

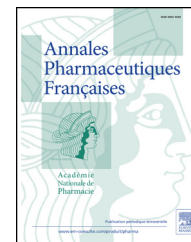


Disponible en ligne sur

ScienceDirect
www.sciencedirect.com

Elsevier Masson France

EM|consulte
www.em-consulte.com



REVUE GÉNÉRALE

Comment mesurer l'efficacité en vie réelle ?



How to measure real-world effectiveness?

R. Launois^{a,*}, J.-B. Trouiller^{a,b}, E. Cabout^a

^a Réseau d'évaluation en économie de la santé (REES), 28, rue d'Assas, 75006 Paris, France

^b Université Grenoble Alpes, Grenoble, 621, avenue Centrale, 38400 Saint-Martin d'Hères, France

Reçu le 26 juin 2018 ; accepté le 6 juillet 2018

Disponible sur Internet le 10 août 2018

MOTS CLÉS

Effectivité clinique ;
Efficacité
expérimentale ;
Vie réelle ;
Étude
observationnelle ;
Biais

Résumé L'utilisation nouvelle des bases de données médico-administratives marque un tournant majeur dans l'évaluation médicale. Face aux limites des essais cliniques randomisés lorsqu'il s'agit de représenter fidèlement l'hétérogénéité et la complexité des prises en charge des patients en vie réelle, de plus en plus d'études observationnelles sont aujourd'hui réclamées par les payeurs comme outils d'aide à la décision. L'absence d'intervention permet de capter les pratiques de routine et de s'approcher de l'efficacité du traitement en vie réelle. En dépit de cette validité externe, les présences de biais et de facteurs de confusion sont autant de freins à la validité interne d'une étude observationnelle. La mise en œuvre de schémas d'étude appropriés et le recours à des techniques statistiques sophistiquées sont nécessaires pour les neutraliser et démontrer l'existence d'une relation causale entre un traitement et un résultat clinique obtenu.

© 2018 Académie Nationale de Pharmacie. Publié par Elsevier Masson SAS. Tous droits réservés.

KEYWORDS

Effectiveness;
Efficacy;

Summary The new use of medico-administrative databases is a major change in the medical assessment field. Given the limits of the randomized clinical trials when it comes to accurately represent the heterogeneity and complexity of medical care in a real-world context, an increasing demand of observational studies comes from healthcare systems as decision tools.

* Auteur correspondant.

Adresses e-mail : launois.reesfrance@wanadoo.fr (R. Launois), jbtrouiller.reesfrance@orange.fr (J.-B. Trouiller), ecabout.reesfrance@orange.fr (E. Cabout).

Real world;
Observational study;
Bias

The lack of intervention generates real-world data and assesses both the care and the outcomes of patients in routine practice. Despite this external validity, the bias and confounding factors are a challenge to the internal validity of observational studies. An appropriate study design and statistical methods are necessary to neutralize them and consider a causal relationship between a treatment and a clinical outcome.

© 2018 Académie Nationale de Pharmacie. Published by Elsevier Masson SAS. All rights reserved.

Introduction

Le choix du bon traitement, pour le bon patient, au bon moment, a été de tout temps au cœur des préoccupations de la communauté médicale, mais les méthodes utilisées pour atteindre cet objectif se sont profondément modifiées au fil des ans en passant d'une « médecine des humeurs », où l'essentiel était de trouver un équilibre entre les quatre principes qui étaient censés gouverner le bon fonctionnement du corps humain : le sang, la pituite, la bile noire et la bile jaune, à une médecine fondée sur les preuves.

Pendant très longtemps, les essais cliniques randomisés (ECR) ont constitué l'étalon « universel » pour comparer l'efficacité des traitements et s'assurer de leur innocuité. La méthode de l'essai contrôlé randomisé consiste à partager en deux groupes, par tirage au sort, les individus recrutés dans la population cible : l'un recevant le traitement (le groupe traité), l'autre ne le recevant pas (le groupe témoin). L'essai est dit « contrôlé » du fait de l'existence d'une population de référence, il est dit « randomisé » puisque la répartition des sujets entre les groupes est le fruit du hasard (aléatoire). De par la loi des grands nombres, les personnes des deux groupes ont en moyenne les mêmes caractéristiques observables et non observables, lorsque la taille des deux populations est grande. Sous cette hypothèse, les différences entre les résultats observés dans les deux groupes sont, toutes choses égales par ailleurs, exclusivement attribuables au traitement.

Les essais randomisés sont construits sur des protocoles mille fois trop limités : trop courts dans leur durée, trop sélectifs dans leur recrutement, trop protocolés dans leurs modalités de prise en charge, si bien que la portée de leurs enseignements, pour le suivi en pratique courante des malades tout-venant, en est réduite d'autant. Le comparateur choisi est bien souvent un placebo, alors que le problème pour les cliniciens n'est pas de savoir s'il faut traiter de préférence à ne rien faire, mais de choisir, parmi les traitements actifs disponibles, celui qui est le plus adapté à son patient.

Après que l'autorisation de mise sur le marché a été accordée, la question n'est plus de savoir si le traitement « peut marcher ». En pratique médicale courante, il s'agit de savoir « s'il marche » lorsqu'il est proposé aux patients dans les environnements familiaux, sociaux et institutionnels très variés. Les méthodes conçues pour mettre en évidence un lien de causalité (les essais cliniques randomisés et contrôlés) ne sont pas nécessairement adaptées à une

telle analyse. La méthodologie des essais cliniques suppose une invariance des comportements et de l'environnement du système de santé. Or, la variabilité des résultats obtenus et les différences de coûts observés entre les traitements prescrits découlent justement de la dynamique des interactions entre les comportements des patients, celui des prescripteurs et les caractéristiques des produits étudiés. Les études de terrain se proposent de vérifier que les résultats annoncés sont bien au rendez-vous, puisque rien ne garantit que les conclusions d'un essai soient transposables à toutes les situations cliniques.

Faute de pouvoir généraliser les résultats des essais randomisés en pratique médicale quotidienne, les médecins de santé publique et les épidémiologistes ont recommandé d'exploiter les bases de données médico-administratives. Ces bases du Système national de santé (SNDS) constituent autant de sources secondaires d'informations, en ce sens qu'elles ont été mises en place pour atteindre des objectifs autres que ceux pour lesquels elles sont actuellement exploitées. La plupart du temps, il s'agit de bases de remboursement (SNIIRAM) ou de financement hospitalier (PMSI) auxquelles on veut faire jouer un rôle, en matière d'étiologie clinique, qui est étranger à leur vocation primitive. Ces bases sont facilement accessibles, mobilisables à très court terme, pour un coût modeste. Elles sont, par ailleurs, représentatives parce qu'exhaustives et leur contenu reflète la réalité des pratiques. Tout biais de mémorisation ou de non-réponse est par définition exclu. En revanche, la démonstration du lien de cause à effet entre le traitement et les résultats obtenus est délicate compte tenu de la présence d'interférences parasites. En l'absence d'attribution aléatoire, le risque de biais est consubstantiel aux études non expérimentales.

Ces enquêtes d'observation sont appelées à être différenciées en fonction de leur contenu selon qu'il s'agit d'études exploratoires ou d'études comparatives.

Les études exploratoires sont conduites dans le cadre d'études non contrôlées monobras qui ne permettent pas de rechercher une relation de cause à effet en l'absence de comparateur. En pharmaco-épidémiologie, une enquête sera alors dite descriptive si l'on peut affirmer que tout se serait passé de la même façon, s'il n'y avait pas eu d'étude (dosage et durée des traitements, modalités de prise en charge et de surveillance, risques encourus, etc.). Le chercheur ne fait qu'observer une population, un phénomène, sans intervenir en quoi que ce soit sur le cours naturel des choses. Il se contente d'identifier les trajectoires

des patients dans le système de soins et de dénombrer leurs contacts avec les professionnels de santé ou les établissements. Dans ces études non-interventionnelles, les participants ne sont assignés à aucune exposition, et le choix du traitement et les modalités de son suivi ne dépendent que des patients et de leurs médecins, reflétant ainsi les pratiques de routine de la prise en charge d'une maladie [1]. L'intérêt de ces études est qu'elles partent des réalités de terrain pour les décrire de la manière la plus fidèle possible. Leur inconvénient majeur est qu'elles ne comportent pas de comparateur. L'absence de plan expérimental rend impossible la recherche des causalités.

Les études comparatives sont conduites dans le cadre de l'épidémiologie analytique (cohorte : exposés non exposés) ou interventionnelle (quasi expérimentales¹ avec comparateurs, mais sans randomisation) dont les caractéristiques rendent possible la recherche d'effet causal. Logiquement pour évaluer le lien de cause à effet qui existe entre un traitement et un résultat de santé pour un individu donné, il faudrait comparer le bénéfice que le sujet a retiré de la mise en place du traitement au résultat qui aurait été le sien s'il n'avait pas été traité ou s'il avait reçu un traitement différent. Une situation réelle et observée serait potentiellement comparée à une situation virtuelle non observée, communément qualifiée de « situation contrefactuelle ». Une telle comparaison est nécessairement une fiction, puisque le même sujet ne peut pas au même instant être bénéficiaire et non-bénéficiaire d'un traitement. Il y a forcément l'une des deux situations qui n'existe pas dans la réalité. Une situation contrefactuelle ne peut être estimée qu'en reconstruisant la situation inobservée à partir de la situation observée chez ceux qui n'ont pas reçu le traitement [2]. Puisqu'il est impossible de mettre en place une expérience personnelle contrefactuelle, les analystes recherchent des clones et constituent un groupe témoin pour disposer d'un schéma d'étude qui soit aussi proche que possible de la situation précédente. La tâche de l'évaluateur est donc de constituer un groupe témoin à partir du groupe des bénéficiaires du nouveau traitement, en sélectionnant des individus qui n'en ont pas bénéficié mais qui présentent des caractéristiques semblables. Autrement dit, on cherche à mesurer ce qui se serait passé en l'absence de prescription du médicament pour mieux en mesurer les effets [3]. Ces études estiment l'effet « propre » du traitement en établissant des différences entre les deux groupes.

La principale distinction entre les deux approches repose sur la présence ou non d'une hypothèse de travail spécifiée a priori. Alors qu'une étude exploratoire se contente de décrire les pratiques ou de veiller au respect des recommandations sur la base d'un calcul d'effectif dont la seule finalité est d'améliorer la précision des intervalles de confiance, une étude comparative s'efforce d'atteindre une rigueur aussi proche que possible que celle qui est obtenue dans un ECR. Il convient donc d'émettre une hypothèse de

la même façon que si les patients avaient été randomisés, ainsi que de définir l'exposition, les résultats, et de calculer le nombre de sujets qu'il est nécessaire d'inclure pour espérer obtenir une valeur p significative sur le critère de jugement principal de l'étude [4].

Niveaux de preuves

Il existe une grande gamme de schémas d'étude disponible pour réaliser l'évaluation d'une intervention de santé. Ces schémas garantissent plus ou moins bien que des facteurs extérieurs ne viennent perturber l'estimation de l'effet d'une intervention.

Standard de référence

Les essais comparatifs randomisés sont considérés par beaucoup de scientifiques comme le « gold standard » pour estimer les effets d'une intervention. Cette supériorité sur les études contrôlées non randomisées a été établie dans un article célèbre publié en 1982 [5]. Les auteurs se proposaient de comparer l'efficacité relative de six traitements en fonction du type de protocole utilisé pour estimer leurs performances : les résultats obtenus au regard des mêmes critères de jugement dans 50 essais randomisés comparatifs et dans 56 études comparatives non randomisées avec groupes témoins historiques ont fait l'objet d'un rapprochement systématique. Les sujets du groupe « traitement testé » obtenaient des résultats similaires quel que soit le schéma d'étude utilisé ; les résultats obtenus dans le groupe témoin des études comparatives non randomisées étaient, en revanche, inférieurs à ceux observés dans les populations homologues des essais. Du fait de ce biais de sélection, la taille des effets des traitements innovants relevée dans les études d'observation était systématiquement supérieure à celle constatée dans les essais randomisés. Les critiques des études comparatives non randomisées dépassent aujourd'hui le cas des études avec témoins historiques, et s'étendent aussi bien aux études comparatives de cohortes qu'aux études cas-témoins [6].

Les défenseurs les plus engagés de « l'evidence-based medicine » ont logiquement proposé de hiérarchiser la qualité des données probantes disponibles en fonction de la robustesse des schémas d'expérience sur lesquels elles s'appuyaient (Tableau 1).

Le système de notation de la preuve de la Haute Autorité de santé (HAS) considère lui aussi, que les ECR de forte puissance fournissent le plus haut niveau de preuve, celui de la « Preuve scientifique établie », tandis que les études comparatives observationnelles n'apportent qu'une « Présomption scientifique » [7] (Tableau 2).

Quel que soit le système de cotations retenu, qu'il soit britannique ou français, les essais comparatifs randomisés se retrouvent en tête du classement dès lors que le critère de validité interne de la preuve est choisi comme base de la hiérarchisation. Au sommet de la pyramide, on trouve les essais contrôlés, au centre, les études comparatives non randomisées, puis les études non contrôlées et non randomisées et enfin les opinions d'experts (cf. Tableaux 1 et 2). La force probante de la preuve est d'autant plus forte qu'elle

¹ D'autres appellations et acronymes recouvrent, à peu de choses près, les mêmes schémas d'études : *Comparative Effectiveness Research* ou CER aux États-Unis (Patient-Centered Outcomes Research Institute, 2012) ; *Relative Effectiveness Assessment* ou REA en Europe (*European Network of Centres for Pharmacoepidemiology and Pharmacovigilance*, 2011).

Tableau 1 Système de notation de la preuve selon le schéma d'étude utilisé (NICE 2005).
Hierarchy of evidence according to the study design (NICE 2005).

Niveau de preuve	Type de preuve
1++	Méta-analyses de haute qualité, revues systématiques des ECR, ou ECR (individuel ou par grappes) présentant un risque de biais très faible
1+	Méta-analyses bien menées, revues systématiques des ECR, ou ECR présentant un risque de biais faible
1– ^a	Méta-analyses, revues systématiques des ECR, ou ECR avec un fort risque de biais ^a
2++	Revue systématique de haute qualité, ou études comparatives non randomisées de haute qualité (études avant–après comparatives, séries temporelles interrompues appariées), cohortes prospectives ou rétrospectives comparatives et analyses multivariées avec un risque très faible de biais de confusion et une très forte probabilité d'imputabilité
2+	Études comparatives non randomisées bien menées (études comparatives non randomisées : études comparatives avant–après, séries chronologiques interrompues comparatives), cohortes comparatives et études multivariées avec un risque faible de biais de confusion et une forte probabilité d'imputabilité
2–*	Études non randomisées (avant–après sans groupe témoin, série temporelle interrompue sans groupe témoin), cohortes non comparatives et études multivariées avec risque substantiel de biais de confusion et un risque significatif de non-imputabilité.
3	Études descriptives (séries d'observations)
4	Opinions d'experts, consensus professionnel

^a Les études avec un niveau de preuve « – » ne devraient pas être utilisées comme base de référence pour établir une recommandation.

Tableau 2 Système de notation de la preuve selon le schéma d'étude utilisé [7].
Hierarchy of evidence according to the study design [7].

Niveau de preuve scientifique fourni par la littérature	Grades des recommandations
Niveau 1 Essais comparatifs randomisés de forte puissance Méta-analyse d'essais comparatifs randomisés Analyse de décision basée sur des études bien menées	A Preuve scientifique établie
Niveau 2 Essais comparatifs randomisés de faible puissance Études comparatives non randomisées bien menées Études de cohorte	B Présomption scientifique
Niveau 3 Études cas-témoin	C Faible niveau de preuve scientifique
Niveau 4 Études comparatives comportant des biais importants Études épidémiologiques descriptives (transversale, longitudinale) Études rétrospectives Séries de cas	

est issue d'études qui se situent au niveau le plus élevé de cette hiérarchie.

Renouveau des études observationnelles

Cependant, ce classement des niveaux de preuves a été remis en question à la fin des années 90. Les auteurs du rapport commandité par le National Coordinating Centre for Health Technology Assessment (NCCHTA) sur ce sujet [8], concluent qu'aucune étude observationnelle, quel que soit le type de schéma utilisé, entraîne une surestimation ou

une sous-estimation systématique de la quantité d'effet du traitement dans les huit interventions étudiées. De la même manière, Concato et al. [9], après avoir fait une méta-analyse des essais comparatifs randomisés et des études comparatives non randomisées réalisées pour évaluer l'impact de cinq interventions différentes, concluent qu'il n'y a pas de différences majeures entre les approches, mais plutôt une convergence entre les estimations des effets du traitement dans les essais comparatifs randomisés, les études comparatives de cohorte et les études cas-témoins. Pour les auteurs, les raisons d'une telle convergence sont

liées à l'amélioration de la qualité des études observationnelles dont les protocoles se rapprochent de plus en plus de la rigueur qui préside à la mise en place des essais cliniques : meilleur choix du comparateur (abandon des groupes témoins historiques et constitution de groupes appariés au sein d'une même étude sur la base des mêmes critères d'inclusion et de non-inclusion), choix d'une date commune d'inclusion pour le recrutement des groupes, analyse en intention de traiter, ou encore ajustement des résultats en fonction des caractéristiques initiales des patients inclus dans chacun des groupes.

Vers une hiérarchie de la preuve moins dogmatique

À la lumière de ces travaux, Barton, dans son éditorial du *British Medical Journal* du 29 juillet 2000 [10] s'est fait le défenseur d'une position plus nuancée. Sans remettre en cause la suprématie des essais dans la hiérarchie de la preuve : « le meilleur essai l'emportera toujours vis-à-vis de la meilleure étude observationnelle disponible », l'auteur affirme in fine que c'est la qualité des études disponibles, aussi bien au niveau de leur conception que de leur réalisation, qui doit déterminer le type de schéma d'étude sur lequel on décide de fonder son argumentaire. La force probante des conclusions d'un essai randomisé de mauvaise qualité sera moindre que celle d'une étude observationnelle bien faite. Les études observationnelles, en analysant des populations plus larges, sont mieux à même de détecter les effets indésirables des interventions et elles peuvent être mises en place dans des situations où les essais comparatifs randomisés seraient irréalisables ou éthiquement inacceptables.

Aux États-Unis, l'US Task Force on Community Preventive Services (TFPCS) a proposé un classement des études en fonction de leur validité interne et de leur plus ou moins grande adéquation par rapport aux objectifs poursuivis [11]. Les schémas d'études qui mesurent l'efficacité d'une intervention par rapport à un groupe témoin sont considérés comme apportant un bon niveau de preuve. Cette graduation place en tête, les études comparatives et prospectives. Cette classification a été élaborée pour permettre une revue des études portant sur des actions préventives. En effet, dans ce domaine, les études randomisées sont rares. Cette classification ouvre la possibilité que les résultats issus d'études non randomisées mais comparatives apportent un bon niveau de preuve.

En France, les financeurs publics sont désormais demandeurs de preuves en vie réelle (PVR) en tant qu'outil d'aide à la décision [12]. Ainsi, des contrats de partage de risque avec les industries pharmaceutiques nécessitent désormais une preuve de l'efficacité en conditions réelles d'usage, que nous désignerons désormais sous le nom d'effectivité clinique, afin de fixer le prix final d'un médicament après sa mise sur le marché. Ils permettent de répondre à ce besoin de PVR, tout en ne freinant pas la diffusion de l'innovation [13,14].

Plusieurs raisons expliquent ce virage. De nombreux facteurs pouvant influencer l'effet du traitement en vie réelle ne sont pas pris en compte dans les essais : les comorbidités et traitements concomitants, l'observance, le suivi

des bonnes pratiques, l'accès au soin, la communication du médecin, et des facteurs socioéconomiques [15]. Des modifications d'exposition comme les changements de traitement font aussi partie des pratiques courantes de prise en charge d'une maladie, et ils ne peuvent être appréhendés que dans le cadre d'une étude observationnelle [4]. De plus, les critères d'inclusion de la population retenus lors d'un essai rendent celle-ci non représentative de la population cible, et son homogénéité reflète mal la diversité des cas en vie réelle. Enfin, le traitement qui est évalué est dans la vraie vie est confronté à un comparateur actif, utilisé en routine et non à un placebo. Bien qu'on puisse s'attendre à une diminution de l'efficacité dans une situation de vie réelle par rapport aux conditions expérimentales, ce n'est pas toujours le cas et des médicaments bien intégrés dans une prise en charge optimisée pourraient même engendrer une quantité d'effet supérieure [16].

Les faits ne sont pas des preuves

L'existence de biais est consubstantielle aux études observationnelles en condition normale d'usage. L'épidémiologiste David Sackett [17], dans un article de 1979, avait identifié 35 types de biais. La dernière publication faite sur le sujet en 2004 en recense 70 [18]. Colliger des données factuelles en situation réelle d'usage n'implique donc pas qu'un lien de cause à effet ait été scientifiquement démontré. Contrairement à un essai clinique randomisé, rien ne garantit que la comparabilité des patients inclus dans chacun des groupes d'une étude observationnelle avec comparateur soit assurée, puisqu'il n'y a pas de tirage au sort de leur affectation.

Le défaut de validité interne, conséquence de l'absence de randomisation, est une des faiblesses majeures des études observationnelles [19]. Bien que les études observationnelles soient généralement utilisées pour déceler les associations possibles entre un niveau d'exposition et un résultat de santé, le but ultime de ces schémas est bien d'appréhender s'il existe ou non une relation de cause à effet entre les degrés d'exposition observés et la survenue d'événements indésirables ou de complications. Il est cependant très difficile d'établir un tel lien causal dans le cas d'études observationnelles puisqu'on ne sait jamais si la survenue des événements est due au facteur de risque identifié, soit à d'autres facteurs qui sont à la fois associés avec ces facteurs de risque et à la fois avec les événements constatés, soit à des facteurs qui majorent ou minorent leurs manifestations.

C'est tout le problème que pose la présence de facteurs modifiants et de facteurs de confusion dans les études observationnelles. Deux concepts qu'il convient de distinguer soigneusement. Un facteur modifiant indique que plusieurs causes interagissent sur le même processus causal entraînant des effets multiplicatifs qui minorent ou majorent ses conséquences, tant au niveau de l'exposition au risque qu'au niveau des effets du traitement [20]. En revanche, un facteur de confusion est le résultat du mélange des effets de plusieurs causes qui agissent de façon indépendante. Ainsi, la cause des événements dans les études observationnelles n'apparaît pas de manière aléatoire mais est toujours imbriquée dans une multitude d'autres facteurs, tant et si bien qu'il est difficile de distinguer quelle cause est à l'origine

de l'évènement. Ces facteurs rendent l'exposition dynamique en ce sens que les conditions d'exposition changent en permanence au cours du temps, ce qui complique leur appréhension et rend leur modélisation difficile. Un vrai défi pour l'analyse des données en vie réelle.

Modificateurs d'exposition et d'effet

En vie réelle, les expositions à un traitement dépendent des comportements à la fois des médecins et des patients. Le suivi des recommandations et l'adaptation du traitement à la spécificité de chaque cas dépendent des réponses du patient au traitement ou des conditions de sa prise en charge par les organismes sociaux, autant d'éléments qui sont susceptibles d'infléchir le comportement des cliniciens et de modifier les conditions dans lesquelles le patient sera amené à bénéficier du traitement, c'est-à-dire à y être exposé. De la même façon, le malade peut être amené à modifier son comportement, vis-à-vis du traitement en fonction de l'efficacité qu'il lui prête, de ses effets indésirables, de la perception qu'il en a, ou encore de son prix et de ses conditions de remboursement. Pour un niveau d'exposition donnée, des différences entre les effets d'une intervention peuvent être observées. Les comorbidités, les coprescriptions, les facteurs démographiques, les facteurs génétiques, ou le style de vie en font partie.

Ces modificateurs d'exposition et d'effet peuvent être spécifiques ou non aux données colligées en vie réelle, et propres ou non aux pays étudiés. Les modalités de traitement et les habitudes de prescriptions du médicament dépendent des pratiques du pays dans lequel ils sont mis en œuvre ; elles doivent être étudiées localement en vie réelle. En revanche, l'administration de coprescriptions ou les utilisations hors AMM sont les phénomènes habituels de ce qui se passe en condition normale d'usage ; ils peuvent donc être étudiés dans le cadre d'enquêtes observationnelles internationales. Les caractéristiques sociodémographiques correspondent à des propriétés intrinsèques de la population du pays étudiée. Leurs analyses ne peuvent être faites que dans le cadre d'études locales : enquêtes de santé, registres, essais cliniques nationaux. En revanche, les effets pharmacologiques peuvent être considérés comme universels et doivent être étudiés dans le cadre d'études internationales. Avant de démarrer l'analyse de données d'une étude observationnelle, il est important d'identifier ces modificateurs et de les caractériser.

Biais de sélection et facteurs de confusion

Ces modificateurs d'exposition et d'effet sont susceptibles de majorer les biais et les facteurs de confusion. Un biais, en épidémiologie, est une erreur systématique qui s'introduit dans une enquête et qui tend à produire une estimation différenciant systématiquement de la vraie valeur et conduisant une interprétation erronée du résultat. En d'autres termes, un biais désigne tout effet qui altère la représentativité des résultats. Un facteur de confusion est un facteur qui perturbe l'association entre l'exposition et le résultat obtenu.

Dans le premier cas, la présence d'erreur systématique fausse l'estimation des données recueillies et invalide le calcul de leurs résumés statistiques. L'estimation des paramètres d'intérêt au niveau de l'échantillon étudié

est une mesure inexacte de leurs vraies valeurs dans la population. Un phénomène courant à prendre en considération lors d'une évaluation de l'effectivité clinique tient à l'existence de biais d'indication [20]. Très souvent celui-ci est observé lorsqu'un nouveau traitement est introduit sur un marché où d'autres traitements efficaces existent déjà. Dans cette situation, le nouveau traitement sera prescrit essentiellement aux patients qui ne répondent pas aux traitements actuels ou aux malades qui sont les plus sévèrement atteints. Un biais de sélection peut donc fausser les conclusions de l'étude en minimisant à tort l'effet du nouveau traitement, alors que ce dernier est prescrit en majorité à des patients à des stades sévères de la maladie ou résistants aux traitements actuels. Cet effet est d'autant plus préoccupant qu'il est particulièrement marqué lorsqu'il survient juste après la mise sur le marché du médicament, période durant laquelle beaucoup d'études observationnelles comparatives sont effectuées. Deux considérations rendent l'estimation du biais d'indication difficile. La première est le manque de variables pertinentes, qui indiqueraient par exemple la sévérité de la maladie ou la résistance au traitement, le biais se trouve alors majoré par les caractéristiques personnelles du patient, la seconde est que la présence de ce biais s'estompe dans le temps, puisqu'il s'atténue au fur et à mesure que s'allonge le temps de présence du médicament sur le marché.

Dans le second cas, c'est la présence de tiers-facteurs qui entraîne des erreurs d'appréciation dans l'interprétation des résultats. Prenons un exemple simple [21] : il existe une forte relation entre le fait d'avoir des cheveux gris et le fait d'être frappé d'un infarctus du myocarde. Le rapport de risque liant ces deux grandeurs est significativement supérieur à 1 (il s'élève à 1,73). En réalité, derrière cette liaison immédiate se cache un troisième facteur, l'âge, qui est étroitement associé à la couleur du cheveu et qui est également lié à la survenue des infarctus. Lorsqu'on introduit cette donnée dans l'analyse en calculant un risque relatif ajusté sur l'âge, l'association entre les cheveux gris et l'infarctus est égale à 0,80, autrement dit il n'existe aucun rapport entre ces deux grandeurs. De même, la guérison d'un rhume banal après l'administration d'un traitement antiviral peut être liée à une résolution spontanée de la maladie beaucoup plus qu'au traitement lui-même. L'évolution du cours naturel de la maladie est un facteur de confusion, c'est-à-dire un facteur qui produit un effet similaire à celui que l'on attend du traitement et qui peut être mis sur le compte du traitement, alors même que celui-ci est dépourvu d'efficacité.

Ceux qui oublient la présence de ces biais et de ces facteurs de confusion risquent d'attribuer à tort l'intégralité de l'effet total du traitement mis en œuvre à l'effet propre du médicament, bien qu'une partie de cet effet total ne lui soit aucunement imputable. Il convient donc distinguer « l'effet propre » du traitement et « l'effet de population » qui résulte des différences préexistantes entre les niveaux de santé initiaux de ceux qui ont bénéficié du traitement, et de ceux auxquels le traitement n'a pas été administré.

Les recherches en vie réelle risquent, lorsqu'elles ne sont pas conduites sur la base d'un protocole préétabli et rigoureux, de ne pas échapper à la tentation du « dragage de données », à savoir la multiplication des analyses jusqu'à ce que l'une d'entre elles fournisse le résultat espéré. Elles

peuvent de surcroît faire l'objet de biais de publication, compte tenu de la tendance des journaux professionnels à ne publier les résultats que s'ils sont positifs [22]. En réponse à ces préoccupations, un rapport de l'International Society for Pharmacoeconomics and Outcomes Research (ISPOR) offre des recommandations quant à la conduite d'une étude observationnelle comparative dans le but d'améliorer son intégrité et la confiance dans les preuves avancées [23].

Comment administrer la preuve de la causalité en vie réelle ?

Schémas d'étude

Le schéma de l'étude est un premier élément à prendre en considération pour améliorer la validité interne des études observationnelles. Chacun de ces schémas présente des avantages et des inconvénients, et ils permettent de neutraliser les biais à des degrés divers.

Essai contrôlé pragmatique

Un essai contrôlé pragmatique (ECP) se situe à mi-chemin entre une étude observationnelle et un essai contrôlé randomisé ; il s'efforce de préserver leurs avantages respectifs. Comme lors d'un ECR, une randomisation est effectuée et chaque patient est affecté à un groupe (intervention ou comparateur), avec pour objectif de réduire les problèmes engendrés par les modificateurs d'exposition et d'effet. En revanche, le reste de l'étude se rapproche du schéma d'une étude en vie réelle. Il s'agit de s'assurer de la valeur pratique d'un nouveau traitement, après que les praticiens l'aient intégré en routine, soit au minimum au bout de 2 ans, dans le cadre « d'essais pragmatiques » de phase IV. Les principes énoncés par leurs promoteurs [24,25] sont les suivants :

- le recrutement des sujets doit être représentatif des malades « tout-venant » susceptibles de recevoir le traitement ;

- les traitements de référence choisis correspondront à ceux qui sont d'usage courant dans l'indication envisagée ;
- les critères de résultat mis en œuvre doivent être très variés, on s'intéressera à l'efficacité, aux effets secondaires, à la qualité de vie, à l'observance, à la survenue d'évènements intercurrents ;
- les conditions de surveillance doivent correspondre à celles que l'on observe en pratique courante, sans qu'aucune contrainte spécifique ne soit imposée ;
- l'analyse des résultats se fera selon une procédure intermédiaire entre l'analyse en intention de traiter, non pragmatique, et l'exclusion systématique des sorties d'essai, source de biais (Tableau 3).

Ainsi, contrairement à un ECR, la population cible sera celle à qui le traitement est prescrit dans la vraie vie, et sera de fait plus représentative de l'hétérogénéité de la population générale. Également, le comparateur sera l'un des traitements alternatifs existants sur le marché. L'utilisation du médicament, l'observance des patients et les changements de traitement ne feront l'objet d'aucune intervention, et refléteront ainsi les pratiques courantes afin d'obtenir des preuves en vie réelle. Le résultat choisi sera de préférence pertinent au niveau clinique mais aussi au niveau de la santé publique (le taux d'hospitalisation, par exemple). Enfin, moins d'interactions auront lieu après la randomisation, de façon à ne pas influencer les pratiques du patient ou sa motivation. Néanmoins, la randomisation d'un ECP est une limite par rapport à l'objectif de représenter le plus fidèlement possible ce qui se passe dans la vraie vie. Le consentement des patients de participer à l'étude ou encore l'information qu'ils reçoivent sont autant d'éléments qui distinguent les patients d'un essai randomisé de ceux de la population générale.

Étude cas-témoin

Les schémas d'étude de type cas-témoin sont une autre façon d'évaluer l'effectivité clinique d'une intervention. Ces études rétrospectives, qui supposent que les résultats soient déjà connus avant qu'elle ne commence, sont particulièrement intéressantes à mettre en œuvre lorsque

Tableau 3 Essai randomisé pragmatique versus essai contrôlé randomisé. *Pragmatic controlled trial versus Randomized clinical trial.*

	Essai randomisé pragmatique	Essai contrôlé randomisé
Objectif	Effectivité clinique (pratique courante)	Efficacité expérimentale (conditions idéales)
Population	Choisir le bon traitement Hétérogène Peu sélectionnée Représentative de la cible	Jugement de causalité Homogène Très sélectionnée « saine »
Comparaison	Protocole de prise en charge	Traitements
Randomisation	Oui	Oui
Aveugle	Non en général (ouvert)	Double insu
Critères de jugement	Multiples ou composite	Souvent critères intermédiaires
Analyse	Intention de traiter (ITT) Choix sans test Le meilleur résultat	Par protocole ou ITT Recherche de la <i>P-value</i> Statistique de test

la variable d'intérêt est un évènement rare. Le principe consiste à comparer la fréquence avec laquelle les sujets ont été exposés dans le passé à un facteur de risque dans un groupe des sujets atteint par la maladie et dans le groupe des témoins qui ne l'ont pas été. Les études cas-témoins sont des études qui peuvent être mises en place rapidement pour un coût relativement modeste et permettent d'obtenir des résultats dans un laps de temps très court. Il ne faut cependant pas oublier que ce type d'étude est celui qui est susceptible de présenter, parmi tous les schémas d'études disponibles, le plus de biais. Leur caractère rétrospectif, empêchant souvent une mesure précise et adéquate des expositions, peut être un frein à leur validité interne [26]. S'ajoute à cela un potentiel biais de mémorisation lorsque les données sont récoltées sur la base d'une enquête, les personnes interrogées se souvenant avec une acuité différente des évènements auxquels ils ont été confrontés selon le groupe de comparaison auquel ils appartiennent. Dans une étude cas-témoins recherchant une association entre médicaments et une malformation congénitale par exemple, les mères ayant donné naissance à un enfant mal formé (les cas) se souviennent, en général, mieux des médicaments qu'elles ont pris au moment de leur grossesse que les femmes qui ont eu des enfants normaux (les témoins). La fréquence d'exposition aux facteurs de risque dans le groupe témoins est sous-estimée et l'*odds-ratio* entre les deux groupes sera surestimé [27].

Étude de cohorte

Un autre schéma d'étude est régulièrement utilisé en pharmaco-épidémiologie, il s'agit de la cohorte. Ce terme est utilisé pour désigner des groupes de sujets suivis au cours du temps. Le début de l'étude peut se situer à un instant donné du présent et la cohorte est suivie de façon prospective dans les années qui suivent l'inclusion, on parle alors de cohorte prospective. Ou l'inclusion peut se faire à partir d'une date choisie dans le passé, et l'on parle alors de cohorte rétrospective ou historique. Dans les deux cas, les cohortes qui font l'objet d'un suivi vont toujours de l'avant comme les légions romaines. On descend le temps tandis que dans les études cas-témoins, on le remontait. Lorsque la variable d'exposition est dichotomique, on parle d'étude exposés/non exposés. Lorsqu'il s'agit de comparer les résultats entre plusieurs groupes exposés à différentes interventions avec comme objectif soit d'en décrire les modalités, soit d'en mesurer l'effet, on préfère parler d'étude longitudinale.

Dans une cohorte rétrospective, les données sont recueillies une fois que les décisions thérapeutiques ont été prises et que le traitement a été administré, de sorte que le recueil n'a aucune incidence sur le déroulement de la prise en charge clinique. On peut alors dresser le profil longitudinal de l'utilisation d'un traitement, ce qui permet de comparer les résultats obtenus sous une option thérapeutique pendant toute la période d'observation, tout comme dans un essai clinique. À cet égard, les études de cohorte se rapprochent d'un schéma expérimental puisque l'exposition y est connue a priori (mais sans tirage au sort), et puisque l'on compare l'incidence ultérieure des évènements entre le groupe exposé et le groupe non exposé dans les deux cas. Autant de caractéristiques communes qui expliquent

le rang élevé des cohortes dans la hiérarchie des preuves scientifiques.

Dans une cohorte prospective, les groupes sont suivis à partir de l'instant où ils sont exposés jusqu'à l'atteinte du résultat recherché, et les sujets recrutés selon leurs expositions ou non à plusieurs traitements alternatifs. Ce type de schéma est souvent pertinent pour répondre à un but de recherche d'effectivité clinique comparative. La collecte de données se fait dans de meilleures conditions que dans un schéma cas-témoin, et l'absence d'intervention (pas d'assignation des patients à un traitement ni de suivi supplémentaire en dehors des visites de routine) offre un suivi proche de la réalité et ainsi une meilleure représentativité des patients que lors d'un essai clinique pragmatique. Les enregistrements de prescriptions, de changements et d'interruptions de traitement reflètent de la meilleure des façons les dynamiques d'exposition dans une prise en charge courante.

L'exemple de l'entrée sur le marché de la rispéridone injectable à libération prolongée dans la prise en charge de la schizophrénie en France illustre bien l'apport potentiellement considérable d'un bon schéma d'étude observationnelle. Cet antipsychotique, le premier de seconde génération à libération prolongée, s'est vu refuser son remboursement dans plusieurs pays européens dû à un manque de preuve d'effectivité clinique comparée à d'autres alternatives. Il était en revanche remboursé en France grâce à un contrat conditionnel entre le laboratoire Janssen Cilag et le payeur français. Une étude de cohorte de 2500 patients a alors permis d'évaluer son effet relativement à d'autres antipsychotiques avec comme résultat le taux d'hospitalisation. Il est à noter que ce résultat, influencé par un suivi régulier et méticuleux des patients, n'aurait pas pu être mesuré de manière pertinente lors d'un ECR. Suite à l'analyse statistique effectuée sur les données de cette cohorte, une diminution du risque d'hospitalisation de 34 % par rapport aux autres traitements a été démontrée [28].

Méthodes statistiques

Du schéma d'une étude observationnelle découle un plan d'analyses statistiques, visant à identifier et limiter les biais inhérents aux données en vie réelle. Du fait des problèmes causés par la présence de facteurs de confusion, la rédaction du plan d'analyse a un rôle encore plus important que lors d'un essai clinique randomisé, pour lequel de simples tests de comparaison suffiront. En vie réelle, les traitements ne sont jamais prescrits au hasard. L'indication du médecin est posée en fonction des caractéristiques observables cliniques, fonctionnelles ou comportementales des patients. Or ces paramètres sont autant de facteurs pronostiques de l'évolutivité de la maladie. Les choix thérapeutiques dépendent également de facteurs de risque plus subtils (comme le désir d'apporter une réponse à une souffrance psychologique ou sociale) qui sont autant de facteurs parasitaires non observables. Par exemple, on peut observer le traitement prescrit à un patient, mais on ne peut pas observer s'il lui a bel et bien été administré. Certains traitements peuvent être plus contraignants que d'autres, et on peut rapporter une bonne observance pour les uns, et une mauvaise pour les autres. Ainsi, l'effet d'un traitement peut être

sous-estimé par sa mauvaise observance, quand un autre ne rencontrera pas ce genre de problème. Une autre raison d'une sous-estimation d'un effet peut être une résistance au traitement, qui touchera seulement certains patients. Cette résistance ne peut pas être mesurée et peut ainsi biaiser les estimations. Il est impératif de distinguer et d'isoler l'effet des caractéristiques observables ou non observables des sujets, pour les distinguer de l'effet spécifique du médicament sur les variables d'intérêt, en mettant en œuvre des études observationnelles avec comparateur et emploi de méthodes statistiques de neutralisation des biais.

Selon le type de facteurs de confusion, les techniques utilisées pour neutraliser les biais qu'ils engendrent ne seront pas les mêmes. Ils peuvent être de natures différentes : observables, non observables et mesurables, ou non observables et non mesurables.

Neutralisation des biais dus aux facteurs observables

Les biais dus aux facteurs observables peuvent se traduire de deux façons : dans un cas, il existe une forte corrélation entre le facteur de confusion et le résultat. Le biais peut alors être contrôlé en ayant recours à une analyse de régression standard (régression linéaire, régression logistique, régression de Cox...) ; dans le second cas, le choix du traitement est fortement corrélé aux caractéristiques de la population des patients sélectionnés dans chacun des groupes. Il peut alors être contrôlé par la méthode du score de propension.

Les méthodes de régression

Les méthodes de régressions multiples permettent de déterminer l'association unique entre un traitement et un résultat, après prise en compte des autres variables explicatives présentes dans le modèle (tels que les caractéristiques du patient), on parle donc de résultats ajustés. Les variables pouvant potentiellement influencer le choix du traitement ou le résultat sont intégrées dans le modèle. Des proxys pertinents de variables manquantes peuvent aussi être utilisés. Dans le cas de variables omises, des estimations biaisées de l'effet du traitement sont à attendre [29].

Le modèle utilisé dépend du type de variables à expliquer. Lorsqu'elles sont continues, la grande majorité des applications effectuent une estimation par la méthode des Moindres Carrés Ordinaires (MCO) dans le cadre d'une régression linéaire. Une variable discrète, le nombre d'hospitalisations par exemple, nécessite quant à elle un modèle de comptage, comme le modèle de Poisson. Dans le cas d'une variable qualitative, le fait de subir un infarctus du myocarde par exemple, une régression logistique est d'usage et l'estimation se fait par la méthode du maximum de vraisemblance. Le résultat représente ici la probabilité de survenue de l'évènement. Dans le cadre de la régression de Cox en revanche, appelée aussi modèle à risques proportionnels, la variable expliquée s'exprime en temps passé avant la survenue d'un évènement. On modélise donc le « quand » au lieu du « si ». Cette régression a l'avantage d'autoriser les variables explicatives à varier dans le temps, de la même façon que le temps avant la survenue de l'évènement varie. Parce que cette caractéristique dynamique est propre aux expositions à un traitement,

la régression de Cox est un outil puissant en pharmaco-épidémiologie. De plus, elle ne suppose pas de distribution particulière pour les temps de survie.

Néanmoins, elle repose sur une lourde hypothèse, celle de la proportionnalité des risques. La régression de Cox autorise les taux de risques à varier dans le temps, mais suppose qu'ils sont proportionnels entre les deux groupes. En d'autres termes, les patients des groupes témoin et intervention pourront avoir des taux de risques différents au cours du temps, mais leurs risques relatifs devront rester constants. Il convient donc de tester cette hypothèse avant d'envisager toute régression de ce type.

Plusieurs méthodes de régressions sont donc disponibles en fonction de la nature de la variable à expliquer. La régression de Cox présente un atout notable lorsqu'il s'agit d'estimer l'association entre une exposition à un traitement et à un résultat, mais repose sur une hypothèse forte. Une autre technique statistique s'est répandue de manière remarquable et est aujourd'hui très populaire, il s'agit du score de propension (SP).

Le score de propension

On peut légitimement imaginer dans le cadre d'une étude observationnelle que ceux qui bénéficient d'une prise en charge spécifique présentent des caractéristiques particulières qui sont autant de facteurs de risque, et que ces caractéristiques auraient joué en leur défaveur, même s'ils n'avaient pas reçu le traitement. Si ce traitement s'avère néfaste, on doit s'attendre à ce qu'ils subissent de son fait un préjudice encore plus grand que celui qui eut été le leur si ce traitement ne leur avait pas été prescrit. Le vrai préjudice de santé est surestimé. Il convient de distinguer et d'isoler l'effet des caractéristiques observables des sujets pris en charge (les covariables ou les facteurs de confusion), de l'effet propre du traitement sur le critère de jugement choisi.

Le score de propension correspond à la probabilité prédite pour un individu, de recevoir un traitement, connaissant ses caractéristiques observables sociodémographiques et cliniques à l'inclusion. Pour le mettre en œuvre, il convient alors d'adopter une démarche en trois temps :

- concevoir en premier lieu, un « modèle d'exposition » indépendamment des effets attendus du traitement, en estimant a priori la probabilité que les sujets en bénéficient sur la base du résumé de leurs « caractéristiques observables », quel que soit le groupe auquel ils ont été affectés ;
- procéder ensuite à « l'appariement » des individus du groupe traité et de ceux du groupe témoin qui ont des caractéristiques proches, en les mettant en correspondance sur la base de leur « score de propension » respectif. Le fait d'être bénéficiaire ou non du traitement devient alors aléatoire (ou tout au moins indépendant de la variable de résultat) comme dans un essai randomisé et le risque de biais est neutralisé aussi bien sur le plan clinique qu'en termes de coût ;
- construire dans un troisième temps, un « modèle d'effet causal » en étudiant la différence de la moyenne observée chez les bénéficiaires du traitement sur la base d'un critère de jugement principal (CPJ), par rapport à celle relevée chez les non-bénéficiaires qui présentent les mêmes scores de propension par quintiles. Le CPJ défini

au préalable se doit d'être simple, pertinent au niveau clinique et de la santé publique, et engendrera un calcul a priori du nombre de sujet nécessaire.

Le score de propension permet d'assurer la comparabilité de deux groupes. Il représente ainsi la probabilité d'être exposé au traitement, conditionnellement aux caractéristiques observables connues à l'inclusion des patients [30]. Lorsque ces caractéristiques individuelles sont résumées par un score unidimensionnel, estimé à l'aide d'une régression logistique et que ce score est équilibré entre individus des deux groupes, l'appariement ne retient en définitive que des individus présentant des caractéristiques personnelles observables semblables. Le fait d'être bénéficiaire ou non du traitement innovant devient alors aléatoire (ou tout au moins indépendant de la variable de résultat) comme dans un essai randomisé. Le risque de biais est neutralisé et seul l'effet directement et exclusivement attribuable au traitement est pris en compte [31].

Appariement versus régression

De nombreuses raisons militent en faveur de l'appariement sur SP de préférence à l'utilisation des techniques de régression [32].

En premier lieu, les distributions des scores de propension des cas traités et des cas non traités permettent d'observer les zones où elles ne chevauchent pas, constituant autant d'ensembles disjoints représentant soit les situations où aucun individu ne bénéficie du traitement innovant, soit celles où tous les individus peuvent en tirer parti. Cette représentation graphique permet de distinguer les patients qui présentent des contre-indications et ceux pour lesquels le traitement s'impose en première intention. Dans ces deux situations, il est impossible d'apparier les membres d'un des deux groupes avec les membres de l'autre groupe. L'élimination de ces patients non comparables restaure une situation d'équilibre, un terme emprunté à la littérature anglo-saxonne désignant l'équivalence qui existe dans l'esprit des prescripteurs entre les diverses options thérapeutiques disponibles [33]. En présence d'une incertitude sur les mérites respectifs des traitements, les médecins choisiront le traitement de manière quasi aléatoire, simulant un ECR. Éliminer les patients pour lesquels le traitement à utiliser va de soi, compte tenu de ses indications ou contre-indications, permet donc de limiter les biais dus à une préférence sélective pour l'un d'entre eux [4].

L'appariement par le SP ne nécessite par ailleurs, aucune hypothèse quant à la forme de la relation entre le score de propension et les résultats, contrairement aux méthodes de régressions linéaires dont les estimations reposent sur l'hypothèse de linéarité entre les variables explicatives et les résultats.

Le SP permet également de travailler sur deux populations dont les caractéristiques socioéconomiques et cliniques observables sont équilibrées. Cette méthode est considérée comme l'une des meilleures qui soit disponible pour parvenir à ce résultat dans des études de cohorte [34]. Il a même été remarqué que ce type d'appariement donne des distributions plus équilibrées entre deux groupes qu'une randomisation [35].

Cette technique d'appariement facilite l'analyse des situations dans lesquelles un petit nombre d'évènements surviennent généralement chez des individus ayant de

multiples facteurs de risque. Dans ces conditions, le SP s'est révélé supérieur aux méthodes de régressions logistiques en donnant de meilleures estimations, moins biaisées et plus robustes [36].

Une analyse stratifiée du SP permet également de mettre en évidence l'hétérogénéité de l'effet d'un traitement [32]. En pharmaco-épidémiologie, l'effet d'un médicament est très dépendant des patients auxquels il est administré, il peut même devenir néfaste pour un patient chez qui il n'est pas ou peu indiqué. En stratifiant les individus selon leurs scores de propension, on peut distinguer l'effet du traitement chez les patients pour lesquels il est fortement recommandé (patients traités avec un SP élevé) et l'effet du traitement chez les patients pour lesquels il l'est moins (patients traités avec un score de propension bas).

Enfin, cette méthode d'appariement permet d'identifier les éléments qui exercent une influence sur le SP, et donc sur la probabilité de recevoir le traitement. Les facteurs de risque d'être traité sont ainsi mis en évidence [32].

D'un point de vue pratique, la construction du score de propension nécessite d'inclure des caractéristiques observables des patients. Ce sont ces variables qui définiront la probabilité d'être exposé au traitement. Il est recommandé d'en inclure le plus possible qui soient reliées aux résultats, peu importe qu'elles ne le soient pas avec l'exposition [2]. En revanche, des variables faiblement reliées aux résultats, même fortement reliées à l'exposition, ne doivent pas être intégrées [37]. L'estimation du score de propension se fait ensuite, dans la grande majorité des applications, par une régression logistique. Il n'a pas été démontré que d'autres approches donnent de meilleurs résultats, et la régression logistique peut apparaître comme plus accessible pour le lecteur [32].

Neutralisation des biais dus aux facteurs non observables

Une des principales difficultés de la neutralisation des biais inhérents aux études en vie réelle est la prise en compte des facteurs de confusion non observables, contrairement aux essais cliniques randomisés et pragmatiques dont la randomisation permet de s'en affranchir. Ces biais « cachés » sont liés à l'existence de variables importantes qui ont été « omises » au départ comme : les plaintes des patients, les habitudes de prescription des médecins, les choix thérapeutiques, ou l'organisation des soins dans un établissement de santé. Ces biais parasitent l'estimation des effets de l'intervention en le majorant ou en le minorant artificiellement. L'enjeu est de pouvoir maîtriser ces biais afin de cerner au mieux l'effet que l'on cherche à mesurer. Le principe de ces méthodes est donc de rendre mesurable ce qui ne l'est pas.

Le score de propension haute-dimension

L'appariement sur SP offre une solution en intégrant des variables servant de substituts aux facteurs inobservables. Ainsi, des variables comme le nombre de consultations, le traitement prescrit ou les hospitalisations sont dépendantes de la facilité d'accès au soin, des habitudes de prescription du médecin, de la gravité de la maladie, ou encore de l'impact du ticket modérateur dans le choix du traitement. Ces variables de substitution permettent d'ajuster le

modèle sur ces facteurs inobservables. Les variables empiriques doivent être incluses, en plus des variables cliniques, dans la construction du score de propension [13].

L'identification de ces consommations dans le cadre d'un score de propension haute-dimension et l'appariement des groupes traités et non traités à l'aide de cet instrument permet de repérer, en fonction des variables empiriques retenues, les différences d'état de santé qui préexistaient, avant traitement, entre les populations. La mise en œuvre de cette méthode dite « des cohortes appariées séquentielles » conçue par Schneeweiss et al. [38] constitue aujourd'hui la méthode de référence à utiliser pour étudier les bases de données médico-administratives de grande dimension.

La publication de Garbe et al. [39] montre l'efficacité de cette méthode qui a été utilisée pour comparer les effets secondaires associés à la consommation de Coxibs par rapport aux anti-inflammatoires non stéroïdiens classiques (AINS). Dans le cadre d'une base de données médico-administrative allemande, 2 242 000 nouveaux utilisateurs ont été étudiés rétrospectivement entre 2005 et 2008, dont 2 104 814 sujets traités par AINS et 37 280 sujets traités par Coxibs. Le risque relatif de complications gastro-intestinales hautes des Coxibs par rapport aux AINS, exprimé en risque relatif brut était égal à 1,21 dans un intervalle de confiance à 95 % de [0,91 ; 1,61]. Ce risque relatif ajusté sur 79 covariables cliniques identifiées par les experts médicaux devenait égal à 0,84 [0,56 ; 1,26]. Lorsque le même calcul était fait dans le cadre d'un score de propension enrichi à partir de 900 variables administratives, le risque relatif n'était plus que de 0,62 [0,43 ; 0,91]. Ces résultats montrent l'importance de la neutralisation des facteurs de confusion dans les bases médico-administratives (Fig. 1).

Une publication française de Blin et al. [40] aboutit aux mêmes conclusions. Elle utilise le score de propension haute-dimension pour comparer un traitement par anticoagulant oral direct ou par anti-vitamine K chez des patients atteints de fibrillation auriculaire non-valvulaire. Les données médico-administratives françaises du SNIIRAM

ont permis d'ajuster le score sur 500 variables réparties en 4 dimensions (affection longue durée, diagnostics d'hospitalisation, médicaments, consultations médicales/paramédicales/biologie). Les résultats ont mis en évidence une nette différence entre les distributions des SP haute-dimension entre les deux traitements, incitant les auteurs à conclure que ce type de méthode permet de construire une variable d'appariement performante pour comparer l'effet d'un traitement entre deux groupes dans le SNIIRAM.

Variables instrumentales

La technique des variables instrumentales, bien que peu utilisée en recherche médicale au contraire de l'économétrie, permet de neutraliser les facteurs de confusion inobservables [41]. Le principe est de recourir à un instrument qui imite le processus d'affectation d'un traitement dans le cadre d'un ECR. La validité de la méthode repose sur le respect des trois conditions rappelées par Ghabri et Launois [42] :

- l'instrument doit être fortement corrélée avec le choix du traitement ;
- l'instrument ne doit avoir aucune influence sur le résultat, autre que celle qu'il exerce indirectement à travers les effets indésirables associés au traitement lui-même (condition d'exclusion) ;
- l'instrument doit être totalement indépendant des éléments non mesurables qui sont susceptibles d'influencer le résultat (condition d'indépendance).

Il faut donc bien spécifier les critères qui président au choix des variables instrumentales pour que celles-ci répondent aux conditions exigées pour avoir un bon instrument.

En général, les bonnes variables instrumentales seront fournies par des « expériences naturelles » qui par définition échappent au contrôle de l'investigateur : on utilise des variations exogènes de l'environnement qui expliquent

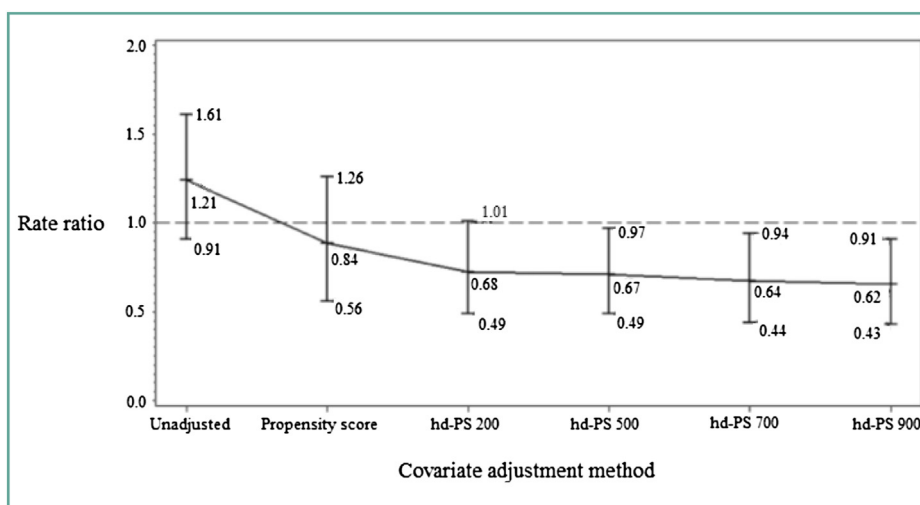


Figure 1. Risque relatif des complications gastro-intestinales hautes Coxibs versus AINS en fonction des méthodes d'appariement utilisées. Hazard ratio of gastro-intestinal complications Coxibs versus NSAID according to the adjustment method.

la probabilité plus ou moins élevée de bénéficier d'un traitement ou de participer à un programme sanitaire.

Chen et Briesacher [43] ont listé quelques exemples d'instruments utilisés dans le domaine de la santé. Pour expliquer le choix de subir ou non une cathétérisation cardiaque, McClellan et al. [44] ont utilisé la distance du patient à l'hôpital le plus proche. Pour expliquer le choix d'un traitement par AINS anti-COX2 sélectifs (comparé à un traitement par AINS non sélectif), Brookhart et al. [45] ont utilisé la dernière prescription d'AINS du dernier patient reçu par le médecin. Enfin, pour modéliser l'effet de l'instrument sur l'exposition au traitement par olanzapine, Zhang [46] a utilisé le laps de temps qui s'était écoulé depuis l'autorisation de sa mise sur le marché par la Food and Drug Administration (FDA). Dans tous les cas, les conditions de validité de l'instrument ont été respectées :

- l'instrument est bien associé au choix de l'intervention ;
- l'instrument et le résultat n'ont pas de cause commune ;
- l'effet de l'instrument sur le résultat passe par l'exposition à l'intervention.

Une fois l'instrument choisi, on compare les utilisateurs d'un traitement, qui l'ont utilisé seulement parce qu'ils ont été exposés à une certaine valeur de la variable instrumentale, au groupe de ceux qui ne l'ont pas utilisé parce qu'ils ont été exposés à une autre valeur de cette même variable. La méthode permet de scinder la population en deux catégories : la population test et la population témoin, comme le ferait un essai randomisé et permet de s'affranchir des facteurs de confusion inobservables qui se retrouvent répartis entre les deux groupes de façon équilibrée.

Une comparaison empirique entre les techniques de régression, le score de propension et les variables instrumentales a été réalisée par Stukel et al. [47] sur l'effet de la cathétérisation cardiaque sur la survie globale de patients ayant subi un infarctus du myocarde. Les résultats démontrent une sensibilité de l'estimation à la technique utilisée. Le risque relatif estimé par la régression de Cox est de 0,51 dans un intervalle de confiance à 95 % de [0,50 ; 0,52]. Le risque relatif estimé par le score de propension est de 0,54 [0,53 ; 0,55]. Enfin, celui estimé par la méthode des variables instrumentales (en utilisant le taux de cathétérisation cardiaque régional comme instrument) est de 0,84 [0,79 ; 0,90]. Afin de juger de la validité de ces résultats, les auteurs se sont aidés des risques relatifs obtenus lors d'essais cliniques randomisés. Ces derniers montrent des risques relatifs entre 0,79 et 0,92. L'estimation produite par les variables instrumentales est donc plus proche de celle produite par d'autres méthodes, d'autant que les résultats des ECR devraient avoir tendance à surestimer l'effet de la cathétérisation cardiaque, étant donné l'expertise technologique des médecins conduisant l'essai. La surestimation de l'effet obtenue par les méthodes de régression et du score de propension est probablement due à des facteurs inobservés omis dans la modélisation, que l'utilisation d'un instrument a permis de neutraliser.

Néanmoins, la technique des variables instrumentales rencontre certaines limites. La principale est l'identification d'un bon instrument. La majorité des instruments ont tendance à être soit faiblement corrélés au traitement, on parle d'instrument faible, soit corrélés aux facteurs inobservables. Et des études ont démontré que l'utilisation

d'instrument faible pouvait engendrer des estimations plus biaisées qu'un simple MCO [48–50]. De plus, même si un instrument est fortement corrélé au traitement, l'hypothèse de non-corrélation aux facteurs inobservables n'est pas empiriquement testable. Il est difficile d'être certain que l'effet de l'instrument sur le résultat est transmis uniquement par le traitement.

En présence d'un instrument parfait, la méthode des variables instrumentales est donc celle menant aux meilleures estimations, dues à la plus forte neutralisation des facteurs inobservables. Il est cependant difficile de trouver ce type de variable, et en son absence, d'autres techniques sont préférables à l'utilisation d'un mauvais instrument.

Différence des différences

D'autres techniques statistiques peuvent aussi être utilisées afin de neutraliser les facteurs non observables. La technique de la double différence nécessite un premier suivi avant l'introduction du traitement, puis un second après [51]. Il devient alors possible de calculer la différence entre les observations avant et après l'introduction du traitement, neutralisant ainsi les facteurs inobservables fixes dans le temps. L'opération est répétée pour les deux groupes, intervention et comparateur, et on peut alors estimer la différence intergroupes des différences intragroupes, représentant l'effet seul de l'intervention.

Recueil de données complémentaires en vie réelle

Néanmoins, d'autres méthodes existent pour limiter les problèmes liés aux facteurs de confusion inobservables dans les études observationnelles. Une étude complémentaire avec plus de variables, dans le but d'être étendue à l'étude principale, en est une [52]. Afin de limiter le biais d'indication, dû en partie à des facteurs inobservables comme la sévérité de la maladie ou la résistance au traitement, Amzal [53] propose aussi une étude complémentaire avec plus de données en vie réelle recueillies, la mise en place d'un essai clinique pragmatique, ou encore l'observation de l'évolution de l'efficacité clinique d'un médicament substitut afin d'en évaluer le biais d'indication, pour l'adapter à l'étude principale.

Modèle observationnel transposable en vie réelle

Le besoin de preuves en vie réelle a modifié les modalités de l'évaluation médico-économique, impactant directement les stratégies de développement pharmaceutique [54]. Les payeurs réclament de plus en plus de preuves de l'efficacité d'un traitement sur les populations stratifiées dans le cadre de leur indication en fonction de leur hétérogénéité. Malgré les nombreux schémas d'études et les méthodes statistiques qui permettent de limiter les biais inhérents aux études observationnelles, le risque de leur présence demeure et d'autres techniques permettant de les maîtriser, émergent. Le recours au modèle observationnel transposable en vie réelle (MOTVR) permet d'appliquer les résultats des essais randomisés ou des méta-analyses à des populations différentes de celles dans lesquelles elles ont été recueillies. Pour tirer pleinement parti de cette possibilité, il convient de s'inspirer des directives l'International Conference on

Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) et de mettre en place une « étude de transition ». Une étude observationnelle complémentaire devra être conduite dans le pays où l'étude multicentrique sera transposée, afin que le promoteur recueille des informations propres à cette zone géographique qui lui permettent de procéder aux ajustements appropriés des paramètres socioéconomiques et cliniques clés, sans qu'il lui soit nécessaire de mettre en place une nouvelle étude ou un nouveau modèle [55]. Une telle approche permet simultanément, de jeter un pont entre l'efficacité expérimentale estimée dans le cadre des essais contrôlés randomisés et l'effectivité clinique évaluée dans les conditions normales d'usage en utilisant toutes les données disponibles et en les intégrant dans des modélisations ad hoc [20]. De cette façon, des résultats obtenus sur une population donnée peuvent être extrapolés à une autre et à long terme en tenant compte des caractéristiques spécifiques de la population ou du pays d'intérêt. Cette possibilité d'extrapolation réduit le nombre d'études à mettre en place pour les industries pharmaceutiques dans les phases de développement du médicament.

La construction d'un MOTVR repose sur trois éléments ; elle nécessite :

- le développement d'un modèle simulant informatiquement (in silico) le devenir de la population étudiée ;
- la mobilisation, dans le cadre de celui-ci, des données issues des essais pivots ou des registres disponibles ;
- la mise en place d'une étude en vie réelle pour documenter les caractéristiques sociodémographiques de la population cible nationale et pour estimer l'effectivité clinique du traitement en fonction des conditions locales de prise en charge.

Les données issues des essais cliniques ou des registres sont nécessaires pour identifier des effets intrinsèques attribuables au produit, quel que soit le pays où il est prescrit. Les données en vie réelle permettent de documenter le rôle que jouent les facteurs modifiant l'exposition et l'effet dans une région ou dans un pays particulier. L'avantage de cette méthode est qu'elle permet de mobiliser toutes les données qui documentent les facteurs modifiant l'exposition et l'effet.

Ces études permettent donc d'utiliser des données épidémiologiques d'un certain pays pour l'adapter au contexte d'un autre. Une quantification de l'incertitude peut aussi être effectuée dans le cadre d'un MOTVR, permettant des analyses de sensibilité. Enfin, ce type d'étude peut être mise en œuvre aussi bien avant la mise sur le marché d'un traitement qu'après, l'intérêt des financeurs et des industries pour ces modèles statistiques est donc grandissant.

Une cohorte prospective de 2 ans sur des patients atteints d'hypercholestérolémie traités par ézétimibe en France (cohorte EZE), réalisée dans le cadre d'un contrat de partage de risques entre le Comité économique des produits de santé (CEPS) et le laboratoire MSD France, illustre bien la prise en compte nouvelle des études MOTVR par les décideurs [56]. Le but de cette cohorte était d'utiliser les données générées dans une modélisation pour fournir des renseignements sur l'effet de l'ézétimibe sur la morbi-mortalité cardiovasculaire sur 5 ans en France, en utilisant toutes les données pertinentes disponibles [57]. La limite de la cohorte

étant que seuls les patients traités par ézétimibe y ont été inclus, le modèle développé a permis d'estimer le nombre d'évènements cardiovasculaire évité par le traitement sans avoir de groupe témoin non exposé. La simulation s'est faite en 3 étapes :

- l'utilisation de l'efficacité expérimentale provenant d'essais cliniques randomisés comme valeur de base ;
- l'utilisation de l'effectivité clinique provenant de la cohorte EZE ;
- le calibrage des équations de risque afin d'ajuster le nombre d'évènements cardiovasculaires prédit par le modèle à ceux observés dans la cohorte.

La modélisation s'est faite à l'aide d'une « Simulation à évènements discrets » (SED). Cette méthode consiste à simuler le comportement ou la performance d'évènements de la vie réelle. Les patients sont modélisés en tant qu'entités indépendantes auxquelles on peut attribuer des caractéristiques. Ces informations peuvent également être modifiées au cours du temps. Cette méthode de simulation complexe est de plus en plus répandue dans le domaine des soins de santé [58]. La spécificité de cette modélisation par SED dans le cadre de cette étude MOTVR est l'utilisation de données en vie réelle alors que la majorité des modélisations utilisent des données provenant d'essais cliniques randomisés.

Les résultats ont ensuite permis de conclure sur une effectivité clinique du traitement en France cohérente avec l'efficacité expérimentale observée lors des essais cliniques, l'avis de la Commission de la transparence a donc considéré que le remboursement initial de 65 % de Ezetrol® (spécialité de l'ézétimibe) n'avait pas de raison d'être modifié.

Conclusion

Chaque type d'étude offre des avantages spécifiques et a sa place dans les différentes phases qui caractérisent le cycle de vie d'une innovation thérapeutique.

Au moment de la mise sur le marché, on ne dispose que de données prévisionnelles quant aux bénéfices médicaux et économiques attendus d'un traitement. Les essais cliniques randomisés de phase III et des méta-analyses sont alors les seules sources d'informations disponibles. Les modèles d'analyse de décision font appel à une somme plus importante de données provenant de plusieurs études. Ils peuvent tenir compte de la réalité du comportement des acteurs dans leur exercice médical quotidien, ce qui permet de généraliser les résultats obtenus dans les essais cliniques contrôlés.

Après que l'autorisation de mise sur le marché a été accordée, les informations prospectives d'utilisation en pratique courante s'accumulent et les études rétrospectives des bases de données médico-administratives permettent d'évaluer les résultats obtenus par le traitement quand celui-ci est mis en œuvre dans un système de santé complexe. Elles offrent alors un moyen efficace de vérifier si les résultats promis sont bien au rendez-vous puisque rien ne dit que les conclusions d'un essai soient généralisables à toutes les situations cliniques et à tous les systèmes de santé. Ces études permettent d'évaluer l'impact du comportement des acteurs et de la diversité des pratiques sur les résultats obtenus. Elles accordent une attention toute

particulière aux interactions qui existent dans un système de santé complexe entre une technologie nouvellement introduite sur le marché, les professionnels qui l'utilisent et les patients qui en bénéficient. Elles s'efforcent de mettre à jour toutes les interférences que les essais cliniques s'efforcent de neutraliser, ce qui requiert des outils et des instruments spécifiques de collecte de l'information. L'impact des comportements sur l'efficacité, la qualité de vie, le coût social de la maladie et des traitements peut dès lors être appréhendé.

Les schémas de ces études conduites en vie réelle n'obéissent pas à un modèle unique. Leurs plans d'expérience sont différents selon les objectifs qu'ils se proposent d'atteindre aussi bien dans leurs définitions des populations d'intérêt que dans celles des fenêtres temporelles à retenir pour leur suivi. Lorsqu'il s'agit de s'assurer que les recommandations de bonnes pratiques établies par les autorités sont respectées, il convient de mettre en œuvre, pendant un laps de temps limité, une étude observationnelle descriptive réalisée sur l'ensemble de la population concernée. Des données à plus long terme seront au contraire mobilisées dans le cadre d'une étude observationnelle sans comparateur pour recenser l'ensemble des effets indésirables, sur toute la population concernée. La recherche d'une efficacité clinique ou d'une efficacité comparative exigera la mise sur pied d'une étude observationnelle quasi expérimentale à court et moyen terme sur un échantillon plus restreint représentatif de la population.

L'intérêt croissant des études observationnelles non comparatives ou comparatives est à la hauteur des challenges auxquels elles doivent faire face. De plus en plus souvent demandées par les décideurs et les financeurs, elles font désormais partie intégrante des processus d'évaluation médico-économique et du développement de molécules innovantes. De nouvelles méthodes de modélisation comme celles utilisées dans les études MOTVR, bien que complexes, favorisent le développement et l'utilisation des études observationnelles comparatives, améliorant ainsi les connaissances sur l'efficacité des traitements récemment introduits sur le marché en conditions normales d'usage. La mise en place des grandes bases de données médico-administratives révolutionne les méthodes d'évaluation. L'automatisation de leurs modalités d'exploitation permettra dans un proche avenir de mettre à la disposition des acteurs du système de soins, des informations comparatives non biaisées sur l'efficacité et l'efficacité, quasiment en temps réel et de façon peu onéreuse.

Déclaration de liens d'intérêts

Les auteurs déclarent ne pas avoir de liens d'intérêts.

Références

- [1] ENCePP. Interpretation of the definition of non-interventional trials under the current legislative framework; 2017.
- [2] Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;127:757–63.
- [3] Ranque B, Mechtouff L, Grabar S. Épidémiologie étiologique : du facteur de risque à la cause. *STV* 2011;23(5):242–52.
- [4] Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand S-L. Prospective observational studies to assess comparative effectiveness: the ISPOR Good Research Practices Task Force Report. *Value Health* 2012;15:217–30.
- [5] Sacks H, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233–40.
- [6] HAS/ANAES. Guide d'analyse critique de la littérature et gradation des recommandations; 2000.
- [7] HAS. Niveau de preuve et gradation des recommandations de bonne pratique; 2013.
- [8] Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;2 [i–iv, 1–124].
- [9] Concato J, Shah N, Horwitz RJ. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–92.
- [10] Barton S. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *BMJ* 2000;321:255–6.
- [11] Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright-De Agüero L, Truman BI, et al. Developing an evidence-based Guide to Community Preventive Services-methods. The Task Force on Community Preventive Services. *Am J Prev Med* 2000;18:35–43.
- [12] Begaud B, Polton D, von Lennep F. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé. Rapport réalisé à la demande de Madame la Ministre de la santé Marisol Touraine; 2017.
- [13] De Pourville G, Mongredien L. L'accès au marché remboursé pour les médicaments : Les contrats de partage de risque fondés sur les résultats. *Banque de Données en Santé Publique*; 2012.
- [14] Launois R, Ghabri S, Navarrete LF, Le Moine J-G, Ethgen O. Paiement à la performance et fixation conditionnelle du prix du médicament. *Rev Fr Aff Soc* 2014;156–78.
- [15] Roche N, Reddel H, Martin R, Brusselle G, Papi A, Thomas M, et al. Quality standards for real-world research. Focus on observational database studies of comparative effectiveness. *Ann Am Thorac Soc* 2014;11:S99–104.
- [16] Haynes B. Can it work? Does it work? Is it worth it?: the testing of healthcare interventions is evolving. *BMJ* 1999;319:652–3.
- [17] Sackett DL. Bias in analytic research. *J Chronic Dis* 1979;32(1–2):51–63.
- [18] Delgado-Rodriguez M, Llorca J. Bias. *J Epidemiol Community Health* 2004;58(8):635–41.
- [19] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence—what is it and what can it tell us. *N Engl J Med* 2016;375:2293–7.
- [20] Amzal B. Comparative effectiveness from real-world evidence, dans Olivier Ethgen et Ulf Staginnus. In: *The Future of Health Economics*. Taylor & Francis Group; 2015.
- [21] Sonis J. A closer look at confounding. *Fam Med Sep* 1998;30(8):584–8.
- [22] Brixner DI, Holtorf A-P, Neumann PJ, Malone DC, Watkins JB. Standardizing quality assessment of observational studies for decision making in health care. *J Manag Care Pharm* 2009;15:275–83.
- [23] Berger ML, Sox H, Willke RJ, ISPOR. RWE Good procedural practice; 2017.
- [24] Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chron Dis* 1967;20:637–48.
- [25] Macrae KD. Pragmatic versus explanatory trials. *Inst J Tech Assess Health Care* 1989;5:333–9.
- [26] Shapiro ED. Case-control studies of the effectiveness of vaccines: validity and assessment of potential bias. *Pediatr Infect Dis J* 2004;23:127–31.

- [27] Werler MM, Pober BR, Nelson K, Holmes LB. Reporting accuracy among mothers of malformed and nonmalformed infants. *Am J Epidemiol* 1989;129:415–21.
- [28] Grimaldi-Bensouda L, Rouillon F, Astruc B, Rossignol M, Bénichou J, Falissard B, et al. Results of the cohort for the general study of schizophrenia (CGS). *Schizophr Res* 2012;134:187–94.
- [29] Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part III. *Value Health* 2009;12:1062–73.
- [30] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
- [31] Schneeweiss S. Developments in post-marketing comparative effectiveness research. *Clin Pharmacol Ther* 2007;82:143–56.
- [32] Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006;98:253–9.
- [33] Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987.
- [34] Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Comput Graph Stat* 1993;2:405–20.
- [35] Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol* 1999;150:327–33.
- [36] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;158:280–7.
- [37] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149–56.
- [38] Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiol Camb Mass* 2009;20:512–22.
- [39] Garbe E, Kloss S, Suling M, Pigeot I, Schneeweiss S. High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications. *Eur J Clin Pharmacol* 2013;69:549–57.
- [40] Blin P, Abouelfath A, Lasalle R, Dureau-Pournin C, Droz-Perroteau C, Moore N. Définition d'un score de propension haute dimension pour comparer des patients initiant un anticoagulant oral direct ou un anti-vitamine K pour une fibrillation auriculaire non-valvulaire d'après les données du Sniiram. *Rev Epidemiol Sante Publ* 2017;65:511.
- [41] Grootendorst P. A review of instrumental variables estimation of treatment effects in the applied health sciences. *Health Serv Outcomes Res Methodol* 2007;7:159–79.
- [42] Ghabri S, Launois R. Évaluation quasi-expérimentale des interventions médicales : méthode des variables instrumentales [Quasi-experimental evaluation of medical interventions: instrumental variables method]. *J Gest Econ Med* 2014;32:371–88.
- [43] Chen Y, Briesacher BA. Use of instrumental variable in prescription drug research with observational data: a systematic review. *J Clin Epidemiol* 2011;64:687–700.
- [44] McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA* 1994;272:859–66.
- [45] Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010;19:537–54.
- [46] Zhang Y. Cost-saving effects of olanzapine as long-term treatment for bipolar disorder. *J Ment Health Policy Econ* 2008;11:135–46.
- [47] Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA* 2007;297:278–85.
- [48] Hahn J, Hausman J. A new specification test for the validity of instrumental variables. *Econometrica* 2002;70:163–89.
- [49] Staiger D, Stock JH. Instrumental variables regression with weak instruments. *Econometrica* 1997;65:557–86.
- [50] Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Stat Assoc* 1995;90:443–50.
- [51] Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. *J Econ Lit* 2009;47:5–86.
- [52] Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* 2005;162:279–89.
- [53] Amzal B. Channeling bias after launch: statistical methods to handle it. Milan: ISPOR Workshop; 2015.
- [54] Eichler H-G, Abadie E, Breckenridge A, Flamion B, Gustafsson LL, Leufkens H, et al. Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. *Nat Rev Drug Discov* 2011;10:495–506.
- [55] ICH. Ethnic Factors in the acceptability of foreign clinical data; 1998.
- [56] Ferrières J, Dallongeville J, Rossignol M, Bénichou J, Caro JJ, Getsios D, et al. Model-observational bridging study on the effectiveness of ezetimibe on cardiovascular morbidity and mortality in France: a population-based study. *J Clin Lipidol* 2016;10:1379–88.
- [57] HAS. Avis EZETROL, Commission de la Transparence; 2015.
- [58] Katsaliaki K, Mustafee N. Applications of Simulation within the Healthcare Context, In: *Operational Research for Emergency Planning in Healthcare*, 2. London: The OR Essentials Series, Palgrave Macmillan; 2016. p. 252–95.