Société Française de Pharmacologie et de Thérapeutique

Groupe de Travail Méthodologie

Livre blanc SFPT

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

Document compagnon

Dossier 9 – Les études observationnelles

14 février 2022

Groupe de rédaction / relecture (Par ordre alphabétique)

- Theodora Angoulvant
- Laurent Bertoletti
- Jean-Luc Cracowski
- Michel Cucherat
- Dominique Deplanque
- Guillaume Grenet
- François Gueyffier
- Behrouz Kassai
- Charles Khouri
- Silvy Laporte
- Bruno Laviolle
- Jean-Christophe Lega
- Clara Locher
- Florian Naudet
- Edouard Ollier
- Antoine Pariente
- Matthieu Roustit
- Tabassome Simon

Table des matières

1 Introduction

Les études observationnelles d'efficacité des traitements tentent d'utiliser l'observation de ce qui se passe dans la pratique médicale courante (vraie vie) pour déterminer l'efficacité des traitements, sans intervenir sur la nature des traitements reçue par les patients.

> *Il existe de nombreuses utilisations des études observationnelles en pharmaco épidémiologie en dehors de l'évaluation de l'efficacité des médicaments comme : décrire les pratiques et leur évolution, rechercher le mésusage, génération de nouvelles hypothèses, détections des EIG rares non détectés dans les essais qui peuvent impacter potentiellement le rapport bénéfice/risque dans des populations plus importantes et avec des durées plus longues, etc. Ces autres usages ne posent pas du tout les mêmes problématiques méthodologiques et ne sont pas concernés par ce qui est abordé dans ce chapitre.*

Les études observationnelles peuvent être utilisées avec d'autres objectifs (décrire les traitements utilisés, recherche du mésusage, etc.), mais ces applications sont hors du champ de ce document. Les études observationnelles considérées ici s'inscrivent dans une tentative d'obtenir des résultats aussi solides que ceux produits par l'approche habituelle, c'est-à-dire obtenir des preuves (*evidences*) (cf. section **Erreur ! Source du renvoi introuvable.**). L'idée est alors de produire des *real world evidence* (RWE), similaires aux *evidences* produite par les études expérimentales randomisés[1,2].

Les études observationnelles peuvent être réalisées de manière prospective (données primaires) ou en analysant de manière *a posteriori* des données déjà existantes, appelées données secondaires pour mentionner qu'il s'agit d'une utilisation secondaire de données. Ces données secondaires peuvent être des données cliniques (dossiers médicaux, *electronic health records*), des données issues de cohortes collections ou de registres, des bases de données administratives, etc.

L'analyse statistique est complexe, car elle cherche à corriger les résultats des biais inhérents à l'approche observationnelle, principalement le biais de confusion. Elle cherche aussi à s'inscrire dans le cadre de l'inférence causale (voire section 2.5) afin de tenter d'établir un lien de causalité entre le traitement étudié et les bénéfices mis en évidence, comme ce qui est obtenu avec l'essai randomisé. Enfin, ces études doivent aussi s'inscrire dans une approche d'émulation d'un essai cible [1] afin de conforter leurs résultats en se rapprochant le plus possible de la méthodologie de l'essai clinique.

La lecture critique de ces études et l'évaluation du degré de certitude des résultats demande une expertise pointue spécifique. Une difficulté supplémentaire apparait du fait d'une recherche méthodologique intense dans ce domaine, conduisant à une rapide évolution des méthodes de référence. L'expertise nécessaire à la lecture de ces études doit donc être continuellement actualisée.

Cette approche est inadaptée à l'évaluation des traitements avant leur commercialisation (ou leur mise à disposition), car l'approche de vraie vie sous-entend que les traitements étudiés sont déjà utilisés dans la pratique médicale courante. Leur utilisation est donc limitée à la confirmation de l'efficacité

---

[1] Le congrès nord-américain a défini des RWE comme *« as data regarding the usage, or the potential benefits or risks, of a drug derived from sources other than traditional clinical trials. FDA has expanded on this definition …»* https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

[2] La FDA ne limite pas les RWE aux études observationnelles. Elle recouvre aussi dans cette appellation les essais randomisés, simples, pragmatiques réalisés à partir de « real world data », c'est-à-dire de données de vraie vie non recueillies spécialement pour l'étude.

des traitements dans la vraie vie en comparant les résultats d'études observationnelles à ceux des essais randomisés (avec une problématique importante de p hacking, car l'analyse est réalisée en connaissant le résultat à obtenir) ou à étudier ce qui se passe sur des durées de suivi plus longues.

Les études observationnelles peuvent aussi être exploitée pour rechercher des hypothèse de repositionnement d'ancienne molécule dans d'autre pathologie que leur usage initial (« *repurposing* ») [2].

Les études observationnelles sont parfois aussi envisagées pour apporter la confirmation du bénéfice après un enregistrement précoce, à la place d'un essai randomisé. L'accès précoce est accordé sans démonstration du bénéfice clinique, avec des études de faible méthodologie, comme des études mono-bras ou des essais sur critères intermédiaires (n'ayant pas valeur de surrogate), à condition qu'une étude de confirmation soit entreprise. Si cette étude ne confirme pas le bénéfice, le traitement devrait être retiré. Cet affaiblissement de la qualité des essais initiaux engendrera une perte considérable de niveau de preuve que les meilleures études observationnelles ne pourront totalement compenser. Le degré de certitude du bénéfice clinique des produits serait bien inférieur à ce qu'apporte la méthodologie classique ou même ce qui est exigé actuellement pour les accès précoces, une confirmation par un essai randomisé (même s'il est vrai que ces études ne sont pas toujours produites. [3, 4, 5, 6, 7], cf. section **Erreur ! Source du renvoi introuvable.**).

Depuis des années, les études observationnelles à promotion industrielle et s'intéressant à l'efficacité des médicaments sont, à de rares exceptions, réalisées dans le cadre de phase 4 sans finalité d'enregistrement. La robustesse méthodologique de ce type d'études est en général faible. [8, 9] et leurs résultats sont principalement exploités à but de communication promotionnelle (ou d'abondement aux scores bibliométriques des auteurs). Pour parer à l'éventualité de résultats opposés à ceux souhaités, ces études aménagent souvent une ambigüité quant à leur objectif : décrire et non pas comparer (voir faire une comparaison descriptive !). Dans ce contexte, ces études ne cherchent pas, en général, à apporter une solution à toutes les problématiques méthodologiques que posent ces études (cf. infra) et ne correspondent absolument pas à ce qui actuellement envisagée comme « *real world evidences* ». S'il existe une possibilité de produire des preuves solides à partir des données observationnelles, de degré de certitude identique à celui apporté par la méthodologique classique, cela passe par une tout autre approche, bien plus complexe et sophistiquée, et il ne faut pas considérer que ces études indigentes méthodologiquement correspondent à ce qui est envisagé pour produire des RWE.

> Par exemple, dans la récente crise de la COVID-19, de nombreuses études « observationnelles » ont été produites pour justifier *a posteriori* des usages compassionnels ou des repositionnements. Ces études reposaient, à de rares exceptions près[3], sur des méthodologies simplistes, comme un ajustement arbitraire non raisonné. Les conséquences ont été particulièrement délétères, car non seulement leurs résultats ne sont pas pertinents, mais plus grave encore, elles ont entravé lourdement la réalisation des essais randomisés bien faits, les médecins préférant prescrire de manière "compassionnel" ces traitements non évalués correctement plutôt que d'inclure les patients dans un essai correct. Ce qui retarde et empêche d'avoir la réponse d'une démonstration d'efficacité ou au contraire de l'échec du médicament à évaluer.

[3] Voir par exemple l'étude de l'ivermectine dans la COVID de Soto-Becerra et al. [10] qui met en œuvre une approche d'inférence causale et d'émulation d'un essai cible et qui ne permet pas de conclure au bénéfice de celle-ci.

2 Problématiques méthodologiques spécifiques et solutions possibles

2.1 Confusion

Une problématique méthodologique importante des études observationnelles réside dans le biais de confusion.

Dans la vraie vie, les médecins ne prescrivent pas les traitements au hasard, mais les choisissent, au cas par cas, en prenant en compte les particularités de leurs patients, et les patients n'accèdent pas aux mêmes traitements selon leurs caractéristiques et leurs comportements de recours aux soins. Certaines de ces caractéristiques, conditionnant l'accès au traitement ou au soin, et le choix du traitement ou la décision de traiter ou de ne pas traiter, peuvent être aussi des variables qui conditionnent en amont le critère de jugement. Dans ces situations, si la méthodologie et les analyses statistiques ne sont pas correctement conçues, une fausse relation entre le traitement et le critère de jugement peut apparaitre du fait de cette relation triangulaire. Il y a confusion entre l'effet de la covariable et l'effet du traitement, conduisant à un résultat biaisé (dans un sens ou dans l'autre). C'est le « biais » de confusion auquel se rattache le biais dit d'indication, le channeling, etc.

Les études observationnelles cherchent à supprimer ce biais de confusion au moment de l'analyse à l'aide de différentes techniques statistiques visant à prendre en compte les différences de caractéristiques pouvant exister entre les patients étudiés (« ajustements », analyse conditionnée) qui peut être réalisé de nombreuses manières. Toutes les approches d'ajustement ne sont pas équivalentes sur d'autres aspects.

La prise en compte statistique de ces différences peu se faire de nombreuses manières (mixables entre elles éventuellement) : restriction, matching, stratification, régression multivariée, pondération ou autres méthodes plus complexes et moins courantes, soit sur les facteurs de confusion eux-mêmes soit sur un score de propension (sorte de condensat des variables sur lesquelles on souhaite ajuster).

En théorie il est possible de corriger complètement un résultat du biais de confusion s'il est possible de prendre ne compte tous les facteurs de confusion à l'origine du biais. En pratique cela suppose une identification raisonnée des facteurs de confusion « potentiel » pour chaque critère de jugement en se basant sur :

- L'identification de tous les facteurs de confusion potentiels par une revue systématique de tous les caractéristiques (prédicteurs ou déterminants) conditionnant le critère de jugement considéré
- L'élaboration d'un réseau de causalité (causal graph) [11] sous la forme d'un graphe orienté acyclique (Directed acyclic graph, DAG) pour identifier les facteurs d'ajustement (et ceux sur lesquels il ne faut pas ajuster, comme les colliders)
- La prise en compte de tous les facteurs de confusion identifié ce qui suppose La disponibilité des données (de bonne qualité) pour chacune de ces variables.

Juger de l'optimalité d'un ajustement *a posteriori* n'est pas aisé même si l'ensemble des points précédents ont été suivis par l'étude. En effet, juger de l'identification complète de tous les facteurs

de confusion pour chaque critère demande une expertise thématique et méthodologique poussée. De plus, dans la réalité, il est rare que l'analyse puisse prendre en compte tous les facteurs de confusion potentiels (non-disponibilité par exemple avec les données secondaires). In fine se pose donc la question d'un **biais de confusion résiduel**.

Plusieurs développements méthodologiques récents donnent des outils pour chercher à répondre à cette question : les contrôles négatifs et positifs et l'analyse quantitative de biais. Si grâce à ces outils il est possible d'exclure formellement un biais de confusion résiduel les résultats pourront être considérés comme fiables sur le plan de la confusion.

Les contrôles négatifs sont soit des critères de jugement que l'on sait non-associés avec le traitement étudié soit des expositions que l'on sait non-associées avec le critère de jugement, mais qui sont susceptibles d'être impactées par les mêmes facteurs de confusion de l'association d'intérêt.

> **Exemples de contrôle négatif.**
> Dans les études des conséquences de l'exposition in utero à un médicament, la prise du médicament longtemps avant ou après la grossesse sont des contrôles négatifs potentiels. Biologiquement il ne peut pas y avoir d'association avec les malformations par exemple, mais cette relation est susceptible d'être affectée par les mêmes facteurs de confusion que la relation d'intérêt (exposition in utero et malformation). Si l'ajustement est insuffisant, une relation apparait sur ce contrôle négatif entrainant la réfutation de l'absence de biais de confusion résiduelle.
>
> Un évènement clinique type effet indésirable médicamenteux, que l'on sait parfaitement exclu avec le médicament d'intérêt, mais dont certains facteurs de risques sont communs avec les facteurs de confusion potentielle (facteurs de fragilité des patients par exemple) peut servir de contrôle négatif.

Les contrôles négatifs ne permettent jamais d'exclure avec certitude un biais de confusion résiduel, car il est toujours possible que les facteurs de confusion persistant après ajustement n'affectent pas en fait les contrôles négatifs considérés et parce que le raisonnement nécessite de conclure à l'absence de relation, ce qui statistiquement est toujours incertain pour des raisons de puissance statistique.

Cependant les contrôles négatifs permettent de réfuter un résultat si on retrouve, pour le contrôle, une association de même ordre de grandeur que celle retrouvée pour l'exposition réelle d'intérêt (ou l'événement réel d'intérêt). Dans cette situation en effet, l'association retrouvée pour le contrôle est à la fois le marqueur l'existence et le quantificateur de l'importance d'un biais de confusion résiduel.

L'analyse quantitative de biais peut prendre plusieurs formes, mais le principe général est de montrer que la taille de l'effet obtenu ne peut pas s'expliquer par des facteurs de confusion qui n'auraient pas été pris en considération. Il s'agit d'une mesure de la robustesse numérique des résultats. Les limites de l'approche résident dans le fait que ces calculs reposent sur des hypothèses sur le nombre et la force des facteurs de confusion oubliés ou débouchent sur une analyse de « tipping point » (point de rebroussement).

2.2 Autres biais

Dans un essai clinique, la combinaison de la randomisation, d'un aveugle (idéalement double) bien construit, et de la standardisation des mesures et du suivi des patients offre une protection systématique contre un grand nombre de biais. Ce n'est pas le cas dans les études observationnelles où les mécanismes de survenue des biais, parfois complexe, peut entraîner de grandes difficultés méthodologiques pour la réalisation des études, ou la nécessité d'une grande expertise pour leur évaluation.

Des outils comme l'échelle ROBINS-I ont été développés pour estimer le risque de biais de manière standardisée. ROBINS-I [12] est maintenant largement adopté (méta-analyses intégrant des études observationnelles, recommandations GRADE [13], etc.). Il permet une évaluation du risque de biais approfondi portant sur toutes les dimensions de biais existants dans une étude observationnelle. Il a été conçu de façon à unifier les biais affectant les études contrôlées randomisées et les études observationnelles. Ainsi son niveau de « low risk of bias » correspond au degré de certitude apporté par un essai randomisé correctement conçu et réalisé [12].

Pour être considérée comme pouvant produire des démonstrations de degré de crédibilité suffisante, comparable à ce que produit un RCT, une étude observationnelle devra être cotée à « low risk of bias » par l'outil ROBINS-I (par définition).

Ces outils doivent dorénavant être utilisés en remplacement des historiques échelles de niveau de preuve qui sont, de façon générale, à considérer comme insuffisamment précises ou caduques.

L'approche d'émulation d'un essai cible (cf. section **Erreur ! Source du renvoi introuvable.**) a été p roposée pour permettre, par une approche systématisée, d'anticiper la survenue de biais lors de la conception puis lors de l'analyse des études observationnelles , en particulier concernant la sélection des patients, la définition de la date index de suivi des patients, le recueil des évènements critères de jugement, et la définition de la population et des modalités d'analyse [1, 14]

2.3 Autres éléments de méthode

2.3.1 Étude exploratoire, étude de confirmation, découverte fortuite

Les études observationnelles, surtout celles réalisées a posteriori à partir de données secondaires, constituent des outils très intéressant pour la génération d'hypothèses / pour es approches exploratoires sans objectif spécifiquement défini a priori. C'est en particulier le cas pour la détection de risques potentiels associés à l'utilisation des médicaments. L'approche exploratoire, conduite par les données (data-driven) et non par des hypothèses préalables, va conduire à une fouille des données sans objectif défini a priori. Cette fouille large expose à un risque important de découverte fortuite, qui peut être contenu, pour une part, par des approches statistiques reposant sur l'utilisation des approches de *false discovery rate*. La possibilité de trouver une explication *a posteriori* au résultat trouvé ne permet pas de renforcer sa robustesse compte tenu du nombre d'explications possibles compte tenu de la complexité des phénomènes biologiques ([15], chapitre 1, inductivism, page 7).

Comme avec les essais cliniques, l'approche exploratoire ne permet en rien de produire des résultats ayant valeur de preuve d'association causale, pouvant faire changer les pratiques, et ne doit pas être utilisée à cette fin.

> *Même pour les questions de sécurité des médicaments, où le principe de précaution prévaut dans la décision par rapport à la recherche de preuve formelle, les approches exploratoires exposent à un risque important de faux signaux et de prise de décision conservative (retrait du médicament, restriction d'utilisation) à tort, d'autant plus problématique que le traitement a montré un bénéfice cliniquement pertinent. Au mieux cette approche permet de générer de nouvelles hypothèses à confirmer dans de nouvelles études (de confirmation). Ces études de confirmation peuvent être des études observationnelles, mais conçues cette fois dans un objectif spécifique et selon des modalités définies a priori au regard de l'hypothèse émise au terme de l'approche exploratoire (hypothesis-driven).*

Dans le contexte de construction des stratégies thérapeutiques, les études exploratoires sont donc non recevables. Les RWE pouvant être considérées pour ces définitions de stratégies thérapeutiques doivent impérativement être issues d'études de confirmation ayant clairement un objectif en phase avec la revendication. Les objectifs compatibles avec l'acceptabilité des résultats pour confirmer la place du traitement d'intérêt dans la stratégie sont alors identiques à ceux des essais randomisés : démontrer la supériorité de N par rapport au traitement de référence sur un critère cliniquement pertinent et dans la population cible du traitement[4].

2.3.2 Études multibases

Même avec une étude de confirmation, une découverte fortuite est toujours possible. Avec les études conduites a posteriori à partir de données secondaires, compte tenu du grand nombre de sources de données disponibles, une même recherche d'association peut être réalisée de multiple fois. Sur le nombre une découverte fortuite peut être faite, qui pourrait en théorie être également le seul résultat publié (cf. biais de publication).

L'utilisation dans la même étude de plusieurs sources de données secondaires (bases de données) permet de limiter les conséquences de cette problématique [16]. L'association d'intérêt est recherchée simultanément avec la même méthode dans plusieurs bases et une conclusion générale ne sera effectuée que s'il y a homogénéité des résultats à travers les bases (après avoir exclu la possibilité d'une hétérogénéité explicable par les différences de populations et des modificateurs d'effet). Le résultat global de l'étude sera la méta-analyse des résultats obtenus sur chaque base. Cette approche multibases permet d'augmenter la reproductibilité des résultats et est maintenant fréquemment utilisé.

2.3.3 P hacking

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.

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*) [17, 18].

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 [19, 20]. 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 à 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 [21].

[4] Nous reviendrons ultérieurement sur les problématiques de pertinence clinique

Cette potentialité peut être aussi illustrée par le concept de vibration des effets [18]. 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.

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 »[5]. On parle aussi de *data-dredging* ou partie de pêche [22, 23].

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. Ainsi aucune adaptation de la stratégie d'analyse ne peut s'effectuer au moment de sa réalisation (sans que cela soit détectable en comparant le plan d'analyse statistique et l'analyse effectivement réalisée).

En pratique il faut bien ici distinguer « stratégie » et « modalités ». Les caractéristiques des variables peuvent amener à modifier les modalités d'analyses dans le respect de la stratégie définie. Les possibilités d'adaptation ou les différents choix qui devront être fait au regard des caractéristiques des variables peuvent tout à fait être spécifié dans le PAS avant que les données ne soient rendues disponibles.

Cependant, pour les études réalisées a posteriori (on parle aussi d'études historiques) sur données secondaires, le SAP sera par définition élaboré alors que les données sont déjà disponibles. Pour donner la garantie de l'absence de tout p hacking (choix post hoc des variables d'ajustements, de la population d'analyse, des définitions des expositions et des critères de jugement), de publication sélective en fonction des résultats, de HARKing 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 [24].

Pour lever ces réserves, ces études doivent donner la garantie qu'elles ont bien procédé à une validation prospective a priori sur des données historiques d'une hypothèse formulée a priori. L'enregistrement des protocoles et des plans d'analyses statistiques, l'utilisation d'algorithme standard de phénotype, la transparence et l'attestation explicite de l'absence de ces pratiques par les investigateurs sont des éléments permettant de lever ces réserves [24, 25, 26].

L'initiative ENCePP et le ENCePP sceal avec dépôt préalable des protocoles d'études façon clinicaltrials.gov pourraient ici être mentionné comme exemple d'initiative permettant de vérifier la concordance entre la démarche finale et la conception initiale de l'étude.

2.3.4 Biais de publication et selective reporting

La problématique du biais de publication est particulièrement prégnante avec les études observationnelles, en particulier avec les études réalisées *a posteriori* sur des données secondaires. Le grand nombre de bases de données (administratives ou autres) disponibles permet de répéter la même étude. Il est ensuite possible de filtrer en fonction de leurs résultats les études (ou les résultats) qui seront exploités pour soutenir une revendication et éventuellement présenter aux autorités.

5 https://en.wikiquote.org/wiki/Ronald_Coase

Associé à une approche exploratoire et au p hacking, la possibilité de biais de publication ou de selective reporting fait que les études observationnelles sommaires ont vite acquis la réputation d'études flexibles à même de produire les résultats attendus.

La solution à cette problématique n'est pas simple. L'enregistrement des protocoles est la première mesure possible, mais se heurte au fait qu'avec les études réalisées a posteriori, il est possible d'enregistrer les protocoles alors que l'analyse a déjà été réalisée. L'enregistrement doit donc est associé à un engagement explicite des investigateurs que le protocole, le SAP et l'enregistrement ont été effectués avant toutes analyses et production de résultats.

2.4 Analyse en intention de traiter

Les études observationnelles sont souvent analysées en comparant des périodes-patients exposées et non exposées. Cette analyse peut impliquer, outre d'être exposée à des biais spécifiques de type biais de sélection, de mesurer un effet du traitement théorique similaire à une analyse per protocole ou un estimé (estimand) « on treament ».

Lorsqu'une étude observationnelle est réalisée dans un objectif d'évaluation d'efficacité et, éventuellement, de recommandation de l'utilisation d'un traitement, le schéma employé doit être choisi différemment, pour permettre la mesure de l'effet de l'initiation d'un traitement (et non pas celui de recevoir un traitement) [27]. Pour documenter cette décision d'instaurer un nouveau traitement et mesurer ce que cette décision induira comme amélioration dans le devenir des patients, une approche d'analyse en intention de traire est nécessaire (ou un estimé de type « treatment policy »).

2.5 Inférence causale (causal inference)

L'inférence causale est une approche récente, encore en plein développement, basée sur une théorie et des hypothèses, des designs et des techniques d'analyse qui permettent de tirer des conclusions de causalité à partir de données observationnelles [1, 28, 29, 30, 31]. 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. Cette approche est donc naturellement la plus appropriée pour des études qui essayent de se passer de l'apport de la randomisation en termes de causalité.

3 Synthèses des problématiques et de leurs solutions

Tableau 1 – Fiche de synthèse récapitulative des problématiques méthodologiques et des solutions attendues afin d'accepter un résultat d'étude observationnelle pour la construction des stratégies thérapeutiques

Problématique méthodologique ***et particularité spécifique à la nouvelle méthodologie.***	**Solution spécifique à apporter avec cette nouvelle méthodologie** ***pour garantir l'obtention du même degré de certitude qu'avec la méthodologie classique***
Nécessité d'un raisonnement contrefactuel pour identifier l'effet propre du traitement et mesurer son importance en raison de la variabilité du vivant (inter et intra sujet)	Étude observationnelle analytique (comparative) type étude de cohorte ou étude cas-témoins intégrant un groupe contrôle contemporain apportant le contrefait, s'inscrivant dans une approche d'émulation d'un essai cible
Biais de sélection Nombreuse possibilité de biais de sélection dans les études observationnelles (temps d'immortalité, ajustement ou restriction sur un collider, biais protopathique, etc.)	Utilisation d'un design d'émulation d'un essai cible, synchronisation des débuts de suivi entre les 2 groupes Ajustement après modélisation du réseau de causalité (DAGs) pour éviter l'ajustement sur les colliders Cotation par ROBINS -I en low risk of bias sur cette dimension
Biais de confusion majeur lié à la nature observationnelle (biais par indication, channeling biais)	Prise en compte de tous les facteurs de confusion dans l'analyse (quel que soit la méthode). Cela nécessite 1) d'identifier tous les facteurs déterminant le critère de jugement considéré pour établir le graphique de causalité (DAGs) et déterminer les réels facteur de confusion et 2) pouvoir prendre en compte toutes ces covariables identifiées Démontrer l'absence de biais de confusion résiduelle (contrôle négatif ou positif, analyse quantitative de biais) L'absence de randomisation ne permettant d'obtenir par design une estimation causale de l'effet traitement, doit être compensée par une approche d'inférence causale.
Biais de réalisation Aucun contrôle par design de ce biais dans les études observationnelles	Cotation par ROBINS -I en low risk of bias sur cette dimension
Biais de suivi Aucun contrôle par design de ce biais dans les études observationnelles	Cotation par ROBINS -I en low risk of bias sur cette dimension
Biais d'attrition Aucun contrôle par design de ce biais dans les études observationnelles	Cotation par ROBINS -I en low risk of bias sur cette dimension
Estimation de l'effet traitement correspond à ce que la recommandation future du traitement produirait comme changement dans le devenir des patients (compte tenu de tout le reste de la stratégie thérapeutique)	Analyse en intention de traiter
Risque de conclure à tort à l'intérêt du traitement du fait de l'erreur statistique alpha (de premier type)	Plan de contrôle du risque alpha global Définition des comparaisons inférentielles (tests qui peuvent conduire à la conclusion à l'intérêt du traitement et donc à la recommandation de son utilisation)
Multiplicité des comparaisons pouvant amener à conclure à l'intérêt du traitement ; multiplicité induisant une inflation du risque alpha global	Plan de contrôle du risque alpha global (non prise en considération de la signification nominale, non-présentation des p values non inférentielles pour éviter les surinterprétations des résultats exploratoires sans contrôle du risque alpha global)

Fraude des investigateurs	Pour les études prospectives : Monitorage de terrain, bonnes pratiques cliniques, recherche systématique de la fraude lors de l'analyse statistique, exclusion des centres en cas de suspicion de fraude Pour les études sur bases : non applicable
Fraude au niveau de l'investigateur principal (sponsor, réalisation de l'étude) possible (cf. Surgisphere)	Système d'assurance qualité (procédure opératoire standard), traçabilité, audit interne et externe (disponibilité des données)
Découverte fortuite, fouille de données (data dredging, data milking)	Réalisation d'étude observationnelle de confirmation respectant pleinement la démarche hypothético déductive avec un objectif fixé *a priori* certifié par les investigateurs. Exclusion des résultats post hoc, ou exploratoire de la prise de décision
Respect de la démarche hypothético déductive	Formulation des hypothèses *a priori*, garantie soit 1) par une démarche prospective ou 2) certifiée par les investigateurs dans le protocole
P hacking (modification de l'analyse statistique jusqu'à l'obtention des résultats voulus) fréquent dans les études observationnelles	Définition d'un protocole fixant les critères de jugement et les grandes lignes de l'analyse Définition d'un plan d'analyse statistique (SAP) précis avant que les données (même parcellaires) soient disponibles et réalisation de l'analyse statistique en stricte conformité avec ce SAP (démarche certifiée par les investigateurs)
Selective reporting (présentation, publication au niveau de l'étude que des résultats positifs permettant de faire la conclusion recherchée)	Protocole établi a priori, enregistrement dans un registre Vérification des résultats mis en avant par rapport au protocole
Biais de publication Risque important avec les études observationnelles rétrospectives compte tenu de la relative facilité de les multiplier	Difficile à exclure. Il faudrait avoir connaissance de toutes les études rétrospectives entreprises sur la même question Garantie apportée par le dossier que toutes les études observationnelles entreprises sont rapportées (solutionnable que par des obligations réglementaires) **Point majeur de dégradation du degré de certitude apporté par les études observationnelles**
Pertinence clinique	Utilisation : • D'un critère de jugement clinique (ou d'un surrogate démontré) • D'un comparateur loyal et représentant le traitement standard du moment où l'étude est analysée pour intégrer le nouveau traitement dans la stratégie thérapeutique • D'une population cible recherchée correspondant à la totalité des patients relevant du traitement évalué
Bénéfice complètement compensé par des effets délétères de manière quantitative ou qualitative	Évaluation de la safety avec la même précision et robustesse que l'efficacité Prise de décision basée sur la balance bénéfice risque = bénéfice clinique (et non pas seulement sur l'efficacité ou sur la sécurité de façon isolée et séparée)
Spin de conclusion (conclusion positive en faveur de l'intérêt du traitement dans une étude en réalité non concluante)	Ne pas lire les conclusions, ne regarder que les résultats et la méthode
Manque de transparence des rapports ou des publications ne permettant pas de voir les limites des résultats	Rapport exhaustif (pas de standardisation actuellement) Guide EQUATOR (STROBE) garantissant l'informativité des publications pour disposer de tous les éléments nécessaires à la l'évaluation critique de l'étude

4 Études de cas, retour sur expérience

Il existe de très nombreux exemples où des bénéfices de traitement suggérés par des études observationnelles n'ont pas pu être retrouvés par des essais randomisés. Ces exemples montrent la fragilité potentielle des résultats des études observationnelles et qu'en l'état actuel des pratiques ces études ne permettent pas de produire des résultats au-delà de tout doute raisonnable.

Lorsque ces études observationnelles sont réalisées après les essais cliniques pour confirmer en vraie vie leur résultat, un constat complètement différent est fait avec très peu d'échecs de confirmation. Cette situation s'explique parfaitement par la problématique connue du p hacking dans les études observationnelles [32]. Quand l'étude observationnelle est réalisée en premier, avant les essais cliniques, les retours d'expériences attirent l'attention sur une faible aptitude à estimer correctement le réel bénéfice des traitements. Tandis que lorsque ces études sont réalisées après les essais alors que le résultat à obtenir est connu, elle réussit presque toujours à retrouver le résultat attendu. Le phénomène connu de p hacking provenant de la possibilité d'adapter les choix d'analyse (covariable d'ajustement ou population d'analyse) en fonction des résultats produits pourrait expliquer la bonne performance des études observationnelles lorsqu'elles sont réalisées alors que le résultat à produire est connu. Un biais de publication n'est pas à exclure aussi dans cette situation.

Cela ne concerne pas les situations particulières où l'étude observationnelle pré-déclarée et bien conduite est de fait la meilleure source de preuve envisageable et, certainement, la meilleure contre-mesure à des développements effectués trop rapidement pour apporter une solution temporaire à des situations d'impasse thérapeutique. La discordance avec des essais réalisés antérieurement est alors, évidemment pas espérée, mais clairement attendue.

4.1 Comparaison de la chlorthalidone et de l'hydrochlorothiazide pour le traitement de l'hypertension

L'étude par Hripcsak et al. donne un exemple complet de la méthodologie sophistiquée pour produire des RW evidence [33]. L'évaluation complète de la méthodologie de cette étude

L'étude est clairement une étude de confirmation (cf. section 2.3.1) dont l'objectif global est de comparer l'efficacité et la sécurité relative de la chlorthalidone et de l'hydrochlorothiazide.

> OBJECTIVE To compare the effectiveness and safety of chlorthalidone and hydrochlorothiazide as first-line therapies for hypertension in real-world practice. [Abstract]

Les critères de jugement sur lesquels était effectuée cette comparaison ont été aussi clairement définis a priori. Les résultats sur lesquels porte la conclusion ont donc été parfaitement défini a priori (et l'étude ne s'inscrit pas une démarche exploratoire qui aurait consistée « à laisser parler les données » pour savoir sur quoi comparer les 2 produits et conclure).

> MAIN OUTCOMES AND MEASURES The primary outcomes were acute myocardial infarction, hospitalization for heart failure, ischemic or hemorrhagic stroke, and a composite cardiovascular disease outcome including the first 3 outcomes and sudden cardiac death. Fifty-one safety outcomes were measured. [Abstract]

L'étude est multibase (cf. section 2.3.2) afin de limiter le risque de découverte fortuite et augmenter la reproductibilité des résultats.

> We included the 3 OHDSI databases that had at least 2500 patients with exposures to each drug who met the eligibility criteria enumerated below. The MarketScan Commercial Claims and Encounters database (CCAE) (IBM Watson Health; 2001 to 2018) database …. The deidentified Clinformatics Data Mart Database (ie, Optum) (OptumInsight; 2001 to 2017) … The Optum deidentified Electronic Health Record Dataset (ie, PanTher) (Optum; 2007 to 2017) database [Methods - Data Sources, pg. E2].

Un « new user design » avec un « active comparator » est employé pour éviter un biais de sélection par déplétion des susceptibles, pour aider au contrôle du biais de confusion et pour synchroniser les suivis dans les 2 groupes.

> We included all patients initiating antihypertensive treatment with chlorthalidone or hydrochlorothiazide, and we defined the index time as the first observed exposure to either drug, including only patients with a prior or concurrent diagnosis of hypertension.

L'étude est conforme à une émulation d'un essai cible, aussi bien en termes d'analyse, de définition des critères de sélection et de traitement

Le début du suivi est parfaitement bien défini et correspond à l'initiation des traitements

> The index event for this study is the first treatment with chlorthalidone or hydrochlorothiazide. It must be taken as a single anti-hypertensive agent, and no other anti-hypertensive agents may precede them. [Supplement §1.8]

Le biais de confusion a été corrigé par un ajustement basé sur un scrore de propension à haute dimension

> Propensity scores (PS) are used as an analytic strategy to reduce potential confounding due to imbalance between the target and comparator cohorts in baseline covariates. [Supplement §1.7.1]

> Each condition, drug, class, etc. is counted as a separate covariate, resulting in over 60,000 covariates per database for this study

Compte tenu de la multiplicité des critères de jugement d'efficacité et de sécurité envisagés, l'inflation du risque alpha a été évitée avec la méthode de Bonferroni, de façon identique à ce qui se fait dans les essais cliniques.

> To address multiplicity concerns, we indicate which estimates remain statistically significant after a Bonferroni correction for 55 hypotheses. However, we report all differences. [Method – Statistical analysis – pg 546]

Des contrôles négatifs et positifs ont été utilisés pour effectuer une recalibration des résultats en prenant des événements type effet indésirable de médicaments mais connu comme n'étant pas liés aux traitements étudiés. Ces événements sont susceptibles d'être relier à des mêmes facteurs de fragilité des patients que l'association d'intérêt.

> We estimated residual bias using 76 negative control outcomes … (ie, outcomes believed to be caused by neither chlorthalidone nor hydrochlorothiazide, which therefore have an assumed HR of 1) identified through a data-rich algorithm, and we augmented the set by injecting events into the negative controls to create synthetic positive controls (ie, outcomes where the true HR is assumed known and greater than 1). We measured how often the true relative risks for controls were inside of their CIs (it should be 95% of the time for 95% CIs), and we calibrated all HR estimates, their 95% CIs, and their 2-sided P values so that approximate 95% coverage was achieved for the controls. [Method – Statistical analysis – pg 546]

> The calibrated and uncalibrated HRs were very close, and this similarity indicated that the 76 negative controls and the synthetic positive controls revealed little evidence of residual confounding (in the form of false-positive or skewed results in the controls) [Results – Effectiveness – pg 547]

4.2 Sécurité cardiovasculaire de l'insuline

La problématique de la fiabilité des études observationnelles se pose aussi pour les questions de sécurité des médicaments. Au premier abord l'approche observationnelle est plutôt séduisante pour les questions d'effets indésirables rares des médicaments. Le nombre de sujets des essais randomisés est calculé pour mettre en évidence avec une certaine puissance l'efficacité. Ce nombre de sujets est souvent insuffisant pour garantir la puissance de la recherche des effets indésirables rares et/ou inattendus. L'évaluation de la sécurité sur les données de vraie vie semble donc une façon de dépasser cette limitation des essais.

Cependant il existe de nombreux exemples d'études observationnelles montrant à tort un effet indésirable. Par exemple chez les diabétiques un surcroit de mortalité et d'évènements cardiovasculaire a été observé avec l'insuline en deuxième ligne par rapport à un inhibiteur de la DDP4 dans une étude observationnelle [34]. Ce résultat n'a pas été retrouvé dans un l'essai randomisé de grande taille ORIGIN dédié à l'évaluation de l'insuline basale [35].

La possibilité de faux positif sur la sécurité des médicaments avec les études observationnelles purement exploratoire est certainement majorée par deux éléments. Le raisonnement en sécurité se base sur le principe de précaution et un simple doute suffit à faire prendre des décisions. Ainsi, les résultats de ces études sont souvent considérés malgré leurs fragilités méthodologiques reconnues. Le 2ème facteur de risque de faux positif est consécutif à la multiplicité de comparaisons présentent dans ces études exploratoires souvent réalisées pour « évaluer la sécurité », sans plus hypothèse construite. Ces limites sont levées par la réalisation de réelles études de confirmation (cf. section 2.3.1 et section 2.3.3).

5 Meta-recherche

La réflexion sur la fiabilité des études observationnelles pour l'évaluation de l'efficacité et de la sécurité des traitements à débuter au début des années 2000 [36, 37, 38].

Plusieurs études de méta-recherche ont étudié la concordance des résultats obtenus par les données observationnelles par rapport à ceux produits par les essais randomisés comparant les mêmes traitements dans la même situation pathologique. Aucune de ces études ne permet de conclure que les études observationnelles donnent systématiquement le même résultat que les essais randomisés [36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53]. Il faut cependant noter qu'il s'agit ici d'une des études telles que réalisées (avec leurs limites et leurs défauts, aussi bien pour les études observationnelles et les essais randomisés) et ce n'est pas une comparaison, stricto sensu, de la science observationnelle à l'approche expérimentale des essais randomisés.

Trois études par Tanen [54], Dahabreh [55] et Lonjon [56] ont été regroupée dans une analyse conjointe [41]. Une méta-analyse de la Cochrane est aussi disponible [57]

Les résultats les plus récents montrent qu'en moyenne les études observationnelles pourraient donner les mêmes résultats que les RCT, mais avec une forte variabilité [57, 58, 59]. Cependant il convient de remarquer que la mesure de concordance en moyenne sur des tailles d'effet n'est pas appropriée, car les surestimations compensent en moyenne les sous-estimations ce qui peut conduire à une moyenne des différences des estimations égale à zéro alors que dans aucun cas les 2 estimations n'ont été identiques. L'écart quadratique (ou la valeur absolue des différences) est un meilleur paramètre, mais rarement utiliser dans ces études de concordances.

De plus, la possibilité dans certains cas d'une erreur importante fait qu'il n'est pas possible de faire confiance par principe aux résultats des études observationnelles et d'une évaluation cas par cas est bien sûr indispensable.

En oncologie, Kumar et al. [43] ont utilisé les données du *National Cancer Database (NCDB)*, un registre de cancérologie, pour reproduire les résultats de survie de 141 essais randomisés. En utilisant des analyses réalisées par score de propension, une concordance des hazard ratio n'a été trouvée que dans 64% des cas et celle des p value dans 45% des cas. La conclusion proposée est que l'approche de *"comparative effectiveness"* à l'aide des données de registre de cancer produit souvent des résultats discordants avec ceux des essais randomisés. Soni et al. [60] évaluent la concordance des hazard ratio de survie entre des études observationnelles publiées et les essais randomisés correspondants. Aucune corrélation n'est retrouvée entre les hazard ratio des deux types d'études et le taux de concordances des hazard ratio n'est pas trouvé supérieur à celui attendu du fait du hasard. Aucune caractéristique des études observationnelles améliorant cette concordance n'a été retrouvée. Dans 9% des cas, des résultats statistiquement significatifs diamétralement opposés sont observés (*Janus effect*) (5% des cas dans le papier par Kumar). Une revue de la littérature portant sur ces résultats en oncologie [39] conclue que l'essai randomisé doit donc rester le standard pour l'évaluation des médicaments en oncologie.

Durant la crise de la COVID-19, une profusion d'étude observationnelle a été publiée (dans des revues ou en préprint). Pour toutes les molécules qui ont échoué ultérieurement à monter un réel intérêt dans des essais randomisés, de nombreuses études observationnelles en faveur de leur efficacité sont disponibles [58], comme par exemple avec les plasmas de patients convalescents [61, 62]. Il faut

cependant noter que dans la plupart de ces études observationnelles la méthodologie était catastrophique.

6 Synthèse, critères d'acceptabilité

Pour positionner ou confirmer la position d'un nouveau traitement dans la stratégie thérapeutique à partir d'une étude observationnelle, il faut :

- Une étude observationnelle de confirmation, hypothético déductive, dont l'objectif défini *a priori* était clairement la comparaison de l'efficacité et de la sécurité du traitement d'intérêt à un comparateur pertinent (comparative effectiveness and safety)
- Une approche d'inférence causale aboutie
- Un design d'émulation d'un essai cible correctement conçu et réalisé
- Une analyse en intention de traiter
- La démonstration de l'absence de biais de confusion résiduelle
    - en montrant que les ajustements ont porté sur la totalité des facteurs de confusion identifiés par une approche formelle (revue systématique des facteurs de risques ou pronostiques des critères de jugement, modélisation des relations de causalité, DAGs), y compris les facteurs de confusion dépendant du temps
    - complété par une démonstration de l'absence de biais de confusion résiduel :
        - contrôles négatifs ou positifs en fonction de la nature du résultat (raisonnablement convainquant de l'absence de biais résiduel en nombre, captation des facteurs de confusion et en résultats), recalibration éventuelle,
        - analyses de biais quantitatif raisonnablement convaincantes (par exemple à l'aide, entre autres, d'approches de type « rule-out approach » ou « array approach » qui consistent à chercher le déséquilibre de prévalence de facteur de confusion hypothétique et la force d'association de celui-ci à l'événement qui seraient nécessaires pour remettre en cause les résultats. Si ces déséquilibre ou forces d'association sont irréalistes, alors la confusion résiduelle ne peut être un argument invoqué pour expliquer les résultats
- Un niveau de risque de biais coté à « low risk of bias » par ROBINS-I
- Une approche permettant d'écarter avec certitude une sélection post hoc des résultats, un p hacking, une approche exploratoire, un HARKing et les limites des approches rétrospectives : démonstration que les résultats n'étaient pas connus (même partiellement) avant l'analyse des données (certifier par le protocole de manière explicite, enregistrement du protocole, SAP daté, etc.)
- Éventuellement, une approche multibase prévue d'emblée et démontrant la reproductibilité des résultats et permettant d'exclure une découverte fortuite
- Comme pour un essai clinique, il faut aussi exiger
    - Des résultats cliniquement pertinents en termes de critères de jugement, comparaison effectuée, contexte de soins contemporain, taille d'effet
    - Une documentation satisfaisante d'une balance bénéfice risque favorable

Les études observationnelles présentent un intérêt mais aussi des limites méthodologiques potentielles qui requièrent un haut niveau d'expertise pour leur conduite comme pour leur interprétation. Des développements récents ont proposé de nouvelles approches méthodologiques et techniques, mais la méta-recherche ne permet pas, par manque d'étude, de connaitre le réel niveau de fiabilité de ces solutions pour l'instant.

Pour ces raisons l'approche observationnelle ne peut pas remplacer actuellement la réalisation d'essais randomisés et ne peut donc être utilisée que dans de rare cas en apportant toutes les garanties nécessaires pour assurer la fiabilité des résultats.

Pour être prise en considération pour l'introduction d'un nouveau traitement dans la stratégie thérapeutique, les études observationnelles doivent avoir mis en œuvre une méthodologie s'appuyant sur l'inférence causale et l'émulation d'un essai cible pour contrôler le biais de confusion, démontrer l'absence de biais de confusion résiduel pouvant avoir expliqué les résultats, présenter une méthode expliquant précisément les modalités retenues pour contrôler les autres biais, et proposer des estimations de l'effet traitement appropriés. Elles doivent de plus s'inscrire dans une démarche de confirmation d'hypothèse et dans une démarche de conduite de recherche garantissant l'absence de HARKing, p Hacking et de publication sélective des résultats.

Références

1 Hernán MA. Methods of Public Health Research - Strengthening Causal Inference from Observational Data. *The New England journal of medicine* 2021 doi:10.1056/NEJMp2113319; PMID:34596980;

2 Park K. The use of real-world data in drug repurposing. *Transl Clin Pharmacol* 2021;29:117–24 doi:10.12793/tcp.2021.29.e18; PMID:34621704;

3 Hatswell AJ, Baio G, Berlin JA, et al. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999-2014. *BMJ open* 2016;6:e011666 doi:10.1136/bmjopen-2016-011666; PMID:27363818;

4 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;

5 Mahase E. FDA allows drugs without proven clinical benefit to languish for years on accelerated pathway. *BMJ* 2021;374:n1898 doi:10.1136/bmj.n1898; PMID:34326042;

6 Naci H, Smalley KR, Kesselheim AS. Characteristics of Preapproval and Postapproval Studies for Drugs Granted Accelerated Approval by the US Food and Drug Administration. *JAMA* 2017;318:626–36 doi:10.1001/jama.2017.9415; PMID:28810023;

7 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;

8 Boyle JM, Hegarty G, Frampton C, et al. Real-world outcomes associated with new cancer medicines approved by the Food and Drug Administration and European Medicines Agency: A retrospective cohort study. *Eur J Cancer* 2021;155:136–44 doi:10.1016/j.ejca.2021.07.001; PMID:34371443;

9 Song F, Zang C, Ma X, et al. The use of real-world data/evidence in regulatory submissions. *Contemporary Clinical Trials* 2021;109:106521 doi:10.1016/j.cct.2021.106521; PMID:34339865;

10 Soto-Becerra P, Culquichicón C, Hurtado-Roca Y, et al. Real-world effectiveness of hydroxychloroquine, azithromycin, and ivermectin among hospitalized COVID-19 patients: results of a target trial emulation using observational data from a nationwide healthcare system in Peru 2020.

11 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;10:37–48 ; PMID:9888278;

12 Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919 doi:10.1136/bmj.i4919;

13 Schünemann HJ, Cuello C, Akl EA, et al. GRADE Guidelines: 18. How ROBINS-I and other tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol* 2018 doi:10.1016/j.jclinepi.2018.01.012;

14 Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19:766–79 doi:10.1097/EDE.0b013e3181875e61; PMID:18854702;

15 Lash TL, VanderWeele TJ, Haneuse S, et al. Modern epidemiology. Philadelphia etc.: Wolters Kluwer 2021 ISBN:1451193289;

16 Schuemie MJ, Ryan PB, Pratt N, et al. Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND). *J Am Med Inform Assoc* 2020;27:1331–37 doi:10.1093/jamia/ocaa103; PMID:32909033;

17 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;

18 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;

19 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;

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

21 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;

22 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;

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

24 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;

25 Orsini LS, Monz B, Mullins CD, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing-Why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Pharmacoepidemiol Drug Saf* 2020;29:1504–13 doi:10.1002/pds.5079; PMID:32924243;

26 Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ* 2018;363:k3532 doi:10.1136/bmj.k3532; PMID:30429167;

27 Hernán MA, Sauer BC, Hernández-Díaz S, et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology* 2016;79:70–75 doi:10.1016/j.jclinepi.2016.04.014; PMID:27237061;

28 Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health* 2005;95 Suppl 1:S144-50 doi:10.2105/AJPH.2004.059204; PMID:16030331;

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

30 Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;60:578–86 doi:10.1136/jech.2004.029496; PMID:16790829;

31 Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 2004;58:265–71 doi:10.1136/jech.2002.006361; PMID:15026432;

32 Belas N. P-hacking in Clinical Trials: A Meta-Analytical Approach ;

33 Hripcsak G, Suchard MA, Shea S, et al. Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension. *JAMA Internal Medicine* 2020 doi:10.1001/jamainternmed.2019.7454; PMID:32065600;

34 Nyström T, Bodegard J, Nathanson D, et al. Second line initiation of insulin compared with DPP-4 inhibitors after metformin monotherapy is associated with increased risk of all-cause mortality,

cardiovascular events, and severe hypoglycemia. *Diabetes Research and Clinical Practice* 2017;123:199–208 doi:10.1016/j.diabres.2016.12.004; PMID:28056431;

35 Gerstein HC, Bosch J, Dagenais GR, et al. Basal insulin and cardiovascular and other outcomes in dysglycemia. *N Engl J Med* 2012;367:319–28 doi:10.1056/NEJMoa1203858; PMID:22686416;

36 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;

37 Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286:821–30 ;

38 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *New Engl J Med* 2000;342:1878–86 doi:10.1056/NEJM200006223422506; PMID:10861324;

39 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;

40 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;

41 Dahabreh IJ, Kent DM. Can the Learning Health Care System Be Educated With Observational Data? *JAMA* 2014;312:129–30 doi:10.1001/jama.2014.4364;

42 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;

43 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;

44 Naudet F, Maria AS, Falissard B. Antidepressant response in major depressive disorder: a meta-regression comparison of randomized controlled trials and observational studies. *PLoS ONE* 2011;6:e20811 doi:10.1371/journal.pone.0020811; PMID:21687681;

45 Oliver S, Bagnall AM, Thomas J, et al. Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technol Assess* 2010;14:1-165, iii doi:10.3310/hta14160; PMID:20338119;

46 Papanikolaou PN, Christidi GD, Ioannidis JPA. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ* 2006;174:635–41 doi:10.1503/cmaj.050873; PMID:16505459;

47 Shikata S, Nakayama T, Noguchi Y, et al. Comparison of effects in randomized controlled trials with observational studies in digestive surgery. *Ann Surg* 2006;244:668–76 doi:10.1097/01.sla.0000225356.04304.bc; PMID:17060757;

48 Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. *PLOS Medicine* 2011;8:e1001026 doi:10.1371/journal.pmed.1001026; PMID:21559325;

49 Kuss O, Legler T, Börgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. *Journal of Clinical Epidemiology* 2011;64:1076–84 doi:10.1016/j.jclinepi.2011.01.005; PMID:21482068;

50 Bhandari M, Tornetta P, Ellis T, et al. Hierarchy of evidence: differences in results between non-randomized studies and randomized trials in patients with femoral neck fractures. *Arch Orthop Trauma Surg* 2004;124:10–16 doi:10.1007/s00402-003-0559-z; PMID:14576955;

51 Edwards JP, Kelly EJ, Lin Y, et al. Meta-analytic comparison of randomized and nonrandomized studies of breast cancer surgery. *Can J Surg* 2012;55:155–62 doi:10.1503/cjs.023410; PMID:22449722;

52 Furlan AD, Tomlinson G, Jadad AAR, et al. Examining heterogeneity in meta-analysis: comparing results of randomized trials and nonrandomized studies of interventions for low back pain. *Spine (Phila Pa 1976)* 2008;33:339–48 doi:10.1097/BRS.0b013e31816233b5; PMID:18303468;

53 Müeller D, Sauerland S, Neugebauer EAM, et al. Reported effects in randomized controlled trials were compared with those of nonrandomized trials in cholecystectomy. *Journal of Clinical Epidemiology* 2010;63:1082–90 doi:10.1016/j.jclinepi.2009.12.009; PMID:20346627;

54 Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *BMJ* 2009;338:b81 doi:10.1136/bmj.b81; PMID:19174434;

55 Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur. Heart J.* 2012;33:1893–901 doi:10.1093/eurheartj/ehs114;

56 Lonjon G, Boutron I, Trinquart L, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Ann Surg* 2014;259:18–25 doi:10.1097/SLA.0000000000000256;

57 Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev* 2014:MR000034 doi:10.1002/14651858.MR000034.pub2; PMID:24782322;

58 Califf RM, Hernandez AF, Landray M. Weighing the Benefits and Risks of Proliferating Observational Treatment Assessments: Observational Cacophony, Randomized Harmony. *JAMA* 2020;324:625–26 doi:10.1001/jama.2020.13319; PMID:32735313;

59 Rush CJ, Campbell RT, Jhund PS, et al. Association is not causation: treatment effects cannot be estimated from observational data in heart failure. *Eur Heart J* 2018;39:3417–38 doi:10.1093/eurheartj/ehy407; PMID:30085087;

60 Soni PD, Hartman HE, Dess RT, et al. Comparison of Population-Based Observational Studies With Randomized Trials in Oncology. *JCO* 2019;37:1209–16 doi:10.1200/JCO.18.01074; PMID:30897037;

61 Klassen SA, Senefeld J, Johnson PW, et al. The Effect of Convalescent Plasma Therapy on COVID-19 Patient Mortality: Systematic Review and Meta-analysis. *medRxiv* 2021 doi:10.1101/2020.07.29.20162917; PMID:33140056;

62 Janiaud P, Axfors C, Schmitt AM, et al. Association of Convalescent Plasma Treatment With Clinical Outcomes in Patients With COVID-19: A Systematic Review and Meta-analysis. *JAMA* 2021;325:1185–95 doi:10.1001/jama.2021.2747; PMID:33635310;