinoma of Poor or Intermediate Risk: The Alliance A031203 CABOSUN Trial. *JCO* 2017;35:591–97 doi:10.1200/JCO.2016.70.7398; PMID:28199818;

175 Gyawali B, Rome BN, Kesselheim AS. Regulatory and clinical consequences of negative confirmatory trials of accelerated approval cancer drugs: retrospective observational study. *BMJ* 2021;374:n1959 doi:10.1136/bmj.n1959; PMID:34497044;

176 The James Lind Library. The trials and tribulations of the University Group Diabetes Program: lessons and reflections. - The James Lind Library 2021. Available at: https://www.jameslindlibrary.org/articles/the-trials-and-tribulations-of-the-university-group-diabetes-program-lessons-and-reflections/ Accessed August 31, 2021.

177 Schwartz TB, Meinert CL. The UGDP controversy: thirty-four years of contentious ambiguity laid to rest. *Perspect Biol Med* 2004;47:564–74 doi:10.1353/pbm.2004.0071; PMID:15467178;

178 Boissel JP. Histoire subjective du souci de la qualité des essais cliniques – Impressionist history of quality concerns with clinical trials ;

179 Stamatakis E, Weiler R, Ioannidis JPA. Undue industry influences that distort healthcare research, strategy, expenditure and practice: a review. *Eur J Clin Invest* 2013;43:469–75 doi:10.1111/eci.12074; PMID:23521369;

180 Sox HC, Rennie D. Seeding trials: just say "no". *Ann. Intern. Med.* 2008;149:279–80 doi:10.7326/0003-4819-149-4-200808190-00012; PMID:18711161;

181 Hill KP, Ross JS, Egilman DS, et al. The ADVANTAGE seeding trial: a review of internal documents. *Ann. Intern. Med.* 2008;149:251–58 doi:10.7326/0003-4819-149-4-200808190-00006; PMID:18711155;

182 Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *JNCI Journal of the National Cancer Institute* 2009;101:1446–52 doi:10.1093/jnci/djp335; PMID:19815849;

183 Granger CB, Alexander JH, McMurray JJV, et al. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med* 2011;365:981–92 doi:10.1056/NEJMoa1107039; PMID:21870978;

184 Chen P, Nirula A, Heller B, et al. SARS-CoV-2 Neutralizing Antibody LY-CoV555 in Outpatients with Covid-19. *N Engl J Med* 2021;384:229–37 doi:10.1056/NEJMoa2029849; PMID:33113295;

Index