Société Française de Pharmacologie et de Thérapeutique

Groupe de Travail Méthodologie

Document de synthèse

Comparaisons à un groupe contrôle externe

Version 1.0 – 24 avril 2026

Comité de rédaction et relecture (par ordre alphabétique)

Jean Luc Cracowski

Michel Cucherat

Dominique Deplanque

Behrouz Kassai

Charles Khoury

Silvy Laporte

Clara Locher

Florian Naudet

Mikail Nourredine

Matthieu Roustit

Table des matières

1 Introduction 9

2 TL ; DR - Guide d'évaluation des comparaisons à un groupe contrôle externe 11

2.1 Validité méthodologique et épistémique 11

2.2 Biais de confusion 13

2.3 Biais de sélection 15

2.4 Autres biais 16

2.5 Autres points : assurance qualité, multiplicité des comparaisons, pertinence clinique 18

3 Les études de comparaison externe, de quoi s'agit-il ? 19

4 Pour quels usages 22

4.1 Utilisation dans le cadre des études monobras 22

4.2 Utilisation avec un essai randomisé 23

4.3 Apporter les preuves nécessaires à la décision 24

4.4 Produire une valeur de référence pour une étude monobras (benchmark) 25

4.5 Utilisation dans les études exploratoires (phase 2) 25

5 Les problématiques méthodologiques soulevées par les comparaisons externes 26

5.1 Solutions potentielles pour les comparaisons externes 27

5.2 Hypothèses des comparaisons indirectes 28

5.3 Solutions générales aux problématiques de l'évaluation du bénéfice clinique des nouveaux traitements 29

6 Les comparaisons externes sont des études observationnelles 35

6.1 Le manque de fiabilité des études observationnelles 35

6.2 Des études montrant des associations, mais ne permettant pas de conclure à la causalité 38

6.3 Différences avec les études de pharmacoépidémiologie 38

7 Position des agences de régulation et de HTA 40

7.1 Documents des agences 40

7.2 Utilisation, évaluation par les agences 44

8 De la nécessité d'avoir des preuves de l'intérêt cliniques des nouveaux traitements 45

9 Les sources de données utilisables 47

9.1 Données historiques, RWD 47

9.2 Groupe contrôle externe prospectif 49

9.3 Sources dédiées 51

9.4 « Données de la baseline » 52

9.5 Recherche et qualification de la source de données 52
9.6 Accès aux données 55
10 Les problématiques liées à l'aspect rétrospectif de ces études 56
10.1 Le risque de HARKing 57
10.1.1 Problématique 57
10.1.2 Solution 58
10.2 Le risque de p-hacking 59
10.2.1 Problématique 59
10.2.2 Solution 59
10.3 Importance du protocole et le plan d'analyse statistique 61
10.4 Analyse de faisabilité 62
11 Rédaction du protocole 64
12 Démarche hypothético déductive 67
13 L'inférence causale et les hypothèses sous-jacentes 69
13.1 Définition 69
13.1.1 Hypothèse de positivité 70
13.1.2 Hypothèse SUTVA 71
13.1.3 Échangeabilité conditionnelle 73
13.2 Petite introduction à l'inférence causale 73
13.3 Association n'est pas causalité 77
13.3.1 DAG générique des comparaisons externes 77
13.4 Effet causal, estimand causal, cible de l'inférence 78
13.4.1 Effet traitement moyen (*average treatment effect)* 78
13.4.2 Analyse en intention de traiter (*as started*) / analyse per protocole (*as treated*) 79
14 Le biais de confusion 82
14.1 Particularité des facteurs de confusion dans les comparaisons externes 82
14.1.1 Modificateurs de l'effet du traitement 84
14.2 La détermination des facteurs de confusion 85
14.2.1 Réseaux de causalité 85
14.2.2 Revue systématique des facteurs pronostiques 87
14.2.3 Lecture critique 89
14.3 Les méthodes statistiques 89
14.4 Les ajustements à éviter car contreproductifs 90
14.5 Sélection de patients 91
15 Les techniques d'analyses statistiques 92
15.1 Les techniques basées sur l'appariement (*matching*) 92
15.2 Le score de propension 93

15.2.1 Définition 93
15.2.2 Le calcul du score de propension 94
15.2.3 L'importance du chevauchement des distributions des scores de propension 95
15.3 L'appariement sur le score de propension 98
15.4 Les méthodes de pondération 101
15.4.1 Principes 101
15.4.2 Pondérations non basées sur le score de propension 103
15.5 La g computation (g formula) 104
15.6 Les méthodes doubles robustes 105
15.7 Les méthodes de régression 105
15.8 Les techniques de maching learning (IA) 107
16 Le diagnostic d'absence de biais de confusion résiduel 109
16.1 Contrôles négatifs 109
16.2 Analyse quantitative de biais, E value 110
16.2.1 Analyse quantitative de biais 110
16.2.2 E value 111
17 Les biais de sélection 114
17.1 Déplétion des susceptibles 114
17.2 Biais lié à un défaut de synchronisation des t0 116
17.3 Groupe contrôle externe non-traité 120
17.4 Fin du suivi 121
17.5 Le biais de sélection vu en termes statistiques : censure à gauche, censure à droite 124
17.5.1 Censures à droite 124
18 Identifications des patients dans la source de données 126
18.1 Aspects chronologiques 126
18.1.1 Détermination du t0 127
18.1.2 Fin de suivi 130
18.2 Représentation graphique du processus d'extraction des données de l'étude 132
19 Biais liés aux données 134
19.1.1 Biais de classification du critère de jugement 134
19.1.2 Le biais de classification de l'exposition 136
19.1.3 Erreur de mesure sur les covariables 137
19.1.4 Précision des données afin d'assurer l'hypothèse de cohérence (*consistency*) (STUVA) de l'inférence causale 137
20 La qualité des données 139
20.1 Exactitudes (accuracy) 140
20.1.1 Généralités 140

20.1.2 Origine des erreurs de classification 140
20.2 Complétudes, exhaustivité 141
20.3 Informativité, pertinence (*relevance*) 142
20.3.1 Critères de jugement 142
20.3.2 Critères d'éligibilité (de sélection des patients de la population visée) 143
20.3.3 Facteurs de confusion 144
20.3.4 Chainage 144
20.4 Origine des données 144
20.5 La validation des données 144
20.6 Recommandations pour la constitution des sources de données 147
20.7 La rwPFS en oncologie 148
21 Les outils d'évaluation du risque de biais 151
21.1 ROBINS-I 151
21.2 APPRAISE 151
22 L'émulation d'un essai cible 153
22.1 Mise en œuvre 154
22.2 Évaluation des performances de l'approche 156
22.3 Méta-épidémiologie 156
23 Le benchmarking et les contrôles positifs 157
24 Analyses de sensibilité, analyses quantitatives du biais 159
24.1 Analyses de sensibilité 160
24.2 Analyse quantitative du biais 161
25 Calcul d'effectif 163
26 Contrôle du risque alpha global 167
27 Pertinence clinique 168
27.1 Critère de jugement 168
27.2 Taille de l'effet 168
27.3 Traitement comparateur 169
27.4 Réalisation, suivi 169
27.5 Balance bénéfice risque 170
28 Méta-épidémiologie et étude de cas 171
28.1 Méta-épidémiologie 171
28.2 Validation empirique 171
28.3 Études de cas 172
28.3.1 Viltolarsen 172
28.3.2 Sodium phenylbutyrate et taurursodiol dans la SLA 172

29 Synopsis - les critères d'acceptabilité des études de comparaisons externes pour la modification des stratégies thérapeutiques 173
30 Annexes 195

1 Introduction

Le recours à une comparaison externe à travers un groupe contrôle externe (*external control arm*, ECA) est fréquemment présenté comme une possibilité pour évaluer un nouveau traitement lorsqu'un essai randomisé n'est pas réalisé[1].

Associée avec les études monobras, elle est la seule approche d'utilisation de données « de vraie vie » (*real world data* RWD) utilisable dans l'évaluation initiale de l'efficacité et de la sécurité d'un nouveau traitement non encore commercialisé.

Cette approche présente de nombreuses limites méthodologiques (cf. section 5) qui, sans solutions satisfaisantes, obèrent la fiabilité des résultats produits et leur utilisabilité au niveau décisionnel (cf. section 5).

Beaucoup de promesses sont habituellement faites sur ce sujet en le présentant de manière superficielle. Le but de ce document est de présenter les problématiques méthodologiques et pratiques liées aux comparaisons à un groupe contrôle externe afin de mettre en évidence les limites que devront surpasser ces études. Pour chacune de ces limites seront présentées des pistes de solutions, qui, si elles peuvent être mises en œuvre de manière satisfaisante, devraient permettre de les dépasser.

Les études de comparaison externe sont par essence des études observationnelles dont elles possèdent les mêmes limitations (cf. section 6). Ainsi, une grande partie des problématiques méthodologiques de ces comparaisons externes se déduisent directement des problématiques qui ont été identifiées par ailleurs pour les études observationnelles classiques, ces problématiques représentent ainsi autant de défis majeurs à relever pour produire des résultats exploitables en termes de prise de décision. Relever ces défis n'est devenu vraiment envisageable que récemment en raison de nombreuses avancées en épidémiologie théorique, comme l'inférence causale, l'émulation d'un essai cible, etc.

La possibilité, grâce à ces solutions, de produire des résultats suffisamment fiables pour être pris en considération au niveau décisionnel reste encore à démontrer par des études de validité empirique à la date de la rédaction de ce document (avril 2026) (cf. section 28). De plus, les travaux de méta-épidémiologie disponibles indiquent que la qualité des études de comparaisons externes publiées jusqu'à présent demeure nettement insuffisante. En conséquence, ces études ne satisfont pas encore les critères requis (cf. section 29) pour être véritablement pris en considération dans l'élaboration des stratégies thérapeutiques.

Actuellement l'emphase est surtout mise sur les données (existence, exploitation, accessibilité, qualité, interopérabilité, etc.). Ces problématiques, essentielles, ne sont cependant pas les seules à considérer dans ces études. Obtenir des données est en soit un défi, mais une fois les données

[1] Voir par exemple :

- https://www.owkin.com/blogs-case-studies/external-control-arms-a-cutting-edge-methodology-to-de-risk-and-accelerate-clinical-trials
- https://www.novainsilico.ai/the-power-of-synthetic-control-arms/
- https://quibim.com/news/synthetic-control-arm-in-clinical-studies/
- https://www.statnews.com/sponsor/2022/11/09/rcts-with-prognostic-digital-twins-overcome-the-limitations-of-external-control-arms/
- https://www.statnews.com/sponsor/2022/11/09/rcts-with-prognostic-digital-twins-overcome-the-limitations-of-external-control-arms/

disponibles, la conception d'études évaluant avec fiabilité l'efficacité des traitements est loin d'être triviale. L'effort investi pour rendre les données disponibles ne pourra porter ses fruits pour l'évaluation des traitements que si les études sont réalisées aux standards méthodologiques modernes et si les données contiennent bien toutes les toutes les variables nécessaires à leur réalisation (critères de jugement, critère d'éligibilité, facteurs de confusion, contrôles négatifs et positifs, etc.), ce qui n'est pas acquis d'emblée si cette utilisation particulière n'est pas envisagée dès la construction des sources de données (cf. section 20.6).

L'objectif de ce document est de présenter les concepts méthodologiques de ces études à un public non-épidémiologiste et d'expliquer comment doit être conçue et réalisée une étude de ce type dont la finalité serait d'introduire un nouveau traitement dans la stratégie thérapeutique. Il n'est pas vraiment orienté vers la réalisation de ces études, mais plutôt vers leur interprétation pour la décision, ou leur conception.

Ce document aborde la question de ces études de comparaisons externes du point de vue très pragmatique de leur utilisation pour décider de l'adoption ou non d'un nouveau traitement. Le regard porté est donc plus celui du pharmacologue, du clinicien, du régulateur, du décideur de santé publique que celui de la recherche méthodologique. Ce regard tient donc compte des conséquences cliniques et sociétales d'adopter à tort un nouveau traitement ainsi que de l'historique des bévues survenues avec l'adoption de traitement sur la base d'études insuffisamment solides (vitalorsen, ataluren dans la maladie de Duchenne, antiarythmiques de classe 1c en post infarctus, etc.)

Sous un angle purement technique et statistique, les comparaisons externes séduisent par la sophistication de leurs méthodes, l'utilisation des nombreuses avancées récentes en épidémiologie théorique et en statistique (comme l'inférence causale) et la forte valorisation qu'elles proposent aux données de vraie vie. Toutefois, il ne faut pas que cet engouement fasse oublier les enjeux fondamentaux de l'évaluation des nouveaux traitements (cf. section 8) et la difficulté rencontrée par ces approches à y répondre de manière adéquate notamment parce que la fiabilité de leurs résultats repose entièrement sur diverses hypothèses qui demeurent souvent difficiles à vérifier et à satisfaire (cf. section 5.2).

Le public visé est composé des lecteurs de ces études au sens large : cliniciens, groupe de recommandations, membres des agences, industriels (médical, *market access*, marketing), sociétés de service ; des décideurs, des étudiants.

2 TL ; DR - Guide d'évaluation des comparaisons à un groupe contrôle externe

Cette section propose un guide pratique d'évaluation (lecture/analyse) d'un travail utilisant une comparaison à un groupe contrôle externe dans le but d'établir le bénéfice clinique d'un nouveau traitement.

La finalité de type de travaux est de permettre une conclusion causale sur l'efficacité et la sécurité du nouveau traitement par rapport au comparateur ad-hoc afin d'apporter la démonstration du bénéfice clinique du traitement, démonstration nécessaire pour le positionner dans la stratégie thérapeutique.

Cette évaluation doit s'effectuer en gardant présent à l'esprit que l'adoption d'un nouveau traitement nécessite, quelle que soit la méthodologie de l'étude pivot, d'obtenir des preuves au-delà de tout doute raisonnable de son bénéfice clinique (cf. section 8).

⚠ Les comparaisons à un groupe contrôle externe présentent de nombreuses limites méthodologiques (cf. section 5) qui, sans solutions satisfaisantes, obèrent la fiabilité des résultats produits et leur utilisabilité au niveau décisionnel (cf. section 5).

Ces problématiques représentent ainsi autant de défis majeurs à relever pour produire des résultats exploitables en termes de prise de décision. Relever ces défis n'est devenu vraiment envisageable que récemment en raison de nombreuses avancées en épidémiologie théorique, comme l'inférence causale, l'émulation d'un essai cible, etc.

Cependant malgré la sophistication de leurs méthodes, l'utilisation des nombreuses avancées récentes en épidémiologie théorique et en statistique (comme l'inférence causale) ces approches peinent à répondre de manière adéquate aux enjeux fondamentaux de l'évaluation des nouveaux traitements (cf. section 8), notamment parce que la fiabilité de leurs résultats repose entièrement sur diverses hypothèses qui demeurent souvent difficiles à vérifier et à satisfaire (cf. section 5.2).

De ce fait, bien que les comparaisons externes puissent être utilisées dans certains contextes, elles ne doivent pas être considérées comme une solution standard pour l'évaluation des nouveaux traitements (cf. section 8).

Afin d'établir si cette comparaison à un groupe contrôle externe peut pallier l'absence d'essai clinique, les points suivants sont à analyser scrupuleusement. Chaque question concerne une problématique méthodologique précise limitant la fiabilité des résultats et interroge sur les moyens, en termes de design ou d'analyses, qui ont été mis en œuvre pour lever cette limitation. Ces questions sont formulées de telle façon qu'une réponse négative corresponde à une limite de l'étude.

2.1 Validité méthodologique et épistémique

La comparaison externe a-t-elle été décidée et conçue avant d'avoir les résultats du groupe traité (monobras ou essai randomisé) ?

La conception de la comparaison externe avant la prise de connaissance des résultats du groupe traité permet d'exclure en partie une conception drivée par les données (cf. section 10). L'anticipation de la comparaison externe débouche sur la réalisation d'un essai contrôlé à contrôle externe (*externally control study*) tel que défini dans ICH E10 (cf. section 3).

Si ce n'est pas le cas, il n'est pas possible d'exclure une conception de la comparaison externe drivée par les données. Une grande transparence dans le processus de sélection (non arbitraire) de la source de données et dans les choix méthodologiques et d'analyse sera alors nécessaire pour écarter HARking et p hacking (cf. sections 10.1 et 10.2).

L'étude est explicitement une étude de confirmation avec une hypothèse thérapeutique clairement définie a priori ?

Tout comme l'essai pivot de phase 3 qu'elle est censée suppléer, cette étude de comparaison externe est une étude de confirmation dont l'objectif était clairement de montrer la supériorité (ou la non-infériorité) du nouveau traitement par rapport au comparateur et non pas une simple étude exploratoire (cf. section 12).

Si ce n'est pas le cas, l'étude ne permet que de générer de nouvelles hypothèses et non pas d'apporter la preuve nécessaire à l'adoption du nouveau traitement dans la stratégie thérapeutique.

L'étude a été conçue (choix de la source de données, identification des facteurs pronostiques, choix des contrôles négatifs et positifs, etc.), a priori, avant toutes analyses inférentielles ?

En raison de la transparence avec laquelle l'étude est rapportée et des diverses garanties et attestations explicitement données, il est possible de conclure que l'étude a bien été conçue (question, choix de la source de données, protocole, plan d'analyse statistique) a priori, sans se baser sur les résultats que les différentes options pouvaient produire. Il est entre autres explicitement attesté que l'étude a été conçue avant toute analyse inférentielle. De ce fait il est possible de garantir l'absence de HARKing et de p hacking (cf. section 10).

Si le groupe contrôle est constitué de manière prospective, ces attentes sont remplies par principe (cf. section 9.2). Par contre se pose la question de la possibilité de réaliser un essai randomisé.

L'étude s'inscrit dans une perspective d'inférence causale

L'enjeu de la comparaison externe étant de suppléer l'absence d'un essai randomisé, une démarche d'inférence causale est obligatoire afin de permettre une conclusion de causalité (cf. section 13).

L'objectif de la comparaison externe correspond à une question causale sur l'efficacité et la sécurité de l'assignement au nouveau traitement. L'estimand causal correspondant à cet objectif a été correctement défini et identifiable avec les données et l'estimateur utilisés. Les hypothèses de l'inférence causale sont plausibles et non rejetées par les données (superposition des distributions des scores de propension par exemple).

La ou les sources de données ont été choisies en fonction de leur adéquation au protocole préétabli ?

La ou les sources de données utilisées pour construire le groupe contrôle externe ont été choisies en fonction des critères prédéfinis dans le protocole (en suivant la démarche PRINCIPLED ou une démarche similaire) (cf. section 9.5). L'étude a ainsi été réalisée avec des données « *fit-to-purpose* » et

non pas avec des données déterminées arbitrairement a priori (afin d'éviter une étude construite autour des données et non pas pour répondre de façon fiable à la question posée, cf. section 11).

Les données utilisées pour constituer le groupe contrôle externe sont-elles adaptée et de bonne qualité ?

L'origine des données est-elle identifiable et traçable (cf. section 9.6) ?

Les données sont-elles pertinentes, rapportent-elles ou permettent-elles de recréer les critères de jugements de l'étude, les critères de sélection, les facteurs pronostiques des critères de jugements, les modificateurs d'effet, les contrôles négatifs nécessaires à la réalisation du protocole tel que préétabli (cf. section 9.5) ?

L'exactitude des données a-t-elle été évaluée (cf. section 20) ? Le degré d'exactitude mesuré (sensibilité, spécificité, VPP, VPN) est-il suffisant pour les besoins de la comparaison externe ?

La qualité globale des données a-t-elle était éprouvée par une approche de benchmarking en montrant qu'il était possible de retrouver des résultats connus en utilisant la méthodologie prévue pour l'étude (design, émulation, ajustement, tec.) (Cf. section 23).

2.2 Biais de confusion

Les facteurs de confusion potentiels sont-ils listés ?

Afin de juger du degré de complétude de l'ajustement effectué, il est indispensable de connaitre l'ensemble des facteurs de confusion potentiels impliqués dans la comparaison externe. Établir cette liste est le pré requis à la construction des ajustements à réaliser pour corriger les résultats du biais de confusion (cf. section 14). Cette liste n'étant pas triviale, il est indispensable qu'elle soit rapportée pour permettre au lecteur d'évaluer la complétude des ajustements réalisés.

Dans les comparaisons externes, les facteurs de confusion potentiels sont soit des facteurs pronostiques des critères de jugement soit des modificateurs d'effet (cf. section 14.1).

Les facteurs de confusion potentiels ont-ils étaient identifiées de façon formelle ?

La détermination des facteurs de confusion affectant une étude n'est pas un problème trivial. Elle doit reposer sur une approche formalisée combinant revue systématique des facteurs pronostiques et analyse du réseau de causalité (par un DAG par exemple).

Les facteurs de confusion pouvant être spécifiques des critères de jugement (cf. section 14.1), leur identification doit être formalisée pour chaque famille de critère de jugement.

Une revue systématique satisfaisante à la recherche des facteurs pronostiques des différents critères de jugement a-t-elle été réalisée ?

La méthodologie de cette revue systématique correspond-elle aux standards attendus (cf. section 14.2.2) ? Est-elle bien spécifique des études pronostiques ?

La problématique des facteurs pronostiques non connus est-elle discutée ?

Les comparaisons à un groupe contrôle externe reposant sur deux échantillonnages distincts, les facteurs pronostiques inconnus peuvent être facteur de confusion en n'étant pas distribués de la même manière dans les deux populations sources dont sont issus les deux échantillonnages (cf. section 13.3.1). Ce point contribue à l'effet étude non réductible. Il doit être discuté.

Un graphique de causalité (DAG par exemple) a-t-il était utilisé pour déterminer les covariables d'ajustement ? Un ajustement sur des collisionneurs ou des médiateurs est exclu ?

Tous les facteurs de confusion identifiés étaient-ils disponibles dans la source de données utilisées ?

La non-prise en compte de certains facteurs de confusion conduit à un biais de confusion résiduel après ajustement obérant la possibilité d'exploiter les résultats pour la construction de la stratégie thérapeutique. Une exploration du biais de confusion résiduel sera d'autant plus cruciale dans cette situation.

Le biais de confusion résiduel a-t-il était évalué ? Par quelles approches (contrôles négatifs, analyses quantitatives de biais) ? Un biais de confusion résiduel est-il exclu ?

Les contrôles négatifs et l'évaluation quantitative de biais sont deux approches, de philosophie différente, mais qui essayent toutes les deux de réfuter la possibilité d'un biais de confusion résiduel persistant après l'ajustement. Bien que très astucieuses, ces méthodes reposent sur des hypothèses et leurs conclusions sont toujours de certitude limitée (cf. section 5.2)

Si cette question est abordée par des analyses de sensibilités, les variations d'analyses explorées par celles-ci permettent-elles vraiment l'évaluation du biais de confusion résiduelle (cf. section 24.1) ?

Les contrôles négatifs utilisés permettent-ils de couvrir la totalité des facteurs de confusion potentiels ?

Est-il raisonnable dépenser que, collectivement, tous les contrôles négatifs utilisés couvrent la totalité des facteurs de confusion affectant les comparaisons d'intérêts (cf. section 16.1) ?

Si ce n'est pas le cas il sera impossible de conclure à l'absence de biais de confusion résiduel et la portée des résultats sera limitée pour la construction de la stratégie thérapeutique.

Les mesures d'association pour les contrôles négatifs sont-elles suffisamment précises pour conclure que l'absence d'association attendue a bien été retrouvée ?

La conclusion à l'absence de biais de confusion résiduelle repose sur une conclusion d'absence d'association qui ne peut être raisonnablement spéculé que si l'estimation est suffisamment précise

Les analyses quantitatives de biais sont-elles à même d'exclure un biais de confusion résiduel ? Dans les analyses quantitatives de biais, les hypothèses de biais nécessaire pour expliquer les résultats sont-elles suffisamment extrêmes pour ne pas être plausibles ?

Si ce n'est pas le cas, il sera impossible de conclure à l'absence de biais de confusion résiduel et la portée des résultats sera limitée pour la construction de la stratégie thérapeutique.

Les modificateurs d'effet des traitements comparés ont-ils été recherchés ?

De la même manière que pour les comparaisons indirectes non ancrées, les modificateurs des effets des traitements comparés sont des facteurs de confusion potentiels des études de comparaisons externes (cf. section 14.1.1). Afin de déterminer si les ajustements effectués sont complets, les modificateurs d'effet doivent avoir été recherchés de manière formelle en se basant sur les analyses en sous-groupes des essais cliniques (indisponibles si le nouveau traitement a été évalué dans une étude monobras) et les mécanismes pharmacologiques connus.

2.3 Biais de sélection

Un biais de sélection peut-il être exclu ?

Un biais de sélection peut être exclu, car l'inclusion (ou la non-inclusion) des patients ou des périodes d'observation par patient dans le groupe contrôle ne pouvait pas dépendre de l'outcome ?

Le biais de sélection est typiquement un biais qui n'est pas dépendant des données, mais qui est introduit principalement par une mauvaise conception de l'étude (cf. section 17). Le cadre conceptuel de l'émulation d'un essai cible cherche à prévenir ces erreurs de construction (cf. section 22).

Un biais de temps d'immortalité est-il exclu ?

La définition des temps de début de suivi (t0) est-elle à même d'éviter un temps d'immortalité (cf. section 18.1.1) ? Les temps de début de suivi utilisés émulent correctement ce qui se passe dans un essai randomisé ? Les résultats (en particulier les courbes de Kaplan Meier) ne suggèrent pas de temps d'immortalité ?

Une approche d'émulation d'essais cibles a-t-elle était utilisée ?

L'émulation d'un essai cible est un cadre de construction des études observationnelles évaluant un traitement qui permet d'éviter une certaine erreur de conception, en particulier celles qui pourraient entrainer un biais de sélection (cf. section 22).

L'analyse de la réalisation de cette émulation amène à se poser les questions suivantes.

L'étude a-t-elle été vraiment conçue suivant cette approche d'émulation ? L'étude de comparaison externe émule-t-elle réellement l'essai cible décrit ?

Le protocole de l'essai émulé est-il satisfaisant ? Correspond-il à un essai adapté pour répondre à la question causale posée (c'est-à-dire le bénéfice clinique du nouveau traitement par rapport au contrôle) ? Cet essai émulé est-il à l'abri des biais (hormis les biais liés à l'absence d'aveugle) ? La question causale à laquelle il va pouvoir répondre correspond-elle à l'objectif de démonter le bénéfice clinique du nouveau produit avec le standard de robustesse des résultats nécessaire pour la construction de la stratégie thérapeutique ?

Les éventuelles adaptations de l'essai effectuées pour coller aux données permettent encore de répondre à la question posée ?

Le groupe contrôle n'a-t-il inclus que des patients incidents (new users design) ?

Un élément de protection par design contre les biais de sélection (déplétion des susceptibles) est d'inclure dans le groupe contrôle externe que des patients incidents pour synchroniser au mieux les t0

des 2 groupes (cf. section 17). Cela n'est pas possible si la comparaison est versus « absence de traitement ».

Les t0 de début de suivi sont-ils correctement synchronisés entre les 2 groupes ?

Le suivi débute-t-il bien dans les 2 groupes à un moment où l'éligibilité est vérifiée et le traitement débuté (cf. section 17.2) ? Ce t0 est-il bien synchronisé entre les 2 groupes ?

En cas de multiplicité des t0 possibles (pathologie chronique, comparaison à un groupe contrôle non traité, etc.), une approche de duplication ou de cloning a-t-elle été utilisée ? l'analyse statistique en découlant est-elle satisfaisante ?

Lorsque la comparaison s'effectue avec un groupe contrôle externe non traité, il existe de nombreuses possibilités de fixer le t0 de début de suivi (il existe plusieurs temps où, simultanément, l'éligibilité est vérifiée et le patient ne reçoit pas de traitement. Ces multiples t0 nécessite de les prendre tous en compte avec des techniques de clonage des patients (cf. section 17.3). Le choix arbitraire d'un de ces t0 serait problématique (par exemple le choix systématique du premier t0 possible augmenterait la durée de suivi des patients et la probabilité d'observer le critère de jugement, favorisant ainsi le traitement étudié).

2.4 Autres biais

Une erreur de classification des expositions (traitement) est-elle exclue ?

Ce type d'erreur est exclu dans le groupe traité (étude monobras expérimentale ou RCT), sauf cas exceptionnel de défaut de réalisation non détecté par le système d'assurance de qualité et les audits de ces études. Ce questionnement est cependant pleinement légitime pour le groupe contrôle externe, car cette erreur de classification peut affecter les sources de données de vraie vie, en particulier lorsqu'il s'agit de traitement pouvant être administré hors dispensation (vaccins, OTC, etc.) ou quand la dispensation peut ne pas être destinée au porteur de l'ordonnance.

Pour un groupe contrôle se voulant être traité avec un certain traitement de référence, ces erreurs consistent soit 1) à la présence de patient non traité (augmentant la fréquence apparente des événements, favorisant la conclusion à la supériorité du nouveau traitement), soit 2) à la présence de patients traités avec un autre traitement actif que le contrôle voulu, conduisant à un biais dans un sens ou l'autre en fonction de l'efficacité relative des traitements effectivement reçus par rapport au traitement attendu.

Pour un groupe contrôle voulu non traité, ces erreurs peuvent conduire à la présence de patients en réalité traités, ce qui conduit à une sous-estimation de la fréquence des événements de ce groupe pouvant conduire à une étude de supériorité négative à tort.

Une erreur de classification des critères de jugement est-elle exclue dans le groupe contrôle ? Si une erreur est possible est-elle aussi présente de façon similaire dans le groupe traité ?

Dans le groupe contrôle expérimental (monobras ou RCT), tout a été mis en œuvre pour assurer l'exactitude des données (data management, procédures opératoires standards, bonnes pratiques cliniques, monitorage sur site, ARC, etc.). Toute erreur dans le groupe contrôle devient donc une erreur asymétrique susceptible d'entrainer un biais (cf. section 19.1.1) qui sera problématique si cette erreur

augmente le nombre d'événements dans le groupe contrôle (faux positif, définition des événements plus large que dans le groupe traité, etc.).

La définition des événements n'est pas sur large que celle utilisée pour le groupe traité ? En cas de données issues d'une base administrative, les algorithmes phénotypiques utilisés couvrent-ils la même définition des critères de jugement que dans le groupe traité ?

Si le résultat est en faveur d'une absence de différence, le raisonnement s'inverse : la définition des événements dans le bras de contrôle n'est pas plus restrictive que celle utilisée dans le bras traité ? Peut-on exclure une détection du critère de jugement dans le groupe contrôle propice aux faux négatifs ?

Existe-t-il des données manquantes sur les critères de jugement ?

Les données manquantes au niveau des critères de jugement sont susceptibles d'induire un biais d'attrition (forme spéciale du biais de sélection). Il est donc important d'avoir une documentation de leur fréquence, de leur répartition entre les deux groupes (elles peuvent aussi survenir dans le groupe traité expérimental).

Les données manquantes au niveau des critères de jugement ont-elles été remplacées de manière conservatrice ?

Compte tenu du biais potentiel qu'elles peuvent induire, les données manquantes des critères de jugement doivent être remplacées par une méthode conservatrice à même de produire un résultat robuste vis-à-vis d ce biais (méthode du biais maximum par exemple). Les méthodes d'imputation basées sur l'hypothèse MAR (*missing at random*, les données manquantes le sont uniquement par hasard) comme les méthodes d'imputation multiple n'effectuent pas un remplacement conservateur.

Existe-t-il des données manquantes sur les covariables ? Ont-elles été imputées de manière satisfaisante ?

Contrairement à l'essai randomisé, les covariables (comme les facteurs de confusion, les contrôles externes, etc.) jouent un rôle primordial dans les comparaisons externes. Les données manquantes à leur niveau sont susceptibles d'induire un biais sur les estimations des effets traitements.

Une imputation basée sur l'hypothèse MAR (*missing at random*) peut être acceptable si cette hypothèse est plausible.

Une différence de traitements concomitants entre les deux groupes est exclue ?

Les patients ont pu bénéficier de la même manière dans les deux groupes des traitements concomitants, des traitements de secours ou des traitements ultérieurs durant toute la période de suivi de l'étude de comparaisons externe ?

Des analyses de sensibilités ont-elles étaient réalisées ? Pour évaluer quel type de robustesse ?

Les analyses de sensibilité réalisée remplissaient-elles leur objectif (cf. section 24.1) ? Les résultats des analyses de sensibilité confirment-ils la robustesse des résultats ?

2.5 Autres points : assurance qualité, multiplicité des comparaisons, pertinence clinique

La qualité de réalisation est assurée aussi bien pour la constitution du groupe contrôle externe que pour la réalisation de comparaison elle-même ?

Un système d'assurance qualité a-t-il été mis en place au niveau de la source de données pour le recueil primaire des données, au niveau de l'extraction du groupe contrôle externe et du traitement de ces données, et pour la réalisation de la comparaison externe elle-même ? Permet-il d'assurer la traçabilité, l'auditabilité, et l'intégrité des données et des analyses ?

Ce point recouvre de nombreux points : respect des cadres de référence qualité des études sur données comme ceux de l'ENCePP ou des agences de régulation (FDA, EMA, etc.), conformité réglementaire (RGPD par exemple), qualité de l'analyse, traçabilité (data trail, changelog, etc., accessibilité aux codes et aux data trials.

Ces principes d'assurance qualité s'appliquent aussi au groupe traité (elles ont été mises en œuvre si ce groupe traité est constitué par une étude monobras ou le bras d'un essai randomisé d'enregistrement).

La multiplicité des comparaisons statistiques est-elle correctement gérée ? un critère de jugement principal unique

La multiplicité des comparaisons statistiques augmente le risque de conclure à tort à un quelconque intérêt du traitement (cf. section 26). La multiplicité doit être gérée soit en utilisant un critère de jugement principal unique (qui ne permettra de faire une seule conclusion) soit une méthode de comparaisons multiples permettant de contrôler le risque alpha global (répartition du risque alpha, hiérarchisation, avec ou sans réallocation).

Les comparaisons en dehors du plan de contrôle du risque alpha global ne permettent pas d'inférer l'effet du traitement et ne peuvent pas être à la base de l'adoption du nouveau traitement dans la stratégie thérapeutique.

La gestion de la problématique des comparaisons multiples est identique à celle employée avec l'essai randomisé.

Les résultats sont-ils cliniquement pertinents ? La balance bénéfice risque est-elle évaluée correctement et est-elle favorable ?

L'analyse de ces points de pertinence est identique à celle effectuée pour juger de la pertinence clinique d'un essai clinique. La pertinence clinique est tout aussi importante que la fiabilité des résultats dans l'utilisation des résultats d'une étude pour changer les pratiques, mais ne peut pas de substituer à cette fiabilité. Un résultat de mortalité peut très bien être dû à un biais malgré la forte pertinence clinique du critère de jugement de mortalité.

3 Les études de comparaison externe, de quoi s'agit-il ?

Un groupe contrôle externe est le groupe contrôle d'une comparaison de traitements, mais dont les patients ne sont pas recrutés dans la même étude que les patients recevant le traitement étudié (groupe expérimental).

> Il s'agit par exemple de la comparaison d'un groupe traité issu de l'étude monobras d'un nouveau traitement avec un groupe contrôle issu d'un registre ou d'une cohorte historique.

Dans un essai contrôlé randomisé, le groupe contrôle est produit par l'essai lui-même. Il s'agit d'un groupe contrôle interne (même si ce terme n'est jamais utilisé). Un groupe contrôle externe provient quant à lui d'une autre dynamique de recrutement ou d'identification de patients, par exemple une source de données de vraie vie. Dans les deux cas, la comparaison sera la même (nouveau traitement N versus traitement contrôle C), mais la nature externe du groupe contrôle soulève une série de problématiques méthodologiques spécifiques, inexistantes dans les essais comparatifs randomisés.

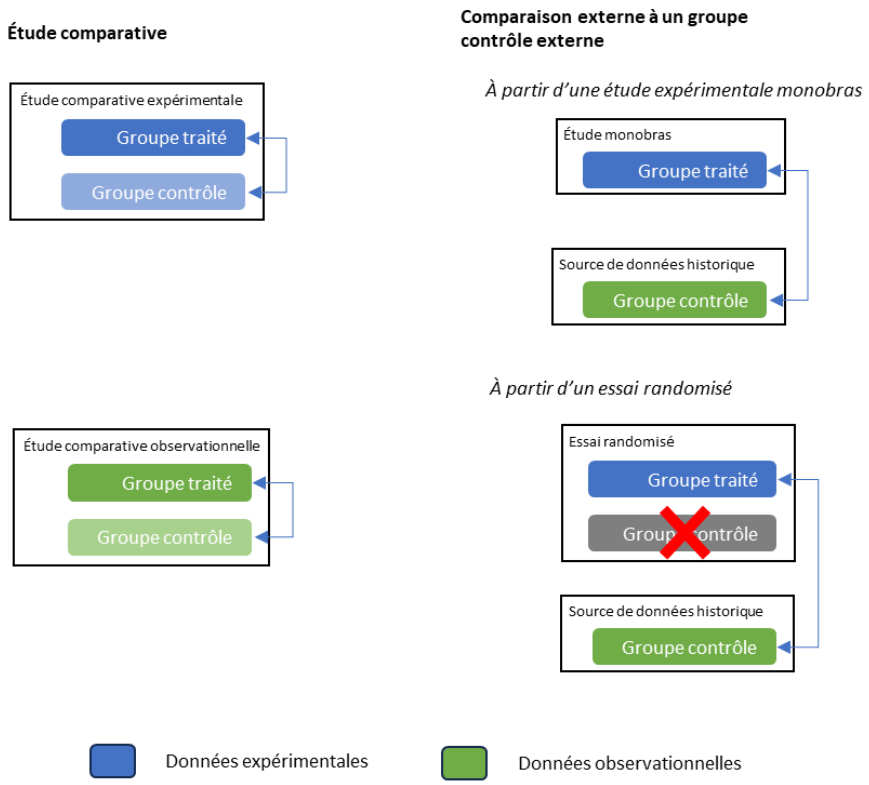


Figure 1 – Illustration graphique d'une comparaison à groupe contrôle externe, comparée à une étude comparative (expérimentale ou observationnelle).

Les études de comparaison externe soulèvent de nombreuses problématiques méthodologiques limitant fortement la fiabilité de leurs résultats et qui est très difficile à solutionner (cf. section 5). La méthodologie des essais randomisés de confirmation pivot a

été élaborée pour proposer une démarche d'évaluation des traitements ne présentant pas ces problématiques et permettant d'assurer la nature causale/la fiabilité des résultats produits.

Pour **ICH E10** [1] " An externally controlled trial compares a group of subjects receiving the test treatment with a group of patients external to the study, rather than to an internal control group consisting of patients from the same population assigned to a different treatment. The external control can be a group of patients treated at an earlier time (historical control) or a group of patients treated during the same time period but in another setting [contemporaneous external control]."

La terminologie n'est pas encore fixée [2] et ce type de comparaison peut aussi être appelée « comparaison indirecte » (par opposition aux comparaisons directes qui sont effectuées directement au sein d'une même étude, avec un groupe contrôle interne) et plus précisément comparaison indirecte non ancrée ». Le terme comparaison indirecte est aussi utilisé pour désigner des approches diverses et variées (comme les comparaisons indirectes ancrées réalisées à partir des résultats de deux essais randomisés ayant le même traitement contrôle). Au sein de toutes les techniques de comparaison indirecte, les comparaisons utilisant un groupe contrôlent externe se distingue, car elles reposent sur les données individuelles pour les groupes (contrairement par exemple aux MAIC qui utilisent seulement une publication pour le groupe contrôle).

Le terme « groupe contrôle synthétique » est parfois utilisé pour désigner les groupes contrôle externe[2]. Cette terminologie est ambiguë, car ce terme désigne depuis longtemps une méthode différente d'analyse des études observationnelles (particulièrement utilisée en sciences sociales et économiques) [3]. Cette appellation est donc à proscrire pour éviter toute confusion.
De plus ce terme est aussi utilisé pour désigner des données artificielles, simulées soit à partir de processus (modèle in silico par exemple) soit à partir d'autres données (génération par IA de données synthétiques) [3].

Le groupe contrôle externe est souvent constitué avec des données observationnelles, dites de « vraie vie » et cette approche est souvent intégrée aux approches dites de « Real World Evidence » (RWE) [4]. De plus, les comparaisons externes engendrent les mêmes problématiques statistiques (biais de confusion) que les études observationnelles et doivent faire appel aux mêmes techniques statistiques pour contrôler ce biais (ajustement, score de propension, etc.).

Le design des études de comparaison à un groupe contrôle externe est celui des études de cohortes (appelés aussi étude exposés/non-exposés).

De façon constante, le traitement étudié a fait l'objet d'une étude expérimentale (interventionnelle) prospective. Autrement si les données concernant le nouveau traitement proviennent aussi de données observationnelles, il ne s'agit plus d'une comparaison à un groupe contrôle externe, mais d'une étude observationnelle, de pharmacoépidémiologie par exemple.

[2] https://servier.com/newsroom/bras-controle-synthetiques-revolutionner-essais-cliniques/, https://www.parissaclaycancercluster.org/post/focus-sur-les-bras-synth%C3%A9tiques-contr%C3%B4les-dans-les-essais-cliniques-en-oncologie
[3] https://en.wikipedia.org/wiki/Synthetic_control_method

Synonymes *– Du fait de la relative nouveauté du concept, la terminologie n'est pas encore fixée et de nombreux vocables différents sont employés pour désigner ces études : comparaison indirecte à un groupe contrôle externe, étude comparative à groupe contrôle externe (*externally controlled study*), bras contrôle externe (*external control arm *ECA) et parfois les termes de groupe contrôle historique, groupe groupe contrôle synthétique [5], de jumeaux numériques recouvrent aussi ces groupes contrôles externes.*

Données artificielles

Il est parfois proposé de ne pas recourir à des patients pour constituer une base de comparaisons et d'utiliser des données artificielles (issues de jumeaux numériques, de prédictions d'IA, etc.) ou obtenues par simulation (à partir de modèles mécanistiques in-silico par exemple). Ces approches sont encore strictement du champ de la recherche et ne sont pas envisageables en routine pour produire des résultats destinés à faire changer les pratiques.

Externally controlled trial

Il existe aussi des études « *externally controlled trial* » qui sont des comparaisons à des groupes contrôles externes où la comparaison externe est prévue d'emblée dans le protocole de l'étude expérimentale du nouveau traitement [6] [7].

Cette approche est mentionnée depuis longue date par l'ICH (ICH E10) [1] et a fait l'objet d'un guide FDA spécifique [6].

De ce fait il ne s'agit plus d'une étude non comparative pour le nouveau traitement étant donné que cette comparaison est prévue dès la construction de cette partie de l'étude, mais d'une réelle étude comparative où les patients contrôles ne seront pas recrutés et observés par l'étude elle-même. Ce design permet de solutionner plusieurs problématiques méthodologiques des comparaisons externes (cf. section 5).

4 Pour quels usages

Pour l'évaluation des nouveaux traitements, les groupes contrôles externes sont envisagés dans plusieurs situations.

- Lorsque le nouveau traitement ne fait l'objet que d'une étude monobras (*single arm study*) afin de fournir un contrefait permettant de déterminer un effet traitement (cf. section 13). Deux variantes sont possibles. Le groupe contrôle externe prévu d'emblée avec la conception de l'étude monobras, il s'agit alors d'un **essai contrôlé à contrôle externe** (*externally controlled trial*). Autrement la comparaison à un groupe contrôle externe est réalisée indépendamment de la monobras, presque toujours de façon post hoc (c'est-à-dire après la réalisation de la monobras et la disponibilité de ses résultats) et est souvent appelée comparaison indirecte.
- Fournir un groupe contrôle supplémentaire à un essai randomisé dans le cas par exemple où celui créé par la randomisation n'est plus satisfaisant en termes de traitement comparateur.
- Pour compléter le groupe contrôle d'un essai randomisé en lui associant des patients externes non issus de la randomisation. Il s'agit alors d'un **groupe contrôle hybride** comportant à la fois des patients issus de la randomisation et de patients externes. Le but étant de limiter le nombre de sujets à inclure dans le groupe contrôle lors de la réalisation de la partie randomisée.

Acceptabilité des comparaisons externes dans le développement des nouveaux traitements

Bien que les comparaisons externes puissent être utilisées dans certains contextes, elles ne doivent pas être considérées comme une solution standard pour l'évaluation des nouveaux traitements (cf. section 8). En effet, ce type de comparaison soulève de nombreuses problématiques méthodologiques qui, en l'absence de solutions adéquates, compromettent la fiabilité de leurs résultats en raison de nombreux biais (cf. section 5). Cette approche ne peut donc être considérée comme une méthode standard utilisable en remplacement de l'essai randomisé. Néanmoins, les avancées en épidémiologie théorique et en inférence causale permettent désormais d'envisager, au moins sur le plan théorique, la production de preuves répondant aux exigences requises dans des situations spécifiques où la réalisation d'un essai randomisé n'est réellement pas envisageable.

4.1 Utilisation dans le cadre des études monobras

Dans le cadre de l'évaluation des nouveaux traitements, le recours à des études de comparaison externe est principalement envisagé lorsque le traitement évalué n'a fait l'objet que d'une étude non comparative à bras unique (étude monobras). En effet, l'absence de groupe contrôle dans une telle étude ne permet pas d'isoler l'effet propre du traitement, à l'exception du cas particulier dit du

0/100% où le traitement change radicalement le devenir d'une maladie[4]; ni de positionner ce traitement dans la stratégie thérapeutique [7].

Le recours à un groupe contrôle externe vise alors à compenser cette absence et à apporter le contrefait de l'étude monobras permettant d'inférer l'effet causal du nouveau traitement (cf. section 13). Sans ce raisonnement contrefactuel il est impossible de déterminer l'effet du traitement (cf. inférence causale section 13).

> Le mobocertinib dans le traitement des cancers du poumon non à petites cellules avancé, porteur de la mutation EGFRex20ins a été seulement évalué par une étude de phase 1-2 monobras (NCT02716116). Un groupe contrôle externe issu de la base Flatiron a été utilisé pour comparer ce produit au standard de soin aux USA [8].

Un groupe contrôle externe peut être utilisé pour apporter le contrefait dans une étude monobras. Il s'agit de l'utilisation la plus fréquente.

4.2 Utilisation avec un essai randomisé

Une étude de comparaison externe peut aussi être envisagée avec un essai randomisé dont le traitement contrôle n'est pas (ou n'est plus) approprié pour la prise de décision et le positionnement du nouveau traitement dans la stratégie thérapeutique. Il s'agit alors d'effectuer la comparaison attendue en utilisant un groupe contrôle externe traité avec le traitement approprié. Il est souvent difficile d'appliquer cette méthode, car les données nécessaires pour constituer le groupe contrôle externe font défaut du fait de sa nouveauté.

Un groupe contrôle externe peut être rajouté à un essai randomisé pour compenser un comparateur inapproprié

> Le plitidepsine (P) a été évalué en association avec une faible dose de dexaméthasone (LD-DXM) chez des patients atteints de myélome multiple en rechute ou réfractaire, dans le cadre de l'essai randomisé de phase III ADMYRE. Faute d'essai randomisé comparant P + LD-DXM à pomalidomide (POM) + LD-DXM, une comparaison directe appariée a été réalisée entre P + LD-DXM (données de l'étude ADMYRE) et POM + LD-DXM, en utilisant des données individuelles issues de plusieurs essais contemporains sur POM + LD-DXM présentant un design similaire [9].

Il est aussi possible d'ajouter des comparaisons externes à un essai randomisé dans de nombreux contextes variés.

[4] Cette situation est, par exemple, celle d'une maladie entrainant invariablement le décès et d'un traitement ayant permis d'éviter cette évolution irrémédiable dans quelque cas.

NEURO-TTRansform (NCT04136184) est un essai clinique randomisé de l'eplontersen versus inotersen (randomisation 6 :1) dans l'amylose à transthyrétine avec polyneuropathies [10]. Cependant la comparaison prévue pour justifier de l'intérêt de l'eplontersen n'était pas la comparaison randomisée, mais une comparaison à un groupe contrôle externe placebo issu d'un essai randomisé précédent NEURO-TTR (NCT01737398) comparant inotersen versus placebo.

L'essai CENTAUR a comparé l'association sodium phenylbutyrate et taurursodiol au placebo dans la sclérose latérale amyotrophique (SLA) [11]. Après les 6 mois de la période en double aveugle, les patients du groupe placebo pouvaient recevoir le traitement étudié (cross over) empêchant donc d'utiliser ce groupe pour évaluer le bénéfice du traitement sur la survie. Une comparaison à un groupe contrôle externe, sans exposition préalable à ce traitement, a été utilisée pour évaluer l'effet de l'association sodium phenylbutyrate et taurursodiol sur la survie à long terme [12]. L'enregistrement accéléré aux USA ainsi que l'AMM conditionnelle européenne de ce produit ont ensuite été retirée en raison de l'échec de l'essai de confirmation.

4.3 Apporter les preuves nécessaires à la décision

Une méthodologie robuste est nécessaire pour donner aux résultats la crédibilité nécessaire pour compenser la non-réalisation d'un essai randomisé approprié.

Dans les situations où un groupe contrôle externe est utilisé pour justifier de l'intérêt d'un nouveau traitement, la comparaison externe vise à compenser la non-réalisation d'un essai randomisé, en produisant à sa place, et d'une autre manière, les preuves nécessaires à l'adoption de ce traitement.

Ainsi, les comparaisons externes sont principalement destinées à apporter des preuves du bénéfice clinique du traitement évalué en l'absence de preuves issues d'un essai randomisé. Pour prétendre à ce statut de preuve au-delà de tout doute raisonnable, ces études doivent contrôler toutes les problématiques méthodologiques qui pourrait conduire à des résultat faux positifs, de la même façon que les essais randomisés pivots de phase 3 (Tableau 2 et section 8).

Leurs enjeux décisionnels (stratégiques) étant identiques à ceux des essais randomisés pivots, ces études doivent être conçues, sur de nombreux points, de manière similaire aux essais randomisés pivots de phase 3 : études de confirmation, pertinence clinique du comparateur et des critères de jugements, aptitude à déterminer la balance bénéfice risque, généralisabilité des résultats à la population visée, contrôle du risque alpha global, etc. Et, contrairement aux essais randomisés, elles doivent aussi solutionner les problématiques méthodologiques qui leur sont propres liées à leur nature observationnelle.

4.4 Produire une valeur de référence pour une étude monobras (benchmark)

Un groupe contrôle externe peut être utilisé parfois, non pas pour effectuer une comparaison formelle avec le groupe traité, mais simplement pour donner un point de repère pour aider à l'interprétation des résultats du groupe traité (benchmark, mise en perspective). Dans certaines études monobras cette information est utilisée pour définir le critère de succès de l'étude.

Bien que couramment évoquée [13] cette simple mise en perspective n'est pas en mesure d'apporter des preuves et ne peut vraiment être utile que dans des situations très particulières (montrer par exemple qu'il s'agit d'une situation de type 0/100%).

4.5 Utilisation dans les études exploratoires (phase 2)

Pour les études exploratoires, de type phase 2 par exemple, le recours à un groupe contrôle externe est parfaitement envisageable. Comme l'étude n'a pas vocation à apporter la preuve du bénéfice clinique du nouveau traitement (étude exploratoire et non de confirmation), les enjeux méthodologiques sont alors peut-être moins prégnants.

Il convient cependant de noter que la terminologie phase 2, phase 3, est fréquemment mal employée. Elle s'applique aux étapes du plan de développement d'un nouveau traitement : les essais de phase 3 sont les essais pivots, destinés à démontrer le bénéfice clinique du traitement, tandis que les phases 2 sont des études préliminaires, destinées à préparer les phases 3, et qui n'ont pas vocation à apporter la preuve de l'intérêt clinique du traitement. Cependant ces dénominations sont assez fréquemment comprises, à tort, comme étant conditionnées par la méthodologie de l'étude : le terme « phase 3 » devenant ainsi abusivement associé à essai randomisé, et « phase 2 » à un design d'étude non randomisé ou non comparatif. Cela provient du fait, qu'assez fréquemment, les études de phase 2, étant donnée leur nature exploratoire, adoptent des designs moins exigeants que les phases 3, alors que ces dernières, compte tenu de leur enjeu de confirmation, sont réalisées avec des essais randomisés. Or rien n'empêche éventuellement de conduire une phase 3 (c'est-à-dire l'étude de confirmation) avec une méthodologie autre que celle de l'essai randomisé si celle-ci permet d'apporter les garanties méthodologies attendues pour une étude de confirmation et de démonstration du bénéfice clinique. C'est le cas, par exemple, avec certaines maladies rares ou certains cancers. Mais fréquemment ces études, qui sont clairement dans le plan de développement du nouveau traitement l'étude pivot, sont dénommées, à tort, phase 2 du fait de leur méthodologie, alors qu'il s'agit de la phase 3 du développement. L'EMA pour contourner cette ambigüité terminologique dans son guide méthodologique sur les études monobras parle de « *single-arm trials submitted as pivotal evidence* » [14].

Ainsi, il est tout à fait possible qu'une étude de comparaison externe soit présentée comme étant une phase 2, simplement du fait de sa méthodologie, bien qu'elle ait la fonction d'une phase 3 dans le plan de développement du nouveau traitement. Dans ce cas, cette phase 2 correspond parfaitement au cadre de ce document (qui est celui des études de confirmation basées sur une comparaison à un groupe contrôle externe) et nécessite impérative la rigueur méthodologique développée dans ce document.

5 Les problématiques méthodologiques soulevées par les comparaisons externes

Les comparaisons externes à des groupes contrôles externes soulèvent plusieurs problématiques (Tableau 3), qui si elles ne sont pas gérées correctement, obèrent fortement la fiabilité des résultats :

- La possibilité de HARKing et de p hacking liée à l'aspect rétrospective de l'approche (cf. section 10)
- L'existence inévitable d'un biais de confusion lié à des différences dans le risque de base des patients du groupe contrôle externe par rapport à ceux du groupe du traitement étudié (cf. section 0)
- La nécessité d'identifier tous les facteurs qui devront être contrôlés (ajustés) pour tenter de supprimer le biais de confusion
- La nécessité d'apporter la preuve de l'absence de biais de confusion résiduelle par des contrôles négatifs ou des analyses quantitatives des biais satisfaisantes
- La difficulté de trouver des sources de données documentant correctement les critères d'inclusion des patients, les critères de jugement, tous les facteurs de confusion potentiels, des variables pouvant servir de contrôles négatifs et de contrôles positifs
- La possibilité de biais de sélection, en particulier de biais de temps d'immortalité, en raison de l'inclusion des patients dans le groupe contrôle dépendante de la survenue du/des critères de jugement et de la difficulté de synchroniser le temps de début de suivi du groupe contrôle externe avec celui de la monobras (cf. section 17)
- La faible qualité des données pouvant conduire à des biais de classification
- Etc.

En effets chacune de ces problématiques est susceptible de produire des résultats suggérant à tort le bénéfice clinique du nouveau traitement. Il est donc indispensable que ces études puissent mettre en place les solutions méthodologiques et statistiques qui soient en mesure de solutionner parfaitement ces problématiques afin de produire des résultats de la même fiabilité que ceux des essais randomisés.

Les comparaisons externes à des groupes contrôles externes soulèvent plusieurs problématiques méthodologiques, qui si elles ne sont pas gérées correctement, obèrent fortement la possibilité d'utiliser les résultats produits pour démontrer le bénéfice clinique d'un nouveau traitement.

Du fait de ces limites laissées sans solution, les études observationnelles produisent fréquemment des résultats qui ne sont pas confirmés par des essais randomisés de confirmation (cf. section 6).

Une partie de ces problématiques sont les conséquences de défauts de conception ou de réalisation de l'étude elle-même (HARKing, p hacking, biais de sélection par déplétion des susceptibles ou par temps d'immortalité, manquement dans l'identification de tous les facteurs de confusion potentiels), tandis que d'autres proviennent de problématiques inhérentes aux données elles-mêmes (erreur de classification et de mesure, données manquantes informatives, facteurs de confusion non mesurés, absence de contrôles négatifs, etc.).

Les problématiques de la première catégorie peuvent être évitées par une meilleure conception ou réalisation. Les cadres conceptuels de l'émulation de l'essai cible (cf. section 22) et des réseaux de causalité (cf. section 13) ont pour objectifs d'éviter ces pièges. En revanche, les problématiques inhérentes aux données sont subies et ne peuvent pas être corrigées au moment de l'étude (la prévention de ces problématiques doit être envisagée au niveau de la constitution et ensuite du peuplement des sources de données).

5.1 Solutions potentielles pour les comparaisons externes

Ces problématiques représentent ainsi autant de défis majeurs à relever pour produire des résultats exploitables en termes de prise de décision. Relever ces défis n'est devenu vraiment envisageable que récemment en raison de nombreuses avancées récentes en épidémiologie théorique (cf. Tableau 3) comme :

- L'émulation des essais cibles
- Les techniques d'inférence causale
- La formalisation de l'identificateur des facteurs de confusion potentiels à l'aide des graphiques de causalité comme les DAGs
- Les contrôles négatifs et l'analyse quantitative de biais
- Les contrôles positifs et les techniques de benchmarking

⚠ Remarque importante

Il faut cependant remarquer que toutes ces solutions reposent sur des hypothèses statistiques ou causales intestables, et donc invérifiables en pratique. L'inférence de l'effet traitement effectuée n'est exacte que si ces hypothèses sont effectivement vérifiées par les données et les conditions de réalisation de la comparaison externe.

Ainsi même si des solutions à ces problématiques semblent apparaître, elles sont imparfaites et n'assurent pas, par principe, l'exactitude des résultats produits, contrairement aux solutions génériques qui ont été trouvées par ailleurs en construisant l'approche basée sur l'essai randomisé (cf. section suivante) [15].

Sans solution aussi robuste que celles développées pour l'approche classique reposant sur l'essai randomisé, les comparaisons externes ne peuvent prétendre se substituer à cette dernière.

> *Par exemple la problématique est de supprimer complètement le biais de confusion comme le fait la randomisation dans un essai randomisé correctement conçu et réalisé. Les techniques d'ajustement actuellement disponibles ne peuvent prétendre que limiter le risque de biais de confusion. Une autre problématique est de pouvoir conclure à l'absence de biais de confusion résiduel. Les techniques disponibles pour cela ne permettent que de renforcer les arguments en faveur d'un faible risque de confusion résiduelle compte tenu de leur limite (cf. section 16).*

5.2 Hypothèses des comparaisons indirectes

L'approche des comparaisons externes repose sur de nombreuses hypothèses fondamentales [16] [17] [18]. Le résultat obtenu ne correspondra à la réalité de l'effet du traitement qu'à condition que ces hypothèses soient vérifiées par les données et la réalisation de l'étude[5]. Dans le cas contraire, les résultats ne seront pas à l'abri des biais et se posera la condition de l'acceptabilité des résultats produits pour la prise de décision. Parmi ces hypothèses fondamentales de validité, la plupart ne sont pas testables et vérifiables, situation pouvant être considérée comme rédhibitoire pour la production des preuves du bénéfice clinique des nouveaux traitements.

En regard de cette situation, la méthodologie de l'essai clinique prospectif, contrôlé, randomisé en double aveugle et analysé en intention de traiter a été développée pour ne dépendre d'aucune hypothèse en dehors de la qualité de sa conception et de sa réalisation (qui est garantie par le système d'assurance qualité de ces études, le respect des bonnes pratiques cliniques et qui est évaluable par les monitorings sur site et les audits de réalisation).

Le Tableau 1 présente les hypothèses des comparaisons externes et montre comment l'essai randomisé classique s'en affranchit.

Tableau 1 – Hypothèses implicites sous-jacentes à l'acceptation des résultats de comparaisons externes comme éléments probant pour la construction des stratégies thérapeutiques.

Hypothèses faites par les comparaisons externes	**Situation dans l'essai randomisé**
Absence de HARKing (pour les études rétrospectives)	HARKing complètement exclut étant donné la nature prospective de l'étude (peut être remis en cause par les amendements en cas d'analyses intermédiaires possibles)
Absence de p hacking	Absence garantie par le plan d'analyse statistique dont la conception avant toute analyse inférentielle est facile à établir du fait de la réalisation prospective de l'essai (peut être remis en cause par les amendements en cas d'analyses intermédiaires possibles)
Hypothèse de positivité	La positivité est garantie par la randomisation
Hypothèses de cohérence	Cohérence assurée par la nature expérimentale
Hypothèses d'échangeabilité (conditionnelle) qui recouvre les hypothèses plus spécifiques suivantes	Échangeabilité garantie par la randomisation : l'assignation à un traitement est indépendante de l'outcome et des caractéristiques des patients. Aucune hypothèse n'est nécessaire sur ce point.
Tous les facteurs pronostiques et les modificateurs d'effets ont été pris en compte par l'ajustement	Ne concerne pas l'essai randomisé (les ajustements effectués dans un essai randomisé ont pour but d'optimiser la puissance et la précision, et en aucun cas de corrigé d'un biais de confusion qui est évité par design grâce à la randomisation. Les ajustements peuvent aussi concerner la non-collapsibilité de l'estimateur de l'effet traitement utilisé)
Les modèles (de traitement, score de propension par exemple ou d'outcome) sont bien spécifiés	
Les contrôles négatifs captent la totalité de la structure de confusion affectant les associations	Sans objet

[5] Ces hypothèses exposant à des biais en cas de non vérification sont aussi parfaitement bien exposées dans le guide EMA sur les études pivots monobras (https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence-marketing-authorisation-application_en.pdf)

recherchées (si des contrôles négatifs sont utilisés pour évaluer le biais de confusion résiduel)	
Absence de biais de sélection	Absence assurée par la randomisation, l'aspect prospectif de l'étude et l'analyse en intention de traité
Absence d'erreur de classification des outcomes entre le groupe traité et le groupe contrôle	Garantie obtenue par le double insu (non-garantie dans les essais en ouvert même en cas d'adjudication des outcomes en aveugle) et le système d'assurance qualité des données de l'essai : data management, monitoring sur sites, etc.
Absence d'erreur de classification de l'assignement à exposition dans le groupe contrôle externe	Garantie par l'analyse en intention de traité et le caractère expérimentale de l'étude
Les données manquantes sur les critères de jugement sont non informatives	Mêmes hypothèses en cas de données manquantes
Les données manquantes sur les covariables sont manquantes au hasard (missing at random, MAR)	Les covariables n'ont pas d'importance dans la fiabilité de l'estimation des effets traitements

5.3 Solutions générales aux problématiques de l'évaluation du bénéfice clinique des nouveaux traitements

Ces problématiques ne sont que le reflet, au niveau des comparaisons externes, des problématiques générales qui surgissent lorsque l'on cherche à évaluer le bénéfice clinique des traitements de façon rigoureuse (Tableau 2).

En effet, de nombreuses problématiques surgissent lorsque l'on veut savoir si un traitement est efficace et quel est le bénéfice clinique qu'il apporte aux patients. Elles ne sont pas anodines, car elles peuvent toutes concourir à donner à tort des résultats en faveur de l'efficacité avec des traitements qui en sont dépourvus. Sans solutions, il serait impossible de conclure sur ces questions d'efficacité et de bénéfice.

Dans le but de pouvoir quand même produire des preuves de l'efficacité des traitements au-delà de tout doute raisonnable malgré ces problématiques, il a été cherché des solutions efficaces et robustes pour empêcher par design la production de résultats positifs à tort. Ces solutions ont progressivement conduit, au fil des années, à la méthodologie standard actuelle d'évaluation des traitements basée sur l'essai randomisé (cf. Tableau 2).

Procéder autrement pour produire des preuves de la même fiabilité nécessite donc de trouver d'autres solutions à ces problématiques consubstantielles à la recherche de l'efficacité d'un traitement.

Le Tableau 2, décrit pour chacune de ces problématiques les solutions apportées par la méthodologie actuelle basée sur l'essai randomisé et les solutions envisageables lorsqu'une approche de comparaisons externes est envisagée à la place d'un essai randomisé.

Tableau 2 – Les problématiques survenant lorsque l'on cherche à savoir ce qu'apporte comme bénéfice aux patients un nouveau traitement, avec en regards les solutions 1) mises en œuvre dans l'approche classique basée sur l'essai randomisé et 2) envisageables pour les comparaisons externes

Problématiques survenant quand on veut évaluer un traitement	**Solution apportée par l'approche classique basée sur l'essai randomisé[6]**	**Solution envisagée dans les comparaisons externes**
Sans évaluation clinique il est impossible d'apprécier le réel bénéfice clinique qu'apporte un nouveau traitement aux patients :		
Limites des raisonnements sur les mécanismes d'action des traitements qui sont peu prédictifs de la réelle efficacité et sécurité des traitements	Vérifier de façon factuelle que le traitement apporte bien le bénéfice clinique escompté en confrontant l'hypothèse à la réalité par une étude expérimentale de confirmation dédiée	Même démarche de vérification par les faits avec une étude de confirmation dédiée, mais qui ne pourra pas être expérimentale par définition et donc tributaire de la disponibilité et de la qualité des données
L'évaluation factuelle de ce que cause un traitement chez des patients débouche sur les problématiques suivantes :		
Facteurs de confusion multiples (effet placebo, évolution naturelle de la maladie, autres traitements, régression à la moyenne, effet Hawthorne) affectant en même temps que le traitement étudié l'évolution des patients et empêchant d'isoler l'effet que cause spécifiquement le traitement	Raisonnement contrefactuel Recours à une comparaison à un groupe contrôle, subissant les effets des mêmes facteurs de confusion, pour isoler l'effet spécifique du traitement parmi toutes les autres influences que subit l'évolution des patients traités ⇨ Essai contrôlé	Recherche d'un raisonnement contrefactuel en utilisant un groupe contrôle externe à l'étude observant l'évolution de patients sous traitement. Inscription de l'étude dans une démarche d'inférence causale
La réalisation d'études comparatives (essais contrôlés) débouche sur les problématiques suivantes :		
Les patients des deux groupes ont un risque de base (pronostic) différent	Randomisation imprévisible (protégeant par design la comparaison d'un biais de confusion et garantissant le respect de l'hypothèse d'échangeabilité de l'inférence causale)	En l'absence de randomisation le traitement reçu n'est pas indépendant des déterminants du critère de jugement entrainant un biais de confusion impliquant de corriger du biais de confusion les résultats en prenant en compte tous les facteurs de confusion par l'analyse
Le critère de jugement est évalué/mesuré différemment entre les deux groupes, favorisant le groupe traité	Double aveugle Standardisation du processus de mesure du critère de jugement	Difficile à prendre en compte (cf. section 19) Éléments constitutifs d'un effet étude probablement irréductible (cf. section 13.3.1)
Asymétrie de prise en charge favorisant le groupe traité	Double aveugle Protocolisation des traitements	
Exclusion d'analyse (biais d'attrition) favorisant le groupe traité : données manquantes et censures informatives (perdus de vue)	Analyse en intention de traiter Remplacement conservateur des données manquantes sur le critère de jugement Gestion conservatrice des événements intercurrents	Idem
Les événements intercurrents peuvent fausser l'évaluation	Définition précise de l'estimand	Idem

[6] Pour une démonstration de supériorité avec un estimand « policy treatment » (analyse en intention de traiter).

Il existe plusieurs estimands causal, population cible pour définir l'effet d'un traitement	Dans l'essai randomisé, tous les estimands causaux ont la même valeur (ATE = ATT = ATC)	L'ATT correspond à la question posée par les comparaisons externes (chercher le contrefait au groupe traité)
Nécessité de comparaisons avec une puissance statistique suffisante pour séparer le signal du bruit	Calcul d'effectifs Recrutement de l'effectif ou du nombre d'événements nécessaire pour garantir le puissance	Idem
Multiplicité des comparaisons statistique induisant une inflation du risque alpha global	Control du risque alpha global	Idem
Début du suivi différent entre les traitements par rapport à l'histoire « naturelle » de la maladie introduisant des temps d'immortalité et autres biais de sélection	Solutionnée par design, le suivi débute à la randomisation. Il est parfaitement synchronisé entre les 2 groupes et cette synchronisation est maintenue par l'analyse en ITT	Recherche d'un t0 de début de suivi émulant ce qui se passe dans l'essai randomisé.
HARKing	Réalisation d'une étude prospective	Réalisation en prospectif Si réalisation rétrospective → nécessité d'apporter des garanties solides d'absence de HARKing
P hacking	Plan d'analyse statistique élaboré a priori	Idem Si réalisation rétrospective → nécessité d'apporter des garanties solides d'absence de p hacking
Disponibilité des données nécessaires pour répondre directement à la question de recherche	Les données nécessaires sont recueillies spécifiquement du fait de la nature prospective	Disponibilité non garantie, Choix de données « *fit-to-purpose* » Appel à des approches indirectes pouvant introduire des erreurs
Indépendamment de l'efficacité, les traitements comportent des risques	Évaluation et décision sur la balance bénéfice risque	Idem
Les effets des traitements sur des critères intermédiaires ne se traduisent pas toujours en bénéfice clinique	Évaluer sur des critères cliniquement pertinents	Idem
Population de l'étude différente de la population cible	Essai pragmatique	Idem (mais les études monobras hypersélectionnent les patients)

En plus, surviennent aussi des problématiques spécifiques de l'approche par comparaisons externes : utilisation de deux échantillonnages indépendants, utilisation de deux recueils d'information indépendants, contrôles non contemporains des patients traités, etc.

Tableau 3 – Problématiques méthodologiques des comparaisons externes et potentielles solutions

De nombreuses problématiques méthodologiques peuvent conduire à conclure à tort à l'intérêt du nouveau traitement avec une comparaison à un groupe contrôle externe. Ce tableau liste ces problématiques et mets en regard les solutions théoriques potentielles.

Problématiques posées par les comparaisons externes	**Mécanisme conduisant à conclure à tort à l'efficacité du nouveau traitement**	**Éléments de solution théorique**
Pour mémoire, la question des comparaisons externes se pose en l'absence de groupe contrôle (ou de contrôle approprié)	*Impossibilité de faire un raisonnement contrefactuel permettant d'isoler l'effet causé par le traitement*	*En l'absence de groupe contrôle interne, utilisation d'un groupe contrôle externe pour s'inscrire quand même dans un raisonnement contrefactuel indispensable pour avancer sur la voie de la causalité*
Possibilité de HARKing et de p hacking liée à l'aspect rétrospective de l'approche	La source de données utilisée pour constituer le groupe contrôle externe a été choisie, car une analyse inférentielle préalable des données montre que ce choix permet d'en tirer la conclusion recherchée	Attestation explicite de l'élaboration du protocole et du plan d'analyse statistique avant toutes analyses inférentielles Protocole daté, signé et enregistré Éventuellement traçabilité des accès aux données Transparence
	Le résultat de la comparaison peut être connu sans réelle analyse inférentielle, car le résultat du groupe traité est déjà connu ainsi que celui du groupe contrôle, car il a déjà publié pour son propre compte (étude de registre par exemple)	Rends inutilisable le groupe contrôle envisagé. Seuls seraient exploitables les nouveaux patients inclus dans le registre après la publication des résultats.
p hacking dans un contexte d'analyse rétrospective	L'analyse statistique a été adaptée en fonction des résultats produits jusqu'à obtenir la conclusion recherchée. Les nombreux tâtonnements ne sont pas rapportés (hidden analyses)	Plan d'analyse statistique (SAP) garantissant qu'il a été élaboré a priori, avant toute analyse inférentielle SAP daté et signé et enregistré
Réalisation d'étude exploratoire	Non-respect de la démarche hypothético déductive. L'étude risque de conclure sur une découverte fortuite purement artéfactuelle et sans réelle existence en dehors du jeu de données particulier analysé.	Réalisation d'étude de confirmation réalisée spécifiquement pour confirmer ou infirmer une hypothèse explicite de supériorité ou de non-infériorité Objectifs clairement spécifiés (et conclusion portant uniquement sur des résultats correspondant à ces objectifs)
Association statistique n'est pas causalité	Une étude observationnelle ne montre qu'une association statistique	Inférence causale par une étude et des données rendant raisonnablement plausible les hypothèses fondamentales de l'inférence causale
Biais de confusion	Existence de différences dans le risque de base des patients du groupe contrôle externe par rapport à ceux du groupe du traitement étudié (biais de confusion à la baseline)	Correction des résultats par l'analyse statistique prenant en compte tous les facteurs de confusion affectant la comparaison externe

Nécessité d'identifier tous les facteurs qui devront être contrôlés (ajustés) pour tenter de supprimer le biais de confusion	La liste des facteurs de confusion ne s'impose pas d'elle-même. Les facteurs de confusion sont à identifier cas par cas en raison de leur association connue avec le critère de jugement et avec le traitement.	Revue systématique des facteurs pronostiques des critères de jugement et établissement d'un réseau de causalité (DAG ou autres) à partir des connaissances et non pas des données
Possibilité de biais de confusion résiduelle	L'ajustement n'a pas pu corriger complètement le biais de confusion, car certains facteurs de confusion n'ont pas été identifiés, ou mesurés, ou en raison d'une mauvaise spécification du ou des modèles statistiques	Tenter de renforcer les arguments en faveur d'un faible risque de confusion résiduelle par des contrôles négatifs et/ou des analyses quantitatives des biais (E value ou autres). Pour certains, le biais de confusion résiduel persiste dans tous les cas, empêchant de considérer ces études comme à faible risque de biais (ROBINS-I).
Possibilité de biais de sélection, en particulier de biais de temps d'immortalité	L'inclusion des patients et/ou les périodes d'observation dans l'étude dépendent à la fois du critère de jugement (ou de facteurs de risque du critère de jugement) et du traitement	Le suivi a débuté dans les 2 groupes lorsque l'éligibilité est vérifiée et le traitement assigné Le t0 du groupe contrôle externe est synchrone de l'inclusion dans la monobras (ou de la randomisation) L'émulation d'un essai cible satisfaisante
Possibilité de biais de sélection par censures informatives (perdus de vue)		Remplacement des données manquantes suivant un scénario du pire
Déplétion des susceptibles ou censures à gauche		New user design (incident ou prévalent) Alignement du temps zéro (top départ du suivie) avec les critères d'éligibilité, l'assignation des traitements. Prévention de la censure à gauche
Biais de classification des expositions	Les patients du groupe contrôle n'ont pas eu le traitement qui définit ce groupe (→ remise en cause de l'hypothèse de cohérence de l'inférence causale)	Étude de validation des données Contrôles positifs et les techniques de benchmarking
Biais d'information (classification de l'outcome)	Le critère de jugement disponible dans la source de données pour constituer le groupe contrôle n'est pas le même ou n'est pas mesuré de la même façon que le critère de jugement du groupe traité	Utilisation du même critère de jugement pour le groupe contrôle que celui du groupe traité s (parfois impossible) Mesure de l'exactitude des données dans une étude de validation → VPP, VPN, Sensibilité, Spécificité Analyse quantitative de biais
Biais de réalisation, asymétrie de prise en charge des patients	Différence dans la fréquence et l'efficacité des traitements de base, concomitants ou post-échec et des prises en charge entre le groupe traité (étude monobras contemporaine) et le groupe contrôle (données historique)	Utilisation de données les plus comptemporain possible et provenant de contexte de soins identiques
Utilisation d'un estimand non approprié	L'estimand doit correspondre à la question causale : quel bénéfice clinique supplémentaire cause la prescription par le	Estimation de l'effet du traitement moyen dans la population cible des patients traités (*ATT average treatment effect among the treated*)

	médecin du nouveau traitement par rapport à la stratégie actuelle chez les patients visés par le nouveau traitement	Analyse en intention de traité avec une stratégie de gestion des événements intercurrents de type « policy treatment » pour les hypothèses de supériorité
Pertinence des données	Les variables nécessaires pour faire l'étude ne sont pas présentent dans la source de données (critère de jugement, facteurs de confusion, contrôles négatifs, variable de sélection des patients, etc.), conduisant à la réalisation d'une étude de méthodologie sous optimale	Choisir une source de données « fit-to-purpose » après l'élaboration du protocole Faire un recueil prospectif des données Compléter les données historiques par une chart review si l'identification des patients est possible Chainer des bases de données entre elles Utiliser des données multibases
Non-contemporanéité des données	Les données du groupe contrôle historique ont été recueillies dans le passé où le contexte de soins, les prises en charge de patients et les traitements subséquents ne sont plus d'actualité. Il y a une tendance séculaire	Prise en compte/modélisation de la tendance séculaire dans l'analyse (introduis une hypothèse de validité supplémentaire) Restriction temporelle des données historiques aux périodes les plus récentes
Effet étude non réductible	Le recueil des données s'effectuant dans 2 études différentes il peut exister des différences entre les 2 études influençant les critères de jugement (par exemple différence de critère de jugement entre les 2 études)	Pas de solution. Éventuellement analyse quantitative de biais À anticiper à la construction de la monobras (*externally controlled trial*) en adoptant le même critère de jugement que celui disponibles dans le groupe contrôle mais soulève d'autres questions
Inflation du risque alpha liée à une multiplicité de comparaisons statistiques	La multiplicité des comparaisons augmente le risque alpha global de trouver un quelconque intérêt au traitement à tort du fait du hasard	Plan de contrôle du risque alpha global (répartition, hiérarchisation, recyclage du risque alpha global)
Selective reporting	Sélection des résultats rapportés dans le rapport ou la publication parmi un grand nombre de résultats produits	Enregistrement du protocole et du SAP
Biais de publication	Plusieurs études de comparaison externe ont été réalisées et leur présentation, publication au sens large dépend de leurs résultats	Enregistrement prospectif des protocoles Réalisation d'études multibases

6 Les comparaisons externes sont des études observationnelles

Lorsqu'un groupe contrôle externe est utilisé avec une étude monobras, cette dernière est expérimentale, mais la comparaison au groupe contrôle externe est de nature observationnelle, car l'étude utilisant ce groupe contrôle n'a conditionné en rien la génération des périodes exposées de celui-ci [19].

> *Une étude observationnelle est une étude qui n'exerce aucune influence sur la prise en charge des patients qu'elle utilisera pour répondre à sa question de recherche. Cette prise en charge a été celle habituelle des mêmes patients dans le soin courant (vraie vie), contrairement aux études expérimentales où la prise en charge des patients est conditionnée par l'étude elle-même et va différer de celle qu'elle aurait été en dehors de l'étude [19].*
> *L'étude observationnelle ne génère pas ses données, mais utilise de manière opportuniste des données issues de l'observation distante de ce qui se passe dans la vraie vie aussi bien en termes de traitements que d'évolution des patients.*

Ainsi une grande partie des problématiques des études de comparaisons externes se déduisent directement des problématiques qui ont été identifiées pour les études observationnelles évaluant l'efficacité et la sécurité des traitements [20] (*non randomized non interventional studies* dans la terminologie de la FDA [4]).

> *Le terme « étude non interventionnelle » recouvre en partie le concept d'étude observationnelle, mais n'est cependant pas du même champ lexical. Il s'agit d'un terme d'ordre réglementaire, administratif et non pas d'ordre méthodologique ou épidémiologique. Si on parle du design de l'étude, il s'agit d'une étude observationnelle et en terme réglementaire elle rentre dans la catégorie administrative des études non interventionnelles. On assiste cependant à un certain glissement lexical avec par exemple l'apparition du terme non interventionnel dans le titre de l'étude et de la publication. La définition réglementaire des études non interventionnelles peut varier en fonction des pays/régions tandis que la définition de l'étude observationnelle est universelle.*

6.1 Le manque de fiabilité des études observationnelles

Jusqu'à très récemment, les études observationnelles étaient considérées comme insuffisamment fiables pour répondre à des questions d'efficacité des traitements [21] [22] [23] [24] [25]. Ceci reposait sur l'existence de problématiques méthodologiques jusqu'à présent insolubles [26] et sur de très nombreux exemples où l'efficacité de traitements mis en évidence par des études observationnelles avait été ensuite infirmée par des essais randomisés de confirmation (cf. encadrés).

> **Traitement hormonal substitutif de la ménopause (THSM) et prévention des événements cardiovasculaires** – L'étude observationnelle de la Nurses' study montrait que l'utilisation du THSM était associée avec une moindre fréquence des événements cardiovasculaires par rapport aux femmes ne l'utilisant pas [27]. L'essai randomisé WHI versus placebo a mis en évidence que le THMS causé une augmentation des événements cardiovasculaires de xx% [28]. Cet exemple fût un des premiers objectivant

la possibilité d'obtenir des résultats faussement positifs sur l'efficacité des traitements avec les études observationnelles et a fortement contribué à la mise en évidence des limites de ces études pour cet usage [29].

Sémaglutide et prévention de la maladie d'Alzheimer – Deux essais randomisés ont échoué[7] à montrer que le sémaglutide permettait de prévenir la maladie d'Alzheimer (EVOKE NCT04777396 et EVOKE+ NCT04777409 [30]). Ces essais ont été entrepris sur la base de 2 études observationnelles montrant un bénéfice potentiel. Une de ces études était construite suivant le cadre de l'émulation d'un essai cible. Malgré cela il est probable que cette étude ait conduit à des résultats biaisés en raison d'un biais de temps d'immortalité comme le témoignent les courbes d'incidences cumulées de la figure rapportées dans la publication [31].

Des études méta-épidémiologiques [32] [33] [34] [35] mettent en évidence ce manque de fiabilité des études observationnelles et le risque encouru de conclure faussement à l'efficacité des traitements. Les études observationnelles de haute qualité peuvent donner des estimations proches de celle des essais randomisés, mais pas de façon systématique, obérant ainsi la possibilité d'en faire un outil standard d'évaluation des bénéfices cliniques [36] [37].

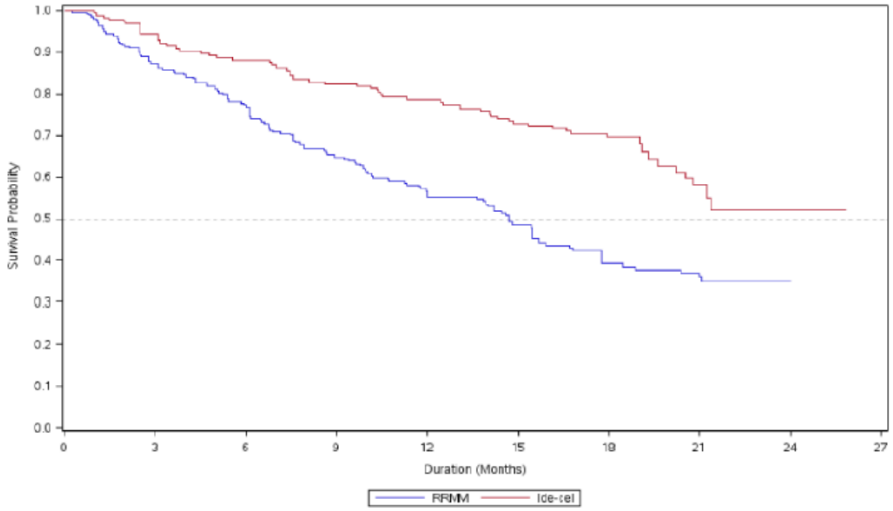
Ces observations, et les réticences qui en découlent, concernent principalement les études observationnelles « classiques », mais peuvent naturellement être étendues aux comparaisons externes qui ne sont qu'un type particulier d'études observationnelles sur l'efficacité des traitements.

On commence d'ailleurs à observer des discordances similaires entre des comparaisons à un groupe contrôle externe et un essai randomisé évaluant tous les deux le même traitement versus le même contrôle dans des situations cliniques similaires (cf. exemple de l'idecabtagene vicleucel ci-dessous).

[7] https://ml-eu.globenewswire.com/Resource/Download/1328a3cb-6359-4bb8-aef1-c7eab58d3016

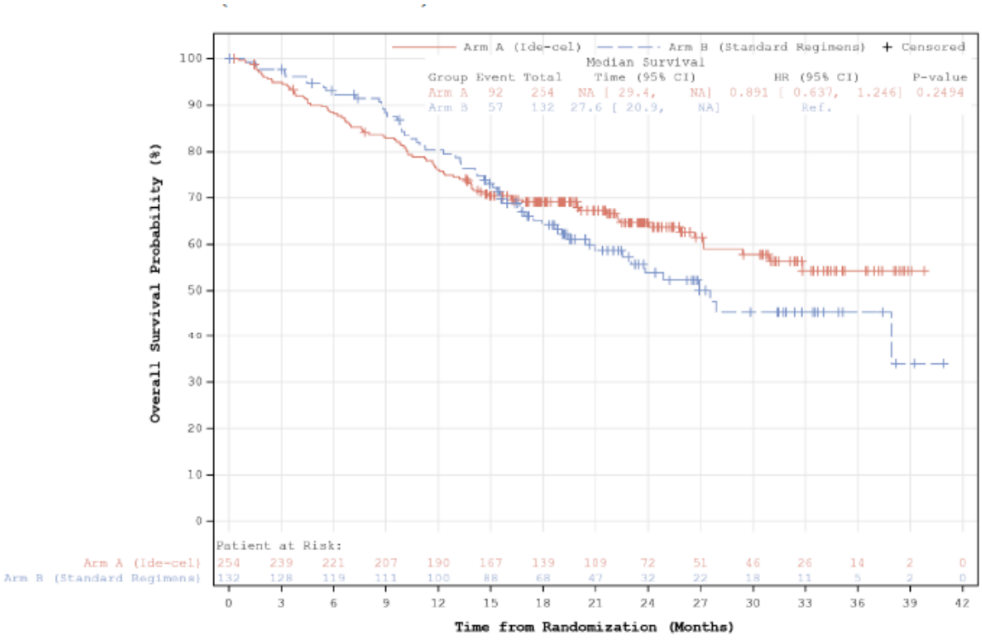
Le CAR T cell idecabtagene vicleucel (Abecma) n'a fait l'objet que d'une étude monobras dans le traitement du myélome multiple réfractaire ou en rechute (RRMM) en 4ème ligne ou plus (L4+). Une comparaison au standard of care a été effectuée à l'aide d'un groupe contrôle externe et elle montre un bénéfice clinique important du CAR T cell par rapport aux traitements standards en termes de survie globale. Cette comparaison n'a pas été publiée et elle n'est disponible que dans l'EPAR du produit (Figure 28 page 97) [38].

Figure 28: Overall Survival Adjusted for Trimmed Stabilised IPTW for the Eligible RRMM Cohort and Ide-cel Enrolled Cohort



Ultérieurement un essai randomisé effectuant la même comparaison chez des patients légèrement différents (2ème ligne ou plus) n'a pas mis en évidence de bénéfice du CAR T celle par rapport au traitement standard. (EPAR variation [39]).
Le résultat de la comparaison externe s'avère donc très discordant par rapport à celui de l'essai randomisé amenant à s'interroger sur sa fiabilité.

Figure 33. KM plot of OS - ITT population at DCO (03-Oct-2022) for the additional interim OS IA2



Pour l'instant il n'existe un nombre réduit d'études méta-épidémiologiques de validité empirique des comparaisons externes, principalement parce que les essais randomisés de confirmation ne sont pas réalisés (cf. section 28.2).

6.2 Des études montrant des associations, mais ne permettant pas de conclure à la causalité

Devant ces constats, les revues biomédicales exigent que les études observationnelles soient rapportées en évitant tout langage de causalité et en mentionnant explicitement qu'elles n'ont mis en évidence que des associations, « association ne faisant pas causalité » [40]. Cette impossibilité est régulièrement rappelée dans les publications de ce type de travaux : « results are likely to be affected by counfounding bias and should be interpreted with caution » ; "… However, because observational studies are prone to confounding and selection bias, causality cannot be affirmed." ; "Because claim databases can be vulnerable to selection and confounding bias, these results are statistical associations but not causal.".

Récemment, cependant certaines revues ont entrepris une réflexion sur la possibilité d'aller au-delà de la mise en évidence d'associations et de conclure à la causalité devant le développement des techniques d'inférence causale [41].

6.3 Différences avec les études de pharmacoépidémiologie

Les groupes contrôles externes sont principalement envisagés pour compléter des études expérimentales[8] (monobras ou essais randomisés). Elles s'inscrivent alors dans un cadre différent de celui, plus classique, des études de pharmacoépidémiologie comme les études de comparative effectiveness (comparaison d'efficience).

Les études de pharmacoépidémiologie ont pour objectif, entre autres, d'évaluer le résultat produit par l'utilisation en pratique des traitements, ce qui est appelé l'effectiveness[9]. Il s'agit d'évaluer l'impact, les conséquences de l'utilisation d'un traitement telle qu'effectuée en pratique courante par les médecins par rapport à l'utilisation d'un autre traitement. La question posée est : en quoi la pratique courante basée sur un certain traitement change le devenir des patients par rapport à une pratique basée sur un autre traitement ; en faisant l'hypothèse que les bénéfices mis en évidence dans les essais randomisés ne reflètent pas directement l'impact que pourront avoir ces mêmes traitements dans la pratique de tous les jours (pour de multiples raisons : utilisation chez des patients différents de ceux inclus dans les essais en termes de comorbidité, de risque de base, défaut d'observance, mésusage, etc.).

À l'opposé, les comparaisons d'études expérimentales à des groupes contrôles externes cherchent à répondre à une tout autre question, qui est celle du bénéfice clinique que le traitement peut apporter au mieux, dans des conditions optimums d'utilisation (patients appropriés, observance optimisée). Il

[8] Les études expérimentales sont les études dans lesquelles le traitement reçu par le patient dépend entièrement du fait qu'il est inclus dans l'étude (à l'opposé des études observationnelles où l'étude n'influence en rien la nature du traitement des patients.)

[9] Le terme effectiveness est ambigüe car la FDA l'utilise pour désigner les résultats des essais randomisés évaluant le bénéfice clinique. Ainsi dans ce contexte ce terme devient synonyme de « bénéfice clinique » et il est utilisé pour faire la distinction avec « l'efficacy », l'efficacité sur des critères intermédiaires.

s'agit de la même question que celle qui est posée dans les essais randomisés de confirmation du bénéfice clinique (essais de phase 3, pivot).

À noter que certaines études de pharmaco-épidémiologie peuvent avoir un objectif similaire d'évaluer l'efficacité intrinsèque du traitement dans des conditions optimales d'utilisation, mais chez des patients non étudiés dans les essais de confirmation (co-morbidité, âge, femmes enceintes, etc.). Ces questions sont inaccessibles avec un groupe contrôle externe, car ces patients n'ont pas été inclus dans l'étude monobras par exemple.

Tableau 4 – Différences entre un essai non-randomisé non-interventionnel utilisant un groupe contrôle externe et une études observationnelles classiques.

	Groupe contrôle externe (utilisé pour émuler un RCT de confirmation)	**Étude classique de pharmacoépidémiologie**
Appellation	Essai non randomisé	Étude observationnelle
Objectif (estimand)	Bénéfice clinique dans des conditions expérimentales	Conséquence de l'utilisation d'un traitement telle qu'effectuée en pratique
	Bénéfice clinique intrinsèque dans des conditions optimales d'utilisation	Impact (conséquences globales) de la mise à disposition aux médecins d'un nouveau traitement
Données	Données expérimentales (groupe traité) + données observationnelles (groupe contrôle)	Toutes les données, traités et contrôles, sont des données observationnelles
Démarche hypothético-déductive	Étude de confirmation (impérativement) : démontrer le bénéfice clinique du nouveau traitement	Étude le plus souvent exploratoires (existe-t-il des différences entre les traitements)

7 Position des agences de régulation et de HTA

Bien que réticentes jusqu'aux années 2020-22 [42], les positions des agences de régulation et de HTA évoluent avec l'apparition de solutions potentielles aux problématiques méthodologiques. Les agences de régulation (FDA [43], EMA) ainsi que plusieurs agences de HTA (NICE, HAS) ont déclaré qu'elles pouvaient envisager, sur le principe, d'intégrer des *real world evidence* (RWE) dans leur décision [4] [44] [45] [46]. La majorité des agences ont émis des documents concernant les études observationnelles et les *real world evidence*, documents de positionnement et guides méthodologiques.

7.1 Documents des agences

FDA, USA

Les comparaisons externes sont explicitement mentionnées dans les utilisations reconnues par la FDA (Food and Drug Administration) des *real world evidence* envisageables pour informer les prises de décision et la construction des stratégies thérapeutiques (cf. Tableau 5) [4]. Elles sont classées dans les « *non randomized interventional study* » compte tenu de l'aspect expérimental (interventionnel) du groupe traité et dénommée « *externally controlled trial* ».

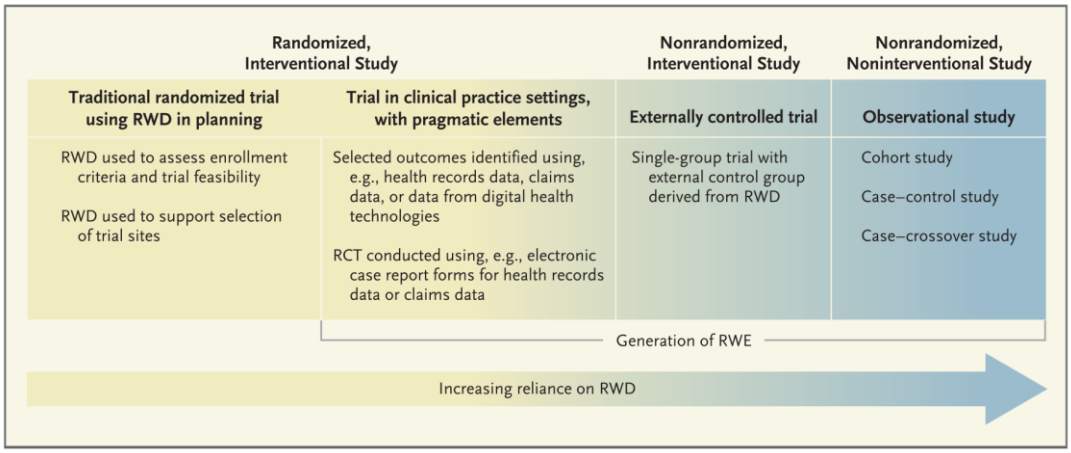


Tableau 5 – Utilisation envisagée des RWE par la FDA [4]

> *Les études de comparaisons externes se distinguent des études observationnelles où les deux groupes comparés sont issus des données de vraie vie («* nonrandomized, non-interventional study *»).*

Globalement la position de la FDA apparait plutôt précautionneuse n'envisageant la possibilité du recours à des comparaisons externes que si la réalisation d'un essai randomisé est impossible. Les guides fixent un haut standard pour la qualité des données, la transparence et le contrôle du biais de confusion (correspondant à ce qui est développé dans ce document) [43] [47]. Leurs attentes en termes de solidité de la démonstration restent identiques quelques soit le design de l'étude

interventionnelle (essais clinique) ou non interventionnelle (études observationnelles) réf. [45] page 4.

Aucune guideline spécifique des comparaisons à un groupe contrôle externe n'a été publiée à la date de rédaction de ce document (décembre 2025). Le plus proche est celui sur les essais contrôlés à comparateur externe (*externally controlled trials*) de la FDA qui concerne bien une comparaison à un groupe externe, mais prévue d'emblée avec l'étude monobras du nouveau produit [6].

Cependant les problématiques méthodologiques posées par ces études et leurs solutions sont identiques à celle des études observationnelles inférentielles du bénéfice du traitement (comme les études de comparatives effectiveness). En l'occurrence les guidelines de ces études s'appliquent aux comparaisons à des groupes contrôles externes comme le récent guide FDA [48].

Les recommandations concernant le recueil des données de vraie vie (registre, bases administratives, dossier médicaux informatisé, etc.) s'appliquent aussi au recueil des données pour le bras contrôle externe [49] [50] [51] [52].

Tableau 6 – Liste des guides FDA traitant des problématiques rencontrées dans les comparaisons externes (entres autres)

Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products [53] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-evidence-considerations-regarding-non-interventional-studies-drug-and-biological-products
Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products [6] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products
Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drugs and biological products [45] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug
Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products [51] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory
Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products [52] https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-registries-support-regulatory-decision-making-drug-and-biological-products

Plusieurs points méthodologiques fondamentaux sont explicitement précisés dans le document chapeau/générique « Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drugs and biological products [45] » publié en aout 2023 :

- Les attentes sont les mêmes, quelle que soit la nature de l'étude (interventionnelle ou non interventionnelle), définies dans 21 CFR 314.126.
- Le protocole et le plan d'analyse statistique doivent être enregistrés

- Le protocole, le plan d'analyse, la source de données n'ont pas été établis ou choisis en fonction des résultats
- Une transparence complète doit être assurée. Un journal d'audit des données (audit trail of data) doit être disponible.
- L'étude doit être monitorée afin de garantir son intégrité scientifique et sa qualité
- Les données et des codes doivent être disponibles et permettre de reproduire l'analyse
- Il est de la responsabilité du sponsor d'assurer l'intégrité scientifique de l'étude, le respect du protocole et du plan d'analyse statistique, etc.

Fin 2025, la FDA a lancé un programme pour 2026 d'incitation à la recherche dans la méthodologie de l'évaluation et de la régulation (*regulatory science*) [10].

EMA, Europe

Au niveau de l'EMA (European Medicines Agency) des informations concernant les comparaisons externes apparaissent dans le document au sujet des études monobras pivots [54].

Le document de réflexion sur l'utilisation des RWD pour générer des RWE destinée à la régulation mentionne son applicabilité aux comparaisons externes (*externally controlled trials*) [55].

Fin 2025, la position de l'EMA ne fait pas état d'une position établie, mais seulement d'une réflexion en cours.

Les comparaisons externes (ECA) apparaissent aussi dans le guide concernant l'évaluation des traitements en oncologie [56].

Tableau 7 – Liste des guides EMA abordant des problématiques rencontrées dans les comparaisons externes

Document spécifique à venir à la suite de la consultation annoncée par le document « Draft concept paper on the development of a reflection paper on the use of external controls for evidence generation in regulatory decision-making juillet 25 » [57] https://www.ema.europa.eu/en/documents/scientific-guideline/draft-concept-paper-development-reflection-paper-use-external-controls-evidence-generation-regulatory-decision-making_en.pdf
Reflection paper on use of real-world data in noninterventional studies to generate real-world evidence for regulatory purposes [55] https://www.ema.europa.eu/en/documents/other/reflection-paper-use-real-world-data-non-interventional-studies-generate-real-world-evidence-regulatory-purposes_en.pdf

MHRA UK

Le MHRA (Medicines and Healthcare products Regulatory Agency) britannique a ouvert une consultation en mai 2025 sur l'utilisation des groupes contrôles externes basées sur les RWD pour

[10] https://sam.gov/workspace/contract/opp/5f582391365645c2b030260aaf922e8b/view.

baser les décisions https://www.gov.uk/government/consultations/mhra-draft-guideline-on-the-use-of-external-control-arms-based-on-real-world-data-to-support-regulatory-decisions [58].

Dans ce projet de document, la position du MHRA apparait plus favorable aux comparaisons externes que celle de la FDA (cf. §11 et §12) avec une approche plutôt pragmatique pour accélérer les enregistrements tous en préservant la vigueur scientifique (mise en place du protocole de la comparaison externe avant le début des inclusions dans l'étude expérimentale par exemple).

Tableau 8 – Liste des guides MHRA abordant des problématiques rencontrées dans les comparaisons externes

Use of External Control Arms Based on Real-World Data to Support Regulatory Decisions (May 2025)
MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions

- ICH

Le concept de groupe contrôle externe apparait dans ICH E10 « Choice of control in clinical trials » [1] où il est conclu : « *the inability to control bias restricts use of the external control design to situations where the treatment effect is dramatic and the usual course of the disease highly predictable*"

ICH (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use) ne propose qu'un guide sur les études observationnelles destinées à l'évaluation de la safety et qui n'aborde pas spécifiquement la question des comparaisons externes [59].

Tableau 9 – Liste des guides ICH abordant des problématiques rencontrées dans les comparaisons externes

ICH HARMONISED GUIDELINE - General Principles on Planning, Designing, Analysing, and Reporting of Non-interventional Studies That Utilize Real-World Data for Safety Assessment of Medicines - M14 https://database.ich.org/sites/default/files/ICH_M14_Step4_Final_Guideline_2025_0905.pdf

Les grandes lignes de ce document sont :

- La nécessité de pré-spécifié un protocole, de l'enregistre avant toute analyse inférentielle afin de prévenir le p hacking
- L'utilisation de données adaptées (*fit-to-purpose*) avec une phase de qualification des données qui démontre leur adéquation à la question posée
- La transparence afin d'assurer la reproductibilité du travail

- ISPOR

L'ISPOR (The Professional Society for Health Economics and Outcomes Research) n'a pas publié de guide spécifique pour les groupes externes (décembre 2025). Cependant plusieurs de leurs autres guidelines, élaborées conjointement avec l'ISPE, concernent les études observationnelles (principalement de comparative effectiveness) [60] [61] [62] [63].

- NICE

Le NICE (National Institute for Health and Care Excellence) britannique mentionne les comparaisons externes dans un document général sur les real-world evidence [64] (https://www.nice.org.uk/corporate/ecd9/chapter/update-information)

- HAS

En France, la Haute Autorité de Santé (HAS) a publié un position paper dans lequel on trouve les grandes lignes de leurs attentes concernant les groupe contrôle externe [44]

- EUnetHTA, JCA

Les comparassions à un groupe contrôle externe sont mentionnées succinctement dans les deux guides JCA EUnetHTA *evidence synthesis* [65] [66].

7.2 Utilisation, évaluation par les agences

Le recours à des comparaisons à un groupe control externe est de plus en plus fréquent dans les dossiers soumis aux agences réglementaires [67].

Plusieurs études de méta-épidémiologie ont analysé les évaluations des agences de régulation et de HTA concernant des comparaisons externes qui leur ont été soumises dans des dossiers [68] [46][69] [70].

Il est possible de conclure que, les agences ne rejettent pas les comparaisons externes (comme les autres RWE) sur le principe, mais rejettent les études dont la méthodologie ne permet pas de produire des preuves solides, la barre des exigences de démonstration des bénéfices des nouveaux traitements étant la même quelle que soit la méthodologie employée pour évaluer l'intérêt clinique d'un nouveau traitement.

8 De la nécessité d'avoir des preuves de l'intérêt cliniques des nouveaux traitements

L'adoption d'un nouveau traitement dans la stratégie thérapeutique d'une condition clinique nécessite des preuves au-delà de tout doute raisonnable de l'intérêt clinique du traitement.

Cette exigence découle de nombreuses considérations éthiques, scientifiques, sociétales, politiques, et déontologiques. Elle assure aussi l'équité entre la collectivité et les industriels des produits de santé. Elle permet d'assurer aux patients, si elle n'est pas dévoyée, qu'ils bénéficient de traitements efficaces dont le bénéfice excède largement les risques encourus. Elle permet aussi aux médecins de respecter leurs engagements déontologiques d'assurer des soins fondés *sur les données acquises de la science (*article R4127-32 du code de la santé publique).

Des preuves au-delà de tout doute raisonnable sont indispensables pour décider de l'adoption d'un nouveau traitement

La méthodologie actuelle d'évaluation, reposant sur l'essai clinique randomisé, a été construite au fil du temps pour produire ces preuves au-delà de tout doute raisonnable (cf. Tableau 2). Cette méthodologie n'est pas un choix arbitraire de faire comme cela et non pas autrement, mais regroupe toutes les solutions aux problématiques méthodologies qui pourraient remettre en cause l'exactitude des résultats obtenus dans l'évaluation des traitements (cf. Tableau 2). Ce sont les principes à mettre en œuvre en termes de conception, réalisation, analyse, monitoring, audit, assurance qualité, transparence pour garantir la fiabilité des résultats produits ; et ces principes vont bien au-delà de la simple randomisation.

La question d'évaluer les nouveaux traitements, avant leur adoption à l'aide d'études reposant sur d'autres types de méthodologie, n'est pas une question de pluralisme, mais une simple question d'aptitude de ces nouvelles propositions méthodologiques, comme les comparaisons externes, à garantir la fiabilité des résultats produits : *Est-il possible de produire des preuves au-delà de tout doute raisonnable autrement qu'avec l'approche basée sur l'essai randomisé de confirmation correctement conçu et réalisé ?*

Avec le standard actuel d'évaluation, il est aussi apparu que les résultats des méthodologies moins abouties, donc moins robustes, produisaient des résultats faussement positifs pouvant faire croire à tort à l'intérêt clinique d'un nouveau traitement alors que celui-ci en était dépourvu, ou pire, aggravait les patients (le Tableau 15 en annexe liste des exemples de nouveaux traitements qui se sont avérés augmentant la mortalité dans leurs essais pivots sans que cela ait été suspecté avec les études préliminaires). Il s'est avéré ainsi qu'environ la moitié des études de phase 3 ne confirme pas les mécanismes d'action, les avis d'experts et les résultats préliminaires des phases 2 qui avaient conduit à la réalisation de ces essais [71] [72]. Ces données empiriques montrent l'impact négatif pour la santé publique, la déontologie, l'éthique, l'équité entre industrie et payeurs, qu'aurait une adoption des nouveaux traitements sur la base de résultats produits par des études ne contrôlant pas correctement toutes les problématiques méthodologiques et donc à risque de produire des résultats faussement positifs.

Il est aussi nécessaire de tenir compte de l'historique des bévues survenues avec l'adoption de traitement sur la base d'études insuffisamment solides comme, les retraits d'enregistrements accélérés par la FDA après non-confirmation du bénéfice de ces traitements lors d'essais de confirmation ultérieurs (cf. Tableau 16 en annexe) ou des molécules adoptées sur la base d'études non comparatives (avec ou sans comparaisons externes) et dont le bénéfice clinique n'a pas été retrouvé par un essai randomisé conduit ultérieurement (Tableau 17 en annexe).

De ce fait, lorsqu'une comparaison externe (ou en général une RWE), qui par définition est une méthodologie non optimale (cf. section 5), est envisagée à la place d'un essai randomisé pour justifier l'intérêt clinique d'un nouveau traitement, celle-ci doit être en mesure d'apporter des preuves au-delà de tout doute raisonnable. Autrement, il y aurait une acceptation implicite d'abaisser la barre des exigences nécessaires pour adopter un nouveau traitement, ce qui n'est pas justifiable.

Les exigences en termes de preuves solides pour adopter un nouveau traitement sont les mêmes, quelle que soit la méthodologie des études utilisées pour apporter cette démonstration

Ce point est parfaitement énoncé dans le guide FDA d'aout 2023 définissant leur cadre général d'acception des études non interventionnelles, intitulé « Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products » [45]. Ce guide mentionne, page 4, section B.1 Overview « *Regardless of a study's interventional or non-interventional design, the evidence submitted by a sponsor in a marketing application to support the safety and/or effectiveness of a drug must satisfy the applicable legal standards for the application to be approved or licensed* ». Ce texte, comme d'autres, confirme que, compte tenu des enjeux, il n'est pas question d'abaisser les exigences lorsque la démonstration de l'intérêt clinique d'un traitement est recherchée par une étude observationnelle/non-interventionnelle.

Le recours aux comparaisons externes ne doit pas s'accompagner d'une baisse des exigences en matière de solidité des preuves

Grâce aux progrès de l'épidémiologie théorique et de l'inférence causale, il devient envisageable, du moins en théorie, de produire des preuves qui répondent à ces attentes très strictes avec les études de comparaison externe (ou d'autres RWE). Toutefois, cela n'est pas encore pleinement garanti (cf. section 5) ni démontré de façon empirique (cf. section 28.2). Il demeure cependant possible de définir les conditions nécessaires à la production de ce type de preuve avec des approches observationnelles [20], ce qui l'objet de ce document.

9 Les sources de données utilisables

Les groupes contrôles externes peuvent être construits à partir de différents types de sources de données historiques ou prospectives.

9.1 Données historiques, RWD

Les groupes contrôles externes sont le plus souvent construits à partir de données historiques de patients déjà traités et suivis dans le passé, et dont les données ont été enregistrées. Les données collectées en routine, aussi appelées données de vraie vie (real world data, RWD) peuvent être utilisées pour cet usage.

Il peut s'agir de données déjà collectées, saisies et stockées sur un support informatique (bases de données par exemple) :

- données d'un essai randomisé ou d'un essai monobras précédent ayant évalué par exemple le traitement choisi pour la comparaison,
- registre de la pathologie (registre maladie rare par exemple comme le registre français de la maladie de Pompe [73]) ou d'un registre traitement (comme le registre DESCART des patients recevant un CAR-T-cells [74]),
- cohorte historique (comme la cohorte Chinoise des cancers du sein [75]),
- entrepôt de données de santé de soin courant (*collections of electronic health records*) constituées entre autres par les données de dossiers médicaux électroniques extraites et structurées sous une forme permettant leur exploitation future,
- base de données administrative, *claims databases*, etc. (comme le SNDS, les données de la Kaiser Permanent, etc.),
- données générées par les patients eux-mêmes ou à travers des dispositifs mobiles.

Dans d'autres cas, les données sont conservées, mais non encore saisie sous forme structurée et nécessiteront un travail de saisie ou extraction spécifique pour la constitution du groupe contrôle :

- Données des dossiers médicaux qui nécessiteront une « *chart review*[11] » afin de remplir un CRF spécifique du groupe contrôle externe
- En cas de dossiers médicaux électroniques, les données nécessaires pour le groupe contrôle externe pourront être identifiées et extraites par des outils basés sur l'IA (ou exploitée par une *chart review*)

En pratique ces sources et moyens peuvent être combinés par exemple en utilisant un registre pour identifier les patients éligibles puis en réalisant une *chart review* sur leurs dossiers médicaux pour compléter les informations manquantes dans le registre. Les données peuvent aussi être complétées par un chainage avec une base administrative par exemple pour les traitements reçus (dispensés).

[11] Le terme « chart review » (ou revue de dossiers médicaux) désigne une collecte de données à partir des dossiers médicaux des patients pour extraire des informations pertinentes à des fins de recherche.

À partir de ces sources de données, le groupe contrôle externe sera créé en extrayant les patients présentant les mêmes critères d'éligibilité que ceux utilisés pour constituer le groupe traité (monobras ou RCT) et ayant été traités par le traitement comparateur voulu.

L'intérêt de recourir à des données historiques ou collectées en routine médicale pour constituer un groupe contrôle externe provient du fait que ces données existent déjà. Cela apporte un gain de temps et de cout. Cependant cette préexistence conduit aux limites des analyses/études rétrospectives qu'il sera nécessaire de gérer spécifiquement (cf. section 10).

- Avantages inconvénients des différentes sources de données secondaires

Chacune des différentes sources de données historiques possibles présente des avantages et des inconvénients.

Les données d'une étude clinique expérimentale précédente (essais randomisés, étude monobras) représentent la meilleure option sur plusieurs plans. Elles peuvent fournir des données sous placebo si celui-ci a été utilisé dans un essai précédent (par exemple pour valider la première génération de traitement de la pathologie), contrairement à toutes les autres sources. La qualité des données est optimale, car enregistrées avec un monitorage des centres et un data management. Les variables sont aussi parfaitement bien définies et recueillies directement par les investigateurs (à comparer par exemple avec les proxys ou les algorithmes phénotypiques qu'il est nécessaire d'utiliser avec les bases de données administratives). Pour une question en oncologie portant sur une tumeur solide, ce sont les seules données qui peuvent documenter les critères de jugement de réponse tumorale (critères RECIST) et de survie sans progression (PFS), ces deux critères étant complètement inaccessibles avec les données recueillies en routine (cf. section 20.7). Par contre, l'accès à ces données est difficile, car appartenant au promoteur de l'étude qui commercialise le traitement standard que le nouveau traitement concurrencera. Cette problématique est plus facile à résoudre s'il s'agit de données issues d'un essai académique précédent. Afin de faciliter l'accès à des données de ce types, plusieurs initiatives académiques ou privées se mettent en place pour regrouper des données d'études précédentes dans des entrepôts (cf. section **Erreur ! Source du renvoi introuvable.**). Ces données sont a ussi souvent limitées sur le plan des covariables (facteurs pronostiques des critères de jugement) et des contrôles négatifs.

Par rapport aux autres sources, les bases de données administratives n'ont pas été initialement conçues dans un but de recherche clinique, mais pour administrer le paiement des institutions et des acteurs de santé. Leur logique et leur finalité sont comptable et non scientifique. Elles ne contiennent pas directement des informations cliniques, biologiques ou d'imagerie, mais seulement des codages suivant une certaine nomenclature des diagnostics, des motifs d'hospitalisation, des actes et des traitements dispensés. Leur remplissage peut aussi être réalisé, non pas dans une logique d'exactitude scientifique, mais d'optimisation financière.

Les informations cliniques nécessaires pour les études nécessitent d'être reconstruites à partir des codages à l'aide d'algorithme phénotypique. Suivant la nature de la source de données, l'information sur les événements cliniques peut être directement présente ou nécessité un travail de reconstruction. C'est le rôle des algorithmes phénotypiques dans les bases administratives qui font cette reconstruction à partir des codages et d'autres informations. Ces algorithmes phénotypiques sont des règles structurées utilisant les données disponibles dans les données de vraie vie (RWD) pour identifier des patients avec une condition clinique, exposition ou événement spécifique.

Parfois il est nécessaire de recourir, à la place de l'entité clinique souhaitée qui est non identifiable en raison des codages utilisés, à un concept lié, appelé proxy, qui lui est accessible.

Ces éléments interrogent souvent sur l'exactitude des données qui peuvent être extraites de ces bases administratives et il est indispensable de les valider par des études spécifiques (cf. section 20.4).

Les autres sources de données sont davantage susceptibles de contenir les données cliniques nécessaires à la réalisation de l'étude de comparaison externe, mais il peut aussi subsister des problématiques de qualités des données en termes d'exactitude et surtout de complétude, la saisie des informations dans les registres se faisant souvent en sus de l'activité clinique des centres participants.

L'utilisation des dossiers médicaux nécessite un travail important d'extraction des informations nécessaires à l'étude. Ce travail peut être supprimé ou réduit par l'utilisation de techniques de traitement naturel du langage, en particulier par IA, mais débouche sur la question de la fiabilité des informations extraites. Se pose aussi la question de données manquantes indétectable.

Les questions de la qualité des données et des problématiques méthodologiques induites sont discutées en détail dans la section suivante (section 9.5).

Au-delà de ces aspects méthodologiques, l'exploitation des données secondaires est une nouvelle discipline qui couvre des aspects très variés, légaux, règlementaires, techniques, informatiques qui ne sont pas abordées dans ce document. De même les développements informatiques très prometteurs d'extraction par traitement automatisé du langage (IA) ne sont abordés.

Les sources de données type registre et cohorte présentent l'avantage d'avoir été connue et mise en place avec un objectif recherche. Elles sont bien plus riches en données médicales, biologiques, d'imagerie que les bases administratives. Une de leur limite réside dans la qualité des données, en particulier des taux importants de données manquantes. Leur utilisation comme groupe contrôle externe peut aussi être compromise du fait des publications qu'elles génèrent pour leur propre compte. En effet, la finalité de ces recueils d'information est de produire des publications indépendamment de leur éventuelle utilisation pour des groupes contrôles externes. Cependant ces publications peuvent révélées les résultats du groupe contrôle externe. En raison du HARKing qui en découle (cf. sections 10 et 10.1), ces sources peuvent ne pas être utilisables quand il s'agit de chercher un groupe contrôle externe pour un groupe traité dont on connait déjà les résultats (choix post hoc d'un groupe contrôle externe pour une étude monobras déjà terminée par exemple). L'exemple de la section 10 illustre le côté rédhibitoire que peut prendre cette problématique.

9.2 Groupe contrôle externe prospectif

Le groupe contrôle externe peut aussi être constitué par un recrutement prospectif de patients présentant les critères d'éligibilités et le traitement voulu. Ces patients seront ensuite suivis prospectivement pour l'acquisition du critère de jugement.

Par rapport aux données historiques, ces données présenteront de nombreux avantages en termes de qualités de données. Toutes les données nécessaires à la comparaison externe pourront être

recueillies (critères de jugement[12], facteurs de confusion, contrôles négatifs) alors que fréquemment certaines de ces variables ne sont pas disponibles dans les données historiques. La qualité des données pourra être assurée via un processus d'assurance qualité reposant sur un data management et un monitorage sur site, qui a souvent été inexistant lors du recueil des données historiques (ce qui se caractérisent par exemple par un fort taux de données manquantes). Un autre avantage est de fournir des données plus contemporaines du groupe traité que les données historiques (en termes de contexte de soins, patients et traitements).

En revanche la durée de ce processus représente son principal défaut. Après la mise en place de l'étude, il sera nécessaire d'attendre le recrutement des patients et la fin de leur suivi longitudinal avant de pouvoir exploiter les données, tandis que les données historiques sont exploitables immédiatement.

De plus, dans le cas des essais monobras, la réalisation prospective d'un groupe contrôle met en évidence que la randomisation était tout à fait possible (aussi bien en termes de nombre de patients mobilisables pour l'étude que de durée de réalisation) montrant, de ce fait, l'aspect questionnable de la réalisation d'une simple étude monobras et du recours à une comparaison externe. En effet, comme un essai randomisé était tout à fait réalisable, recourir à un design moins fiable pour établir un nouveau traitement soulève la problématique éthique de courir le risque de conclure à tort à l'intérêt d'un traitement et de le faire utiliser à tort alors qu'une évaluation fiable aurait été tout à fait accessible. La plupart des agences de régulation ont en premier dans leurs critères d'acceptabilité des études monobras l'impossibilité réelle de réaliser un essai comparatif randomisé. [44] [58].

Ce type de recueil de données nécessite de mettre en place une « étude » spécifique qui recrutera des centres investigateurs (services hospitaliers ou médecins suivant le cas). Ces investigateurs recueilleront ensuite les données nécessaires pour tous les patients éligibles qu'ils prendront en charge dans leur pratique quotidienne durant la durée de ce recueil d'information. Étant observationnelle (non interventionnelle) cette étude n'interviendra en aucune manière sur la façon dont les investigateurs prendront en charge les patients. Seul un recueil supplémentaire d'information (ou greffé sur le système d'information habituel utilisé par ces investigateurs) sera demandé par l'étude et ne concernera que des informations habituellement produites par la pratique des investigateurs (pas d'examen complémentaire supplémentaire réalisé uniquement pour les besoins de l'étude).

À ce niveau il convient de bien distinguer le recueil d'information nécessaire pour constituer le groupe contrôle externe et l'étude de comparaison externe pour laquelle ce recueil est effectué. La logistique mise en place pour la constitution du groupe contrôle externe est souvent appelée « étude » et protocolisée de façon indépendante sous la forme d'une étude descriptive de real world evidence. Dans cette approche la réelle finalité de ce recueil est quasiment éclipsée, apparaissant au mieux comme objectif secondaire de cette étude descriptive. Méthodologiquement cette approche est problématique, car la comparaison externe doit être une étude à part entière compte tenu de l'importance de ses enjeux (et non une simple analyse exploratoire de données). Elle doit impérativement faire l'objet d'un protocole spécifique. Le recueil d'information n'est qu'un outil pour réaliser cette comparaison externe (cf. section 11).

Il est aussi possible, à cette occasion, d'initier la mise en place, de novo, d'un nouveau registre qui servira dans un premier temps pour la présente étude et pour d'autre étude dans le futur.

[12] Sauf si ceux qui nécessiteraient de changer la pratique médicale courante (comme la PFS suivant les critère RECIST) car, dans ce cas, le recueil d'information observationnel deviendrait interventionnel.

9.3 Sources dédiées

Devant l'importance prise par les études monobras dans le développement de certains nouveaux traitements, en oncologie en particulier, des bases dédiées ont été créées à la construction de groupe contrôle externe ont faites leur apparition. La plus importante est la base commerciale Flatiron (https://flatiron.com/real-world-evidence/real-world-data).

La société MEDIDATA commercialise pour la création de groupes contrôles externes des données historiques issues de fichiers d'essais cliniques (https://www.medidata.com/en/clinical-trial-products/medidata-ai/real-world-data/synthetic-control-arm/).

Transcelerate biopharma inc. propose le même type de partage de données historiques d'études cliniques (https://www.transceleratebiopharmainc.com/initiatives/historical-trial-data-sharing/)

Quelques initiatives se mettent en place pour regrouper aux mêmes endroits des données d'études précédentes ou de registres dans le but de simplifier la création de groupe contrôle externes. L'initiative la plus aboutie en avril 2026 concerne le glioblastome [76].Ces registres peuvent être constitués soit prospectivement soit en regroupant des sources de données existantes.

Tableau 10 – Comparaissons des différences sources de données historiques

	Avantages	**Inconvénients**
Registres, cohortes	Effectifs importants Suivi longitudinal Standardisation du recueil des données Diversité des patients Biais de sélection type 2	Limitation aux données collectées Manque souvent des variables clés (critères de jugement, facteurs de confusion, contrôles négatifs Données manquantes
Dossiers médicaux, EDS, electronic health records	Richesse en variables cliniques Effectifs importants Exhaustivité des variables patients Données biologiques et d'imagerie	Inconsistance dans le recueil des données (non standardisées,) Données non structurées nécessitant un pré-processing important Données manquantes
Chart review	Spécifique à la question de recherche Information spécifique Recueil de données ciblées Capture les nuances médicales, biologiques, etc.	Couteux en temps et en argent Effectifs plutôt réduits Nécessite un consentement patient Non représentatif de la population
Bases de données administratives	Disponibilité Grand nombre de patients Conservation des historiques des traitements et des actes	Finalité très éloignée de la recherche clinique Pas de données cliniques
Essais cliniques précédents	Qualité optimale des données Variables bien définies et mesurées directement Disponibilité des critères de jugement spécifiques à l'évaluation (PFS par exemple)	Données propriétaires, difficilement accessibles en cas d'études industrielles de traitements qui seront concurrencés par le nouveau traitement N'existe pas toujours (maladies rares)

9.4 « Données de la baseline »

En remplacement d'une comparaison à un groupe contrôle externe, certaines études monobras utilisent une comparaison intra-patient du changement de la valeur du critère de jugement par rapport à la baseline, en invoquant le principe du patient étant son propre témoin. Le changement entre la fin de l'étude et la valeur de baseline (« *change from baseline* », on parle aussi de « *change score* », est alors proposé comme mesure de « l'effet causal » du traitement. Cependant il est montrable que ce changement ne permet pas de faire de l'inférence causale [77].

> *Le principe du patient étant son propre témoin est exploitable dans l'essai randomisé en cross-over, justement en raison de la randomisation des séquences de traitement sur les 2 périodes de l'essai. Cette randomisation permet de se défaire de l'influence de l'évolution temporelle.*

En dépit de cette limite qui est mal connue, cette approche simpliste est fréquemment utilisée. Elle a été retrouvée dans 33% des soumissions hors oncologie à la FDA incluant un « contrôle externe » [78].

Le principe du patient étant son propre témoin peut être exploité dans les études observationnelles, sous la forme d'étude de type « série temporelle interrompue » ou autres designs autocontrôlés (*self controlled*) [79] [80] [81] [82] [83] [84]. Dans ces études, le critère de jugement fait l'objet de nombreuses mesures au cours du temps avant et après l'introduction du traitement. Les mesures répétées avant le traitement permettent de modéliser l'évolution « naturelle », sans traitement, de chaque patient. De même, les mesures répétées après l'introduction u traitement permettent de modéliser l'évolution sous traitement. L'effet du traitement est recherché en comparant les modèles avant et après.

La comparaison des évolutions avant et après permet dans une certaine mesure d'isoler l'effet traitement de l'évolution naturelle du critère de jugement au cours du temps (à condition que le modèle soit bien spécifié).

La forme la plus aboutie de cette approche, appelée « *difference in differences* », DID [85], utilise un vrai groupe contrôle externe comme référence de l'évolution temporelle sans traitement sur la même période. Mais il existe de nombre méthode ou design d'étude « *self-controlled* » cherchant à exploiter le sujet comme son propre témoin [86].

Ce sujet étant très différent des groupes contrôle externe (qui par définition repose sur une comparaison inter-patient), nous ne rentrerons pas plus dans le détail de ces méthodes.

9.5 Recherche et qualification de la source de données

La recherche de la source de données s'effectue après avoir élaboré le protocole de l'étude de comparaison externe, car celui-ci conditionne les données nécessaires pour l'étude [87]. Choisir arbitrairement une source de données avant de connaitre les variables dont il va être nécessaire de disposer pour les critères de jugement, les critères de sélection des patients, les facteurs de confusion, les contrôles négatifs, etc. expose à ce que l'étude présente, in fine, de nombreuses limites invalidant son interprétation causale. De même, choisir une source de données, sans connaitre le nombre de sujets nécessaire pour garantir la puissance aux comparaisons à réaliser, expose à réaliser l'étude avec une source de données insuffisante et, par conséquence, à obtenir des résultats non statistiquement

significatif conduisant à une étude non concluante ne pouvant apporter la preuve recherchée du bénéfice clinique du traitement étudié.

> *Toutes les sources de données ne permettent pas de faire une étude donnée. Si une étude est forcée d'être réalisée avec des données inadaptées, insuffisantes ou de mauvaise qualité, le protocole qui aurait permis d'obtenir des résultats fiables ne pourra pas être réalisé et l'étude ne pourra pas produire des résultats fiables et recevables pour démontrer le bénéfice clinique du traitement étudié. Au total la réalisation de cette étude aura été un gaspillage des ressources.*

Si cette comparaison externe est effectuée de manière post-hoc après que les résultats de l'étude monobras (ou du RCT) soient connus, il est indispensable de montrer que le choix de la source de données a été effectué, indépendamment des résultats produits, uniquement sur la base de son adéquation avec l'étude qu'il était nécessaire de réaliser.

Ainsi survient, après la construction du protocole, une étape d'inventaire des sources de données potentiellement adaptées pour le groupe contrôle externe envisagé. Cette étape dresse une liste, la plus exhaustive possible suivant une démarche prédéfinie de revue systématique de la littérature et des autres ressources à même d'identifier des sources de données [64] [45]. Puis chaque source de cette liste est méthodiquement passée en revue afin de déterminer son adéquation à l'étude en termes d'accès aux données, de qualité, de disponibilité des variables nécessaires, etc.

Il est extrêmement important que ce processus de qualification de la source des données soit soigneusement tracé/transparent pour pouvoir garantir que le choix définitif n'a pas été effectué en fonction des résultats produits, mais bien uniquement sur l'adéquation des données à l'étude envisagée [88]. L'enjeu est de pouvoir faire accepter aux agences, comme l'HAS [44] ou le NICE [64], un groupe contrôle construit a posteriori.

Indépendamment de ce point, l'élaboration du protocole et du plan d'analyse statistique avant le choix des données permet de garantir l'absence de HARKing et de p hacking indispensable pour faire accepter les résultats (cf. section 10).

Lorsque les données adaptées au protocole et pouvant procurer le nombre suffisant de patients sont identifiées, le protocole peut être complété pour intégrer les parties spécifiques à ces données : description de leur origine, de la validation de leur qualité, des algorithmes phénotypiques, etc.

La question du stade auquel le protocole doit être enregistré se pose ici (cf. section 10.3). L'enregistrer avant le choix des données oblige à procéder à sa modification par amendement lorsque les données auront été choisies, mais cette démarche est la plus transparente et donc adaptée pour garantir l'absence de HARKing et de p-hacking.

Cette démarche de qualification des données est maintenant bien codifiée par différents textes de recommandation comme PRINCIPLED [87], the Structured Process to Identify Fit--For--Purpose Data (SPIFD, SPIFD-2) [89] [90], le framework RWE (ECD9) du NICE (section « Choosing fit-for-purpose data ») [64] ou un guideline plus ancien publié par Hall et al. [91].

Durant cette phase de qualification de la source de données, des analyses sont souvent nécessaires, par exemple, pour connaitre le nombre de sujets potentiellement éligibles, le taux de données manquantes, etc. Il est impératif qu'aucune analyse inférentielle ne soit réalisée durant ces analyses [87] et que cela puisse être prouvé (cf. section 10).

Tableau 11 – Mentions concernant la recherche et la qualification des données dans les principaux guides des agences

Agence	Document	Page	Extrait
EMA	Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence - Scientific guideline	-	NA
MHRA	MHRA draft guideline on the use of external control arms based on real-world data to support regulatory decisions	§38	If the ECA is identified after the clinical trial has finished, it will be important to adequately justify the choice of external control and how it was chosen from amongst all possible other data sources. In particular, as the results of the clinical trial are already known, it will be important to demonstrate that the decision was not result driven, i.e., a dataset with particularly poor results was chosen for comparison when another dataset exists for the same population with better results.
FDA	Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products	5	Sponsors should describe in the study protocol, or as an appendix to the protocol, the data sources evaluated when designing the study, including results from feasibility evaluations or exploratory analyses of those data sources. Sponsors should provide a justification for selecting or excluding relevant data sources from the study. Sponsors describe how the choice of the final data sources, study design elements, and analytic approaches aligns with the research question of interest and that the data sources, study design elements, and analytic approaches were not selected to favor particular study findings.
FDA	Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products Guidance for Industry	§D, page 6	Sponsors should demonstrate the appropriateness of the proposed data source(s) to address specific hypotheses and research questions. Given that data sources used in a non-interventional study design are often generated for purposes other than research, it is important that sponsors understand the potential limitations of such data sources and determine whether those limitations can be addressed or if another data source should be pursued.
FDA	Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products	4	access to and evaluation of relevant data sources or databases are important steps in designing a control arm for externally controlled trials and in evaluating the trial's feasibility. Sponsors should document and describe in the trial protocol all data sources accessed when designing the control arm of the trial and the results of any feasibility evaluations or exploratory analyses. Sponsors should provide a justification for selecting or excluding relevant data sources and demonstrate that the choice of a final analytic dataset for the control arm aligns with the research question of interest and was not chosen to favor particular study results.
HAS	Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health [44]	2	To avoid a post hoc selection, the choice of an external control must be done prior to conducting the trial, after a well--performed systematic search. It must fit the standard of

			care. The retained external source of data must be chosen because it fits best the research question and not because it would arbitrarily favour the treatment of interest.
NICE	NICE real-world evidence framework [64]	35	We encourage developers to identify candidate data sources through a systematic, transparent and reproducible search, including: …

9.6 Accès aux données

Des difficultés pour accéder aux données peuvent apparaitre pour des raisons de confidentialité et d'anonymat par exemple. À ce niveau, deux solutions techniques peuvent contribuer à lever ces difficultés.

Le calcul réparti ou fédéré (*federated learning*) permet d'utiliser des données issues de plusieurs hôpitaux ou bases de données sans jamais centraliser les données individuelles. Les données ne sont pas regroupées et les calculs sont répartis entre les centres qui les détiennent. Seuls les résultats intermédiaires circulent. De telles techniques ont été développées par exemple pour l'IPTW et les données de survie [92].

Les données synthétiques pourraient être aussi une solution pour éviter de transmettre des données sensibles [93].

L'utilisation de ces données devrait permettre à des études de produire les mêmes résultats que si les données réelles avaient été utilisées [94] [95] [96].

Cependant la fiabilité de cette approche et son aptitude à produire des résultats d'études fiables est encore débattue [97].

> *Le terme synthétique peut aussi prêter à confusion dans ce domaine. Les groupes contrôles externes constitués à partir de données de patients sont parfois appelés « groupe contrôle synthétique ». Dans ce vocable, le terme synthétique n'implique pas le recours à des données synthétiques ou à d'autres types de données artificielles.*

10 Les problématiques liées à l'aspect rétrospectif de ces études

Bien qu'il soit tout à fait possible de construire un groupe contrôle externe de manière prospective, ces études sont quasi exclusivement des analyses rétrospectives utilisant des données historiques (cf. section 8).

La démarche rétrospective de ces études offre la possibilité de construire l'étude et son analyse en fonction des résultats produits (HARKing et p hacking) en explorant différentes sources de données, fenêtres temporelles d'extraction, jeux de variables d'ajustement, méthodes d'analyse et en ne retenant que l'analyse donnant les résultats les plus intéressants. Ainsi les résultats produits n'auraient qu'une valeur exploratoire insuffisante pour changer les pratiques.

La possibilité de HARKing et de p hacking représente l'une des plus grosses limites méthodologiques des études rétrospectives, mettant ainsi en périls la validité des comparaisons à un groupe contrôle externe historique

Étude/analyse rétrospective. Le terme rétrospectif a été utilisé en épidémiologie avec de nombreuses acceptions toutes différentes. Actuellement il désigne des études (pour les puristes, des analyses) conçues alors que les données existent déjà. Devant la polysémie de ce terme, certains, comme les recommandations de reporting STROBE [98], recommandent de ne plus l'utiliser, d'employer éventuellement le terme « historique » pour qualifier les données et surtout de bien décrire la chronologie entre la génération des données (génération des périodes de temps exposées) et la décision d'utiliser ces données pour l'analyse considérée.

Pour considérer que les résultats d'une analyse rétrospective sont de nature confirmatoire, il faut pouvoir exclure avec certitude HARKing et p-hacking. Cela implique d'avoir la garantit que l'analyse a été unique et réalisée d'après un protocole et un plan d'analyse statistique (SAP) établi a priori. Le protocole et le SAP doivent être datés et signés, l'étude doit avoir été enregistrée avant sa réalisation dans un registre comme clinicaltrials.gov (cf. section 10.3) et, surtout, les investigateurs doivent attester explicitement dans le rapport et la publication que l'étude a été conçue avant toutes analyses inférentielles [87].

La réalisation d'une **analyse de faisabilité** (cf. section 10.4), avant de décider de réaliser l'étude, est problématique sur ce plan et doit impérativement garantir que la faisabilité a été évaluée sans disposer des données sur les critères de jugement et sans réalisation d'aucune analyse inférentielle [87].

La FDA a rejeté une comparaison externe[13] pour laquelle le risque de HARking était important [99]. La comparaison externe avait été décidée et conçue après que les résultats du groupe traité avaient été connus. Les résultats du groupe contrôle choisi étaient aussi déjà connus, car ayant fait l'objet d'une publication. Et cette publication avait déjà était utilisé pour faire une comparaison indirecte de type MAIC

[13] https://download.open.fda.gov/crl/CRL_NDA210862_20251104.pdf

avec les résultats du groupe traité. Dans son avis la FDA rappelle l'intérêt des *externally controlled* trial qui évite l'ensemble de ces problèmes (cf. section 3).
Globalement cette comparaison externe présentait tous les défauts de conception liés au caractère rétrospectif de ces analyses et pouvant conduire au HARKing. Non seulement les données, mais aussi les résultats étaient disponibles au moment de la conception de la comparaison externe. Le groupe contrôle était finalement la récupération d'un travail antérieur de RWE réalisé avec son propre objectif. Cette situation est très fréquente, car les registres et les cohortes ont presque toujours des objectifs de description qui conduise par principe à faire des publications de ce type. Il apparait donc que les objectifs initiaux des registres et cohortes compromettent leur utilisabilité pour produire des groupes contrôles externes acceptables pour l'évaluation des nouveaux traitements. Une solution serait de n'exploiter que les nouveaux patients inclus dans les registres depuis leur dernière publication.
De plus, ces mêmes données avaient déjà été exploitées dans une MAIC réalisée avec la publication du groupe contrôle ce qui est équivalent à une analyse inférentielle de la comparaison externe réalisée avant la conception de l'étude de comparaison indirecte et faisant décider de sa réalisation.
Il est complètement impossible d'exclure que cette analyse n'a pas été décidée uniquement du fait qu'elle conduisait au résultat attendu. Elle ne peut être considérée comme un test loyal de l'hypothèse de la supériorité du nouveau traitement à son contrôle et n'a donc aucune valeur de preuve.

10.1 Le risque de HARKing

10.1.1 Problématique

Les approches rétrospectives présentent plusieurs limites méthodologiques directement liées au fait que les données sont disponibles au moment où l'étude est envisagée. En effet, cette préexistence offre la possibilité de « préanalyser » les données avant de décider de les utiliser pour cette question de recherche ou avant de finaliser la question de recherche et d'établir le protocole et le plan d'analyse statistique (SAP). Dans ces deux cas, l'étude aura été préconditionnée par construction pour donner le résultat attendu. Elle ne sera plus une confrontation loyale d'une hypothèse à la réalité, mais une sorte de démarche tautologique, toujours gagnante, ou l'on utiliserait des données pour tester si une hypothèse est « confirmée » ou infirmée par la réalité alors que l'on connait par avance le résultat. On parle de HARKing (hypothesizing after the results are known) [100]. Les bases de la démarche scientifique hypothético-déductive ne sont donc pas respectées (cf. livre blanc, Dossier 1 – La démarche hypothético-déductive et les résultats post hoc[14]).

Cette situation peut exister même sans véritable « pré analyse » des données par les investigateurs de l'étude (cf. exemple précédent). Les résultats potentiellement obtenables en choisissant une certaine source de données ou une certaine façon de les analyser peuvent transparaitre à travers des publications antérieures basées sur les mêmes données. Par exemple pour les maladies rares, il existe de nombreux registres académiques potentiellement mobilisables pour la constitution de groupes contrôles externes, mais ces registres font l'objet de publications descriptives purement académiques régulières. Ces publications révèlent dans une certaine mesure les résultats qui seraient obtenus dans le groupe contrôle externe et peuvent orienter a priori le choix vers la source de données qui conduira au meilleur résultat voulu pour la comparaison à cette source.

[14] https://sfpt-fr.org/livreblancmethodo/part4/file_0.htm

10.1.2 Solution

La solution à cette problématique n'est pas simple. Elle passe par différents éléments de méthode qui pourront finalement relativement assurer l'absence de HARKing et donc la fiabilité des résultats.

Un protocole et un plan d'analyse statistique doivent impérativement avoir été élaborés a priori, avant toutes analyses inférentielles des données.

> *Une **analyse inférentielle** est une analyse qui produit les résultats pour lesquels on réalise l'étude, comme les résultats d'efficacité et de sécurité dans le cas de l'évaluation d'un traitement.*

Pour attester de cela, le protocole doit être versionné et daté. Il doit être aussi enregistré.

Mais aucune de ces mesures ne donne une garantie absolue, car un protocole, un SAP peut être écrit et enregistré après avoir fait l'analyse des données (étant donné qu'elles existent déjà).

Des garanties supplémentaires peuvent être parfois apportées quand les données ne sont accessibles que sur autorisation de leur gestionnaire, mais cela ne concerne qu'un faible nombre de sources de données. Même dans ce cas cette autorisation n'est pas une garantie absolue, car les mêmes données ont pu être disponibles antérieurement dans le cadre d'un tout autre projet.

L'inscription du travail de recherche que représente la comparaison externe dans le cadre d'un « Registered Report » augmentera notablement le degré de confiance vis à vis des résultats qui seront produits ultérieurement.

> **Registered report**
>
> Un *registered report* est un format de publication scientifique dans lequel le protocole de l'étude, incluant la question de recherche, la méthodologie et le plan d'analyse statistique, est soumis à une revue par les pairs et enregistré auprès d'une revue scientifique avant la collecte ou l'analyse des données. Ce processus vise à garantir la transparence et l'intégrité scientifique, en évitant que les résultats de l'étude n'influencent la conception ou l'analyse, ce qui limite les risques de HARKing (formulation d'hypothèses après connaissance des résultats) et de p-hacking (multiplication d'analyses pour obtenir des résultats significatifs). Une fois le protocole accepté, la revue s'engage à publier les résultats quelle que soit leur nature, à condition que l'étude soit menée conformément au protocole enregistré.

Pour les registres, les cohortes collections, etc. il est fréquent que les études soient réalisées par l'équipe gestionnaire de la source de données.

Dans ces conditions, il est finalement attendu que les auteurs **attestent explicitement** dans le protocole, le SAP, les rapports et les publications que l'étude a été entièrement construite avant toutes analyses inférentielles. Si au moment de l'analyse, des choix préétablis s'avèrent inadéquats, une section « déviation par rapport à l'analyse prévue » permettra de les décrites et de les justifier.

Cette attestation n'est pas non plus une garantie absolue, mais la faire à tort devient un acte de fraude scientifique (et non plus de méconduite scientifique).

Dans le cadre du recours à groupe contrôle externe pour une étude monobras, le choix du groupe contrôle dans le protocole de l'étude monobras apporte une garantie absolue d'absence de HARKing, car l'étude monobras étant prospective, le choix du groupe contrôle est alors forcément antérieur à toute analyse inférentielle.

10.2 Le risque de p-hacking

10.2.1 Problématique

Au moment de l'analyse des données, un problème peut aussi provenir de l'exploration de nombreuses façons d'analyser les données en faisant varier les méthodes, les covariables, les critères de sélection des patients, les définitions des populations d'analyses, les critères de jugement, les temps de mesure, etc. Ces différentes options pouvant donner des résultats très différents (on parle de vibration des effets en fonction de l'analyse effectuée). Il pourrait être alors possible de ne retenir que l'analyse donnant les résultats les plus proches de ceux souhaités. On parle alors de p hacking [101] [101] [102].

10.2.2 Solution

Pour rendre acceptable à but décisionnel une approche rétrospective, il est nécessaire de garantir l'absence de p hacking. Cela passe par l'élaboration d'un plan d'analyse statistique a priori, avant toutes analyses inférentielles (directe ou indirecte, cf. problématique des publications descriptives des sources de données) et par son strict respect lors de l'analyse effective des données. Cela permet de garantir que les résultats sur lesquels reposera la prise de décision proviennent d'une analyse unique des données, définies indépendamment des résultats qu'elle produit. Il existe maintenant un guide décrivant le contenu de ce document [103]

Cependant l'établissement d'un SAP ne donne pas une garantie formelle d'absence de p hacking, car ce plan peut avoir été établi après une analyse non déclarée. Il en est de même pour l'enregistrement du protocole ou du SAP. Ainsi la garantie d'absence de p-hacking ne peut passer que par une attestation explicite dans le rapport et la publication de l'étude que l'analyse présentée a été unique et établie a priori. Bien que purement déclarative, cette attestation explicite donne des garanties fortes, car, si en réalité cela n'a pas été respecté, cette mention relèvera du domaine de la fraude scientifique. Ce point est abordé dans les guidelines de plusieurs agences (cf. Tableau 12) sous la forme du besoin d'une transparence forte sur la réalisation de l'étude [13].

> *On peut remarquer que la problématique des analyses cachées et du p hacking concerne aussi les études prospectives, comme les essais cliniques randomisés. La solution a été apportée par le plan d'analyse statistique, mais dans les études prospectives la date de ce document, antérieure à la disponibilité des données, donne une garantie formelle qu'il a bien été établi a priori, indépendamment des données. Mais cette garantie formelle est impossible à apporter avec une analyse rétrospective, car la date d'établissement du SAP est forcément postérieure à la disponibilité des données par définition.*

Études ou analyses rétrospectives ? Compte tenu de ces limites des études rétrospectives, certains ne parlent que **d'analyses** et non pas d'études rétrospectives, pour mettre l'accent sur le fait que les garanties méthodologiques apportées par les **études** prospectives ne peuvent pas être apportées dans la démarche rétrospective.

Tableau 12 – mention des problématiques liées aux études rétrospectives dans les guides des agences

Agence	Document	Page	Extrait
EMA	Draft Concept Paper on the Development of a Reflection 5 Paper on the Use of External Controls for Evidence Generation in Regulatory Decision-Making	3	Prospectively planned external control comparisons vs comparisons conducted when results are already available (either trial data, external control or both)
EMA	Reflection paper on use of real-world data in noninterventional studies to generate real-world evidence for regulatory purposes	-	Problématique non abordée dans ce document
MHRA	MHRA draft guideline on the use of external control arms based on real-world data to support regulatory decisions		§23. The protocol for the trial should be of the same standard, style and level of detail as would be expected for a traditional RCT intended to support a regulatory submission, including prespecification of the objectives, data to be collected, primary and secondary endpoints and analysis methods. §26 It is important that the protocol is finalised before enrolment begins e.g. the design and analysis methods should be sufficiently specified in the protocol and there should be no amendments planned to fill in important missing details. §47 The precise method to be used for the analysis and the covariates to be included should be pre-specified and justified in the protocol.
FDA	Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products	5, section 2	2. Transparency Regarding Data Collection and Analysis FDA must be confident (based on corresponding documentation) that particular data sources or databases were not selected, or that specific analyses were not conducted, to favor a certain conclusion. The protocol and SAP should be finalized and shared with FDA prior to conducting the prespecified analyses described in the protocol and SAP. In addition, any revisions to the protocol should be date-stamped, and the rationale for each change should be provided. To ensure transparency regarding their study design, sponsors should post their study protocols on a publicly available website, such as ClinicalTrials.gov13 or the web page for the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) for post-authorization studies.
FDA	Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products Guidance for Industry	7	The prespecified SAP should address the specific study objectives and detail the primary 216 analysis and any secondary analyses.
FDA	Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products	NA	-
HAS	Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health [44]	2	To avoid a post hoc selection, the choice of an external control must be done prior to conducting the trial, after a well--performed systematic search. It must fit the standard of care. The retained external source of data must be chosen because it fits best the research question and not because it would arbitrarily favour the treatment of interest

NICE	NICE real-world evidence framework [64]	125	Planning studies before conduct improves the quality of studies and can reduce the risk of developers performing multiple analyses and selecting those producing the most favourable results. Developers should aim to pre-specify as much of the study plan as possible When decisions will be driven by the data, these should be clearly described and planned approaches justified. Pre-specifying analysis plans is especially important for studies of comparative effects. For such studies, we encourage publishing the study protocol on a publicly accessible platform, with any changes to the protocol registered and justified.

Il faut aussi noter que le principe de détermination a priori de la méthode d'analyse est quelque peu antinomique avec les principes et pratiques de la modélisation statistique [104]. Dans cette approche il convient de trouver le modèle le mieux spécifié, c'est-à-dire qui mathématiquement représente le mieux les données. Pour cela il est nécessaire de tester plusieurs méthodes (modèle traitement, modèle d'outcome, méthode double robuste) et plusieurs modèles pour chaque approche. Le choix du modèle final s'effectuant en comparant les prédictions données par chaque modèle. Théoriquement le choix ne se fait pas sur l'effet traitement obtenu, mais celui-ci est forcément calculé et donc révélé aux investigateurs durant ce processus.

10.3 Importance du protocole et le plan d'analyse statistique

L'établissement d'un protocole et d'un plan d'analyse statistique (SAP) est indispensable, non seulement pour protocoliser la démarche de l'étude, mais surtout, dans le cas d'études rétrospectives, pour assoir la garantie d'absence de HARKing et de p hacking.

Pour apporter cette garantie, le protocole et le plan d'analyse statistique doivent avoir été établis a priori avant toutes analyses inférentielles.

> *Le récent guide FDA sur l'utilisation des RWD et RWE pour les dispositifs médicaux mentionne :* « Therefore, to mitigate potential bias, careful study design is needed, and a study protocol and analysis plan should be created prior to accessing, retrieving, and analyzing RWD, regardless of whether the RWD are already collected (retrospective) or if they are to be collected in the future (prospective design). *» page 8 [50]*
>
> *Le document NICE real-world evidence framework est lui aussi explicite [64] : "Planning studies before conduct improves the quality of studies and can reduce the risk of developers performing multiple analyses and selecting those producing the most favourable results.", "Pre-specifying analysis plans is especially important for studies of comparative effects. For such studies, we encourage publishing the study protocol on a publicly accessible platform, with any changes to the protocol registered and justified"*

Dans cette perspective le protocole doit être daté, signé, et enregistré dans un registre d'études.

Cet enregistrement peut être effectué dans clinicaltrials.gov qui comprends maintenant un descripteur « observationnal study ». Il permet d'obtenir un numéro NCT d'identification unique de l'étude.

Exemples d'enregistrement NCT d'études de comparaison externe :
https://clinicaltrials.gov/study/NCT05796726
https://clinicaltrials.gov/study/NCT04697446
https://clinicaltrials.gov/study/NCT00230243
https://clinicaltrials.gov/study/NCT05236257
https://clinicaltrials.gov/study/NCT06973161
https://clinicaltrials.gov/study/NCT06504524

Il existe plusieurs autres registres possibles comme le « HMA-EMA Catalogue of real-world data studies » de l'EMA (https://catalogues.ema.europa.eu/) ou les plateformes d'open science comme Center for open science (https://www.cos.io/).

Une initiative de l'ISPOR cherche aussi à promouvoir la transparence dans les études observationnelles : *Real-World Evidence Transparency Initiative* (https://www.ispor.org/strategic-initiatives/real-world-evidence/real-world-evidence-transparency-initiative). Leur registre est hébergé sur le site de science ouverte OSF (https://osf.io/registries/rwe).

La publication sur une plateforme d'open science comme OSF (https://osf.io/), Zenodo (https://zenodo.org/), etc., permet de rendre publics les documents de l'étude (protocole et SAP) et assure une totale transparence sur l'étude. Cette publication peut être effectuée à la place ou en complément d'un enregistrement sur un registre d'études plus classiques. En effet, les informations collectées par les registres d'études sont en général succinctes et s'avèrent insuffisantes pour apporter toute la transparence nécessaire pour garantir la fiabilité des résultats de l'étude une fois réalisée.

L'enregistrement du protocole et du SAP est explicitement mentionné dans plusieurs guides d'agences : NICE [64], HAS [44], comme dans ceux de plusieurs sociétés savantes comme l'ISPOR [60] [62] [105], l'ESMO [106].

Cet enregistrement est aussi souhaité par presque tous les journaux, comme le JAMA [107] ou le BMJ [108]. Il est aussi considéré comme primordial dans les guidelines de l'ISPOR [60] [61] [62] (cf. section 11).

Bien entendu l'existence d'un protocole et d'un plan d'analyse statistique ne suffit pas à elle seule pour garantir l'absence de HARKing ou de p-hacking. Encore faut-il que ces documents donnent des éléments factuels et transparents comme quoi ils ont bien été élaborés à priori et avant tout accès aux données des critères de jugement, c'est-à-dire avant toute analyse inférentielle. Cela doit être justifié en attestant de cela de façon explicite et non ambiguë dans le protocole, le SAP, les rapports de l'étude et la ou les publications (cf. section 10).

10.4 Analyse de faisabilité

HARKing et p hacking peuvent aussi provenir de la réalisation d'analyse de faisabilité avant de construire définitivement le protocole et de réaliser l'étude. En effet, la faisabilité de ces comparaisons n'est pas toujours acquise et il est fréquent qu'elles ne soient pas réalisables pour des raisons de

qualité des données, d'absence des données indispensable comme le critère de jugement, les covariables, etc. Il y a donc dans ces études un instant où les données envisagées doivent être investiguées pour vérifier qu'elles permettront de réaliser l'étude envisagée (mais il faudra impérativement garantir que cette analyse n'a pas permis de vérifier que leur analyse donnera le résultat escompté !). On parle souvent d'étape de qualification des données.

Pour ne pas remettre en cause l'intégrité scientifique de l'étude, il est impératif que cette étape garantisse que cette qualification a été effectuée indépendamment de toute analyse inférentielle (c'est-à-dire une analyse correspondant à l'objectif de l'étude), par exemple en garantissant que les valeurs des critères de jugement n'étaient pas disponibles dans le fichier ayant servi à la qualification des données [87].

Dans une démarche prospective, l'étude ne peut pas être décidée ou conçue en fonction des résultats qu'elle donnerait, car les données n'existent pas encore au moment où la question de recherche est formulée et le protocole et le plan d'analyse statistique établis (à condition de les respecter dans l'analyse).

11 Rédaction du protocole

La rédaction d'un protocole et d'un plan d'analyse statistique avant la réalisation d'une étude de comparaison externe est indispensable pour garantir la valeur méthodologique du travail (cf. section 10).

Le but de ces comparaisons externe étant de fournir des preuves (RWE) pour justifier de l'intégration du nouveau traitement dans la stratégie thérapeutique, il s'agit d'une étude à part entière et non pas d'une simple analyse rétrospective de données.

Le protocole doit être centré sur la comparaison et non pas sur la constitution du groupe contrôle externe. L'objectif primaire de l'étude est de faire avant tout une comparaison externe. La constitution du groupe contrôle n'étant qu'un moyen et non une finalité.

La démarche qui sera utilisée pour constituer le groupe contrôle externe doit évidemment être aussi protocolisée de manière rigoureuse, mais il ne s'agit qu'un des éléments de l'étude de comparaison externe. En effet la finalité de cette partie est de fournir des données adaptées à la comparaison qui est envisagée, mais non de faire de cette comparaison. Ce point peut paraitre trivial, mais on rencontre assez fréquemment des protocoles qui prévoient la comparaison externe d'intérêts comme simple objectif secondaire, ou accessoire, d'un protocole d'une étude de RWD dont l'objectif primaire était la description d'une cohorte traitée avec le traitement standard. La démarche ne peut être celle d'une simple étude de « RWE » dont l'objectif est de décrire une cohorte de patients traités dans la vraie vie avec le traitement standard et qui, en objectif secondaire ou accessoire, prévoit de faire la comparaison d'intérêt (en précisant souvent qu'il s'agit d'une comparaison exploratoire) comme par exemple : https://clinicaltrials.gov/study/NCT07028489, https://clinicaltrials.gov/study/NCT05842486, https://clinicaltrials.gov/study/NCT06973161.

Une comparaison externe ne peut pas reposer entièrement et uniquement sur un simple plan d'analyse statistique (SAP) élaboré, par exemple, en marge d'un registre existant. À nouveau, ces comparaisons sont des études à part entière, visant à produire des preuves, et non pas de simple analyse de données. Le SAP ne peut se substituer à un protocole, il n'est qu'une mise en forme opérationnelle de l'analyse prévue au protocole.

Une comparaison externe étant une étude à part entière, un protocole et un plan d'analyse statistique établi a priori sont obligatoires

De même le protocole général de constitution d'un registre ne peut être considéré comme étant le protocole d'une comparaison externe. Celui-ci ayant été établi pour encadrer le recueil de données sans objectif précis d'étude, il n'aborde aucun aspect de la comparaison externe réalisée avec ses données (comme par exemple https://clinicaltrials.gov/study/NCT04328298).

Les études conçues comme un essai comparatif à contrôle externe (« *externally controlled trial* ») au sens proposé par la FDA dans le guide « *Considerations for the Design and Conduct of Externally Controlled Trials*» [6] reposent sur un seul protocole. En effet ces études sont conçues d'emblée

comme étant une comparaison externe d'un groupe traité avec le nouveau traitement et d'un groupe contrôle (souvent historique). Cette comparaison externe est donc prévue avant le recueil des données des patients traités. Un seul protocole va codifier la réalisation de la partie expérimentale avec le nouveau traitement, la constitution du groupe contrôle externe et de la comparaison de ces 2 groupes.

À l'opposé, lorsque la comparaison externe prend place après la réalisation d'une étude monobras, le protocole ne concernera que la constitution du groupe contrôle externe et la comparaison. Au total l'étude reposera sur deux protocoles : celui de la comparaison externe et celui antérieur de la monobras.

Dans cette situation il est aussi possible de séparer le protocole de la constitution du groupe contrôle externe de celui de la comparaison externe. Apparaissent alors trois protocoles au total. L'individualisation du protocole de constitution du groupe contrôle externe correspond survient souvent lorsque le groupe contrôle n'est pas extrait d'une base de données cliniques déjà structurée, mais nécessite un travail approfondi d'identification des données comme une extraction de dossiers médicaux, une *chart review*, une extraction d'une base de données administratives, etc.

Plusieurs guides définissant le contenu du protocole des études observationnelles sont disponibles [105] [98] [103] [109] [110]. ISPOR propose un modèle de protocole standard dénommé HARPER [105]. Bien que non spécifique des études de comparaisons externes, ces guides plus généraux, s'appliquent aussi à ces études moyennant quelques adaptations mineures.

L'étude doit être construite avant le choix des données, car c'est l'étude qui conditionne les données nécessaires à sa réalisation en termes de variables de sélection des patients, de critères de jugements, de facteurs de confusion, et de contrôles négatifs, etc.

L'étude doit être construite avant le choix de la source de données, car c'est l'étude qui conditionne les données nécessaires et non pas les données qui drive la construction de l'étude.

Choisir les données en premier introduit des contraintes fortes dans la construction de l'étude qui peuvent conduire à une étude méthodologiquement inappropriée pour répondre à la question de recherche ; par exemple en utilisant un critère de jugement insuffisamment cliniquement pertinent, car le critère adapté à la question de recherche n'était pas disponible dans les données choisies ou effectuant des ajustements insuffisants pour contrôler tous les facteurs de confusion.

Toutes les données disponibles ne permettent pas de faire une étude donnée. Les données doivent donc être choisies en fonction de leur d'adéquation à l'étude envisagée et de leur qualité. Après la construction de l'étude et la rédaction du protocole, la première étape de réalisation d'une étude de comparaisons indirectes est de chercher les données adaptées à l'étude. On parle de qualification des sources de données, ou de « *fit-to-purpose assessment* » [89] [87].

Il est d'ailleurs possible que de telles données n'existent pas et qu'il s'avère donc impossible de répondre à la question de recherche par la voie d'une comparaison externe. En effet, faire à tout prix

une étude avec des données insuffisantes, inadaptées ou de mauvaise qualité conduira immanquablement à une étude qui ne pourra pas être exploitée pour la prise de décision et donc à un travail inutile. Dans ce cas la seule solution envisageable sera celle d'un groupe contrôle prospectif.

Bien entendu une fois les données identifiées, le protocole devra être complété avec les aspects spécifiques à des données. Ainsi la construction de ces études s'effectue en plusieurs étapes :

1. la construction de la partie méthodologique, la définition de l'essai à émuler, l'identification des facteurs pronostiques et des modificateurs des effets, etc. qui débouchent sur le cahier des charges que devront remplir les données pour permettre de réaliser cette étude
2. la recherche des sources de données potentiellement adaptées à l'aide de la démarche décrite dans le protocole, suivie de la sélection de la source de données adaptée à la réalisation de l'étude. Cette sélection peut conduire à des analyses des données. La plus grande transparence sur cette étape est nécessaire pour garantir l'acceptabilité méthodologique de l'étude étant donné le risque de choisi de la source en fonction des résultats produits.
3. la finalisation de protocole avec tous les aspects techniques et spécifiques liée aux données retenues

Avec cette séquence se pose la question du moment de l'enregistrement du protocole (cf. section 10.3). Si celui-ci est effectué à l'issue de l'étape 1, la finalisation du protocole après avoir trouvé les données appropriées fera l'objet d'un amendement du protocole. Une autre option est de n'enregistrer le protocole qu'une fois les données choisies et le celui-ci finalisé. [111] [87]. Cette approche présente l'inconvénient de conduire à un enregistrement du protocole alors que le travail a déjà débuté avec la recherche et la qualification des données adaptées.

Si la recherche des données est entreprise de manière post-hoc, après l'obtention des résultats de l'étude monobras (ou du RCT) la possibilité d'un choix du groupe contrôle en fonction des résultats produits devient une limite rédhibitoire si elle n'est pas correctement gérée (cf. section 10)

12 Démarche hypothético déductive

Les comparaisons externes à un groupe contrôle externe sont utilisées pour apporter la preuve du bénéfice clinique d'un traitement en remplacement d'un essai randomisé pivot qui n'a pas été réalisé. De ce fait il s'agit d'étude de confirmation comme les phases 3 pivots du médicament (ICH E9) [28].

> *Dans le cadre des essais cliniques, ICH E9 [28] fait clairement la distinction entre les essais de confirmation et les essais exploratoires, ces derniers étant insuffisamment robustes, du fait de leur approche purement inductive, pour constituer des études pivots prises en compte dans la décision. Cette distinction étant d'ordre épistémologique, elle s'applique de facto aux études observationnelles [29, 30].*

Pour cela les comparaisons externes doivent spécifier explicitement leur objectif en termes d'hypothèse, le but de la comparaison étant de confirmer ou infirmer l'hypothèse que le traitement évalué apporte un bénéfice clinique tout comme l'aurait fait un essai randomisé de confirmation.

> *De manière générale, nous verrons à plusieurs reprises que ces études de comparaisons externes ont tout avantage à s'aligner avec la démarche et la méthodologie de l'essai randomisé qu'elles visent à remplacer. C'est l'idée de l'approche de l'émulation d'un essai cible (cf. section 22).*

Elles s'inscrivent dans la démarche hypothético-déductive assurant la solidité scientifique de leur résultat. À l'inverse les études exploratoires qui ne testent pas spécifiquement d'hypothèse thérapeutique ne peuvent pas produire de preuve pour démontrer le bénéfice d'un traitement en raison des limites du raisonnement inductif sur lesquelles elles reposent.

L'explicitation d'une hypothèse permet d'inscrire l'étude dans le cadre de la démarche hypothético-déductive et assure donc la solidité épistémologique du résultat.

Ainsi, les études avec comparaison externe doivent être de réelles études de confirmation, s'inscrivant pleinement dans une démarche hypothético-déductive avec l'objectif explicite de confronter à la réalité une hypothèse de recherche latéralisée (supériorité, non-infériorité). Ces études de comparaison externe s'inscrivent donc dans le cadre plus général des études observationnelles « Hypothesis Evaluating Treatment Effectiveness » (HETE) telles que définies par l'ISPOR/ISPE [61] [62].

Cela signifie que ces études peuvent être « négatives », non concluantes, à la différence d'une étude exploratoire qui conclura toujours. Les résultats produits par ces études qui ne sont pas en lien direct avec leur objectif seront donc post hoc et purement exploratoires.

L'objectif doit ainsi être du type « montrer la supériorité de N par rapport à C sur les événements cardiovasculaires », « montrer la non-infériorité de N par rapport à C en termes de qualité de vie mesurée par x », etc. Autrement l'étude ne sera qu'exploratoire et impropre à amener une preuve formelle de l'intérêt clinique du traitement. Des objectifs du type « comparer N et C », « décrire le devenir des patients sous N et C », etc. ne cherchent pas à confirmer ou infirmer une hypothèse sur le bénéfice clinique du traitement et donnent donc des études seulement exploratoires et non pas de confirmation. La pré-formulation de l'hypothèse est indispensable dans un cadre de comparaison de 2

traitements. Sans cela l'étude ne peut jamais être « négative », non concluante, ne cherchant à réfuter aucune hypothèse. La conclusion à la supériorité d'un traitement par rapport à l'autre (ou à son infériorité ou à la non-différence entre les 2 traitements) reposera alors uniquement sur un raisonnement inductif qui est faible, car non logiquement contraint (modus tollens).

Une étude **exploratoire** comparant un nouveau traitement N versus un traitement contrôle C pourra faire trois conclusions N est supérieur à C, N est inférieur à C ou l'étude ne montre pas de différence. L'étude exploratoire n'est jamais négative, non concluante, car elle conduit toujours à une conclusion. Ne pouvant pas être non concluantes, ces études ne peuvent pas produire de preuve que seule la démarche hypothético-déductive peut fournir.

Par contre la même étude **de confirmation**, dont l'hypothèse est que N est supérieur à C, pourra être non concluante et réfuter cette hypothèse, d'où sa valeur épistémique.

13 L'inférence causale et les hypothèses sous-jacentes

13.1 Définition

L'inférence causale est une approche théorique, mathématique [112] [113] [114] qui a permis de déterminer les conditions que devait vérifier une association statistique pour avoir la valeur de relation causale. Cette approche permet aussi de clarifier la notion de question causale et d'estimand causal qui permettent de comprendre quelle analyse est nécessaire pour répondre à quelle question.

En pratique, dans une étude observationnelle, l'inférence causale vise à estimer, autant que possible, l'effet causal d'un traitement malgré l'absence de randomisation

Ainsi, pour permettre une conclusion causale, quatre hypothèses fondamentales doivent être vérifiées :

- L'hypothèse de positivité
- L'hypothèse de non-interférence
- L'hypothèse de cohérence (*consistancy*)
- L'hypothèse d'échangeabilité

L'inférence causale (conclure à un effet causal) nécessite donc la validité de ces quatre hypothèses appelées conditions d'identifiabilité (*identifiability conditions*).

L'inférence causale montre aussi que toutes méthodes d'estimation d'une différence entre les groupes n'estiment pas forcément un effet causal du traitement. De plus les termes "effet du traitement" ou "effet causal" sont extrêmement ambigus, car il est possible de définir plusieurs effets traitements (effets causaux). Tout dépend de la question causale dont on part [115].

Ainsi cette réflexion théorique [41] montrer l'importance fondamentale :

- De bien définir **l'effet traitement causal** que l'on cherche (on parle de **question causale**)
- D'en déduire la quantité, l'estimand qui correspond à cette question causale, c'est-à-dire **l'estimand causal**.
- D'utiliser une méthode statistique qui estime bien l'estimand causal (estimand statistique c'est-à-dire ce que mesure la méthode d'estimation est bien l'estimand causal auquel on s'intéresse)
- De vérifier que les données utilisées vérifient bien les hypothèses de l'inférence causale (on parle d'identification de l'estimand causal par les données) et les hypothèses de la méthode statistique utilisée pour l'estimation

eFigure 1: Schematic of the Relationship Between Causal Estimands, Statistical Estimands, and Statistical Analysis Methods Applied to Data

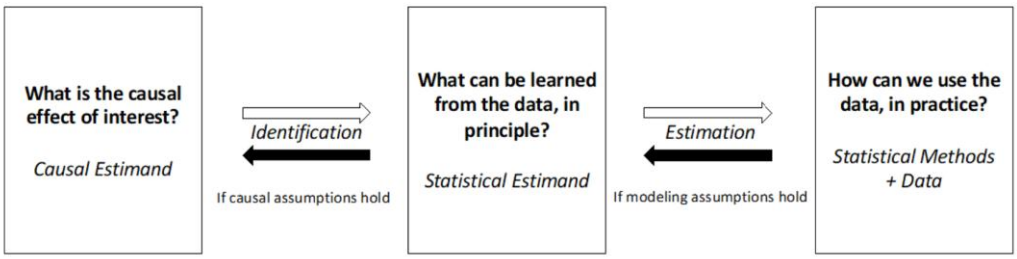


Figure 2 – Estimand causal, estimand statistique et estimation (supplément de la réf [41])

Une simple comparaison avant-après n'est pas une comparaison contrefactuelle acceptable [77]. Il est donc impossible de conclure à un effet causal à partir d'une simple étude monobras même si le critère est un changement avant après qui donne l'impression que le patient est son propre témoin et que le raisonnement contrefactuel est respecté (cf. section 9.4).

13.1.1 Hypothèse de positivité

L'hypothèse de positivité (*positivity assumption*) implique que chaque patient avait une probabilité non nulle (strictement positive) de recevoir l'un ou l'autre des traitements comparés. C'est-à-dire, qu'aucun patient n'avait une impossibilité structurelle de recevoir l'un des 2 traitements.

> *Cette problématique survient, par exemple, en cas de contre-indications d'un des deux traitements qui ne sont pas des contre-indications de l'autre.*

Cela signifie qu'il n'existe aucune caractéristique des patients (ou combinaison de caractéristiques) qui empêchent de recevoir l'un des deux traitements comparés (comme des patients ayant une contre-indication formelle à l'un des traitements, mais pas à l'autre). En d'autres termes, à caractéristiques identiques, il doit y avoir une diversité dans les traitements donnés par les médecins dans la vraie vie.

$$Pr(T = t|X = x) > 0$$

Habituellement cette hypothèse est explorée en comparant les distributions des caractéristiques entre les 2 groupes. Lorsqu'un score de propension est utilisé, l'hypothèse de positivité sera vérifiée lorsque les distributions des scores de propension des deux groupes se chevauchent (cf. section 15.2.3).

> *On parle aussi d'hypothèse de positivité conditionnelle (*conditional positivity*)*

En cas de non-respect de l'hypothèse de positivité, les estimations seront instables, exposée aux biais et les résultats difficilement généralisables à une population plus large.

Dans les comparaisons externes, plusieurs scénarios peuvent conduire à une rupture (partielle) de l'hypothèse de positivité.

Une de ces situations est constituée par un traitement standard ciblé sur un biomarqueur (et qui n'est utilisé que chez les patients biomarqueurs positifs) et un nouveau traitement qui n'a pas cette

restriction. Le nouveau traitement a été évalué dans une étude monobras incluant donc des patients biomarqueurs positifs et négatifs. Le groupe contrôle externe sera constitué uniquement de patients biomarqueurs positifs bien évidemment. Les patients biomarqueurs négatifs présents dans le groupe du nouveau traitement ont une probabilité nulle de recevoir le traitement standard et il y a rupture de la positivité. Ce non-respect de la positivité ne peut pas être solutionné par restriction de population du nouveau traitement au biomarqueur positif, car ce biomarqueur n'a pas été mesuré dans l'étude monobras (pour laquelle il était inutile). Ce point illustre l'importance de concevoir la comparaison externe au moment de la conception de l'étude monobras. Dans ce cas cela aurait permis de percevoir l'importance de mesurer ce biomarqueur dans cette étude même s'il est sans utilité pour le nouveau traitement.

Cette situation pourrait avoir une solution à la condition qu'il soit parfaitement bien démontré que le biomarqueur n'est pas un facteur pronostique, démonstration difficile à produire et qui en général n'est pas disponible. Cette condition rajoute une hypothèse forte, souvent intestable, limitant davantage cette approche pour les comparaisons de ce type.

Si malgré tout c'est le cas, et que le résultat de cette comparaison externe est acceptable, cette comparaison externe ne documentera que le bénéfice du nouveau traitement chez les patients biomarqueurs positifs, car ce n'est qu'uniquement chez ces patients que la comparaison versus traitement ce traitement standard à un sens médical. Donc dans une population différente de l'indication recherchée par le nouveau traitement. À ce niveau se pose aussi la question du choix de l'estimand, ATC (*average treatment effect among control*) correspondant plus à la question posée que l'ATT (*average treatment effect among treated*) qui est pourtant l'estimand logique des comparaisons à un groupe externe en dehors de cette situation.

Comme le nouveau traitement s'adresse aussi aux patients biomarqueurs négatifs, il va être aussi nécessaire de réaliser une autre comparaison externe versus le traitement standard actuel de ces patients. La même problématique de non-vérification de l'hypothèse de positivité se posera.

Une autre situation de non-vérification (partielle) de l'hypothèse de positivité sera la situation en miroir avec un nouveau traitement ciblé et un traitement standard non ciblé.

13.1.2 Hypothèse SUTVA

L'hypothèse SUTVA (*Stable Unit Treatment Values Assumption*) recouvre en fait deux hypothèses : celle de non-interférence et celle de cohérence.

- Non interférence

La non-interférence est l'hypothèse que le traitement de certains patients dans la population ne modifie pas le couple d'outcome potentiel $\{Y^{a=1}, Y^{a=0}\}$ des autres. Dans le domaine de l'évaluation des traitements, cette situation est rare et n'affecte principalement qu'un seul domaine celui des vaccins. En effet, dans une population, la vaccination d'un grand nombre d'individus bénéfice aussi à ceux qui ne sont pas vaccinés (immunité collective, *herd immunity*). Avec un groupe contrôle externe historique, cette hypothèse dépendra du contexte d'où est issu le groupe contrôle.

- Cohérence (*consistency*) /

L'hypothèse de cohérence (*consistency*) stipule que la variable traitement est parfaitement bien définie, c'est-à-dire qu'il n'y a pas de multiples versions du même traitement caché derrière chaque modalité de la variable traitement et qui correspondrait donc à des outcomes potentiels différents (regroupés à tort sous la même dénomination de la variable traitement). L'outcome potentiel est le même pour un même traitement. Par exemple il n'y a pas plusieurs modalités d'administration du

traitement (seul, en association à la chimiothérapie, etc.) ou s'il y a plusieurs modalités, celles-ci ont le même effet et donc l'outcome potentiel est le même.

Il convient de remarquer que le traitement dont il est question est le traitement prescrit, initialement attribué aux patients. Des adaptations de doses, pour défaut de tolérance, ou même des arrêts de traitement peuvent ensuite survenir dans le déroulé du traitement et du suivi du patient, mais cela ne conduit pas à des versions différentes du traitement. Il y a bien qu'une seule version du traitement initialement prescrite par les médecins, mais qui chez certains patients subit des modifications en cours de route du fait de la réaction du patient (ce qui est inévitable, bienvenu et qui se reproduira à l'identique dans le futur si le traitement est utilisé en pratique). Il ne faut pas comprendre que cette hypothèse de cohérence implique une analyse per protocole de l'étude (« *as treated* »), ne devant prendre en considération que des périodes exposées où le traitement administré/reçu est strictement le même (dans toutes ses modalités).

En revanche, si les médecins prescrivent différentes doses ou durées de la même molécule et que ces différentes versions sont toutes regroupées derrière la même appellation dans la variable traitement, cela représentera une rupture de l'hypothèse de cohérence.

> Cette hypothèse implique que l'issue observée (outcome) pour un individu sous le traitement qu'il reçoit est égale à son outcome potentiel[15] sous ce traitement. Cela nécessite qu'il n'existe pas de multiples versions du traitement évalué associées chacune à un outcome potentiel différent. En effet, l'issue observée chez un patient recevant apparemment le traitement A alors qu'en réalité il reçoit le traitement B ne correspondra pas à son outcome potentiel avec le traitement A.
>
> Cela peut aussi survenir en cas d'erreur de classification de l'exposition dans le groupe contrôle (cf. section 19.1.4). Par exemple dans un groupe contrôle non traité, si des patients en réalité traités sont inclus par erreur, leur issue observée ne correspondra pas à leur outcome potentiel sans traitement.
>
> Avec un groupe contrôle traité, identifier à partir d'une base de données administrative de dispensation, des patients inclus dans ce groupe (considérés traités par le traitement contrôle) pourront en réalité être non traités (la dispensation n'impliquant pas forcément la prise du traitement) et leur l'issue observée ne correspondra pas à leur outcome potentiel quand ils sont traités avec ce traitement contrôle.

Compte tenu de la finalité de ces études qui est de déterminer l'utilité médicale d'un nouveau traitement (en quoi l'utilisation par le médecin de ce nouveau traitement causera un changement dans le devenir des patients), l'outcome potentiel d'un patient donné est sa valeur avec le nouveau traitement tel qu'il a été nécessaire de l'adapter chez lui en fonction de sa tolérance et de son évolution clinique (**policy treatment estimand**). Ce n'est pas la valeur « potentielle » qui aurait été obtenu avec le traitement administré pendant toute la durée du suivi suivant le schéma posologique idéal et sans événement intercurrent (**hypothetical estimand**).

[15] Compte tenu de la finalité de ces études qui est de déterminer l'utilité médicale d'un nouveau traitement (en quoi l'utilisation par le médecin de ce nouveau traitement causera un changement dans le devenir des patients), l'outcome potentiel d'un patient donné est sa valeur avec le nouveau traitement tel qu'il a été nécessaire de l'adapter chez lui en fonction de sa tolérance et de son évolution clinique (policy treatment estimand). Ce n'est pas la valeur « potentielle » qui aurait été obtenu avec le traitement administré pendant toute la durée du suivi suivant le schéma posologique idéal et sans événement intercurrent (hypothetical estimand).

13.1.3 Échangeabilité conditionnelle

L'hypothèse d'échangeabilité conditionnelle signifie, qu'après ajustement, échanger les traitements entre les deux groupes conduits aux mêmes résultats (au niveau du groupe) pour chacun des traitements. C'est-à-dire, si les patients qui ont eu A avaient eu B et vice versa, les résultats obtenus pour le groupe A et le groupe B seraient les mêmes.

La question de l'échangeabilité conditionnelle correspond à la problématique du biais de confusion

Cela nécessite que le résultat d'un groupe (fréquence ou moyenne du critère de jugement) ne dépende que du traitement et d'aucune variable ; c'est-à-dire que l'ajustement prend bien en compte tous les facteurs influençant le critère de jugement. Dans un ajustement par appariement, cela signifie que les deux groupes sont identiques sur toutes les variables influençant le critère de jugement indépendamment du traitement. C'est pour cette raison que cette hypothèse est aussi appelée hypothèse NUC (*no unmeasured confounder*). Il n'existe alors pas de biais de confusion résiduel.

> *Il existe d'autres vocables pour exprimer cette hypothèse : No unmeasured confounding, no selection on observables, no omitted variables, exogeneity, conditionally exchangeable.*

En d'autres termes, cela signifie que l'outcome potentiel avec le traitement A des patients est le même aussi bien pour les patients ayant reçus A dans la vraie que pour ceux qui ont reçu B (et de façon identique pour l'outcome potentiel avec B). En effet s'il existe un déséquilibre entre les groupes sur des variables influençant le critère de jugement, l'outcome potentiel avec le traitement A n'est pas identique entre les 2 groupes, car l'outcome potentiel dépend à la fois du traitement et des autres variables influençant le critère de jugement.

La prise en compte des facteurs de confusion dans l'analyse a pour objectif d'assurer l'échangeabilité conditionnelle

13.2 Petite introduction à l'inférence causale

La mise en évidence directe de l'effet causal d'un facteur donné nécessite de pouvoir comparer l'état d'un système en la présence et en l'absence de ce facteur. L'effet de ce facteur sera donné par la différence entre ces 2 états. Ainsi la détermination d'un effet causal nécessite de pouvoir observer le même système **avec et sans** le facteur étudié.

Soit *Y* l'état du système et *a* le facteur étudié qui a 2 modalités *a=0* ou *a=1*.

L'effet causal produit par le facteur *a* sera déduite de :

$$Y^{a=1} - Y^{a=0}$$

Où $Y^{a=1}$ désigne l'état du système quand *a=1* et $Y^{a=0}$ son état quand *a=0*.

Si l'on veut transposer cette formalisation de la causalité aux traitements, Y correspond au critère de jugement (l'outcome) et *a* au traitement avec *a=0* l'absence de traitement et *a=1* la prise du traitement.

Pour un patient *i*, l'effet causal du traitement est donc la différence entre la valeur du critère de jugement quand le patient i est traité et sa valeur quand il n'est pas traité :

$$\delta_i = Y_i^{a=1} - Y_i^{a=0}$$

Mais contrairement à beaucoup de systèmes physiques, il va être impossible d'observer le même patient *i*, au même moment de sa maladie, à la fois avec traitement et sans traitement. Il est donc impossible d'obtenir la valeur du critère de jugement Y dans ces 2 états, car soit le patient est traité soit il n'est pas traité. Il est donc impossible d'observer l'effet du traitement δ_i chez un patient.

$Y_i^{a=1}$ et $Y_i^{a=0}$ sont appelés outcomes potentiels (*potential outcome*), le terme potentiel signifiant que ces deux valeurs existent potentiellement (ont une vraie signification) mais qu'ils ne sont pas accessibles toutes les deux pour le même patient. En effet, soit le patient est traité, soit il n'est pas traité. Un seul de ces 2 outcomes potentiels est donc observable par les patients (mais il est facile de convenir que si un patient est traité, le critère de jugement aurait eu une autre valeur sans traitement et vice versa).

Pour tenir compte de la variabilité inter sujets, l'estimation de l'effet d'un traitement doit être envisagée au niveau populationnel, statistique. Cela ne change rien à la notion de causalité seulement celle-ci est une causalité stochastique, appréhendable qu'à un niveau populationnel sous la forme d'un effet traitement moyen (ce qui est le cadre classique de l'évaluation des traitements).

Y est alors une variable aléatoire, c'est-à-dire une variable qui est susceptible de prendre une valeur différente pour chaque patient (sans que cela ne soit prévisible d'où l'assimilation à un phénomène aléatoire) mais dont la distribution des valeurs peut parfaitement bien se caractériser au niveau populationnel (par exemple par une moyenne et un écart type pour une variable aléatoire distribué normalement).

Ainsi l'effet traitement moyen (*average treatment effect*, ATE) sera alors

$$ATE = E\left(Y_i^{a=1} - Y_i^{a=0}\right)$$

Où $E(\quad)$ désigne l'espérance mathématique (la moyenne pour une variable continue ou la fréquence pour une variable binaire). Ainsi l'effet traitement moyen populationnel sera la moyenne de la différence des outcome potentiel de chaque patient avec et sans traitement.

Mais cette valeur est inaccessible, car les outcomes potentiels ne sont pas tous les 2 mesurables simultanément chez un même patient. Une seule mesure est possible par patient, *Y*, qui est soit la valeur de l'outcome avec le traitement *a=1* soit sans le traitement *a=0*. Le tableau (ci-dessous) illustre la différence entre l'outcome mesuré et les outcomes potentiels dans une situation hypothétique où l'on pourrait connaitre les valeurs des 2 outcomes potentiels de 5 patients. T_i désigne le traitement réellement appliqué au patient *i* et Y_i la valeur de son critère de jugement. Quand le traitement du patient est *T=1*, *Y* prend la valeur de l'outcome potentiel avec traitement $Y^{a=1}$ et quand *T=0, Y* est égal à l'outcome potentiel sans traitement.

	$Y_i^{a=1}$	$Y_i^{a=0}$	Y_i	T_i
Patient 1	210	220	220	0

Patient 2	168	178	168	1
Patient 3	196	205	196	1
Patient 4	187	197	197	0
Patient 5	182	190	190	0

En pratique on peut donc mesure avec une série de patients traités ou non traités

$E(Y|T = 1)$ et $E(Y|T = 0)$, c'est-à-dire la moyenne des valeurs de l'outcome *Y* chez des patients traités (T=1) (la notation $(Y|T = 1)$ signifie valeurs de la variable *Y* quand la variable *T* est égale à 1).

$E(Y|T = 1)$ et $E(Y|T = 0)$ sont donc les moyennes des valeurs de deux groupes de patients l'un traité et l'autre non traité. La comparaison de ces 2 valeurs est de l'ordre de la recherche de l'association statistique, mais pas de la causalité var ne correspond pas à l'expression de l'effet causal, car E($Y \mid T$=1) n'est pas forcément égale à $\mathrm{E}(\mathrm{Y}^{\mathrm{a}=1})$ et $\mathrm{E}(\mathrm{Y}|\mathrm{T} = 0)$ à $\mathrm{E}(\mathrm{Y}^{\mathrm{a}=0})$.

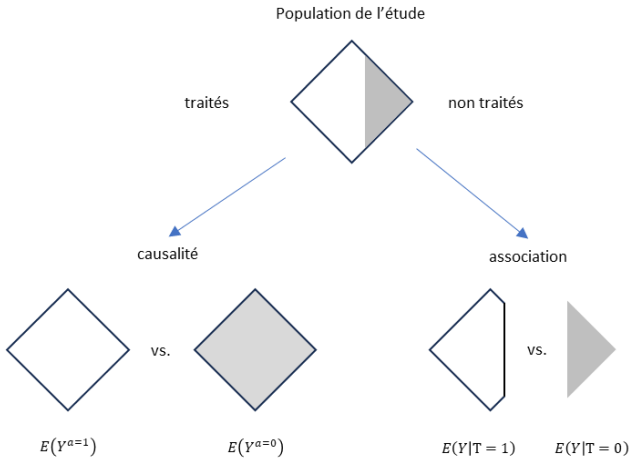


Figure 3 – Représentation graphique de la différence entre association et causalité (reproduit de réf. [112]).

Ainsi on se retrouve dans une situation où l'on peut seulement, à partir des données d'une étude, connaitre $E(Y_i|T_i = 1) - E(Y_i|T_i = 0)$ alors qu'il faudrait pouvoir connaitre $E\left(Y_i^{a=1} - Y_i^{a=0}\right)$ pour pouvoir conclure à la causalité. Il est possible de passer de la première expression à la dernière en faisant un certain nombre d'hypothèses. Il est ainsi possible de montrer que l'association statistique a valeur d'effet causal quand ces hypothèses sont vérifiées.

L'association statistique observée correspond à

$$E(Y_i|T_i = 1) - E(Y_i|T_i = 0)$$

Qui peut être réécrite en faisant l'hypothèse de positivité (1 ci-dessous) et l'hypothèse de cohérence (*consistency*) (2 ci-dessous)

$$= E\left(Y_i^{a=1}|T_i = 1\right) - E\left(Y_i^{a=0}|T_i = 0\right)$$

En faisant l'hypothèse d'échangeabilité (3 ci-dessous), cette expression peut se réécrire

$$= E\left(Y_i^{a=1}\right) - E\left(Y_i^{a=0}\right)$$

Et comme l'opérateur $E(\cdot)$ est un opérateur linéaire

$$= E\left(Y_i^{a=1} - Y_i^{a=0}\right)$$

qui est la définition dans le cadre des outcome potentiels de l'effet causal ■

(1) Si l'hypothèse de positivité est vérifiée alors $\Pr(\mathrm{T_i} = \mathrm{t}) > 0$, cela garanti que $E(Y_i|T_i = t)$ est identifiable dans les deux groupes, sans positivité un de ces termes conditionnels peut ne pas exister. La quantité $E(Y_i|T_i = t)$ n'est définie que si $Pt(T_i = t) > 0\ \forall t$. La positivité assure que le contrefait d'un traitement sera identifiable (si aucun patient n'a reçu un traitement, on ne peut donc pas apprendre quoi que ce soit sur leur outcome potentielle sous traitement, même par extrapolation).

(2) Si l'hypothèse de « Stable Unit Treatment Value Assumption (SUTVA)" est vérifiée alors $\mathrm{Y_i^{a=t}} = \mathrm{Y_i}$ si $\mathrm{T_i} = \mathrm{t}$

Cela nécessite deux conditions :

- Cohérence (consistency) : Il n'y a pas de variante dans les traitements qui pourrait conduire à diffèrent outcomes potentiels, $\mathrm{T_i} = 1$ signifie la même chose pour toutes les unités statistiques
- Non interférence : Aucune influence du traitement des autres unités statistiques

(3) Si l'hypothèse d'échangeabilité est vérifiée, les outcomes potentiels des patients sont les mêmes qu'ils aient reçus un traitement ou l'autre :

$$\mathrm{E(Y_i^{a=1}) = E(Y_i^{a=1}|T_i = 1) = E(Y_i^{a=1}|T_i = 0)}$$

et

$$\mathrm{E(Y_i^{a=0}) = E(Y_i^{a=0}\big|T_i = 1) = E(Y_i^{a=0}\big|T_i = 0)}$$

Cependant si les traitements sont donnés en fonction du pronostic des patients :

$$\mathrm{E(Y_i^{a=0}\big|T_i = 1) \neq E(Y_i^{a=0}\big|T_i = 0)}$$

(Des patients de pronostic de base différents ne peuvent pas avoir la même valeur de l'outcome potentiel sans traitement vu que cet outcome potentiel sans traitement correspond au pronostic de base).

L'effet traitement identifié en inférence causale est un effet moyen sur une population (espérance mathématique = moyenne des « effets individuels » de chaque patients *i*. Cet effet moyen (*average treatment effect*) dépend donc de la population cible de l'inférence causale. Cette propriété conduit à la définition de différents effets traitements (ATE, ATT, etc.) (cf. section 13.4)

La meilleure façon s'assurer le respect des hypothèses de l'inférence causale est de réaliser un essai randomisé. La randomisation (comme toute autre méthode de fixation vraiment arbitraire du traitement) assure la positivité et l'échangeabilité. L'hypothèse de cohérence est assurée par la réalisation en double aveugle et en ITT de l'essai.

13.3 Association n'est pas causalité

L'association observée dans une étude observationnelle simple ne peut pas permettre de conclure à la causalité, car le design de l'étude n'empêche pas l'existence d'interprétations alternatives, ce que fait le design de l'essai randomisé. En effet de nombreux mécanismes peuvent conduire à des associations statistiques en dehors de toute causalité et ainsi révéler des associations non causales[16].

Les études qui permettent de conclure à la causalité bloquent la possibilité d'interprétation alternative devant les associations qu'elles mettent en évidence. Principalement grâce à la randomisation, au double aveugle, à l'analyse en intention de traiter dans l'essai randomisé correctement conçu et réalisé ou grâce à une approche formalisée d'inférence causale et de contrôle/correction des biais pour les études observationnelles correctement conçues et disposants de données appropriées au contrôle de tous les biais.

Association statistique versus inférence causale

En d'autres termes, une association ne permet pas de conclure à la causalité, car elle ne fait que montrer que l'état des patients est différent entre les patients recevant le traitement et ceux ne le recevant pas. La causalité nécessite de montrer en quoi l'état des patients est changé par le traitement par rapport à ce qu'il aurait été sans ce traitement.

La recherche de la causalité nécessite donc de montrer, pour les patients traités, qu'il y a une différence entre l'état du patient avec ce traitement et celui-ci qu'il aurait eu s'il n'avait pas eu le traitement. Pour les patients non traités, cela revient à montrer qu'il existe aussi une différence entre l'état du patient non traité et celui qu'il aurait eu s'il avait été traité.

L'état du patient qu'aurait eu le patient avec l'autre modalité de traitement que celle qu'il reçoit est appelé le contrefait.

L'inférence causale offre un cadre pour formuler mathématiquement la recherche de la réponse à ces questions : ce que les choses auraient dû être si … (prédiction du contrefait) à partir de ce que les choses sont.

13.3.1 DAG générique des comparaisons externes

Le réseau de causalité générique des comparaisons externes peut être représenté de la façon suivante [18].

[16] https://catalogofbias.org/2019/03/05/association-or-causation-how-do-we-ever-know/

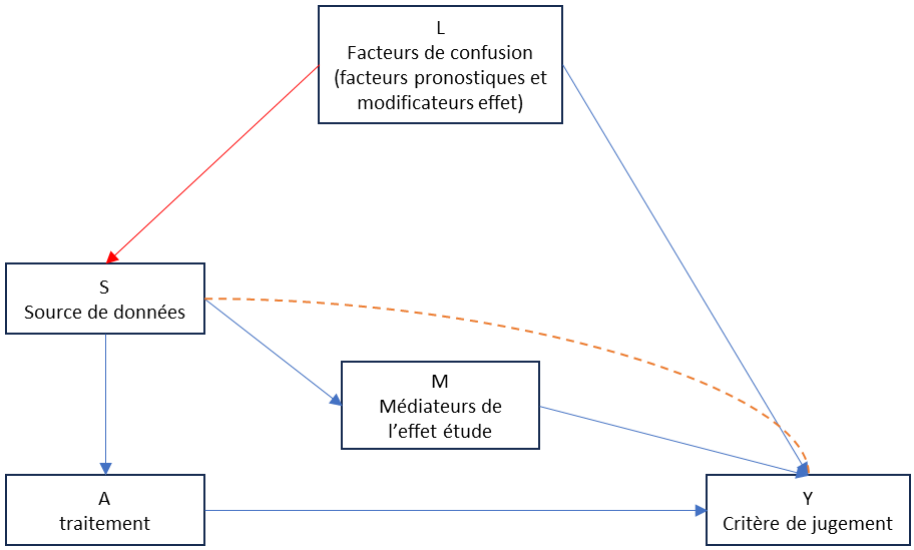


Figure 4 – DAG générique des comparaisons externes à un bras contrôle externe.
La source de données (S) désigne les 2 groupes constituant cette comparaison (étude monobras d'un côté et groupe contrôle externe de l'autre)

Apparait aussi l'existence d'un effet étude direct sur le critère de jugement (ligne pointillée orange). La valeur du critère de jugement Y dépend, indépendamment du traitement et des patients inclus, de l'étude dans laquelle il est mesuré. Les médiateurs de cet effet sont facilement imaginables. [54] [16] : différence de définition, mesure du critère de jugement entre les 2 études, différences de suivi, de contexte de soins, de prise en charge des patients, de paramètres populationnels influençant le niveau de santé (comme le niveau de vie, la structure en classes socio-économiques, socio-éducatives des populations sources des deux études, etc.).

Comme il est impossible d'ajuster sur l'étude, le seul moyen permettant de supprimer cet effet-étude direct serait de connaitre ses médiateurs et de les prendre en considération dans le design et l'analyse. Mais pour beaucoup de ces variables, l'étude va être une variable instrumentale du traitement impossible à prendre en compte par ajustement (par exemple PFS dans l'étude monobras et rwPFS dans le groupe contrôle externe). Ce point représente un véritable défi qui sans solution satisfaisante obère la possibilité d'estimer correctement l'effet des traitements par des comparaisons externes.

13.4 Effet causal, estimand causal, cible de l'inférence

13.4.1 Effet traitement moyen (*average treatment effect)*

L'estimand causal naturel dans la comparaison externe est l'ATT (*average treatment effect among the treated*). La méthode statistique utilisée devra donc nécessairement permettre d'estimer cet estimand (appariement des patients du groupe traité sur les covariables ou sur le score de propension, pondération avec les poids adaptée, g-computation adaptée)

L'estimand causal naturel dans la comparaison externe est l'ATT (average treatment effet among the treated).

L'ATT consiste à chercher l'effet du traitement dans une population similaire à celle qui a été incluse dans l'essai monobras. Cet estimand correspond donc à celui d'un essai randomisé incluant les mêmes patients que l'essai monobras (en effet dans un essai randomisé correctement conçu et réalisé, l'ATE = ATT = ATC).

Population cible de l'inférence causale

Le concept d'effet causal (effet traitement moyen) n'est pas unique. Plusieurs effets causals peuvent être dérivés en fonction de la cible de l'inférence causale (population ciblée par l'inférence causale).

ATE (Average Treatment Effect) : L'ATE mesure l'effet moyen du traitement dans l'ensemble de la population cible, c'est-à-dire la différence moyenne attendue entre les résultats si tous les individus recevaient le traitement par rapport à s'ils ne le recevaient pas. C'est l'estimand le plus général, pertinent lorsqu'on souhaite connaître l'impact global d'une intervention.

ATT (Average Treatment Effect among the Treated) : L'ATT s'intéresse uniquement à la population qui a effectivement reçu le traitement. Il estime l'effet moyen du traitement chez ceux qui ont été traités, ce qui est particulièrement pertinent dans le cadre d'une comparaison externe où l'on souhaite savoir ce que le traitement apporte à ceux à qui il a été assigné.

ATC (Average Treatment Effect among the Controls) : L'ATC estime l'effet moyen que le traitement aurait eu chez les individus du groupe contrôle s'ils l'avaient reçu. Cet estimand est utile pour évaluer l'impact potentiel d'un traitement dans une population qui n'y a pas encore eu accès.

ATO (Average Treatment Effect among the Overlap population) : L'ATO cible l'effet moyen du traitement dans la sous-population présentant le plus grand recouvrement de caractéristiques entre les groupes traité et contrôle. Il s'agit d'un estimand qui vise à maximiser la comparabilité entre les groupes et à limiter les biais liés au manque d'équilibre des covariables, mais dont la population d'inférence est plus compliquée à imaginer.

Chacun de ces effets causaux répond à une question d'intérêt différente et le choix dépend du contexte de l'étude et de l'objectif causal recherché.

13.4.2 Analyse en intention de traiter (*as started*) / analyse per protocole (*as treated*)

Pour la même raison l'estimand devra s'intéresser à l'effet d'être assigné à un traitement (et non pas de le recevoir), c'est-à-dire que l'estimand devra être dans la logique de l'analyse en intention de traiter (ITT, *as started*). Il s'agit d'évaluer l'effet causal de décider de recourir, et donc d'initier, le traitement évalué par rapport à une stratégie de prise en charge reposant sur l'initiation du traitement contrôlé. Ainsi les patients du groupe contrôle seront des sujets pour lesquels une décision de recourir au traitement contrôlé aura été prise dans la vraie vie et qui seront suivis pour enregistrer la survenue du (ou des) critère de jugement indépendamment de la prise effective du traitement et la durée de celui-ci. En d'autres termes, le suivi doit se prolonger après l'arrêt du traitement dans le groupe contrôle jusqu'à la date de point virtuelle choisie de la même manière que ce qui se passe dans un essai randomisé suivant le principal de l'analyse en intention de traiter (ou de l'estimand *policy treatment*).

Pour permettre une analyse en intention de traiter, les patients du groupe contrôle doivent être inclus sur la base d'une intention d'initier le traitement contrôle chez eux et suivi après l'arrêt du traitement.

En pratique il sera très difficile et même impossible d'identifier les patients dans la source de données suivant ces termes, car l'information sur l'intention d'initier un traitement n'est pas enregistrée dans les sources de données. Cela nécessiterait entre autres d'avoir la prescription du médecin, ou, pour les pathologies le nécessitant, le compte rendu de la réunion pluridisciplinaire décidant la modalité de prise en charge des patients. Ces informations sont rarement colligées[17] et les patients sont principalement identifiés par la notion de dispensation ou de prise plus ou moins effective du traitement [116].

Au-delà de la question de la définition de l'estimand causal, la disponibilité de l'information de l'intention d'initier un traitement est importante pour la synchronisation des débuts de suivi (t0) nécessaire à la prévention des biais de sélection temporels. En effet, ne pas pouvoir inclure dans le groupe contrôle des patients pour lesquels il y a eu intention de recourir au traitement contrôle, mais qui ne l'on pas reçu est susceptible d'introduire un biais, car dans le groupe traité (de nature expérimentale) les patients inclus dans l'étude, mais qui n'ont pas reçu le traitement étudié sont pris en compte ainsi que leurs événements. L'approche de l'émulation de l'essai cible permet de comprendre et prévenir les biais dépendants de cette problématique (cf. sections 22 et 17).

> *Un estimand de l'effet causal du traitement lorsqu'il est effectivement pris peut aussi être utilisé, mais celui-ci correspond en termes d'émulation d'un essai cible à une analyse per protocole et n'apporte donc pas l'information de base nécessaire à l'évaluation d'un nouveau traitement et à la décision.*

Petite présentation rapide des différentes stratégies de gestion des événements intercurrents

Il existe plusieurs façons de concevoir que qu'est l'effet d'un traitement en particulier en fonction de la façon de gérer les événements intercurrents (arrêts du traitement pour effet indésirable, arrêts du traitement sans raison connue, recours à un traitement de secours ou concomitant, etc.). Couramment on distingue 3 stratégies de gestion différentes (mais il en existe d'autres). Ce n'est pas 3 façons de mesurer la même chose mais bien trois concepts différents, trois façons de concevoir ce que peut apporter un traitement.

Policy treatment (correspond à l'ITT) : Quel est l'effet de prescrire le traitement A par rapport à la prescription du traitement B, indépendamment de ce qui se passe ultérieurement ? Les changements, interruptions et traitements concomitants additionnels sont considérés comme attendus et habituels dans la prise en charge des patients en vie réelle. C'est souvent la question la plus pertinente pour la prise de décision « que se passe-t-il si l'on recommande d'utiliser A en pratique ? », à condition que le temps zéro et les critères d'éligibilité soient clairement définis.

[17] Les dossiers médicaux permettent de retrouver les prescriptions initiales des médecins. Cette information peut être présente dans les registres. Cependant elle n'est pas accessible dans les bases de données administratives ou de dispensation.

Hypothetical (correspond au per protocol) : Quel serait l'effet si les patients initiaient puis adhéraient à A à la place de B comme prévu pendant la durée spécifiée ? Cette option est souvent séduisante du point de vue du traitement. Elle requiert des hypothèses fortes et une définition claire de ce que signifie « adhérer comme prévu » et ce qui est considéré comme un écart au protocole (durées, modifications de dose, traitements additionnels autorisés, etc.).

While on treatment (suivi jusqu'à arrêt/changement) : Quel est l'effet de A par rapport à B pendant la période où les patients restent sous leur traitement initial, le suivi prenant fin à l'arrêt ou au changement de traitement ? Cette approche peut être utile, mais elle n'a une interprétation causale claire que si l'on prend en compte le fait que l'arrêt ou le changement de traitement est souvent lié au pronostic, à la tolérance, à la réponse précoce, au jugement du clinicien. En d'autres termes, la censure est probablement informative.

14 Le biais de confusion

Pour permettre d'isoler, par comparaison, l'effet spécifique du traitement, le groupe contrôle ne doit différer du groupe traité que par le traitement appliqué. L'hypothèse d'échangeabilité de l'inférence causale doit être vérifiée. En d'autres termes les patients inclus dans le groupe contrôle ne doivent pas être différents des patients du groupe traité au niveau de leur pronostic, de leur risque de base de faire l'événement critère de jugement. Autrement, une différence au niveau du critère de jugement pourrait ne pas refléter l'effet du traitement, mais simplement la différence de risques de base entre les 2 groupes. Cette différence au niveau du critère de jugement due à la différence de risque de patients des patients entre les 2 groupes pourrait ainsi être confondue avec l'effet du traitement, d'où le nom de biais de confusion.

Le biais de confusion représente l'une des limites importantes des études observationnelles (pas seulement des comparaisons à un groupe contrôle externe), leur empêchant souvent de conclure à la causalité.

Ce biais va pouvoir être en partie supprimé par l'analyse (et aussi par la construction de l'étude, mais de façon limitée).

De nombreuses techniques d'analyse sont disponibles pour prendre en compte les facteurs de confusion d'une étude et corriger le résultat du biais de confusion qu'ils induisent : analyse ajustée à l'aide d'un modèle de régression multivariable, appariement à l'aide d'un score de propension, pondération, g-computation, méthodes double-robustes (comme le TMLE ou l'AIPW), etc. D'une façon générale, on parle d'analyse conditionnellement au facteur de confusion, d'analyse contrôlée et par abus de langage d'analyse « ajustée » (stricto sensu les termes ajusté, ajustement ne concernent que les techniques de régression).

La suppression du biais de confusion par l'analyse nécessite impérativement un modèle statistique bien spécifié, c'est-à-dire contenant la totalité des facteurs de confusion et dont la forme mathématique capte correctement les relations entre ces facteurs et le traitement et le critère de jugement (il ne contient pas de variables superflues et les hypothèses statistiques sont vérifiées). Dans ce cas l'hypothèse d'échangeabilité de l'inférence causale sera vérifiée conditionnellement à l'analyse effectuée, on parle alors d'échangeabilité conditionnelle.

Juger si l'analyse conditionnelle d'une étude particulière a conduit à des résultats corrigés du biais de confusion est une question centrale

> *La question de la confusion dans les comparaisons à un groupe contrôle externe est un peu différente de celle des études observationnelles classiques où les 2 groupes proviennent de la même source de données. Dans la comparaison externe il existe un « effet étude » (cf. section 13.3.1) qui ne dépend pas entièrement des facteurs pronostiques et des modificateurs d'effets, mais unique de l'étude (partie expérimentale et groupe contrôle externe), rendant impossible sa correction par analyse.*

14.1 Particularité des facteurs de confusion dans les comparaisons externes

Classiquement les facteurs de confusion sont des facteurs qui causent à la fois le critère de jugement et le traitement reçu par les patients. Il s'agit donc, dans les études observationnelles classiques où les

patients traités et les patients contrôles sont issus de la même source de données, des facteurs pronostiques du critère de jugement qui ont aussi été pris en considération par les médecins pour choisir le traitement des patients.

Dans les comparaisons à un groupe contrôle externe, ce mécanisme de choix du traitement n'existe pas vraiment. Personnes ne « choisi » qu'un patient soit dans le groupe contrôle plutôt que dans l'étude expérimentale (monobras par exemple). Mais il est tout à fait possible que les patients de ces 2 groupes soient différents, car recrutés dans deux populations sources différentes. Même si le mécanisme est différent, le résultat sera le même, avec des différences sur des variables pronostiques entre les deux groupes pouvant entrainer une différence sur le critère de jugement qui pourrait être confondue avec l'effet du traitement. [117].

> *Dans les études observationnelles classiques, il n'y a qu'un seul échantillonnage de patients traités et contrôles provenant de la même source de données (Figure 5). Les études de comparaison externe reposent sur deux échantillonnages distincts : les patients traités sont issus d'un premier échantillonnage dans une certaine population source et les patients contrôles proviennent d'un autre échantillonnage dans une population différentes*

Pour cette raison, dans les comparaisons externes, tous les facteurs pronostiques peuvent être facteurs de confusion, car leurs distributions peuvent être différentes entre les deux « populations sources », y compris les facteurs non connus : facteurs génétiques, environnementaux, socioéconomiques et qui ne pourront jamais être pris en compte. Pour ces raisons la possibilité d'obtenir l'échangeabilité conditionnelle est encore moins probable que dans les études observationnelles classiques.

> *Il y a endogénéité, car il existera une corrélation entre des facteurs pronostiques inconnus ou non mesurés/mesurables et le traitement (c'est-à-dire l'étude). On parle d'endogénéité lorsqu'une variable explicative (exposition, traitement, facteur de risque) est corrélée avec l'erreur du modèle, c'est-à-dire avec des facteurs non observés qui influencent aussi le résultat étudié.*

Bien qu'amenant à la même situation, cette différence de mécanisme dans le phénomène de confusion va impacter la façon identifier les facteurs de confusion potentielle. En l'absence de processus de choix raisonné sous-jacent du traitement, il devient difficile d'identifier parmi les facteurs pronostiques des critères de jugement, ceux qui pourraient ne pas être distribués de la même manière entre les 2 groupes. De ce fait tous les facteurs pronostiques des critères de jugements sont des facteurs de confusion potentiels, car rien ne permet de savoir sur quels plans les deux populations sources diffèrent. En effet ces deux populations sources ne sont pas accessibles et il est donc impossible de les comparer ou de disposer de connaissances externes sur les différences de distributions des facteurs pronostiques entre ces deux populations.

La construction d'un graphe de causalité (par un DAG par exemple) permettra d'aider à la compréhension des relations entre facteurs pronostiques et donc à la détermination du « *minimally sufficient adjustment sets*».

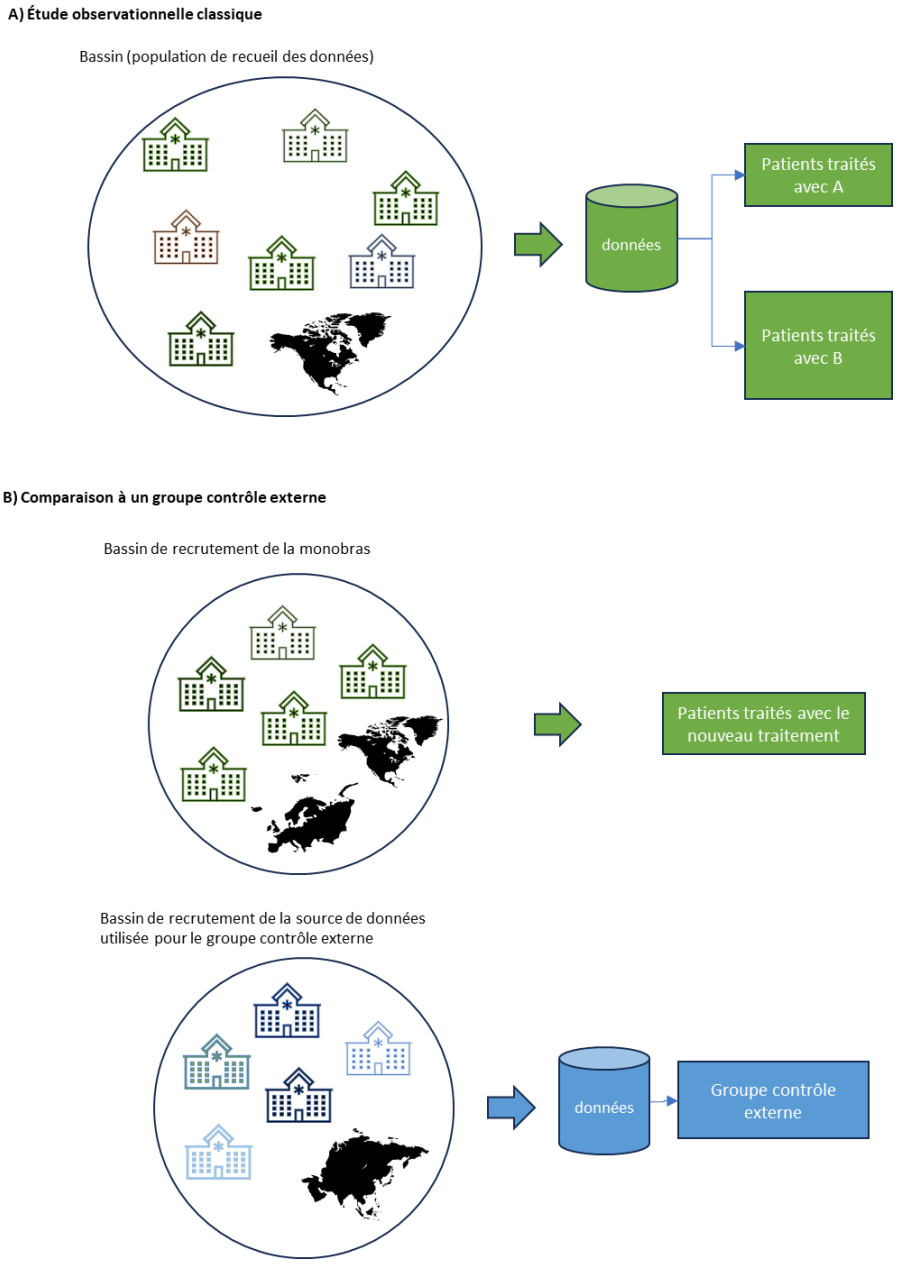


Figure 5 – Différence dans le processus d'échantillonnage d'une étude observationnelle classique (A) et d'une comparaison à un groupe contrôle externe (B). dans une études observationnelles classiques (A) l'échantillonnage s'effectue dans un seul « bassin de recrutement », constitué de différents hôpitaux ou médecins, mais où les deux traitements comparés sont susceptibles d'être utilisé. Dans les comparaisons externe, l'échantillonnage s'effectue dans deux « bassins de recrutement » différents, où les hôpitaux et les médecins peuvent être systématiquement différents par traitement (tous comme les aires géographiques).

14.1.1 Modificateurs de l'effet du traitement

Les modificateurs d'effet peuvent aussi être facteur de confusion à côté des facteurs pronostiques. Ce point parfaitement connu pour les comparaisons indirectes non ancrées type MAIC [118] et plus rarement mentionné lorsqu'il s'agit de comparaison à un groupe contrôle externe.

Se pose aussi la question des modificateurs de l'effet des facteurs pronostiques quand le groupe contrôle provient d'une population source très différente (par exemple géographiquement) de l'étude expérimentale, soulevant des questionnements sur la transportabilité.

Une même variable peut ne pas avoir la même valeur pronostique dans ces deux populations sources. Dans ce cas sa prise en compte dans l'ajustement ne permettra pas de corriger du biais de confusion induit par une distribution différente de cette variable entre les 2 groupes comparés.

14.2 La détermination des facteurs de confusion

La prise en compte de tous les facteurs de confusion (le contrôle sur tous les facteurs de confusion) est indispensable pour corriger les résultats du biais de confusion. L'identification de ces facteurs de confusion est donc un point fondamental de validité des études observationnelles et à pour but de pouvoir obtenir l'hypothèse d'échangeabilité conditionnelle de l'inférence causale, ou hypothèse NUC *no unobserved confusion*).

La prise en compte de tous les facteurs de confusion est indispensable pour vérifier l'hypothèse fondamentale d'échangeabilité conditionnelle de l'inférence causale

Or la liste de ces facteurs de confusion ne s'impose pas d'elle-même. L'établir nécessite un travail à part entière suivant un processus formalisé garantissant que tous les facteurs de confusion affectant une étude ont été en mesure d'être identifiés correctement.

Les facteurs de confusion doivent être déterminés en se basant sur les connaissances externes concernant les facteurs influençant le critère de jugement et non pas sur ce qui pourrait transparaitre de l'analyse des données de l'étude. Même si cette dernière option a été pratiquée dans le passé en épidémiologie, il a été montré qu'elle est contreproductive (introduit des biais) et insuffisante pour identifier les réels facteurs de confusion.

Ces techniques sont encore proposées parfois et des méthodes plus sophistiquées de sélection automatique des variables d'ajustement sont aussi recherchées [119].

Les déterminants des critères de jugements (facteurs pronostiques, facteurs de risques) peuvent varier d'un critère à l'autre. Par conséquent, les variables d'ajustement doivent être définies spécifiquement pour chaque critère de jugement et ne sauraient être considérées comme uniques à l'échelle globale d'une étude.

> *Par exemple les déterminants de la mortalité totale en oncologie (overall survival) peuvent inclure les facteurs de risques des décès d'autres causes que le cancer (pour les cancers à relativement bon pronostic comme le cancer du sein précoce) qui ne seront pas des déterminants des critères plus spécifiques du cancer comme la réponse tumorale ou la PFS.*

14.2.1 Réseaux de causalité

Des techniques ont été développées pour formaliser la recherche des facteurs de confusion dans les études observationnelles classiques comme les réseaux de causalité (Figure 6) [120] [121] [122]. Il s'agit d'approches graphiques (intégrant éventuellement des relations quantitatives) qui représentent les interrelations existantes entre les variables, le traitement et le critère de jugement, dans le but

d'aider à l'identification des variables à prendre en compte dans l'ajustement (*sufficient set of covariates*). Ces réseaux ou graphiques de causalité sont souvent représentés sous la forme de graphique acyclique dirigé (*directed acyclic graph DAG*) [123] [124] [125] ou de SWIGs [126].

Ces graphiques sont établis à partir des relations entre variables rapportées dans la littérature (principalement pour les facteurs de risque/facteur pronostique du critère de jugement) et de la connaissance concernant les critères de choix des traitements qui auraient prévalu dans la pratique captée par les données. L'élaboration de ces graphiques nécessite un travail collaboratif entre des épidémiologistes, statisticien, des cliniciens et des personnes connaissant bien la base de données.

La réalisation des graphiques de causalité (DAG) pour déterminer la liste des facteurs de confusion est maintenant mentionnée dans la plupart des recommandations abouties concernant les études observationnelles [48] [64].

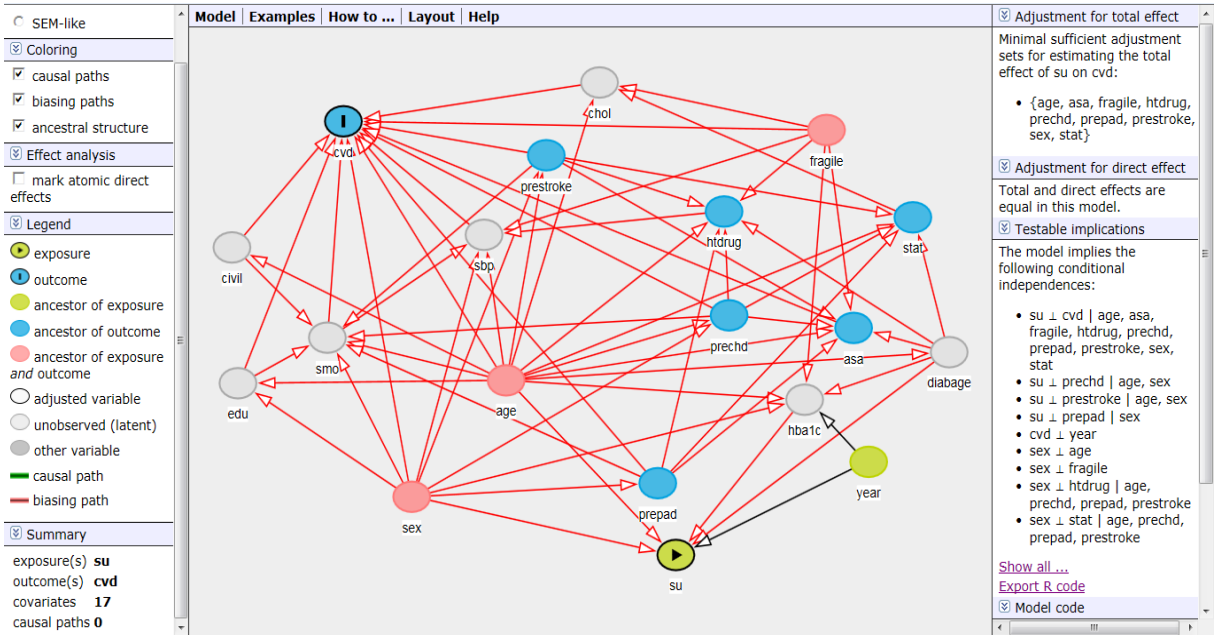


Figure 6 – Exemple de graphique de causalité établi par un DAG utilisé dans une étude observationnelle pour déterminer la liste des facteurs de confusion et le « sufficient set of covariates » [127].

> *Il convient de noter que ces graphiques type DAG sont des outils pour déterminer les facteurs de confusion et non pas une simple représentation graphique des variables d'ajustement qui auraient été choisis de manière plus ou moins arbitraire.*

Compte tenu de l'aspect particulier des facteurs de confusion dans les comparaisons externes où finalement il n'y a pas eu de choix par les médecins dans la vraie vie entre les traitements comparés, le diagramme de causalité est un peu particulier. Les facteurs associés avec le critère de jugement seront facteur de confusion lorsqu'ils sont aussi associés avec la population dans laquelle a été échantillonné le groupe traité qui est différent de la population de laquelle provient le groupe contrôle externe. En effet les patients du groupe traités proviennent d'une certaine population dans laquelle a été échantillonnée l'étude monobras et les patients contrôles proviennent d'une autre population qui a été captée dans la source de données utilisée pour faire le groupe contrôle externe. Il est possible

que les facteurs associés avec le critère de jugement ne soient pas distribués de la même manière dans ces deux populations d'origine des groupes (échantillons).

14.2.2 Revue systématique des facteurs pronostiques

L'identification de tous les facteurs influençant les critères de jugements de l'étude est l'étape initiale fondamentale de l'identification des facteurs de confusion potentiels d'une étude de comparaison externe. Ne seront ensuite retenus, parmi tous ces facteurs à valeur pronostiques, que ceux qui sont potentiellement différents entre les 2 groupes (un facteur de confusion est un facteur associé à la fois avec le traitement et le critère de jugement).

L'identification de tous les facteurs pronostiques, facteurs de risques, connus, sera effectuée à l'aide d'une revue systématique afin d'atteindre cette exhaustivité et de ne pas sélectionner de manière arbitraire ces facteurs.

La revue systématique nécessaire pour cette recherche des facteurs pronostiques est une revue systématique d'un type particulier, différent de celui des revues systématiques couramment utilisées dans le domaine de l'HTA et qui se focalisent sur les études d'efficacité des traitements (conduites pour la réalisation d'une méta-analyse en réseau par exemple).

La méthodologie des revues systématiques des études de facteurs pronostiques est maintenant bien connue et codifiée, en particulier sous l'impulsion d'un groupe Cochrane dédié (The Cochrane Prognosis Methods Group) [128] [129] [130]. différents outils ont été développés pour leur réalisation : QUIPS pour l'évaluation spécifique du risque de biais de ces études [131], CHARMS-PF pour l'extraction des informations à rapporter dans la revue systématique.

Il existe aussi des formalismes spécifiques pour formuler la question de recherche comme le PICOTS : Population, Index prognostic factor, Comparator prognostic factors, Outcome, Timing, Setting [130].

Les revues systématiques proposées pour identifier les facteurs de risque d'une comparaison externe pourront être évaluées de manière critique avec l'outil AMSTAR-PF afin de déterminer son acceptabilité [132].

Les déterminants (facteurs pronostiques) des critères de jugement peuvent être différents suivant les critères (élément O Oucome du PICOTS). La revue systématique devra donc couvrir différentes questions de recherche (PICOTS) afin d'identifier la liste des facteurs pronostiques spécifique de chaque (type de) critère de jugement et conduira certainement à différents jeux de covariables à prendre en compte en fonction du critère de jugement.

Au niveau réglementaire, la mention d'une revue systématique pour l'identification des facteurs de confusion apparait dans les guides suivants

Agence	Document	Page	Extrait
EMA	Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence - Scientific guideline	10	The chosen approach for the identification of known confounders has to be clearly described. They should be systematically identified and clearly stated at the design stage, and the design should attempt to minimise their impact on the results. Potential confounders (risk factors for the outcome of interest) should be identified from various sources (e.g., disease knowledge and previous studies identified through systematic literature search) to plan the data collection or extraction for the variables to be accounted for
MHRA	MHRA draft guideline on the use of external control arms based on real-world data to support regulatory decisions	NA	-
FDA	Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products Guidance for Industry	NA	-
FDA	Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products	NA	-
HAS	Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health [44]	Box 1	Well-performed systematic review identifying relevant prognostic variables, confounders and effect modifiers
NICE	NICE real-world evidence framework [64]	124	Identify potential confounders (including time-varying confounders) using a systematic approach and clearly articulate causal assumptions. Key prognostic variables were prospectively identified by a systematic review.

14.2.3 Lecture critique

L'existence d'un processus formalisé d'identification des covariables à prendre en compte pour assurer l'échangeabilité conditionnelle est aussi un élément indispensable pour l'évaluation ou la lecture critique d'un travail de comparaison externe.

En effet, à ce niveau, il convient de s'assurer que tous les facteurs de confusion potentiels ont été contrôlés. Cela nécessite de disposer de la liste des facteurs de confusion potentiels. La simple liste des contrôles pris en considération ne permet pas de juger de cela car rien ne garantit que cette liste soit complète. L'évaluateur ou le lecteur n'a pas forcément l'expertise complète permettant de répondre à cette question. Ainsi à la place de donner, ex abrupto, une liste, les investigateurs doivent documenter la démarche formalisée qu'ils ont utilisée pour arriver à cette liste. Pour l'évaluateur ou le lecteur, la tâche consistera alors à valider cette démarche. Si elle est correcte, la liste qui en découle devient acceptable de facto. Suivre ce raisonnement nécessite cependant une très grande transparence de la part de l'étude.

L'IQWIQ recommande l'utilisation l'approche proposée par Pufulete et al. [133] pour l'indentification des facteurs de confusion.

14.3 Les méthodes statistiques

Comme dans toute études observationnelles, la finalité de l'analyse des données va être de corriger les résultats du biais de confusion (de supprimer la confusion).

> *On parle souvent d'analyse ajustée ou d'ajustement sur les facteurs de confusion. Pour les puristes le terme ajustement ne désigne que les méthodes de régression. Comme il existe des d'autres méthodes que les régressions, il conviendrait plutôt de parler d'analyse conditionnelle, ou de prise en compte des facteurs de confusion, ou d'analyse contrôlant les facteurs de confusion.*

De nombreuses techniques sont disponible pour effectuer cette correction [134] [135]. Elles fonctionnent toutes correctement lorsque leurs hypothèses fondamentales sont vérifiées. Certaines sont plus robustes (moins sensibles à un écart aux hypothèses) que d'autres, mais en général il est difficile de dire qu'une technique est plus adaptée qu'une autre.

La validité des résultats réside principalement dans la prise en compte de la totalité de facteurs de confusion (cf. section 14.2). En effet, si certains facteurs de confusion ne sont pas pris en compte, la méthode, quelle qu'elle soit, y compris une méthode d'IA type machine learning, ne pourra supprimer le biais de confusion induit par ces facteurs oubliés et le résultat sera affecté d'un biais de confusion résiduel.

Comme il est toujours possible que certains facteurs de confusion n'est pas été identifié ou pris en compte, certains estiment qu'il n'est jamais possible de conclure à l'absence de biais de confusion résiduel. C'est par exemple le parti pris par l'outils d'évaluation du risque de biais des ROBINS I (cf. section 21.1).

La section 15 propose une description succincte de des différentes techniques statistiques utilisables pour prendre en compte les facteurs de confusion.

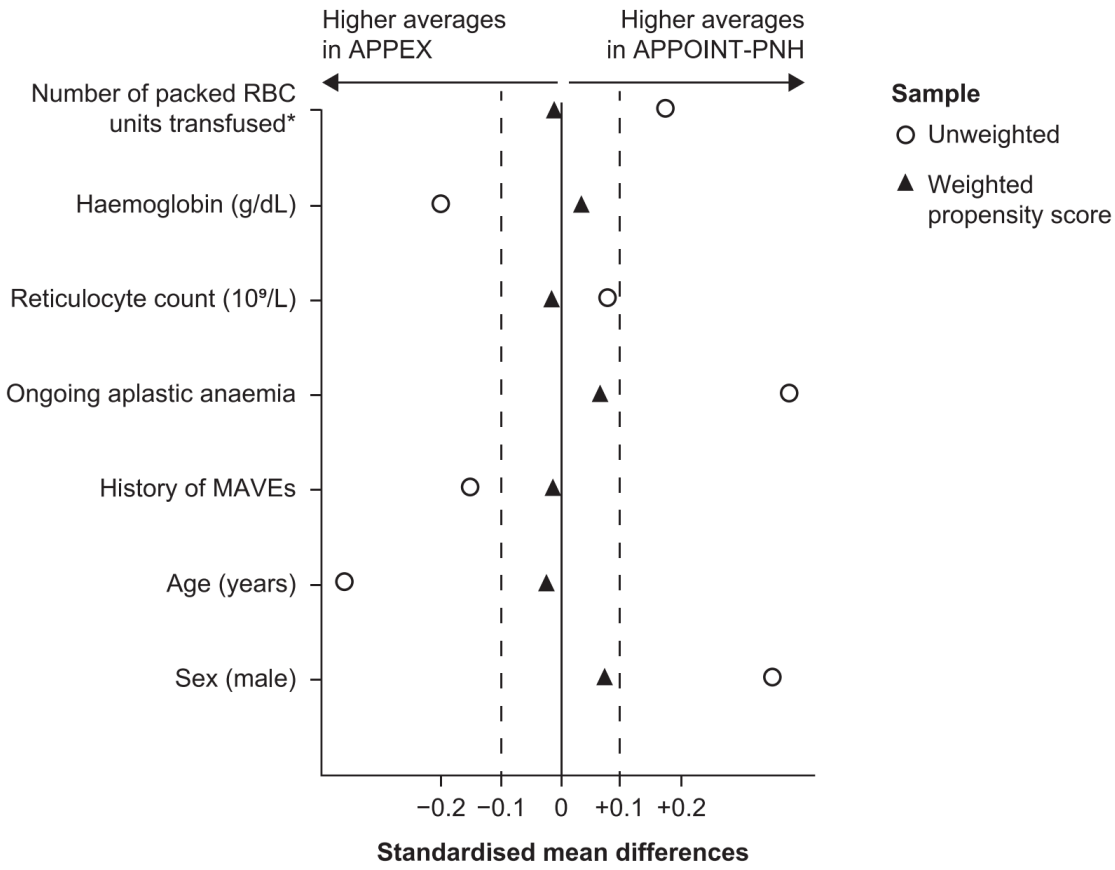


Figure 7 – Exemple de représentation graphique des SMD avant et après pondération à partir d'un score de propension [136].

14.4 Les ajustements à éviter car contreproductifs

L'ajustement sur certaines covariables peut être contreproductif en introduisant un biais. Il s'agit des intermédiaires (médiateurs) et des collisionneurs [124]. Les variables d'ajustement doivent donc être soigneusement choisies pour éviter ces écueils. Les graphes de causalité type DAG permettent d'identifier ces types de covariables afin de ne pas les prendre en compte dans l'ajustement.

Les approches non raisonnées de détermination des covariables d'ajustement ne sont donc pas souhaitables, car elles exposent au risque d'ajustement malencontreux sur ce type de variables.

Les intermédiaires (médiateurs) sont des facteurs intermédiaires sur le chemin causal allant du traitement au critère de critère de jugement. C'est un médiateur de l'effet du traitement (comme la baisse de LDL est un médiateur d'une partie de l'effet des statines sur les évènements cardiovasculaires). Ajuster sur un médiateur coupe le chemin causal et fait disparaitre l'effet du traitement sur le critère de jugement si celui-ci passait entièrement par son intermédiaire. Peut rester un effet passant par d'autres chemins causals.

14.5 Sélection de patients

La première mesure qui vise à limiter le biais de confusion dans une comparaison externe est la sélection des patients du groupe contrôle externes dans la source de données à l'aide des mêmes critères d'éligibilité (critères d'inclusion et de non-inclusion) que ceux utilisés pour l'étude expérimentale du nouveau traitement (monobras ou essai randomisé) [13].

Cependant l'application des mêmes critères de sélection ne garantit pas que le groupe contrôle sera comparable au groupe traité sur les facteurs pronostiques. Des différences dans la distribution de ces variables existeront. En effet, sélectionner des sujets de plus de 18 ans dans 2 populations différentes ne garantit d'obtenir la même distribution d'âge dans les 2 groupes (même moyennes, mêmes intervalles interquartiles, etc.) car la distribution de l'âge peut être différente dans les 2 populations sources. Cependant pour d'autres facteurs qui sont binaires, cette sélection attendra son but. Par exemple si le bras expérimental du nouveau traitement n'a inclus que des patients diabétiques, la sélection dans la source de données externe que des patients diabétiques (avec la même définition) permettra d'obtenir un groupe comparable aux groupes traités (avec 100% de diabétique de même définition dans les 2 groupes).

L'application des mêmes critères de sélection des patients entre le groupe traité et le groupe contrôle est donc un moyen nécessaire, mais pas suffisant. Il va être nécessaire de tenter de corriger les résultats des conséquences de ces différences inter-groupes entre les distributions des facteurs de confusion potentiels à l'aide de l'analyse statistique (cf. sections 14.3 et 15).

15 Les techniques d'analyses statistiques

De nombreuses méthodes statistiques sont disponibles pour prendre en compte les facteurs de confusion dans l'analyse, afin de tenter de corriger le résultat du biais de confusion induit par les facteurs de confusion affectant l'étude.

Les méthodes basées sur le score de propension (appariement, pondération, double robuste) sont les plus populaires, mais existent aussi les classiques méthodes de régression multivariées, la g computation, l'appariement sur les covariables elles-mêmes, etc. Ces méthodes sont basées sur différents paradigmes d'estimation/inférence statistique, mais apparaissent aussi des approches basées sur l'IA type machine learning (comme le TMLE) qui se substituent aux algorithmes classiques d'estimation au sein de ces méthodes, mais qui ne change pas leurs principes généraux.

Le Tableau 13 récapitule les principales méthodes disponibles envisageable avec les groupes contrôles externes.

> *D'autres approches comme le score de propension à haute dimension (*HD propensity score*) ou la pondération avec stratification fine (*fine stratification weights*) nécessite un grand nombre de patients ou de variables et ne sont donc pas utilisables dans la problématique des comparaisons à un groupe externe d'une étude monobras ou d'un essai clinique.*

Tableau 13 – Récapitulatif succinct des méthodes de prise en charge des facteurs de confusion dans l'analyse dans le but de tenter de corriger les résultats du biais de confusion induit par les facteurs de confusion affectant l'étude

Restriction
Stratification
Appariement • Sur les covariables directement • Sur le score de propension
Redressement d'échantillon par pondération • Basée sur le score de propension (IPW, IPTW, etc.) • Autres poids (*entropy balancing*)
G computation (*g formula*)
Approche type *causal machine learning*
Méthode double robuste (AIPW, TMLE, etc.)

15.1 Les techniques basées sur l'appariement (*matching*)

Le principe général des techniques basées sur l'appariement (*matching*) est d'associer chaque patient du groupe traité à un patient contrôle comparable. Le groupe contrôle sera donc constitué de tous ces patients. Par construction il s'agira d'un groupe comparable au groupe traité et la comparaison de ces deux groupes produira une estimation de l'effet du traitement sans biais de confusion si l'appariement a porté sur tous les facteurs de confusion, car les deux groupes seront comparables, hormis pour le traitement, sur tous les facteurs influençant le critère de jugement.

À chaque patient est associé (appareillé) un patient contrôle similaire sur toutes les variables sur lesquelles portent l'ajustement.

> Si l'appariement est fait sur l'âge, le sexe et le stade de la maladie. Un patient du groupe traité (46 ans, homme, au stade 1 de la maladie sera appareillé avec un patient contrôle du même âge, du même sexe et au même stade de la maladie.

La limite principale de l'appariement réside dans la difficulté de trouver un patient similaire pour chaque patient traité. Pour certains patients il est impossible de trouver un patient contrôle ayant les mêmes valeurs sur toutes les covariables. Cela entraine une réduction d'effectif dans le groupe traité, car les patients non appareillés sont exclus de facto de l'analyse. De même cela conduit à ne pas exploiter la totalité des patients contrôles disponibles dans la source de données utilisée. Pour limiter ce dernier point des appariements un pour deux (1 :2), 1 :3 ou autres peuvent être utilisés s'il y a suffisamment de patients contrôles potentiels.

En général il est déraisonnable pour cette raison d'essayer d'apparier les patients sur plus de 3 ou 4 variables. C'est en raison de cette limite qu'ont été développées les techniques d'appariement sur le score de propension. Le score de propension (cf. section 15.2) permet de synthétiser plusieurs covariables en une seule valeur : le score de propension.

Cependant des techniques récentes augmentent la faisabilité de l'appariement direct sur un grand nombre de covariables. Il s'agit, par exemple, des méthodes d'appariements basées des algorithmes génétiques (*evolutionary algorithm*) [137][137][137]. La description technique de ces approches dépasserait le cadre de ce document.

15.2 Le score de propension

15.2.1 Définition

L'idée de base du score de propension (*propensity score, PS*) est de résumer en une seule valeur plusieurs autres variables. Il permet de synthétiser tous les facteurs de confusion identifiés en une valeur entre 0 et 1, le score de propension, qui lorsqu'elle est utilisée comme unique variable « d'ajustement » produit le même effet (sauf cas exceptionnels) que l'ajustement sur toutes les variables qui rentrent dans sa composition.

Le score de propension (SP) peut être utilisé dans les techniques d'appariement, de pondération, de stratification, de régression.

> *Au niveau mathématique, le score de propension est la probabilité qu'un individu reçoive le traitement étudié conditionnellement à ses caractéristiques.*

> Introduit par Rosenbaum et Rubin en 1983 [138], il permet de réduire un problème d'ajustement multivarié à un ajustement sur une dimension, via l'appariement, la stratification, la pondération, ou l'ajustement sur le SP.

Bien que très populaire, le score de propension n'est pas le seul outil statistique permettant de corriger du biais de confusion. Son succès provient principalement de son utilisation en appariement qui

produit deux groupes de caractéristiques comparables, très similaires en apparence à ce qu'aurait produit un essai randomisé (le tableau des caractéristiques à la baseline post-matching présentée dans ces publications).

15.2.2 Le calcul du score de propension

Le score de propension est déterminé par une modélisation du traitement (*treatment model*) à l'aide d'un modèle cherchant à prédire le traitement des patients en fonction de leurs caractéristiques [139] [140]. Plus précisément, il s'agit d'un modèle prédisant la probabilité de recevoir le nouveau traitement (traitement étudié) en fonction des caractéristiques des patients.

Bien qu'il s'agisse structurellement d'un modèle prédictif du traitement (la variable à expliquer est le traitement), sa construction, c'est à dire le choix des variables explicatives, ne s'effectuera pas avec la logique habituelle de construction des outils prédictifs [141]. En fait ce modèle traitement n'a pas pour fonction d'être un vrai outil de prédiction comme l'est, par exemple, une équation de risque cardiovasculaire. Un tel outil prédictif n'aurait d'ailleurs pas beaucoup d'utilité en pratique. Ce « modèle traitement » n'est en fait qu'un « intermédiaire calculatoire » permettant, in fine, par différentes méthodes, d'équilibrer les groupes sur les facteurs de confusion. Les variables à prendre en compte dans ce modèle ne sont pas les prédicteurs du traitement, mais les variables de confusion qui sont, avant tout, des prédicteurs de l'outcome. Il existe une littérature fournie sur la nature des variables à prendre en compte dans le score de propension et les conceptions ont évolué au cours du temps. Le consensus actuel peut se résumer de la façon suivante [141] [142] :

Les variables associées avec l'outcome (critère de jugement) doivent toujours être introduites dans le score propension. Si elles ne sont pas des facteurs de confusion (car elles ne sont pas aussi associées au traitement) elles ne contribueront pas à réduire le biais, mais elles permettront de réduire la variance de l'estimation de l'effet traitement recherché et donc d'améliorer précision et puissance statistique.

Pour les petits effectifs, il est cependant conseillé de ne pas retenir les variables fortement reliées au traitement et faiblement reliées à l'outcome.

Les variables strictement associées avec le traitement (appelées « variables instrumentales » ou simplement « instruments ») ne doivent pas être introduites dans le score de propension (ce qui parait paradoxal au premier abord quand le score de propension est présenté, de façon simpliste, comme étant un modèle prédictif du traitement, cf. discussion ci-dessus).

Les collisionneurs (*collider*) doivent aussi être tenus en dehors du score de propension tout comme les médiateurs (variables intermédiaires dans le chemin causale entre le traitement et l'outcome).

De ces règles découlent tout l'intérêt de la revue systématique à la recherche des facteurs pronostiques des critères de jugement (cf. section 14.2.2) et la réalisation des graphiques de causalités (à l'aide de DAG par exemple, cf. section 14.2.1) pour identifier les variables à ne pas introduire dans le score de propension (collisionneurs et variable instrumentale).

Les variables prédictives de l'outcome DOIVENT être dans le modèle
Une variable qui n'est pas prédictive de l'outcome Ne DOIT PAS être dans le modèle même si elle est prédictive du traitement

Une fois le modèle estimé à partir des données de l'étude, il est utilisé pour calculer, pour chaque patient de l'étude, son score de propension à partir de ses valeurs des variables qui sont prises en considération par le modèle.

Le calcul du score de propension est donc une procédure à deux étapes :

1. Construire un modèle à partir des données de l'étude prédisant la probabilité de recevoir l traitement étudié en fonction des facteurs de confusion
2. Utiliser numériquement ce modèle pour calculer pour chaque patient son propre score de propension à partir de ses valeurs sur les facteurs de confusion (pris en compte dans le modèle)

Au niveau calculatoire le modèle traitement est le plus souvent établi à l'aide d'une régression logistique, mais d'autres techniques en particulier de machine learning peuvent être utilisées (cf. section 15.7).

Le modèle peut être adapté itérativement (en retirant ou rajoutant des variables) en fonction des résultats qu'ils donnent en termes d'équilibrage des facteurs de confusion entre les groupes, à la condition expresse, cependant, que cette adaptation ignore complètement les résultats produits en termes d'estimation de l'effet du traitement. Autrement, ce processus s'apparenterait au p hacking. Il est donc indispensable que cette adaptation se fasse avant toute analyse inférentielle et ne porte que sur l'équilibre des facteurs de confusion[18]. Ces itérations de mises au point du modèle doivent être rapportées en toutes transparences.

15.2.3 L'importance du chevauchement des distributions des scores de propension

La distribution des scores de propension dans les deux groupes permet d'explorer le respect de l'hypothèse de positivité de l'inférence causale (cf. section 13.1.1).

Cette hypothèse stipule que tous les patients, quelles que soient leurs caractéristiques étaient susceptibles de recevoir les deux traitements. Il n'existe pas de patients qui compte tenu de leurs caractéristiques avaient une probabilité nulle de recevoir l'un des deux traitements, c'est-à-dire qu'ils ne pouvaient pas recevoir l'un des deux traitements.

Comme le score de propension est un condensé de caractéristiques des patients, l'hypothèse de positivité sera non vérifiée si pour certains scores de propension existent que des patients traités avec l'un ou l'autre des deux traitements. Ce point peut se vérifier en comparant la distribution des scores de propension dans les 2 groupes (Figure 8, Figure 9, Figure 10 et Figure 11). Cette hypothèse sera vérifiée quand les deux distributions se chevauchent parfaitement.

[18] De manière générale plusieurs méthodes de prise en compte peuvent être essayer afin de déterminer celle qui conduit à la meilleure comparabilité des groupes après ajustement. Cela ne peut être accepté que s'il est certain qu'aucune analyse inférentielle n'a été réalisée durant ce process.

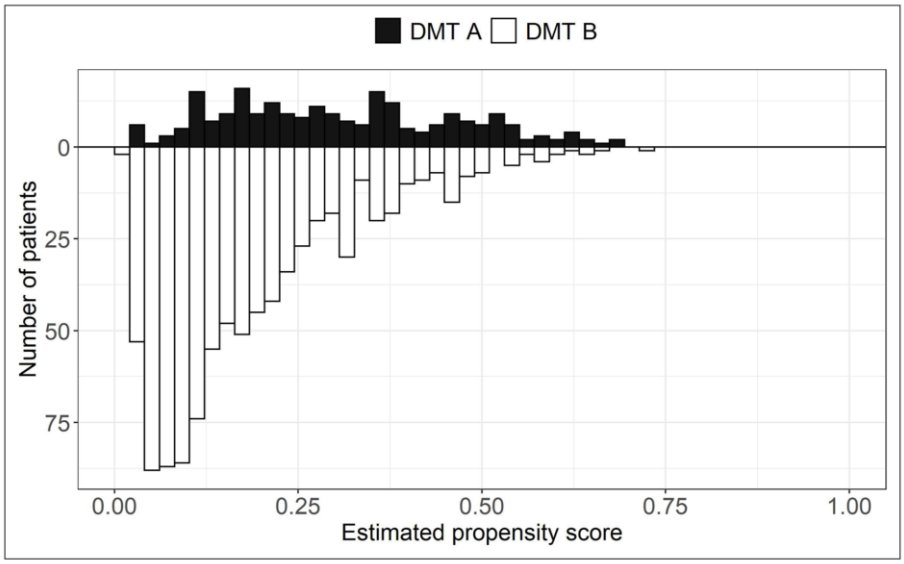


Figure 8 – Exploration du chevauchement des distributions du score de propension dans les deux groupes de traitement (DMT A et DMT B). les deux distributions se chevauchent très largement permettant de conclure que l'hypothèse de positivité n'est pas violée (d'après ref [143]).

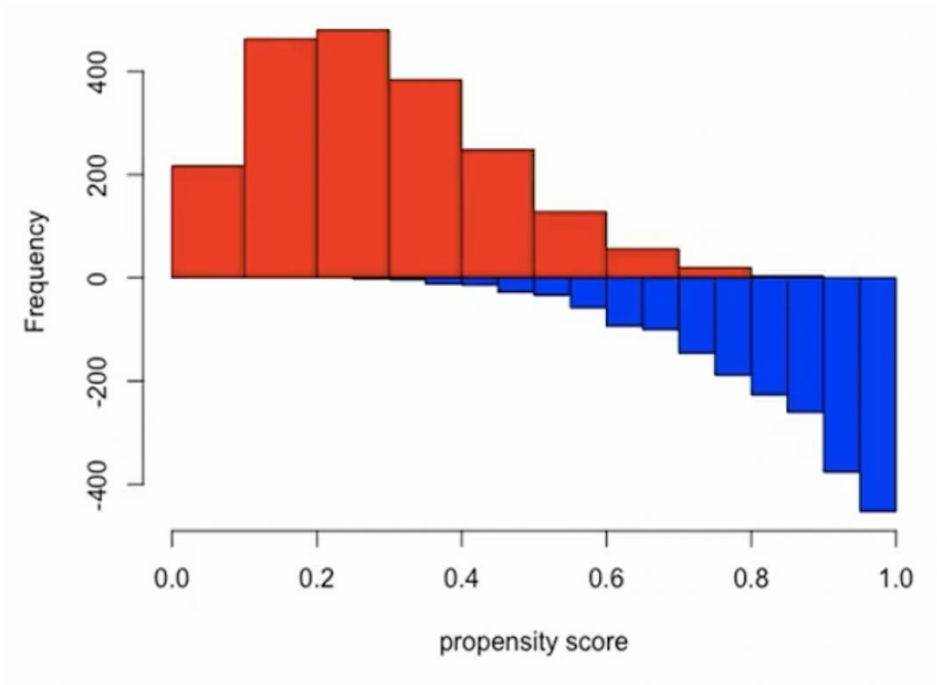


Figure 9 – Situation caractéristique d'un non-respect de l'hypothèse de positivité. Les deux distributions ne se chevauchent pas. Pour un score de propension inférieure à environ 0.25 les patients avaient une probabilité nulle de recevoir le traitement bleu et pour les scores supérieurs à 0.8, les patients avaient une probabilité nulle de recevoir le traitement rouge.

En cas de non-vérification de l'hypothèse de positivité du fait d'un non-chevauchement des distributions, il est parfois proposé de faire un trimming, c'est-à-dire de ne retenir que la plage de score de propension où les distributions se chevauchent. Dans ce cas et pour les études observationnelles classiques l'estimand change est devient l'estimand ATO (*average treatment effect in the overlap population*) qui peut s'interpréter comme l'effet causal du traitement évalué chez les patients pour lesquels les médecins sont encore ambivalents quant au choix du traitement (il y a équipoise pour ces patients) [144]. Cependant si cette approche améliore les performances statistiques elle limite l'interprétation et la généralisabilitée des résultats. L'ATO est l'estimand d'un effet traitement dans une population difficile à définir de façon clinique. Son utilisation dans les comparaisons externes est à exclure.

Dans le cas des comparaisons à un groupe contrôle externe l'interprétation de l'estimand obtenu après trimming est moins évident. D'autres solutions en cas de non-positivité sont en cours de développement [145].

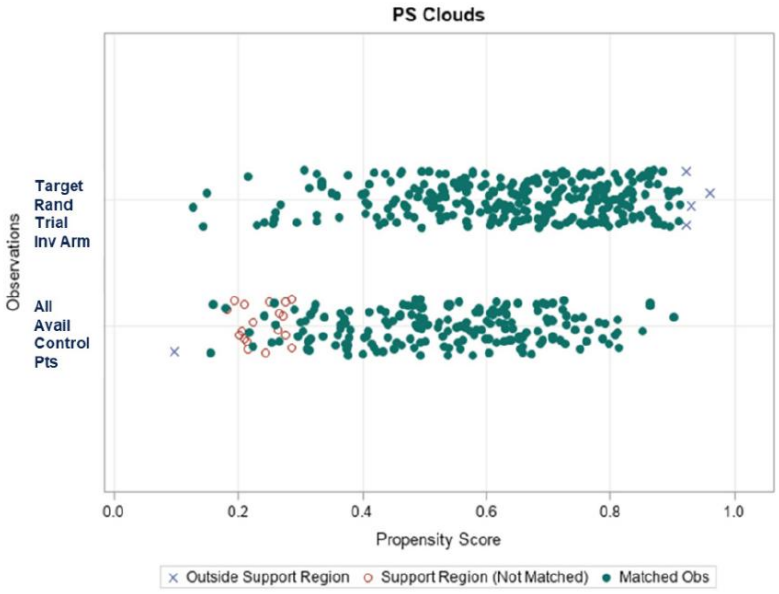


Figure 10 – Représentation en nuage de points des scores de propension des deux groupes comparés.

Apparaissent sur les patients (points) des deux groupes qui sont en dehors de la région commune (*common support*) et ceux du groupe contrôle qui n'ont pas pu être appariés.

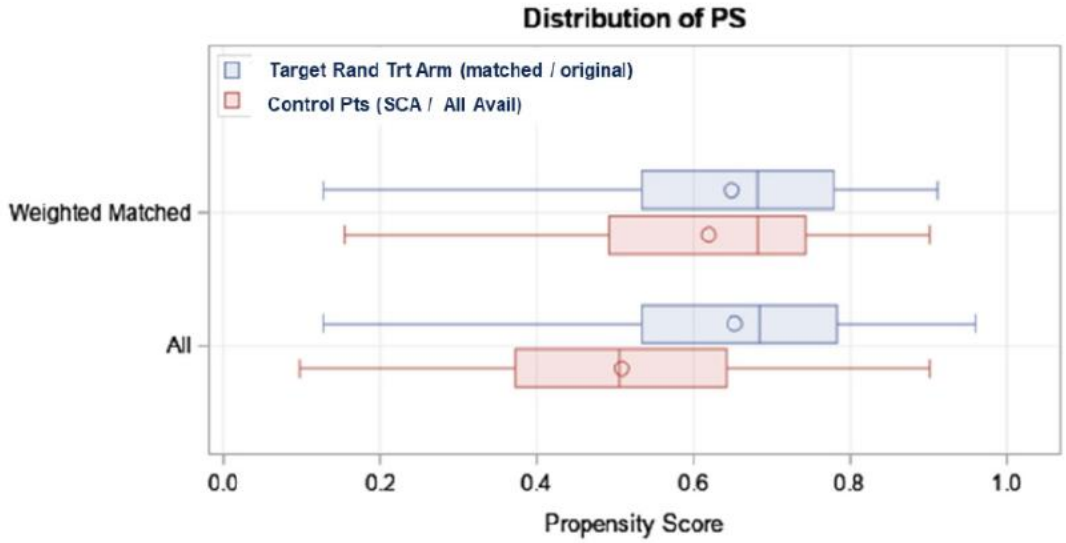


Figure 11 – Représentation en boxplot des distributions des score de propension avant et après appariement

15.3 L'appariement sur le score de propension

Le score de propension peut être utilisé comme variable d'appariement unique. À chaque patient du groupe traité est associé le patient contrôle ayant le score de propension le plus proche de celui du patient (plus proche voisin), ce qui empêche d'avoir des patients impossibles à apparier contrairement à l'appariement sur les covariables elles-mêmes. Cependant cela pourrait conduire à apparier des patients ayant des valeurs de score de propension très différentes. Pour l'éviter, une contrainte est ajoutée n'autorisant l'appariement de 2 patients que si leurs scores ne sont pas trop différents dans une certaine tolérance appelée *caliper*. Il est donc possible de ne pas pouvoir apparier certains patients, car aucun patient contrôle n'a de score de propension suffisamment proche dans cette tolérance.

L'intérêt de l'appariement sur le score de propension est qu'il permet de constituer deux groupes de sujets comparables sur toutes les variables constituant le score de propension (sauf cas assez exceptionnel). Cette propriété a fait le succès de cette méthode et explique pourquoi elle est largement utilisée même si elle n'est pas optimum et si de meilleures méthodes sont disponibles (pondération). En effet, après appariement (*after propensity score matching*), le tableau de description des caractéristiques de baseline des patients montre des valeurs presque identiques entre les 2 groupes appariés pour toutes les variables comprises dans le score de propension. Ce tableau ressemble alors comme à ce qui aurait été obtenu avec un essai randomisé. Les effectifs sont aussi identiques entre les deux groupes si un appariement 1 :1 a été utilisé, de la même façon que dans un essai randomisé. La comparaison à l'essai randomisé s'arrête ici, car, si des facteurs de confusion n'ont pas été pris en compte dans le score de propension, la fiabilité du résultat n'est pas identique à celle d'un essai randomisé où la randomisation prend en compte, automatiquement, tous les facteurs de confusion sans qu'il soit obligé de les identifier et de les mesurer.

Le tableau de l'exemple encadré ci-dessous illustre cet apport de l'appariement sur le score de propension. Les trois premières colonnes comparent les données du groupe traité et du groupe

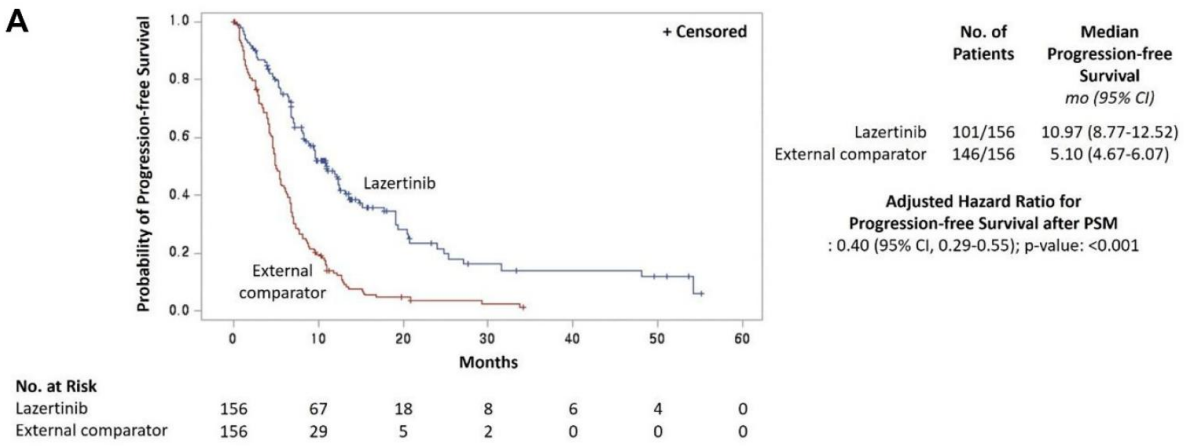
contrôle avant appariement (*before PS matching*). Les trois dernières colonnes rapportent les caractéristiques dans les groupes appariés (*after PS matching*).

Une étude de comparaison externe compare le lazertinib à une chimiothérapie à base de sels de platine chez des patients avec un cancer du poumon non à petites cellules EGFR positifs après échec des inhibiteurs de l'EGFR [146]. Le groupe traité par lazertinib regroupe initialement 200 patients. Un groupe contrôle externe est réalisé à partir d'une source d données de patients traités par chimiothérapie comprenant 334 patients. Un appariement par score de propension a été utilisé pour corriger les résultats du biais de confusion.
La partie gauche du tableau ci-dessous décrit les caractéristiques des patients de ces 2 groupes (before PS matching).
La partie droite (after PS matching) décrit les deux groupes obtenus par appariement. L'effectif de ces groupes est inférieur à l'effectif initial témoignant que certains patients du groupe traité n'ont pas pu être appariés à des patients contrôles, car aucun patient contrôle n'avait un score de propension suffisamment proche (compte tenu du caliper utilisé).

	Before PS Matching			After PS Matching		
	Lazertinib (*n* = 200)	External Comparator (*n* = 334)	aSD	Lazertinib (*n* = 156)	External Comparator (*n* = 156)	aSD
Age, median (Q1–Q3)	64 (56–71)	63 (56–70)	0.11	63.5 (55–70.5)	62.5 (57–69)	0.03
Female sex, *n* (%)	114 (57.0)	191 (57.2)	0.00	85 (54.5)	87 (55.8)	0.03
No history of smoking, *n* (%)	123 (61.5)	213 (63.8)	0.05	94 (60.3)	94 (60.3)	0.00
ECOG performance status, *n* (%)						
0	41 (20.5)	19 (5.7)	0.45	16 (10.3)	17 (10.9)	0.02
1	159 (79.5)	315 (94.3)		140 (89.7)	139 (89.1)	
Adenocarcinoma tumor histology, *n* (%)	194 (97.0)	320 (95.8)	0.06	151 (96.8)	150 (96.1)	0.03
Brain metastasis, *n* (%)	76 (38.0)	90 (27.0)	0.24	56 (35.9)	56 (35.9)	0.00
EGFR mutation status, *n* (%)						
L858R	37 (18.5)	145 (43.4)	0.56	27 (17.3)	67 (43.0)	0.58
Exon19Del	80 (40.0)	179 (53.6)	0.27	68 (43.6)	83 (53.2)	0.19
Others (G719X, L861Q, others, unknown)	83 (41.5)	19 (5.7)	0.35	61 (39.1)	8 (5.1)	0.90
Previous lines of systemic therapy, median (min–max)	1 (1–10)	1 (1–5)	0.45	1 (1–5)	1 (1–5)	0.06
Previous lines of EGFR-TKI treatment, median (min–max)	1 (1–10)	1 (1–5)	0.36	1 (1–3)	1 (1–5)	0.03
Type of previous EGFR-TKI treatment, median (min–max)						
Gefitinib	120 (60.0)	117 (35.0)	0.52	94 (60.3)	54 (34.6)	0.53
Erlotinib	25 (12.5)	77 (23.1)	0.28	15 (9.6)	43 (27.6)	0.47
Afatinib	61 (30.5)	193 (57.8)	0.57	46 (29.5)	89 (57.1)	0.58
Time from immediate previous EGFR-TKI treatment, *n* (%)			0.58			0.15
<30 days	140 (70.0)	167 (50.0)		108 (69.2)	117 (75.0)	
≥30 days	47 (23.5)	165 (49.4)		45 (28.9)	37 (23.7)	
No	13 (6.5)	2 (0.6)		3 (1.9)	2 (1.3)	
Duration of previous EGFR-TKI treatment, *n* (%)			0.45			0.14
<6 months	8 (4.0)	59 (17.7)		6 (3.9)	11 (7.1)	
≥6 months	192 (96.0)	275 (82.3)		150 (96.2)	145 (93.0)	

PS, propensity score; SD, standardized difference; ECOG, Eastern Cooperative Oncology Group; EGFR, epidermal growth factor receptor; EGFR-TKI, epidermal growth factor receptor-tyrosine kinase inhibitor.

Les eux groupes appariés sont ensuite comparée de manière classique pour déterminer l'effet du traitement :



Un indice, appelé SMD (*standardized mean difference*), permet de mesurer la différence entre les 2 groupes au niveau d'une variable. Cet indice prend des valeurs entre à et 1. Zéro indique que les 2 groupes ont la même valeur sur cette variable. Plus la valeur de SMD est grande, plus il y aune différence importante entre les 2 groupes. Habituellement on considère qu'une valeur supérieure à 0.1 témoigne d'un écart important. Le SMD est parfois exprimé en pourcentage, dans ce cas la valeur seuil est 10% et il peut aussi avoir un signe positif ou négatif indiquant le sens de la différence. Il peut aussi s'appeler aSD (*adjusted Standard Difference*) à ne pas confondre avec la SD (standard deviation qui est l'écart type).

Dans l'exemple précédent, la SMD (aSD) avant appariement pour le score ECOG est de 0.45, supérieure au seuil de 0.1 et témoignant ainsi d'une différence notable. Après appariement la différence devient négligeable avec un SMD (aSD) en dessous de 0.10 à 0.02 [147].

Des p value sont parfois donnée à la place de la SMD. La p value n'a pas beaucoup d'intérêt dans ce cas (manque de puissance, non adapté pour montrer l'absence de différence et valeur dépendant de la taille de l'échantillon) et ne devrait pas être utilisé [19].

De nombreuses variantes d'appariement par score de propension sont possibles.

Des ratios d'appariement supérieurs à 1 :1 peuvent être utilisées, augmentant la précision, mais exposant à une diminution du contrôle de la confusion et à un manque de transparence [148] [149] [150].

L'appariement peut être effectué avec ou sans remise. En cas d'appariement avec remise, un patient du groupe contrôle peut-être apparié plusieurs fois avec des patients différents du groupe traité. Cela peut être utile en cas de petits effectifs.

L'appariement peut aussi se faire avec ou sans caliper. Le caliper désigne la tolérance de différence de score de propension acceptable pour apparier deux patients. En pratique, on fixe une valeur seuil (par exemple, 0,2 fois l'écart-type du score de propension) : seuls les patients contrôles dont le score de propension se situe à l'intérieur de cette fenêtre autour du score du patient traité sont éligibles à l'appariement. L'utilisation d'un caliper permet de réduire le risque d'apparier des patients trop

[19] La SMD mesure une taille d'effet, pas une significativité et ne dépend pas de la taille de l'échantillon

différents, améliorant ainsi la comparabilité entre les groupes. Sans caliper, l'appariement se fait avec le plus proche voisin. Cela évite d'avoir des patients non appariés, mais autorise l'appariement de patients ayant des scores de propension très différents.

Il existe aussi des techniques qui ne vont pas travailler patient après patient (*greedy matching*), mais qui vont chercher la combinaison globale optimale pour minimiser la somme des différences de score de propension intra-paire (optimal matching).

- Limites de l'appariement sur le score de propension

L'appariement par score de propension, bien qu'apprécié pour sa simplicité et sa compréhension intuitive, présente plusieurs limites. Tout d'abord, il entraîne fréquemment une réduction significative de la taille des échantillons, ce qui peut diminuer la puissance statistique de l'étude. Ensuite, le calcul de la variance des estimations obtenues après appariement reste complexe et souvent approximatif, rendant l'interprétation des résultats moins fiable. Pour ces raisons on lui préfère actuellement la pondération (mais qui a aussi des limites).

15.4 Les méthodes de pondération

15.4.1 Principes

Les méthodes de pondération s'apparentent à des techniques de redressement d'échantillon qui permettent d'obtenir deux échantillon « virtuel » comparables en appliquant des poids bien choisis aux patients d'une étude [151].

Dans le cas des comparaisons à un groupe contrôle externe, les poids sont seulement appliqués au groupe contrôle afin de le rendre comparable au groupe traité (cf. ci-dessous) et d'obtenir l'estimand ATT attendu (cf. section 13.4).

On peut présenter l'intuition qui est derrière cette approche de la façon simplifiée suivante. Imaginons que le groupe contrôle à un âge moyen plus faible que celui du groupe traité, car il ne comprend qu'un seul patient de 80 ans contre 4 dans le groupe traité. L'idée générale est d'augmenter virtuellement le nombre de patients de 80 ans dans le groupe contrôle en surpondérant (en surreprésentant) celui qui est présent. Pour cela il lui sera donné un poids supérieur à 1, 4 par exemple. Cela conduira à avoir 4 patients virtuels de 80 ans issus de ce patient. Bien sûr, le même poids sera donné au critère de jugement de ce patient. S'il était décédé, cela conduira à 4 décès dans le groupe virtuel. S'il n'était pas décédé, cela ne rajoutera pas de décès au groupe virtuel. Évidemment la méthode réellement employée est bien plus complexe que cela, car les poids appliqués aux patients doivent tenir compte des déséquilibres initiaux sur toutes les variables pour lesquelles l'équilibrage est recherché.

Pour obtenir une comparabilité sur plusieurs variables, la pondération utilisera le score de propension qui repose sur ces variables. De manière générale on parle de techniques de IPW (*inverse probability weighting*), de IPTW (*inverse probability of treatment weighting).*

> Une étude monobras évaluant le monocertinib a été comparée à un groupe contrôle externe traité par le « standard of care » issu de la base de données Flatiron [8]. Une méthode de pondération basée sur le score de propension a été utilisée pour corriger les résultats du biais de confusion.

Table 1
Baseline demographics and clinical characteristics in the mobocertinib group and the RWD group before and after weighting.

Baseline Characteristics	Mobocertinib	Unweighted RWD	Weighted RWD
No. of patients	114	50	109 [a]
Age (years)			
Mean (SD)	59.6 (11.53)	64.3 (10.26)	60.8 (16.20)
Sex, n (%)			
Female	75 (65.8)	34 (68.0)	72 (65.8)
Male	39 (34.2)	16 (32.0)	37 (34.2)
Time since initial diagnosis (months)			
Mean (SD)	23.8 (27.92)	17.2 (20.29)	20.9 (34.70)
Brain metastasis, n (%)			
Yes	40 (35.1)	17 (34.0)	42 (38.2)
No	74 (64.9)	33 (66.0)	67 (61.8)
History of smoking, n (%)			
Yes	33 (28.9)	21 (42.0)	30 (27.2)
No	81 (71.1)	29 (58.0)	79 (72.8)

RWD, real-world data; SD, standard deviation.
After weighting, all variables were well balanced, with $P > 0.05$.
[a] weighted sample size.

Le groupe contrôle externe est composé de 50 patients peu comparables à ceux du groupe traité (il manque les SMD dans ce tableau). Après pondération, un groupe redressé « virtuel » de 109 patients est obtenu qui s'avère davantage comparable au groupe traité (l'article utilise à tort des p values pour analyser la comparabilité des groupes, l'utilisation de SMD aurait été plus appropriée, cf. ci-dessus).

Le critère de jugement est le taux de réponses objectives (ORR), cf. table 3 de la publication ci-dessous. Initialement, avant pondération (unweighted), ce « taux » était de 14%. Avec les poids appliqués aux patients, le taux redressé (weighted) devient 11.9%., conduisant à un odds ratio de 3.75 (contre 3.32 avant pondération, avec les données brutes).

Table 3
Confirmed overall response rate in the mobocertinib group and the RWD group before and after weighting.

Outcome	Mobocertinib	RWD		Odds ratio of mobocertinib vs RWD (95 % CI)	
		Unweighted	Weighted	Unweighted	Weighted
cORR,% (95 % CI)	35.1 (26.4, 44.6)	14.0 (5.8, 26.7)	11.9 (5.8, 18.0)	3.32 (1.68, 6.58) $P < 0.01$[a]	3.75 (2.05, 6.89) $P < 0.01$[a]

CI, confidence interval; ORR, overall response rate; RWD, real-world data.
[a] *P* value was determined using Cochran-Mantel-Haenszel test.

Si tous les facteurs de confusion affectant l'étude ont été pris en considération dans le score de propension produisant les poids utilisés et si le groupe contrôle pondéré obtenu peut être considéré comme similaire au groupe traité, l'odds ratio de 3.75 serait corrigé du biais de confusion affectant la valeur brute (non pondérée).

De nombreuses variantes calculatoires existent en fonction de la façon dont sont calculés les poids à partir du score de propension. Chacune de ces variantes correspond à un effet causal différent (ATE, ATT, ATC, etc.). Dans les comparaisons externes, l'effet causal d'intérêt est *l'average treatment effect among treated* (ATT). Le poids utilisé doit donc correspondre à cet effet causal. La méthode utilisant le poids approprié est parfois appelée SMRW *standardized mortality ratio weighting* (à la place de l'appellation générique IPTW ou IPW) [152].

Comme avec le matching, il n'est pas attendu que ces méthodes augmentent la comparabilité sur des covariables non incluses dans le score de propension. Des améliorations (ou des aggravations) de la comparabilité peuvent cependant être observées sur de telles variables du fait de leur interrelation avec celle incluse dans le score de propension.

La précision des estimations de l'effet traitement peut être améliorée en utilisant des poids stabilisés. Cela est obtenu en multipliant les poids par les probabilités marginales d'être traité ou non (dans l'échantillon global).

- L'effective sample size (ESS)

L'effective sample size (ESS) est une mesure qui quantifie la taille d'échantillon "effective" après pondération. Ce n'est pas vraiment un effectif (décompte de sujets), mais une mesure statistique. L'ESS ne soit pas être confondu avec l'effectif après appariement des études utilisant un matching sur le score de propension.

L'ESS doit être impérativement rapportée dans les études utilisant une pondération (IPW).

L'ESS n'est pas la somme des poids, mais une mesure dérivée de l'hétérogénéité des poids entre eux. L'ESS est égal à l'effectif initial avant pondération si tous les poids sont égaux. Si ce n'est pas le cas (hétérogénéité des poids), l'ESS sera inférieur à l'effectif avant pondération. L'ESS permet d'évaluer la perte d'information due à la variabilité des poids et à l'instabilité des estimations.

L'ESS donne une idée de la « taille utile » (de la taille réelle ») de l'échantillon après pondération. Avec un ESS de 60, la précision d'un résultat peut être vue comme étant celle qu'aurait donnée un échantillon de 60 patients. Les petits ESS laissent présager d'une faible précision/puissance. Une forte réduction de l'ESS par rapport à l'effectif avant pondération témoigne d'une forte perte d'information liée à la pondération et témoigne d'une différence importante entre les caractéristiques du groupe contrôle externe et du groupe traité.

- Distribution des poids, hypothèse de positivité

La comparaison de la distribution des scores de propension entre le groupe traité et le groupe contrôle permet d'explorer la plausibilité de l'hypothèse de positivité (cf. section 13.1.1). Avec les techniques de pondération (IPW) il est aussi recommandé d'analyser les distributions des poids. Cependant, avec l'utilisation de l'ATT dans les comparaisons externes, la comparaison de la distribution des poids entre les deux groupes n'a pas de sens étant donné que les poids du groupe traités sont, par principe, tous égaux à 1. Seule compte alors la distribution des poids du groupe contrôle.

Des poids proches de zéros dans le groupe contrôle évoque un non-respect de l'hypothèse de positivité pour ces patients.

Une distribution des poids (stabilisés) très asymétrique conduit à donner un poids très important à quelques patients très atypiques. Pour éviter cela un élagage (trimming) des poids peut être effectuée.

15.4.2 Pondérations non basées sur le score de propension

La pondération peut être effectuée avec des poids différents du score de propension.

Il s'agit, par exemple, des méthodes d'appariements basées sur l'entropie (*entropy balancing*) [153] [154] [155]. La description technique de ces approches dépasserait le cadre de ce document.

La selpercatinib a été évaluée dans le cancer du poumon non à petites cellules RET+ dans une étude monobras LIBRETTO-1 (NCT03157128). Une comparaison externe a été employé pour évaluer le bénéfice clinique par rapport aux standards de soins [156]. Une méthode d'entropy balancing a été utilisée pour obtenir la même distribution des caractéristiques considérées dans les 2 groupes

Table 1. **Baseline characteristics of patients treated with selpercatinib (LIBRETTO-001 trial) and the real-world control, before and after entropy balancing**

Characteristic	Selpercatinib cohort (LIBRETTO-001)	First-line setting, before entropy balancing			First-line setting, after entropy balancing[a]		
		Real-world control Analytic Strategy 1 (*RET* fusion positive)[a]	Real-world control Analytic Strategies 2 and 3	Real-world control sensitivity analysis	Selpercatinib cohort (LIBRETTO-001)	Real-world control Analytic Strategies 2 and 3	Real-world control sensitivity analysis
N	48	29	2791	985	48	2791	985
Sex, *n* (%)							
Female	29 (60.4)	13 (44.8)	1132 (40.6)	369 (37.5)	29 (60.4)	1686 (60.4)	595 (60.4)
Male	19 (39.6)	16 (55.2)	1659 (59.4)	616 (62.5)	19 (39.6)	1105 (39.6)	390 (39.6)
Age, mean (SD)	62.2 (14.1)	65.6 (11.0)	68.8 (9.5)	68.1 (9.3)	62.2 (14.1)	62.2 (13.5)	62.2 (14.1)
Body weight, mean kg (SD)	71.7 (18.1)	75.8 (16.3)	75.6 (19.3)	75.8 (18.6)	71.7 (18.1)	71.7 (19.7)	71.7 (20.0)
ECOG performance status, *n* (%)							
0	20 (41.7)	7 (24.1)	733 (26.3)	282 (28.6)	20 (41.7)	1163 (41.7)	410 (41.7)
>0	28 (58.3)	15 (51.7)	1556 (55.8)	528 (53.6)	28 (58.3)	1628 (58.3)	575 (58.3)
Missing	0	7 (24.1)	502 (18.0)	175 (17.8)	0	0	0

15.5 La g computation (g formula)

Le principe de base de la *g-computation* est de modéliser la survenu du critère de jugement en fonction du traitement reçu et des covariables [135]. On parle alors de modèle d'outcome (*outcome model, Q model*). Sans rentrer dans les détails, ce modèle va ensuite être utilisé pour prédire le critère de jugement pour chaque patient à partir de ses valeurs sur ces covariables. L'effet traitement sera déduit de la comparaison des valeurs moyennes du critère de jugement prédites dans chaque groupe : valeurs moyennes prédites sous traitement pour les patients du groupe traité et valeurs moyennes prédites sans traitement pour ces mêmes patients du groupe traité ce qui permet d'obtenir l'ATT (*average treatment effect among treated*, cf. section 13.4). Ainsi la *g-computation* permet de trouver le contrefait du groupe traité (qui sera comparé non pas à la valeur moyenne observé du critère de jugement du groupe trait émis à la valeur prédite par le modèle sous traitement ceci pour des raisons mathématiques qu'ils ne nous aient pas possibles d'expliciter ici)

> *Cette approche s'inscrit directement dans la logique du raisonnement contrefactuel de l'inférence causale. L'effet causal est la différence entre les 2 outcomes potentiels de chaque patient, c'est-à-dire la valeur du critère de jugement sous traitement et sans traitement de chaque patient. Bien sûr les valeurs de ces 2 outcomes potentiels ne sont pas accessibles simultanément pour un patient donné (qui a été soit traité, soit non traité). Elles sont cependant estimables sous certaines conditions à l'aide du modèle d'outcome qui permet d'appréhender ce que devraient être ces 2 outcomes potentiels avec et sans traitement en fonction des caractéristiques des patients.*

De nombreuses variantes sont possibles. Un seul modèle prédisant l'outcome en fonction du traitement et des covariables peut être construit à partir de toutes les données de l'étude, ou deux modèles, un pour chaque traitement.

Dans le cadre des comparaisons à un groupe contrôle externe, afin d'obtenir l'estimand ATT, l'outcome model est appliqué aux patients du groupe traité, mais en forçant l'absence de traitement. Cela permet de prédire l'outcome potentiel des patients traités s'il n'avait pas reçu ce traitement (mais le traitement contrôle). Il s'agit bien du contrefait du groupe traité.

Le modèle d'outcome est en général construit en utilisant la régression logistique, mais comme il s'agit avant tout d'un modèle prédictif rien n'interdit l'utilisation d'autres techniques, en particulier, celles de machine learning (IA) : xgboost, réseau neuronaux, SVM, random forest, ou superlearner.

La *g-computation* est aussi appelée *g-formula*, *g-estimation*, *marginal structural models*.

La *g-computation* permet de calculer indirectement un effet marginal à partir d'une approche de régression grâce à une étape de standardisation.

La *g-computation* est aussi utile pour gérer des facteurs de confusion évoluant au cours du temps (*time-varying confounding*), ou des traitements changeant au cours du temps. Elle est aussi utile quand les facteurs de confusion à la baseline sont influencés par des traitements précédents.

15.6 Les méthodes doubles robustes

Les méthodes doublement robustes (double robust) prennent en compte les facteurs de confusion à travers deux modèles : un modèle de traitement (exposition) comme le score de propension et un modèle d'outcome (d'où la notion de double). L'intérêt est qu'il suffit qu'un des deux modèles soit correctement spécifié (exact) pour que l'estimation soit valide (d'où la notion de robuste). Si les deux modèles sont mal spécifiés, le résultat restera baissé.

L'AIPW (*Augmented Inverse Probability Weighting*) est une des méthodes doublement robustes largement utilisée dans les études observationnelles. Le TMLE est une autre méthode de cette catégorie plus flexible car utilisant des techniques de machine learning (superlearner par exemple)

15.7 Les méthodes de régression

Les facteurs de confusion peuvent être pris en compte dans l'analyse par les classiques méthodes de régression multivariables comme la régression logistique, le modèle de Cox, etc.

Cette approche est puissante et flexible, permettant une modélisation fine des données. Le modèle comprend la variable traitement et des variables d'ajustement. Le but n'étant pas de chercher les déterminants indépendants du critère de jugement (utilisation habituelle de ces modèles en recherche clinique), seul a un intérêt le coefficient de la variable traitement. Les autres variables étant dans le modèle pour ajuster l'analyse et non pas pour documenter leur relation avec le critère de jugement.

Le score de propension peut aussi être utilisé dans une approche de régression comme unique covariable d'ajustement.

> *Avec leurs variantes pondérées, ces méthodes de régression sont l'instrument technique permettant de réaliser les méthodes de pondération (IPW, IPTW,) par exemple avec un Cox pondéré. Dans ce cas la régression en comporte qu'une seule variable explicative le traitement, les covariables intervenant par l'intermédiaire des poids attribués à chaque observations (cf. section 15.4)*

Les méthodes de régression peuvent aussi être employées après un appariement pour ajuster sur des covariables qui reste déséquilibrée (SMD >0.10) ou sur toutes les variables incluses dans le score de propension [157] mais cette approche est encore discutée [158]. Une alternative serait de recourir à une méthode double-robuste.

Les covariables prédictives du critère de jugements sont à mettre dans le modèle surtout si elles sont facteur de confusion. Les variables prédictives non liées au traitement ne corrigeront pas du biais de confusion, mais permettront de gagner en précision. Par contre, les variables purement prédictives du traitement (et non liées au critère de jugement) ne doivent pas être introduites dans le modèle, comme les intermédiaires et les collisionneurs. Les règles de choix des variables pour le score de propension sont identiques à ces principes de modélisation (cf. section 15.2).

Il est déconseillé aussi de mettre toutes les variables pré-traitements disponibles avec le risque d'ajuster sur des collisionneurs ou des médiateurs. La réalisation d'un diagramme de causalité (DAG) permet d'éviter cette situation.

Il a été montré que la régression pouvait apporter un même contrôle de la confusion que les techniques basées sur le score de propension dans les études de cohorte à la condition d'un nombre suffisant d'événements pour pouvoir introduire la totalité des covariables nécessaire à ce contrôle (contrainte que n'a pas la modélisation du score de propension, le nombre de patients traités avec le traitement d'intérêts étant souvent bien plus grand que le nombre d'événements). Cependant la régression estime un effet conditionnel différent des effets traitement moyens (ATE, ATT, ATC) estimables par les approches basées sur le score de propension. Il s'agit d'un effet défini localement pour un niveau donné des covariables fixées, et non pas après intégration sur leur distribution. C'est un effet local (CATE), pas un effet moyen sur une population. Il s'agit d'un effet conditionnel qui n'est pas l'estimand causal final. Cette estimation fait une hypothèse supplémentaire qui est celle de la constance de l'effet à travers les valeurs des covariables. Les méthodes de type IPW ou g computation ne reposent pas sur cette hypothèse étant donné qu'elles estiment un effet marginal.

La correction du biais de confusion avec les approches de régression nécessite que le modèle soit bien spécifié, c'est-à-dire adapté aux données dans sa forme fonctionnelle (forme mathématique, interaction entre variables, etc.). En cas de mauvaise spécification du modèle, un biais de confusion peut perdurer même si tous les facteurs de confusion ont bien été pris en compte. L'utilisation de modèle flexible (comme les techniques de machine learning) permet de réduire le risque de modèle mal spécifié, mais augmente le risque de surdétermination (overfitting). La construction des modèles n'est pas un point trivial et nécessite une expertise appropriée.

Les techniques de régression sont mal adaptées aux situations où les événements sont rares, car cette situation limite le nombre de covariables pouvant être pris en compte. Une approche utilisant le score de propension peut alors permettre la prise en compte d'un plus grand nombre de facteurs de confusion si les deux groupes de traitement sont de tailles assez similaires.

Un désavantage de la régression par rapports à l'appariement ou la pondération par score de propension est de ne pas permettre une visualisation de la comparabilité des groupes obtenue. Il s'agit d'un désavantage uniquement en termes de présentation des résultats car le contrôle de la confusion est obtenu par un autre mécanisme calculatoire.

Les approches de régression ne mesurent pas le même effet traitement que les autres méthodes comme l'appariement, la pondération ou la *g-computation*.

Les méthodes de régression estiment typiquement un **effet conditionnel** (*conditional average treatment effect*) tandis que les méthodes d'appariement, de pondération ou de g computation donne un **effet marginal**.

L'effet marginal dépend de la structure de la population, c'est un effet moyen dans la population (effet global). L'effet conditionnel est l'effet du traitement quand les autres covariables sont fixées (effet local dans des strates de patients tous identiques). L'effet conditionnel répond à "quel est l'effet pour des individus similaires ?", tandis que l'effet marginal correspond à "quel est l'effet au niveau de la population ?".

Les deux effets peuvent être différents suivants des données et les modèles utilisées.

Effet conditionnel	Correspond à l'effet moyenné à travers des sous population présentant les même caractéristiques
Effet marginal	Effet moyen de la population (différence entre deux mondes hypothétiques où dans l'un tout le monde est traité et dans l'autre personnes n'est traités

L'effet conditionnel est égal à l'effet marginal uniquement avec les modèles linéaires en l'absence d'interaction entre le traitement et des covariables. Avec les modèles non-linéaires (comme la régression logistique) les eux effets sont différent. Cette question est connexe à la discussion de la non collapsibilité des métriques utilisées pour mesurer la taille de l'effet (indices d'efficacité) et sort du cadre de ce document.

15.8 Les techniques de maching learning (IA)

Les techniques d'IA, et plus exactement les techniques de *causal machine learning*, sont utilisables à plusieurs niveaux dans l'analyse statistique des études observationnelles et donc des comparaisons externes [159].

Ces techniques permettent de construire des outils de prédiction. L'utilisation d'un modèle prédictif apparait dans plusieurs méthodes utilisées pour prendre en compte les facteurs de confusion.

Le score de propension est un modèle de prédiction du traitement étudié en fonction de covariable. Classiquement ce modèle est construit avec la régression logistique (qui est d'ailleurs un outil d'IA), mais toutes techniques permettant de construire un modèle (algorithme) de prédiction à partir de données sont potentiellement utilisables. Cependant les méthodes d'IA ne permettent pas de modéliser facilement la variance, mais des solutions existent maintenant (TMLE). Ainsi de nombreuses approches peuvent être utilisées en remplacement de la régression logistique pour la construction du modèle (apprentissage) puis le calcul du score de propension de chaque patient (inférence) comme par exemple des modèle paramétrique ou LASSO ou des méthodes plus flexibles comme les *random forests*, le *support vector machine*, *gradient boosting* (XGBoost), les réseaux neuronaux (neural network, deep learning) ou des approches plus classiques de régulation (*L1-regularized regression*) [160].

L'approche des *SuperLearner* mixe plusieurs méthodes de bases pour augmenter les performances prédictives.

La g computation se base sur un modèle prédictif de l'outcome qui peut être construit avec une de ces techniques de *machine learning*.

Un inconvénient de ces outils prédictifs basés sur le *machine learning* est qu'ils ne permettent pas en général de calculer la précision de leur estimation. Cette problématique a été solutionnée dans des approches comme l'AIWP ou le TMLE [161] [162] [163], approche qui s'applique aux comparaisons externes [117].

Ces approches ne représentent pas une solution magique garantissant automatiquement la suppression du biais de confusion. Elles permettent peut-être d'obtenir des modèles mieux spécifiés, car plus flexibles. Cependant cette flexibilité peut être contreproductive en augmentant la variance des estimations nécessitant de trouver un équilibre en biais et variance. De plus, même avec ces techniques, la problématique reste la prise en compte (et donc la mesure) de tous les facteurs de confusion affectant l'étude.

L'analyse des performances intrinsèques des différentes méthodes et de leur cas d'usage dépasse le cadre de ce document.

16 Le diagnostic d'absence de biais de confusion résiduel

Il est rare que l'approche mise en œuvre pour corriger les résultats du biais de confusion soit complètement satisfaisante (souvent, car certains facteurs de confusion potentiels sont absents des données). Se pose alors la question d'un biais de confusion résiduel, c'est d'un biais de confusion qui perdure malgré l'analyse ajustée qui est censée le supprimer.

L'importance de ce biais résiduel doit être systématiquement explorée à l'aide d'outils appropriés comme les contrôles négatifs (si l'objectif est de mettre en évidence une supériorité) et/ou l'analyse quantitative de biais [164].

16.1 Contrôles négatifs

Les contrôles négatifs sont une approche de falsification des résultats. L'objectif est de mesurer dans l'étude certaines associations particulières dont on sait qu'elles n'existent pas, mais qui sont influencées par les mêmes facteurs de confusion que les associations d'intérêt. Si l'analyse de l'étude ne prend pas en compte tous les facteurs de confusion, une association sera retrouvée au niveau de ces contrôles négatifs, reflétant, non pas la réalité, car on sait qu'il pas d'association à ce niveau, mais simplement le biais de confusion restant du fait de la prise en compte incomplète des facteurs de confusion affectant cette étude [165] [166] [167]. En revanche, si l'absence d'association est bien retrouvée, cela confortera l'idée de l'absence de biais de confusion résiduel.

> **Exemple** – Une étude observationnelle compare deux antiagrégants plaquettaires dans les syndromes coronariens aigus, un nouveau le ticagrelor et un ancien de clopidogrel. Les critères de jugement d'intérêts sont les événements cardiovasculaires ischémiques et les hémorragies. Un troisième critère est utilisé comme contrôle négatif, les pneumopathies. En effet il n'est attendu aucune différence de fréquence des pneumopathies entre ticagrelor et clopidogrel (absence d'association entre l'antiagrégant et le risque de pneumopathies). Cependant les facteurs déterminants de ces 3 pathologies sont en partie communs (tous les facteurs liés à la fragilité des patients comme l'âge, le diabète, les comorbidités, etc.). Ainsi les pneumopathies vérifient les conditions pour être un contrôle négatif dans cette étude. Si la prise en compte des facteurs de confusion est optimale, aucune association ne sera trouvée entre antiagrégant et pneumopathie.
> Dans l'analyse brute (sans prise en compte des facteurs de confusion), un hazard ratio de 0.60 est trouvé pour la comparaison ticagrelor versus clopidogrel sur les pneumopathies, témoignant d'un biais de confusion important. Les médecins ont donné le ticagrelor à des patients globalement en meilleure santé que ceux à qui ils réservent le clopidogrel, donc à plus faible risque de pneumopathies. Après prise en compte des facteurs de confusion identifiés, le hazard ratio devient 1.01 et retrouve l'absence d'association attendue. L'absence de biais de confusion résiduel (après prise en compte des facteurs de confusions identifiés) est donc fortement suggérée par ce résultat. Cependant rien ne garantit que les pneumopathies soient affectées par tous les facteurs de confusion des comparaisons d'intérêts sur les événements cardiovasculaires ou les hémorragies. Cette démonstration n'est donc pas totalement convaincante. Il aurait été souhaitable que plusieurs contrôles négatifs soient employés.

Bien que séduisante, cette approche présente plusieurs limites. Les contrôles négatifs doivent présenter les mêmes facteurs de confusion que l'association étudiée. Cette condition est rarement

vérifiée. Il convient donc de multiplier les contrôles négatifs en sélectionnant des associations dont les facteurs de confusion recoupent ceux de l'étude, afin d'augmenter les chances de couvrir l'ensemble des facteurs de confusion pertinents.

L'absence de variables pouvant servir comme contrôle négatif dans les sources de données nécessite de recourir à du chainage entre sources de données.

16.2 Analyse quantitative de biais, E value

Pour la question du biais de confusion résiduel, les techniques d'analyse quantitative des biais [168] [164] [169] [170] [171] [172] [173] consistent à voir si le résultat de l'étude pourrait quantitativement être expliqué par des facteurs de confusion non pris en considération, c'est-à-dire, être principalement dû à la non-prise en considération de ces facteurs[20]. Il s'agit soit des facteurs que l'on sait explicitement non pris en compte (ils ont été identifiés, mais ils n'ont pas été mesurés dans l'étude ou le groupe contrôle externe par exemple), soit de facteur hypothétique. Dans ce dernier cas, il s'agit de déterminer quel devraient être les caractéristiques numériques (cf. ci-dessous) d'un facteur de confusion oublié pour que cela invalide numériquement le résultat (on parle de nullification de résultat).

Ces analyses quantitatives de biais, comme la E value ou d'autres approches de *tipping value [173]*, possèdent aussi leur limite. La conclusion à l'absence de biais de confusion résiduelle se base sur le jugement subjectif de plausibilité de l'hypothèse conduisant à la négativation du résultat. Cette interprétation dans la nuance est sujette aux biais cognitifs et conduits souvent à des conclusions différentes en fonction du lecteur.

Malgré ces limites il est indispensable qu'une ou plusieurs approches de ce type permettent d'écarter un biais de confusion résiduel patent. Cependant, même dans ce cas, il est difficile d'exclure un tel biais résiduel. Ainsi l'outil de risque de biais ROBINS-I V2 [174] n'offre que la possibilité de conclure à un risque de biais de confusion faible indépendamment d'un risque de biais de confusion résiduel : « *Low risk of bias (except for concerns about uncontrolled confounding)* ». Cela étant dû à l'impossibilité d'exclure un biais de confusion non contrôlé dans les approches observationnelles. Cette impossibilité représente une limite intrinsèque non réductible pour la démonstration des bénéfices cliniques des nouveaux traitements par les approches observationnelles.

16.2.1 Analyse quantitative de biais

L'analyse quantitative de biais (*quantitative bias analysis*, QBA) vise à apprécier numériquement l'impact potentiel des biais, notamment ceux liés à la non-mesure ou au déséquilibre de certains facteurs de confusion, sur l'estimation de l'effet du traitement donnée par l'étude. Cette démarche implique de modéliser la manière dont les biais et les facteurs non pris en compte pourraient fausser le résultat observé, en s'appuyant sur des hypothèses ou des estimations issues de la littérature et en tenant compte de l'incertitude statistique [164] [170] [171].

Cette approche s'applique à toutes études observationnelles comparatives et donc aux comparaisons externes [168][168] [169] [175] [176] [177] [172] [173][136].

[20] Ces techniques ont aussi applicables pour éprouver la robustesse des résultats vis-à-vis des autres biais (sélection, mesure, etc.). Nous ne les envisagerons pas ici.

Concrètement, la QBA permet de déterminer, pour chaque facteur possible, la force de son association avec le critère de jugement et le degré de déséquilibre entre les groupes, puis d'évaluer si ces paramètres pourraient suffire à annuler le résultat. Cette analyse est souvent représentée graphiquement par des frontières de nullification, illustrant les combinaisons de paramètres qui rendraient l'effet observé non significatif.

Ainsi, l'analyse quantitative de biais pour un groupe contrôle externe consiste à simuler et quantifier l'effet potentiel de biais résiduels, afin d'aider à interpréter la validité des résultats en présence d'incertitudes sur certains facteurs confondants ou sur la qualité des données externes. Cela permet de mieux appréhender la plausibilité que des biais non contrôlés puissent expliquer l'effet observé, et donc de renforcer ou nuancer la confiance dans les conclusions de l'étude.

16.2.2 E value

Le but de la E-value [178] est de caractériser numériquement par une seule valeur les propriétés que devrait avoir un facteur de confusion non pris en compte pour expliquer à lui seul le résultat de l'étude. Cette valeur permettra d'apprécier la robustesse de ce résultat vis-à-vis d'un potentiel biais de confusion résiduelle.

Deux paramètres rentrent en ligne de compte dans le calcul de la E -value : 1) la fréquence du facteur oublié (noté U) dans les 2 groupes ou, plus précisément, l'importance de la différence entre les deux groupes de cette fréquence (quantifiée par exemple par un risque ratio, noté RR_{EU}) et 2) la force de liaison de ce facteur avec le critère de jugement (noté D), c'est-à-dire de combien ce facteur multiplie la fréquence du critère de jugement (mesuré aussi par un risque ratio, noté RR_{UD}).

L'approche de l'E-value donne les valeurs minimales de ces deux paramètres que devrait avoir un facteur de confusion oublié pour invalider (nullifier) le résultat obtenu. Par exemple, si ces valeurs (RR_{EU}, RR_{UD}) sont égales à (2, 2), il faudrait que le facteur oublié soit 2 fois plus fréquent dans le groupe traité que dans le groupe contrôle et qu'il multiplie par 2 la fréquence du critère de jugement. Si ces valeurs sont réalistes, la robustesse du résultat obtenu est remise en cause, car il n'est pas suffisamment important en taille pour ne pas pouvoir provenir uniquement du facteur de confusion non pris en compte. En revanche, si l'E-value est trop élevée pour être réaliste, le résultat est robuste, il ne peut pas être entièrement expliqué par une confusion résiduelle. Le résultat permet de conclure à un effet non nul du traitement, même si la taille de l'effet peut être surestimée par le résultat produit (certaines méthodes permettent de corriger le résultat, mais nécessitent de faire des hypothèses numériques sur le ou les facteurs oubliés).

En réalité il n'y a pas qu'un seul couple de valeurs de (RR_{EU}, RR_{UD}) qui nullifie le résultat, mais une infinité. Un facteur oublié, très lié au critère de jugement, peut fortement biaiser le résultat même si sa distribution est peu déséquilibrée entre les 2 groupes, et, à l'inverse, un faible déterminant du critère de jugement peut tout autant biaiser le résultat en cas d'une grande asymétrie de sa fréquence entre les 2 groupes. Il existe ainsi une frontière de nullification dans le plan des paramètres (RR_{EU}, RR_{UD}) comme le représente la Figure 12.

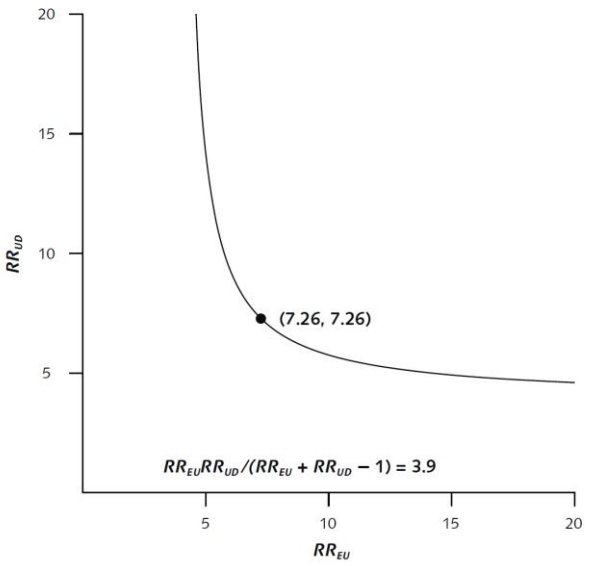
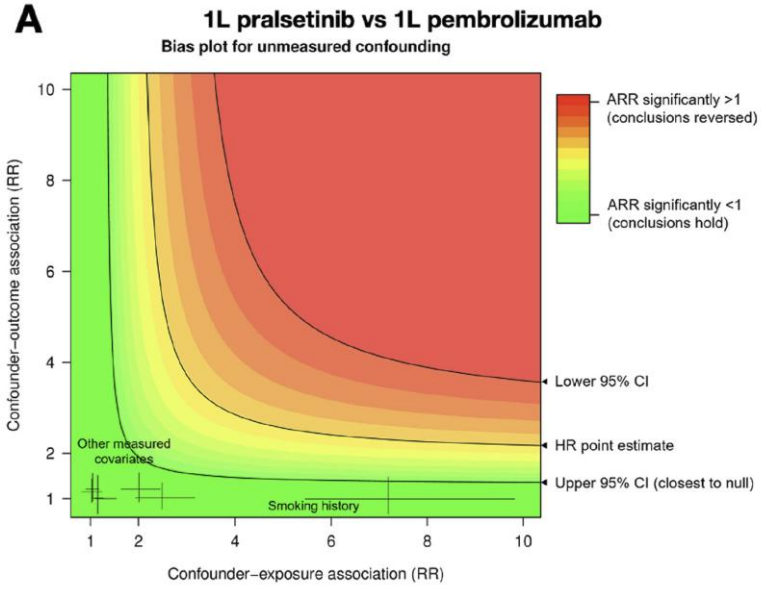


Figure 12 – La E-value est l'intersection avec la bissectrice de la courbe matérialisant la perte de l'association observée dans une étude en cas d'oubli d'un facteur de confusion lié avec le critère de jugement avec un risque relatif RR_{UD} (ordonnée) et déséquilibré entre les deux groupes suivant un rapport RR_{EU} (abscisse) (extraite de la référence [179]).

L'analyse de la robustesse du résultat s'effectue alors de la même façon, mais en appréciant globalement la plausibilité de toutes les valeurs délimitées par cette frontière de nullification.

Pour prendre en compte l'incertitude statistique, ces calculs ne doivent pas être réalisés pour chercher à nullifier l'estimation ponctuelle, mais bien la borne péjorative de l'intervalle de confiance du résultat obtenu par la comparaison indirecte.

Dans une comparaison externe du pralsetinib au pembrolizumab [175], le hazard ratio sur la survie globale (OS) est HR=0.36 [0.21, 0.64]. La figure 2A reproduite ci-dessous donne les résultats de l'analyse quantitative de biais pour les facteurs de confusions non pris en compte :



Sur ce schéma apparait la limite de nullification de l'estimation ponctuelle (HR point estimate) à partir de la laquelle sont déterminée les zones de couleur qui représente la valeur du risque ratio corrigé du biais (ARR adjusted RR). Cependant figure aussi la limite de nullification de la borne péjorative (la plus proche de l'absence d'effet, ici la borne supérieure, upper 95% CI) qui est la limite sur laquelle se joue la perte de la signification statistique. Sont aussi représentés (les croix) les facteurs de confusions pris en compte présenté par une croix centrée sur le RR de leur association avec le critère de jugement (en ordonnée) et le RR de leur déséquilibre entre les 2 groupes en abscisse. La largeur et hauteur de croix représentent les intervalles de confiance. Cette information permet d'apprécier l'ordre de grandeur de l'association avec le critère de jugement et celui du déséquilibre des facteurs connus et mesurés afin d'aider à l'appréciation de la plausibilité de l'existence de facteur correspondant à la limite de nullification. Il faut cependant noter que ces valeurs sont celles trouvées dans le jeu de données analysé. Il serait préférable qu'il s'agisse, pour la force de liaison avec le critère de jugement, d'estimations provenant de la littérature (des études dédiées à l'estimation des facteurs pronostiques et des facteurs de risques). Pour le déséquilibre, il n'est pas non plus certain que le niveau de déséquilibre des facteurs mesurés soit similaire à celui des covariables non mesurées.

Cette publication propose aussi d'autre analyse quantitative de biais pour apprécier les conséquences des données manquantes, de la fréquence des métastases cérébrales, et d'une éventuelle mauvaise performance du comparateur dans ces données de vraie vie.

17 Les biais de sélection

Le biais de sélection est un biais complexe. De plus, son nom est ambigu conduisant ce terme à être parfois utilisé de manière erronée pour désigner un biais de confusion ou pour faire référence à un défaut de représentativité de l'échantillon de l'étude (problème de validité externe).

Le terme biais de sélection ne fait pas référence à un défaut de représentativité des patients de l'étude

Dans une étude ou analyse comparative, un biais de sélection survient si la sélection des sujets ou la sélection des périodes observée dépend à la fois de l'outcome et du traitement. Cette sélection passe souvent par la non-inclusion de certains patients en fonction du traitement et du critère de jugement (par exemple dans le biais de sélection par déplétion des susceptibles) ou par la non-observation (non-inclusion dans le suivi) dépendant du traitement et du critère de jugement de certaines périodes de temps (comme dans les biais de sélection par temps d'immortalité [180] [181] ou le biais de sélection dû aux perdus de vue).

> *Par exemple, les patients perdus de vue introduisent un biais de sélection lorsqu'ils quittent l'étude en raison de l'évolution de la maladie (raison liée à l'outcome) et de façon différente entre les groupes (raison liée au traitement). Dans ce cas il y a exclusion d'un temps de suivi en fin d'étude qui dépend de l'outcome et du traitement.*

17.1 Déplétion des susceptibles

Un groupe contrôle où des patients traités de longue date seraient inclus sans que le suivi disponible enregistré dans la source de données remonte jusqu'à l'initiation de leur traitement peut induire un biais de sélection par déplétion des susceptibles (*outcome susceptible*). On parle dans ce cas de patients prévalents car le traitement est déjà instauré quand le patient commence à être suivi par l'étude. La définition de la baseline n'est donc la même pour les patients incidents et les patients prévalent.

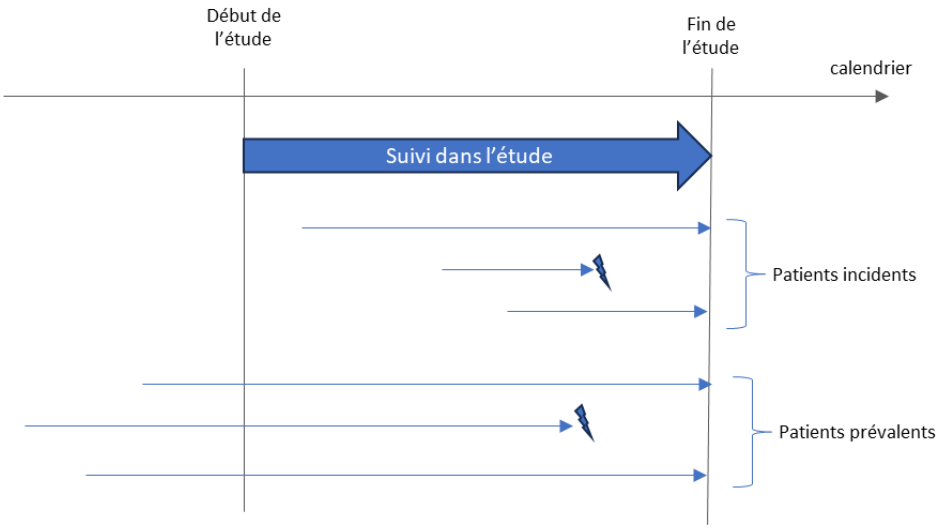


Figure 13 – Définition des patients prévalents et incidents
Le début des flèches représente l'initiation du traitement. Seuls les événements survenant durant le suivi dans l'étude seront enregistrés.

Par exemple, il pourrait apparaitre que le nouveau traitement induit plus d'effets indésirables que le traitement contrôle sans que cela soit vraiment le cas en réalité.

Avec des anticoagulants, par exemple, il pourrait faussement apparaitre que le nouveau traitement est plus pourvoyeur d'hémorragies que le contrôle. En effet, dans une étude monobras, le traitement des patients avec le nouveau traitement est initié dans l'étude. Les sujets sont donc suivis depuis l'initiation du traitement. On parle de patients incidents. Tout effet indésirable (toute hémorragie) survenant précocement chez ces patients est observé durant le suivi dans l'étude et colligé.

En revanche, dans un groupe contrôle de patients prévalents, le suivi ne débute pas à l'initiation du traitement, mais alors que les patients prennent déjà le traitement depuis un certain temps. Il se peut alors que des patients ayant initiés dans le passé le traitement contrôle aient présenté avec la même fréquence les mêmes effets indésirables (hémorragies) que les patients recevant le nouveau traitement. Cependant, la survenue de l'effet indésirable ayant mis fin au traitement, ces patients ne seront pas inclus dans le groupe contrôle car non traité au moment où le groupe contrôle sélectionne ses patients dans la source de données (fenêtre de sélection[21]). Ainsi en prenant des patients déjà traités, il y aura déplétion des patients susceptibles de faire un effet indésirable.

> *Ainsi ce phénomène de déplétion des susceptibles répond bien à la définition générale du biais de sélection (cf. supra) car l'exclusion de l'étude des patients susceptibles de faire l'événement dépend bien du traitement (elle n'a lieu que dans le groupe contrôle est constitué de patients avec un traitement prévalent) et de la survenue du critère de jugement. On peut expliquer ce biais en disant qu'il y a sélection de patients non susceptible de faire l'événement critère de jugement ne survient que dans le groupe contrôle et que cette sélection dépend du critère de jugement, mais de façon passive,*

[21] Dans un groupe contrôle externe acceptant des patients prévalents, les patients du groupe contrôle sont sélectionnés dans la source de données à partir du moment où ils reçoivent le traitement contrôle durant la fenêtre temporelle d'inclusion. Seuls des patients ayant débuté le traitement et ne l'ayant pas arrêté précocement pour effet indésirable arriveront toujours traités dans la fenêtre temporelle/calendaire d'inclusion.

indirecte, car ne sont considérés pour l'inclusion dans le groupe contrôle que des patients qui ont pu poursuivre sans encombre un traitement initié précédemment.

Pour éviter cela, le groupe contrôle ne doit être constitué que de patients initiant le traitement contrôle. Dans ce cas leur suivi débutera bien dès l'initiation du traitement et les effets indésirables précoces seront captés de la même manière qu'ils seraient captés dans une étude monobras. On parle de *new users design [182] [183] [184]*. Cette approche permet de raisonner sur une « *inception cohort* » où tous les participants sont recrutés à un point de départ commun dans l'évolution de leur pathologie, typiquement le tout début de la maladie ou de l'exposition étudiée. Ils sont ensuite suivis à partir de ce point zéro commun à tous les patients.

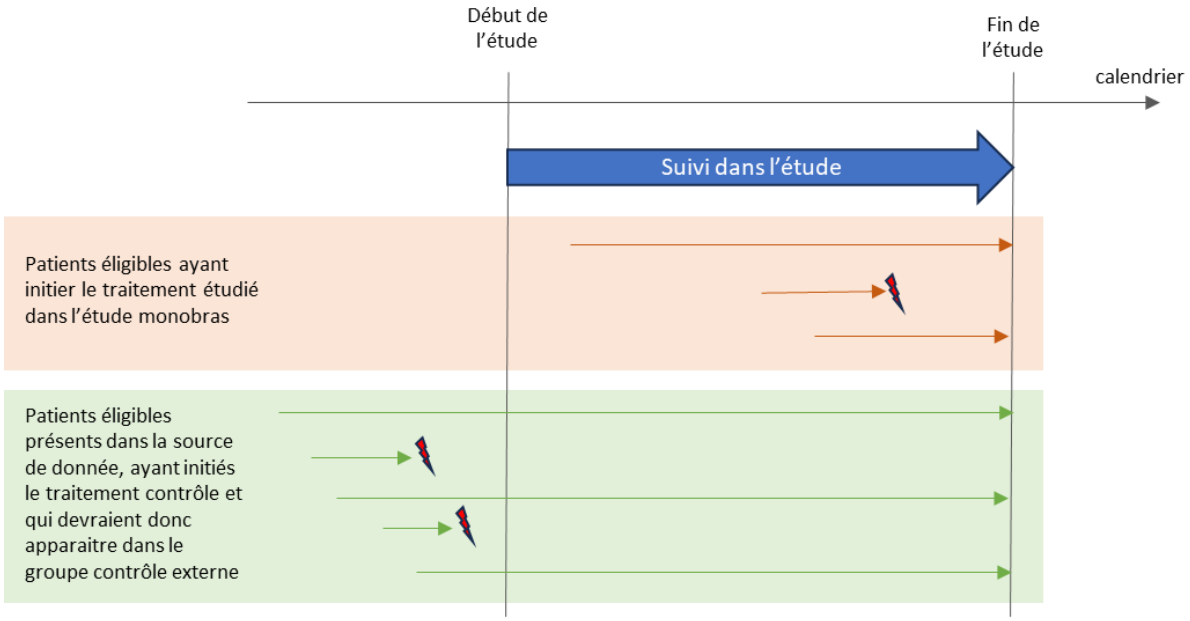


Figure 14 – Mécanisme du biais de sélection par déplétion des susceptibles.
Aucun événement n'est observé durant le suivi de l'étude dans le groupe contrôle composé de patients prévalant, car les patients traités avec le traitement contrôle et qui « devait » présenter l'événement l'on fait précocement, avant le début du suivi dans l'étude. Donc, parmi les patients ayant initier le traitement contrôle, seuls arrivent dans l'étude ceux qui « n'ont pas de susceptibilité » à faire l'événement. Il y a eu déplétion des susceptibles dans le groupe des patients ayant initiés le traitement contrôle avant que le suivi de ce groupe dans l'étude débute.

17.2 Biais lié à un défaut de synchronisation des t0

La question du temps zéro (t0) de début de suivi des patients est une question cruciale pour un groupe contrôle externe (voir aussi la section 18.1.1). De la justesse de sa définition et de sa similitude avec celui du groupe traité dépendra directement la validité des résultats produits. Sur ce point aussi, l'émulation d'un essai cible permet d'éviter les bévues [185].

Suissa et al. montre que dans une comparaissons à un groupe contrôle externe d'une étude monobras du blinatumomab dans la leucémie aigüe lymphoblastique réfractaire ou en rechute, le choix du t0 conduit à des résultats soit en faveur d'un bénéfice soit non concluants [186].

Dans les études expérimentales (essais randomisés, études monobras) le début du suivi (t0) correspond à un instant où :

1. Le patient est éligible (il vérifie les critères d'inclusion et d'exclusion) ;
2. Le traitement lui a été assigné (par la randomisation dans un essai randomisé ou par son inclusion dans une étude monobras) ;
3. Le suivi longitudinal d'observation du patient dans le but de recueillir les critères de jugement débute (sous forme de visites programmées).

Dans ces études, cet instant est parfaitement bien défini et parfaitement bien identifiable dans le fichier de données. Toute la difficulté va être de synchroniser le t0 du début du suivi des patients dans le groupe contrôle externe avec celui du groupe traité.

Cette synchronisation des t0 est indispensable pour éviter les biais, en particulier le biais par temps d'immortalité. L'exemple suivant illustre le mécanisme des biais qui peuvent être induits en cas de mauvaise synchronisation des t0 entre le groupe traité et le groupe contrôle externe.

> Dans une hémopathie maligne, un nouvel anticorps bispécifique a la même cible qu'un CART-T cells utilisé en standard. Le groupe traité est représenté par une étude monobras de cet anticorps bispécifique (il s'avérait impossible de financer un essai comparatif compte tenu du coût du CART T cells). Une comparaison externe avec un groupe contrôle recevant le CAR-T cells est envisagé.
> Le groupe contrôle est extrait d'un registre de patients ayant bénéficié d'un CAR-T cells. Une particularité du traitement par un CAR-T cell est qu'il s'écoule un délai entre la décision de recourir à ce traitement et le moment où le traitement est prêt pour être administré. Ce délai est le temps nécessaire pour réaliser la leucaphérèse et la transformation des cellules avant leur injection aux patients. Il s'avère que certains patients vont malheureusement décéder durant cet intervalle ou progresser ou ne plus être éligible à l'administration du traitement en raison d'une dégradation de leur état général. Il arrive aussi que la préparation des cellules échoue. Ces patients n'ayant pas reçu le traitement n'apparaitront pas dans le registre constitué de patients ayant reçu un CAR-T cells. Pourtant ces mêmes patients, s'ils avaient été inclus dans l'étude monobras, auraient été comptabilisés et leur devenir pris en compte pour les critères de jugement (survie sans progression, survie globale).
> Ainsi, la comparaison à ce registre introduira un biais lié au t0 (si le début du suivi ne commence pas au même moment dans les 2 groupes, des événements ne seront pas observés dans le groupe contrôle leur qu'ils l'auraient été dans le groupe traité) et à un temps d'immortalité si l'analyse fait débuter le suivi des 2 groupes au diagnostic afin d'avoir le même t0 dans les 2 groupes (dans ce cas aucun événement ne sera observé durant la première partie du suivi des patients CART-cell, correspondant au temps de préparation du traitement, comme si les patients étaient « immortels » initialement).
>
> Le recours à l'émulation d'un essai cible (cf. section 22) permet de comprendre ces 2 phénomènes. Dans un essai randomisé où l'anticorps bispécifique aurait été directement comparé au CAR-T cells, le suivi des patients aurait débuté dès la randomisation dans les 2 groupes et non pas à l'administration dans le bras CAR-T cells. Les progressions et décès avant administration auraient été pris en compte dans ce bras (tout comme les mêmes événements précoces survenant dans le bras bispécifique). De plus, l'essai cible aurait été analysé en intention de traiter et les patients alloués au bras CAR-T cell qui n'aurait pas pu recevoir l'injection du traitement aurait été maintenu dans le bras et l'analyse (avec leurs événements) tandis qu'ils ont tout bonnement disparu du registre et donc du bras contrôle externe.
>
> Pour éviter ces biais il serait nécessaire d'avoir dans la source de données servant à constituer le groupe contrôle tous les patients pour lesquels l'indication de CAR-T cells a été portée (qu'ils aient pu ou non-recevoir ensuite l'injection du traitement). Cela nécessite donc d'avoir un registre de pathologie plutôt que traitement, où les données des décisions de traitement sont collectées, par exemple, avec la présence des comptes rendus des réunions de concertation pluridisciplinaire. En pratique il s'avère fréquemment

que les informations indispensables pour émuler correctement un essai cible et éviter ces biais ne sont pas disponibles dans les sources de données, car ce besoin n'a pas été anticipé. Cela renvoie à la nécessité que le recueil des données dans les sources de données mobilisables pour ce type d'études soit construit après une réflexion concernant leurs usages (cf. section 20.6)

L'évaluation d'un nouveau traitement dans un essai randomisé de supériorité repose sur l'analyse en intention de traiter. En termes d'émulation cela nécessite que le suivi dans le groupe contrôle externe débute dès l'assignement du traitement au patient, effectué immédiatement après la vérification de l'éligibilité. L'événement le plus proche de cela dans les données observationnelles est la prescription du traitement par le médecin. Dans le cas de traitement non médicamenteux, cette prescription correspond à la décision (du médecin ou collégiale dans une RCP) de proposer une certaine stratégie thérapeutique (chirurgie, intervention, CAR T cell, etc.) au patient. La date de prescription (ou de décision de recourir à un certain traitement) permet ainsi une émulation satisfaisante de l'analyse en ITT.

Temps d'immortalité - définition

Le temps immortel est une période de suivi prise en considération dans l'analyse, mais où par construction un patient ne peut pas présenter l'événement d'intérêt, car les patients qui ont présenté l'événement durant cette période ont été exclus par construction de l'étude.

Par exemple dans le cadre d'un traitement en prévention primaire, la sélection dans le groupe traité de patient ayant eu au moins 6 mois de traitement entrainera qu'aucun événement ne pourra survenir durant ces 6 mois. En effet si un événement moins de 6 mois après l'instauration du traitement, celui-ci devient non éligible. Si bien que l'éligibilité dans ce groupe devient : avoir initié un traitement de prévention primaire et ne pas avoir fait d'événements d'intérêts avant 6 mois. Comme le suivi de ces patients débute à l'initiation du traitement, aucun événement ne surviendra durant les 6 premiers mois du suivi, ces 6 premiers mois seront un temps immortel.

NB : si la question clinique d'intérêt est une comparaison de durées de traitement, les groupes devront être constitués sur la base de la durée de la prescription initiale et non sur la durée effectuée. Si la prescription ne précise pas la durée anticipée, la réponse à cette question devra faire appel à une tout autre approche. [187]. À noter cependant que la question de l'évolution de l'effet traitement au cours du temps est une question exploratoire ne correspondant pas à un estimand causal d'évaluation de l'intérêt clinique de la prescription d'un traitement.

La disponibilité de cette date dans la source de données utilisée pour le groupe contrôle n'est pas garantie. Elle sera très certainement absence dans une base administrative où seulement la date de dispensation du traitement sera présente. Elle peut ne pas être présente aussi dans des sources de données orientées recherche comme des registres. Si les données sont extraites de dossiers médicaux, celle-ci pourra être extraite des dates des ordonnances, des lettres de liaison, de prise de rendez-vous auprès du service allant appliquer le traitement (chirurgie, radiothérapie, etc.) ou des comptes rendus de réunion de concertation pluridisciplinaires (RCP) lors de la phase d'abstraction/extraction des données (manuelle ou automatique, spécifique à l'étude ou non). La prise de conscience de ce point par les gestionnaires de données est indispensable pour augmenter l'utilisabilité des sources de données qui se constituent pour ce type d'étude (cf. section 20.6).

Il faut aussi que le suivi pour le groupe traité débute lui aussi au moment de la décision d'utiliser le nouveau traitement (date d'inclusion dans une monobras ou date de randomisation).

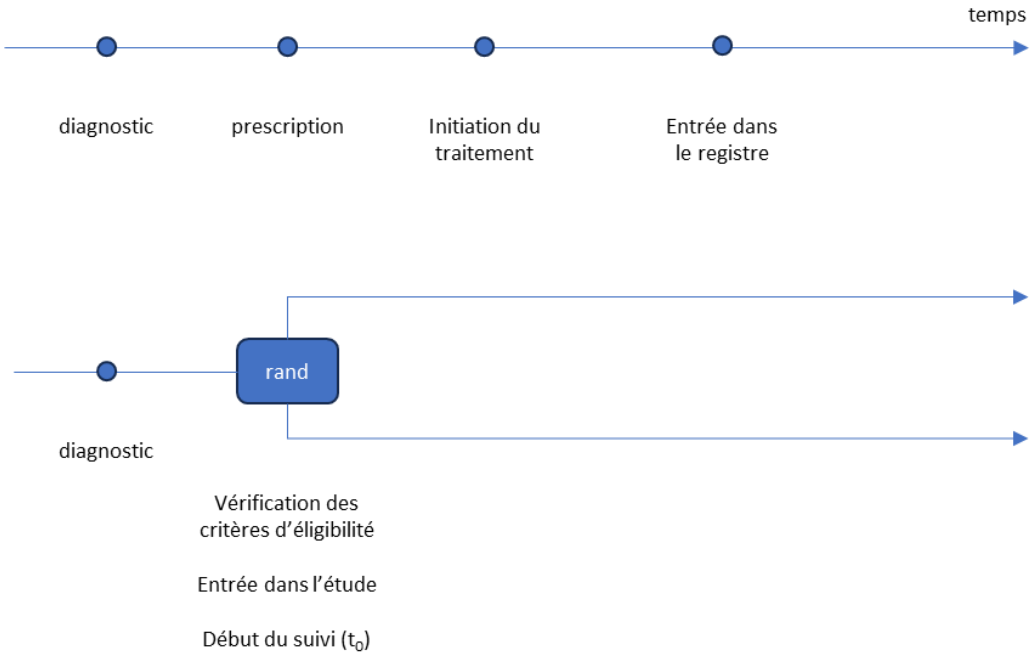


Figure 15 – Comparaison de la chronologie des événements entre un essai randomisé (en bas) et un recueil de données de routine (en haut).

Dans certains cas, la date de début de traitement peut être employée comme t0 du groupe contrôle sans trop de risque de biais de sélection, par exemple quand il est raisonnable de penser que le traitement a été débuté très peu de temps après la prescription et/ou quand il est peu probable que des événements critère de jugement puissent survenir dans cet intervalle.

Dans tous les cas, la réflexion par rapport à ce qui se passerait dans un essai randomisé équivalent (évaluation d'essai cible) réduit le risque de défaut de conception conduisant à un biais de ce type (cf. section 22).

Situations pouvant créer un biais de sélection :

- Exiger un minimum de suivi pour les patients du groupe contrôle (la sélection pourra dépendre de la survenue du critère ou d'une censure informative). Pour assurer un minimum de suivi, il est possible d'arrêter le recrutement historique des patients à une date garantissant un suivi minimum pour le dernier patient inclus (par exemple 365 jours avant la date de point retenue). Tous les patients ayant une date de début de suivi avant cette date de fin d'inclusion seront retenus, quelle que soit leur durée de suivi effectif (des perdus de vue et des suivis censurés doivent pouvoir survenir).
- Exiger un minimum de durée de traitement (pour des raisons identiques). Si la prescription est disponible, il est possible de ne retenir que des prescriptions d'une certaine durée (par

exemple au moins 12 mois d'anticoagulant pour le traitement d'un épisode aigu de thrombose veineuse profonde).

- Exclure des patients qui présenteraient des critères de non-inclusion (exclusion) après le début du suivi. Par exemple si l'étude monobras a inclus des patients diabétiques sans maladie rénale, il ne faudra pas retenir pour déterminer la présence de maladie rénale et donc l'inclusion du patient dans le groupe contrôle une information mesurée après le début du suivi potentiel du patient s'il été inclus dans le groupe contrôle. La tentation est grande, car il est difficile dans les données de vraie vie d'avoir la mention explicite de l'absence d'une situation morbide (comme une maladie rénale) et ne pas trouver la mention de l'absence d'une maladie ne signifie pas que le patient ne l'a pas. Ainsi si cette notion apparait à un moment donné dans le suivi du patient il est tentant de faire comme si cette information s'applique rétrospectivement au moment de l'inclusion du patient dans le groupe contrôle. D'où l'intérêt de représentation graphique des fenêtres temporelles utilisée pour définir la Baseline (cf. section 18.2).
- Débuter le suivi avant la date de début du traitement (cf. explication dans le texte ci-dessus).
- Faire débuter le suivi de certains patients à distance (postérieurement) au début du traitement.

La granularité temporelle doit être suffisamment fine pour éviter une *reverse causation* liée à une mesure de l'association de type transversale et non pas longitudinale. En effet, les dates des événements, des mesures, des examens ou des traitements ne sont parfois pas directement connues (colligées) mais dérivées des dates des comptes rendus de consultations, de visites pour les registres, d'hospitalisations, etc. Cette incertitude temporelle peut très bien conduire à considérer que deux informations sont séparées dans le temps alors, qu'en réalité, elles étaient synchrones, ou même, de temporalité inverse.

> *Un délai de grâce est parfois proposé comme solution à un éventuel biais par temps immortel dans les études observationnelles classiques. Cette solution n'est pas vraiment adaptée aux groupes contrôles externes, car elle conduirait à tronquer aussi le suivi du groupe traité issu d'une monobras ou d'un essai randomisé.*

17.3 Groupe contrôle externe non-traité

La détermination du t0 pour un groupe de patients non traité est difficile, car, par définition, rien ne définit le début du non-traitement.

Le moment où le patient vérifie les critères d'éligibilité, est souvent non unique offrant la possibilité d'avoir de nombreux t0 pour un même patient.

Cette situation survient par exemple avec les maladies rares d'évolution lente. Dans l'étude expérimentale (monobras) lorsque celle-ci est initiée, les patients sont recrutés dans une file active existante. Il n'y a pas nécessité d'attendre de nouveaux patients primo diagnostiqués. Tous les patients de la file active des centres investigateurs qui vérifient les critères d'éligibilité sont inclus. Il en résulte l'inclusion de patients qui sont à des moments différents dans leur histoire de la maladie (mais qui sont tous dans le même état clinique étant donné que la maladie est d'évolution lente).

Lorsqu'il s'agit de constituer un groupe contrôle, la difficulté va être de reproduire la même distribution de temps de début de suivi des patients en termes de temps dans l'histoire de leur maladie.

Ce point est crucial, car si c'est systématiquement le premier temps où sont remplis les critères d'éligibilité qui est utilisé comme t0, le groupe contrôle sera certainement constitué de patients étant plus précoces dans leur évolution naturelle de la maladie que le groupe traité et/ou ayant un suivi plus long [188] [189] [190] [191] [192].

17.4 Fin du suivi

La définition de la fin du suivi peut aussi entrainer un biais de sélection en excluant des périodes de suivi en fin de suivi en fonction du critère de jugement. Il s'agira de censures informatives du suivi liées par exemple à un changement de traitement. Comme le changement de traitement (comme l'arrêt du traitement initial) est potentiellement lié à une évolution de la maladie, censurer le suivi à cet instant introduit une censure informative. Si le groupe traité est suivi jusqu'à une date de point indépendamment des arrêts de traitement, cette asymétrie dans l'estimand utilisé introduira un biais de sélection. La prévention de ce biais nécessite que l'alignement de l'estimand de prise en compte des événements intercurrents entre les deux groupes. Là aussi l'émulation d'un essai cible permet d'éviter des erreurs de design ou d'analyse (cf. section 22). Si le groupe traité utilise l'estimand « policy treatment », avec un suivi systématique et un décompte de tous les événements jusqu'à une date de point arbitraire, un suivi du même type doit être réalisé dans le groupe contrôle.

Dans les groupes traités utilisés dans les comparaisons externes (études monobras ou bras traités d'un essai randomisé) la fin du suivi longitudinal des patients peut être définie principalement de 2 façons. Le suivi s'interrompt simultanément pour tous les patients à la même date calendaire, la date de point (data cut-off) ou le suivi est fixe et identique pour tous les patients, par exemple 52 semaines.

Dans les sources de données, la notion de fin de suivi n'existe pas (dans la quasi-majorité des cas, sauf s'il s'agit par exemple d'un précédent essai randomisé ou d'une précédente monobras). Dans ce cas plusieurs options se présentent.

Une première option est de prendre en compte la totalité du suivi disponible pour les patients sélectionnés pour le groupe contrôle externe. Cela conduira à des courbes de survie tracées sur un temps plus long pour le groupe contrôle que pour le groupe traité (cf. Figure 16).

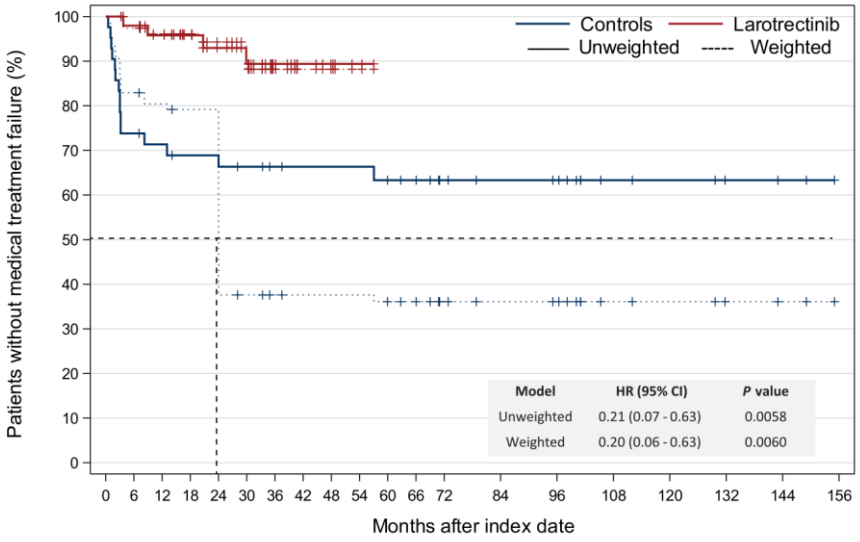


Figure 16 – Exemple de durée de suivi asymétrique entre le groupe traité et le groupe contrôle [193]

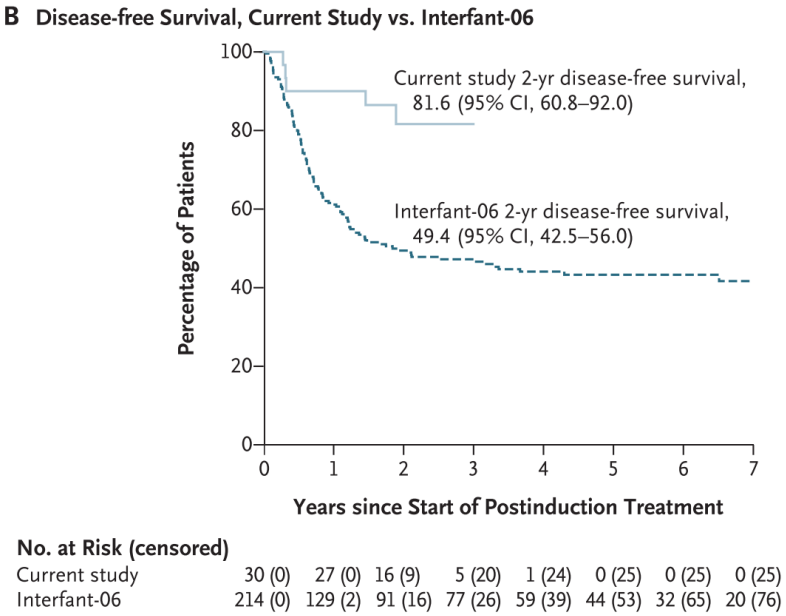
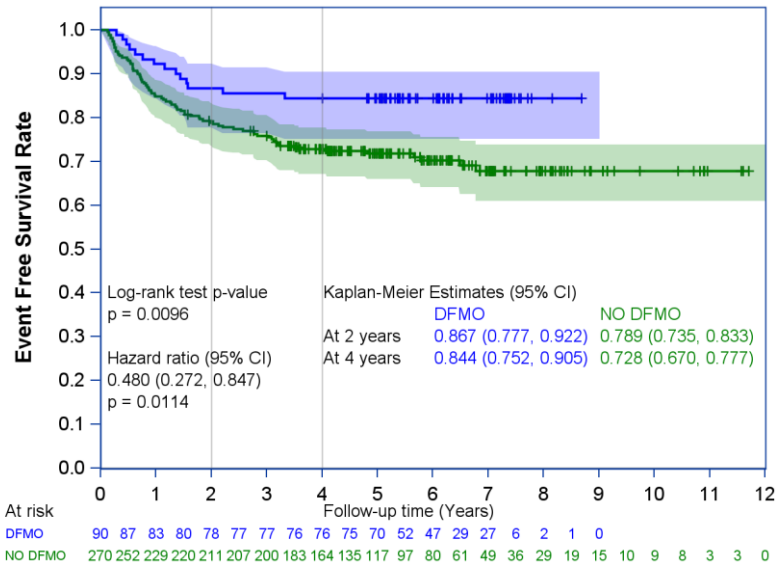


Figure 17 from ref [194]

Parfois le suivi dans le groupe contrôle pourra être inférieur à celui du groupe traité. Cela survient dans les domaines à renouvellement rapide des standards de soins (SOC) comme l'oncologie par exemple. Le SOC actuel auquel veut se comparer le nouveau produit peut être récent, quelques mois par exemple. De plus la source de données peut avoir un délai d'actualisation et ne pas contenir par exemple la dernière année écoulée. La conjonction de ces 2 points peut conduire à un suivi très court pour les patients et leur information disponibles pour le groupe contrôle externe, inférieur à celui de l'étude monobras.

Pour éviter cette différence de suivi entre les deux groupes, il peut être procédé à une troncature des suivis dans le groupe contrôle, en utilisant, par exemple, le suivi maximal du groupe traité. Dans ce cas la distribution des censures liées à la fin du suivi (censures liées à la date de point aussi appelée censures administratives) ne sera pas comparable entre les deux groupes. Dans le groupe traité, ces censures seront étalées dans le temps sur la fin des courbes compte tenu de l'étalement des inclusions dans l'étude d'origine. Dans le groupe contrôle externe, l'étalement de cette censure dépendra uniquement de la position du délai de troncature par rapport à la distribution des suivis disponible dans la source de données. Il est alors possible que tous les patients du groupe contrôle aient la même durée de suivi si le délai de troncature utilisé est inférieur au suivi le plus court disponible dans la source de donnée (et donc il n'existera pas d'étalement des censures de fin de suivi en fin de courbe). Quoi qu'il en soit cette asymétrie de distribution des censures ne pose pas de problème en terme statistique (en termes d'inférence) car ces censures sont non informatives.

Dans le dossier FDA d'évaluation de eflornithine dans les neuroblastomes pédiatriques à haut risque [195], il existe une grande asymétrie des durées de suivi entre l'étude monobras de l'eflornithine et le groupe control externe.



Des analyse de sensibilité ont é été faites avec une censure administrative réalisée à différentes échéances ([195] page 235).

Table 32. Analysis of EFS and OS using Administratively Censored Data – Propensity Score Matched (3:1)

	Event-free Survival			Overall Survival		
	DFMO Events/N	NO DFMO Events/N	HR (95% CI)	DFMO Events/N	NO DFMO Events/N	HR (95% CI)
Primary Analysis	14/90	79/270	0.48 (0.27, 0.85)	7/90	57/270	0.32 (0.15, 0.70)
Follow-up						
2-year administrative censoring	12/90	57/270	0.60 (0.32, 1.11)	1/90	19/270	0.16 (0.02, 1.16)
3-year administrative censoring	13/90	65/270	0.56 (0.31, 1.02)	4/90	34/270	0.34 (0.12, 0.95)
5-year administrative censoring	14/90	75/270	0.51 (0.29, 0.91)	6/90	47/270	0.34 (0.14, 0.79)

Pour les cas où le délai de suivi est fixe et identique pour tous les patients, l'émulation dans le groupe contrôle est simple. Se pose cependant le problème de l'irrégularité des temps de mesure du critère de jugement dans les données de vraie vie. La valeur la plus proche du délai voulu est alors utilisée si l'écart n'est pas trop grand (à fixer dans le protocole de l'étude). Cela entrainera une variabilité du temps de mesure pouvant avoir des conséquences en termes de pertinence clinique et des données manquantes sur le/les critères de jugement (si aucune mesure du critère de jugement proche du délai voulu n'est disponible).

17.5 Le biais de sélection vu en termes statistiques : censure à gauche, censure à droite

17.5.1 Censures à droite

Les censures à droite correspondent aux habituelles censures des analyses de survie (time to event analysis). Il existe deux types de censures : celles liées à la date de point (*data cutoff*), appelées aussi censures administratives, et les censures précoces liées à des interruptions prématurées du suivi du patient avant la date de point. Ces dernières censures correspondent donc à des patients perdus de vue.

Les censures de date de points n'entrainent pas de biais. Cependant les censures précoces sont susceptibles d'introduire un biais (le biais d'attrition). En effet ces censures sont potentiellement informatives, c'est-à-dire prédictives de la survenu du critère de jugement (ou liées à un facteur de risque du critère de jugement). Des censures asymétriques entre les 2 groupes entrainent un biais de sélection. En effet elles rendent inobservable par l'étude des périodes de suivi des patients et cela en étant associées au critère de jugement et au traitement.

Dans une comparaison externe, les censures informatives peuvent aussi bien survenir dans le groupe traité (monobras, RCT) que dans le groupe contrôle externe.

La problématique des censures informatives est très souvent négligée dans les études en général et en donc aussi dans les études de comparaison externe partant du principe que les techniques d'analyse de survie (type Kaplan Meier) prennent en compte les censures. Effectivement, intègre les suivis censurés, mais en faisant l'hypothèse que ces censures sont purement aléatoires, non informatives. Dans le cas de censures informatives, le résultat de l'analyse sera biaisé.

La meilleure façon de gérer les censures précoces est de les considérer comme informatives et de faire une analyse suivant un scénario du pire (*worst case*), par exemple sous l'hypothèse d'un biais maximum. Si sous ce scénario les résultats restent inchangés, ils pourront être considérés comme robustes vis-à-vis de ce problème et pourront être exploités pour la décision. Cette approche est donc une analyse quantitative de biais dédiée aux censures précoces (non dues à la date de point).

Une autre possibilité est de corriger les résultats par une technique de pondération similaire à celles mises en œuvre pour corriger du biais de confusion (IPW, IPTW). Cette approche est dénommée IPCW *inverse probability of censuring weighting*.

Elle consiste, dans un premier temps, à la construction d'un modèle explicatif des censures précoces à partir des caractéristiques des patients. Il s'agit de comprendre quels sont les facteurs qui augmentent la probabilité de survenue d'une censure précoce chez un patient. Ensuite les patients sont pondérés de telle façon que les patients similaires aux patients censurés, mais qui ont un suivi non censuré soient surreprésentés pour compenser les patients dont on ne connait pas le devenir [196].

Cela revient donc à gonfler les patients qui ont un suivi complet et qui ont un profil de risque de censure comparable aux patients effectivement censuré dont on ne connait pas le réel devenir. Cette approche fait l'hypothèse que le devenir inconnu des patients censurés est comparable à celui des patients ayant le même profil de risque de censure, mais qui ne l'on pas été.

L'IPCW corrige du biais induit par les censures informatives en les rendant indépendantes du pronostic (comme l'IPTW rend le traitement indépendant des covariables).

La performance de cette méthode repose sur la capacité à modéliser adéquatement le risque de censure. Il est nécessaire que les censures soient de type MAR (*missing at random*), ce qui implique qu'elles puissent être expliquées par des variables présentes dans la base de données. Cette condition n'étant pas toujours vérifiée, cette approche, bien qu'intéressante, ne permet pas de garantir l'élimination du biais introduit par des censures informatives.

Une autre difficulté propre aux comparaisons externes est qu'il est assez plausible que les mécanismes de censure informative soient différents entre le groupe expérimental et le groupe contrôle externe qui lui est observationnel. La modélisation des censures par un seul modèle ne pourra pas prendre cette hypothèse en compte. Il conviendrait donc peut être de modéliser indépendamment les censures dans les deux groupes, ce qui reste encore du champ de la recherche en méthodologie.

18 Identifications des patients dans la source de données

18.1 Aspects chronologiques

Dans un essai randomisé, l'inclusion et le suivi d'un patient s'effectuent suivant la chronologie suivante.

Schématiquement, l'éligibilité du patient est vérifiée lors d'une visite d'inclusion où sont relevées les caractéristiques de base du patient (baseline) et où le traitement est déterminé par la procédure de « randomisation ». Le traitement est alors débuté le plus rapidement possible après la « randomisation » et le patient suivi régulièrement jusqu'à une visite/date de fin d'étude. Cette fin d'étude est définie soit par un suivi fixe pour chaque patient (52 semaines par exemple) ou par une date de point identique pour tous les patients (et déterminée quand le nombre d'événements nécessaire pour garantir la puissance de la comparaison est atteint).

> *Bien sûr il existe des variantes comme une sélection s'étalant sur une certaine période pour vérifier la stabilité des critères d'éligibilité entre une visite de présélection et la visite d'inclusion. L'inclusion dans un essai peut être envisagée lors du diagnostic de la maladie, éventuellement en urgence (essai d'un traitement de l'AVC) ou lors d'un événement qualifiant émaillant l'évolution d'une maladie (comme une progression d'un cancer impliquant le passage à un traitement de ligne ultérieur). Dans le cas de maladie stable (maladie chronique comme le diabète, maladie rare stable) l'inclusion peut être envisagée de manière ponctuelle, sans définition chronologique particulière, lorsqu'une consultation de routine par exemple.*

Au niveau de la chronologie de l'essai, l'effectif nécessaire à l'essai n'est pas recruté instantanément, mais les inclusions des patients s'étalent au cours du temps depuis une date de début des inclusions jusqu'à l'obtention du nombre de sujets voulu par le protocole. L'étude se terminera ensuite soit après la fin de suivi du dernier inclus dans le cas d'une durée de suivi fixe pour tous les patients soit à une date de point déterminée de façon arbitraire ou lorsque le nombre d'événements nécessaire pour garantir la puissance de la comparaison est atteint (dans ce cas chaque patient aura une durée de suivi différente en fonction de sa date d'inclusion).

Le principe d'émulation d'un essai cible (cf. section 22) voudrait que l'on reproduise à l'identique ce schéma d'inclusion et de suivi dans la constitution et le suivi du groupe contrôle externe avec :

- Un point temporel clair où l'éligibilité est confirmée, les caractéristiques de baseline recueillies et le traitement initié.
- Un point temporel clair où le suivi et le recueil des critères de jugement sont arrêtés.

Cependant dans presque toutes les sources de données utilisables pour constituer un groupe contrôle externe, ces points temporels ne sont pas identifiables en tant que tels.

Par exemple dans un registre constitué à partir des consultations ou hospitalisation de routine, les valeurs des variables de baseline ne seront pas toutes mesurées à la consultation où un nouveau

traitement est initié. Il sera nécessaire de les reconstituer en cherchant des informations dans des consultations antérieures.

> Par exemple dans un cancer, un changement de traitement décidé devant une progression, les données de l'imagerie seront antérieures (comme dans un essai randomisé), mais le type exact de cancer, les comorbidités, etc. seront à rechercher dans des enregistrements de consultation précédente plus ou moins éloignée dans le passé.

Aussi bien pour la détermination de l'éligibilité que pour les valeurs de baseline et afin d'éviter les biais de sélection, il n'est pas possible de prendre des valeurs mesurées après le début de suivi/traitement du patient (cf. section 17).

18.1.1 Détermination du t0

Dans le groupe traité, où les patients sont inclus prospectivement, leur éligibilité est vérifiée à la date de la visite de sélection (des variantes existent, mais ne changent pas le schéma général décrit ici). Si le patient s'avère éligible et consent à être inclus, ses caractéristiques de baseline sont mesurées à cette visite. La date de la baseline est donc parfaitement bien définie. Souvent cette éligibilité et/ou les caractéristiques de base reposent sur une imagerie qui a pu être faite quelques jours au paravent, mais ces informations sont rattachées à cette date. La date d'inclusion, à laquelle est mesurée la baseline est donc clairement définie.

Dans la constitution d'un groupe contrôle (cohorte) à partir de données historiques, la situation est plus complexe, car il n'y a pas de visite d'inclusion [188] [189].

Pour reproduire au mieux ce qui se passe dans un essai randomisé, le suivi devrait débuter à la date où le patient vérifie les critères d'éligibilité et reçoit une prescription[22] du traitement d'intérêt [197]. Dans beaucoup de sources de données, la notion de prescription n'est pas disponible et elle sera remplacé par la dispensation ou le début du traitement. Cette approximation peut cependant introduire un biais de sélection par temps d'immortalité dans certaines situations (cf. section 17.2).

> Dans une étude monobras, des patients hospitalisés pour une exacerbation de BPCO reçoivent, à la sortie de l'hôpital, un nouveau traitement dont le but est de prévenir les réhospitalisations pour récidive. Pour comparer ce nouveau traitement aux corticoïdes inhalés, un groupe contrôle externe est constitué à partir d'une base de données administrative comprenant les dispensations. Sont inclus dans ce groupe contrôle externe des patients qui ont une dispensation de corticoïdes inhalés à la suite d'une hospitalisation pour BPCO. Ce design entrainera certainement un biais de sélection, car les patients pourront mettre quelques jours après leur sortie de l'hôpital pour aller chercher leur traitement à la pharmacie. Durant ce délai certains d'entre eux pourront être réhospitalisés pour des rechutes très précoces et ne seront pas inclus dans le groupe contrôle, car n'ayant pas eu de dispensations de corticoïdes inhalés bien qu'ayant eu une telle prescription sur leur ordonnance de sortie. Par contre, dans la monobras, les mêmes patients rechutant rapidement seront des comptabilités dans le groupe traité.

Parfois ces deux critères (vérifier l'éligibilité et être assigné à un traitement) se vérifient à plusieurs moments (même si dans le cas des groupes contrôles externes d'une étude expérimentale cette

[22] La notion de prescription correspond à celle d'être assigné à un traitement par la randomisation.

situation est moins fréquente qu'avec les études observationnelles classiques en raison de la définition d'un estimand plus strict).

> *Dans une étude monobras en oncologie, les patients étaient éligibles quelle que soit la ligne de leur traitement. Dans cette étude se côtoient donc des patients de première ligne (L1), de deuxième ligne (L2) ou plus (L2+). Une comparaison externe versus traitement standard est réalisée. Dans la source de données utilisée pour constituer ce groupe contrôle externe certains patients sont suivis depuis leur prise en charge initiale en L1 et l'historique de leur ligne successive est aussi enregistré. Ces patients vont présenter à plusieurs reprises le doublet : vérification des critères d'éligibilité et débuter un traitement (lors de leur L1 puis lors de leur passage en L2, L3, etc.). En d'autres termes ils pourraient servir de contrôle plusieurs fois : pour un patient inclus dans la monobras en L1, pour un patient inclus en L2, etc.*

Différentes options sont possibles pour ces situations : tirage au sort parmi ces multiples possibilités, inclure plusieurs fois le même patient dans le groupe contrôle en tenant compte de cette réplication au niveau statistique, réalisation d'émulation séquentielle de plusieurs essais puis pooling de ces essais [198] [199] [185], clonage de ces patients puis prise en compte de la réplication engendrée ou time-conditional PS.

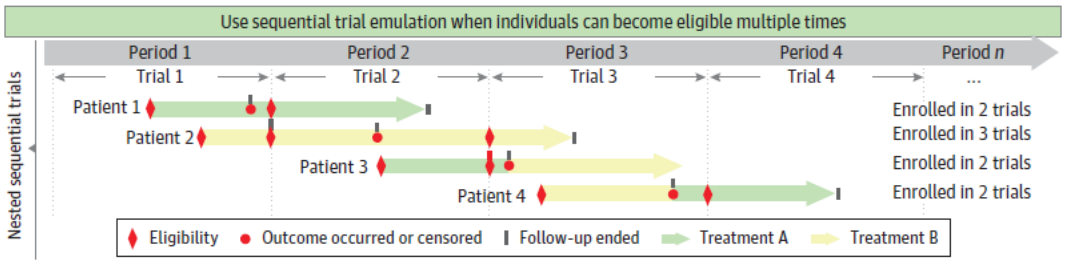


Figure 18 – principe de l'émulation séquentielle d'essais [200].

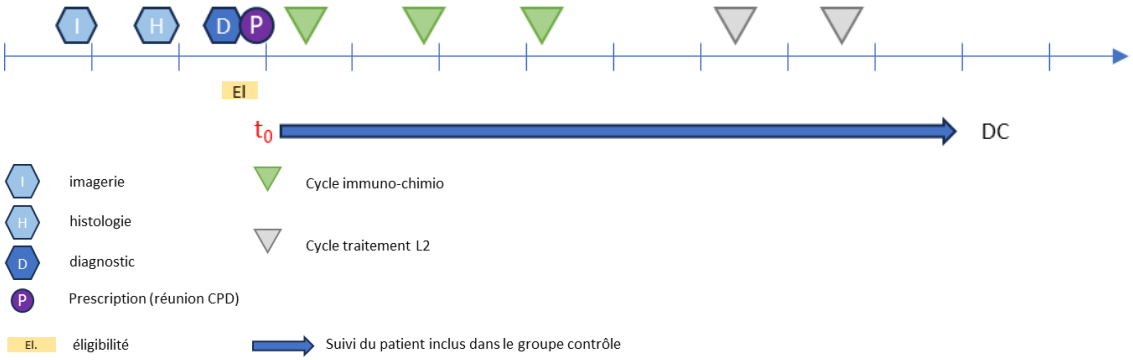
Pour les **groupes contrôles non traités** , le suivi débutera à partir du moment où le patient vérifie les critères d'éligibilité et ne reçoit pas de traitement (avec une tolérance temporelle pour prendre en compte des délais en prescription et dispensation par exemple pour capter une prescription ou un début de traitement à cet instant alors que l'information apparaitra à une date légèrement différée du fait du temps de latence lié par exemple à la dispensation : temps pour le patient d'aller à la pharmacie pour récupérer son traitement par exemple).

> *Pour ces groupes contrôles non traités, il est indispensable que l'éligibilité soit déterminée à un instant donné sans tenir compte de ce qui se passera dans le futur de cet instant pour éviter le risque de biais de sélection (cf. section 17). Pour cela il est dangereux de ne retenir que des patients qui ne recevront jamais le traitement durant la période où il serait éligible. Si un patient reçoit le traitement au-delà de l'instant où il vérifie les critères d'éligibilité pour être dans un groupe non traité, cela ne doit pas entrainer son exclusion de ce groupe. En termes d'émulation cette situation correspondra à celle du recours à un traitement de secours dans un essai randomisé ou à un écart au protocole ou un retrait volontaire du patient de l'essai. Ainsi l'inclusion dans le groupe contrôle externe non traité d'un patient*

qui recevra après le t0 le traitement étudié est conforme à ce qui peut se passer dans le groupe contrôle d'un essai randomisé. Cela conduira donc à des mises sous traitement dans le groupe contrôle externe tout comme cela arrive dans un essai clinique où ces patients sont bien entendu maintenu dans l'analyse en intention de traité

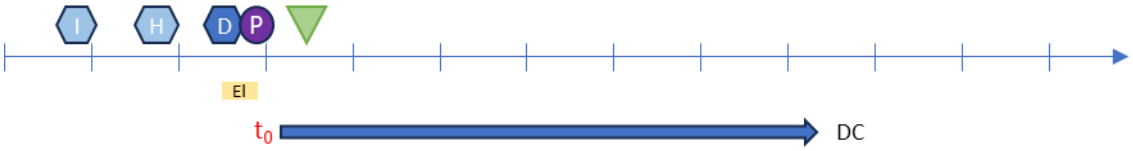
Un essai monobras d'un nouveau traitement IV inclus des patients avec un cancer du poumon non à petites cellules nouvellement diagnostiqués (non traité auparavant). Les patients sont inclus dans l'étude monobras (t0) quant à une consultation (visite 0) le diagnostic de certitude peut être porté sur la base de l'imagerie et de l'histologie et que les autres critères d'éligibilité sont remplis. Et le traitement est débuté dès qu'un rendez-vous en hôpital de jour est obtenu (ne devant pas excéder 1mois d'attente).
Le groupe contrôle de patients traités par l'immunochimiothérapie standard de ce cancer issu d'un registre de la pathologie.

Dans la vraie vie, le diagnostic (D) est porté lors d'une consultation lorsque les résultats de l'imagerie (I) et de l'histologie (H) sont disponibles. Le traitement est décidé (prescription P) rapidement après lors d'une réunion de concertation pluridisciplinaire et le premier cycle administré quelques jours après.

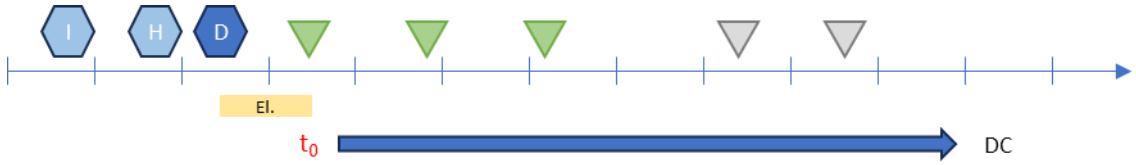


Le t0 correspondra à la date de cette réunion de concertation (P) qui coïncide avec le fait que le patient est devenu éligible (El.) depuis la disponibilité de l'imagerie et de l'histoogie, et qu'il est assigné au traitement contrôle (prescription/réunion de concertation). Il se peut d'ailleurs que le diagnostic soit établi que lors de cette réunion et donc établissement de l'éligibilité et assignation au traitement coïncide.

Le suivi se poursuivra jusqu'au décès du patient ou à la date de point que le traitement soit poursuivi, interrompue, changer ou que le patient progresse et passe en deuxième ligne (analyse en intention de traiter, correspondant à un estimand du type : « effet d'être assigné en première ligne à un traitement de type immunochimiothérapie ».



Si la notion de prescription ne transparait pas dans la source de données, le t0 sera la date de l'administration du premier cycle du traitement.

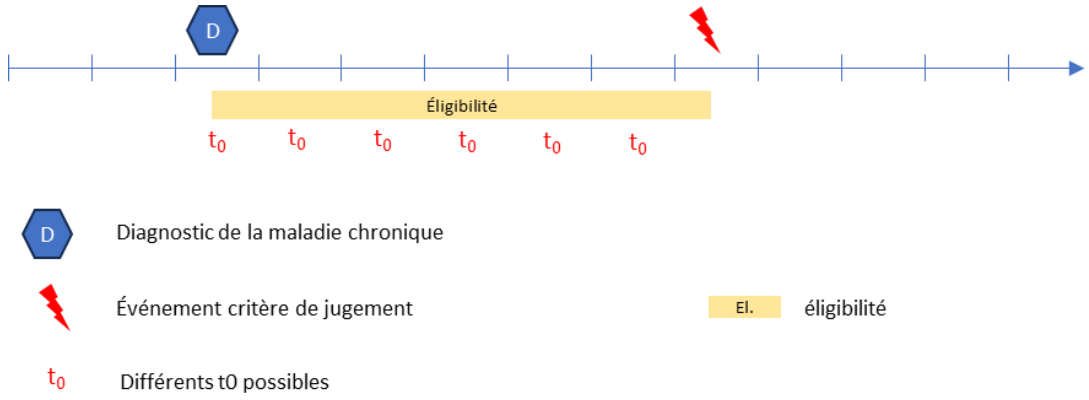


- Cas particulier du groupe contrôle non traité

Certaines comparaisons externes impliquent un groupe contrôle non traité si aucun traitement n'est disponible pour l'indication visée par le nouveau traitement. C'est par exemple le cas dans les maladies rares lorsque la comparaison se veut versus l'histoire naturelle de la maladie.

Dans le cas du traitement d'un événement aigu, la date de ce dernier servira de t0. Mais pour les maladies chroniques, où le traitement étudié est débuté durant la maladie, indépendamment d'un événement aigu, se pose alors la question de déterminer la date de début du non-traitement.

Dans cette situation, l'éligibilité est vérifiée à de nombreux temps sans qu'il soit possible d'en privilégier un. De nombreuses options sont possibles [201] [189] [188] [192] [191] [197].



Une approche consiste à émuler une série séquentielle d'essais, par exemple tous les mois de la fenêtre temporelle considérée, ce qui permet de prendre en compte tous les temps durant lesquels un patient est éligible [202] [203]. Ensuite tous les essais émulés générés sont poolés ensemble pour donner le résultat de l'étude. Un même patient est pourra donc participer à plusieurs essais émulés avec plusieurs t0. Cette redondance des patients doit être prise en compte dans l'analyse statistique. L'avantage de cette méthode, qui semble s'imposer par rapport aux autres, est d'éviter de choisir un t0 parmi plusieurs possibles.

18.1.2 Fin de suivi

L'émulation de la fin du suivi est en général plus directe. En cas de suivi fixe, la date de fin d'étude est clairement définie pour chaque patient en fonction de sa date de début de suivi/traitement. Dans un registre elle correspondra à la première consultation après cette date théorique (éventuellement avant comme c'est souvent aussi possible dans un suivi prospectif. À ce niveau les mêmes tolérances que celles de la monobras pourront être utilisées. Pour un suivi à date de point, l'émulation procédera de la même façon, mais avec une date de fin d'étude identique pour tous les patients.

Plusieurs choix sont possibles pour la fixation de la date de point de l'étude. La date de dernière intégration de données dans la source de données peut être utilisée. Elle a l'avantage d'exploiter au maximum les suivis des patients retenus. Elle peut aussi être fixée quand un certain nombre d'événements est atteint (cf. section sur le calcul de puissance section 25).

- Captation des patients perdus de vue

Il est important que la définition du suivi puisse faire apparaitre les patients perdus de vue. En effet, ces études utilisent souvent tout le suivi disponible pour un patient dans la source de données, c'est-à-dire jusqu'à l'événement critère de jugement ou la date de dernières nouvelles disponibles sans définir une date de fin de suivi théorique/attendu par patient. En procédant ainsi, les données paraissent parfaites (sans censures liées à des suivis interrompus précocement, donc sans perdus de vue) et contrastent avec les données du groupe traité où il peut exister de telles censures. Dans ce cas il ne sera pas possible de s'interroger sur la possibilité d'un biais d'attrition de fait de suivi censuré précocement et potentiellement informatif dans le groupe contrôle et qui existe en réalité.

> Par exemple le patient est décédé à domicile ou dans un autre hôpital/service et le service/médecin assurant le remplissage du registre n'a pas eu l'information ou celle-ci n'a pas été saisie, car la saisie ne s'effectuant que l'on des consultations.

La sélection des patients sur la base de la disponibilité d'un suivi minimal expose à un biais de sélection, car l'inclusion/non inclusion dans le groupe contrôle pourra dépendre de la survenu du critère de jugement (cf. Figure 19). En effet un patient n'ayant pas le suivi minimal voulu peut être un patient perdu de vue rapidement après le début de son traitement. L'exclure conduira à méconnaitre le fait qu'il y a un patient perdu de vue pouvant conduire à sous-estimer la fréquence/taux de l'événement dans le groupe contrôle. L'inclure ne permettra pas de savoir s'il a fait l'événement ou pas (il est perdu de vue de toute façon) mais alertera sur le fait qu'il existe des patients perdus de vue pouvant biaiser le résultat de la comparaison. Le biais sera plutôt conservateur sauf si l'objectif et de ne pas mettre en évidence de différence (non-infériorité, safety).

Pour éviter cette problématique, les patients inclus dans le groupe contrôle doivent être identifiés uniquement sur le fait qu'ils vérifient les critères d'éligibilité dans une fenêtre temporelle prédéfinie et qui leur assure un suivi minimal compte tenu de la date de point retenue pour l'analyse (ou la date d'extraction des données de la source ou la date de dernière mise à jour de la source de données). Cette procédure (cf. Figure 19) permettra d'inclure des patients indépendamment de leur durée de suivi et inclura par exemple des patients dont le suivi s'est terminé prématurément avant la date de point retenue. Elle permet de recréer dans l'histoire passée des patients de la source de données le même processus d'inclusion et de suivi que dans une étude prospective monobras ou essai randomisé.

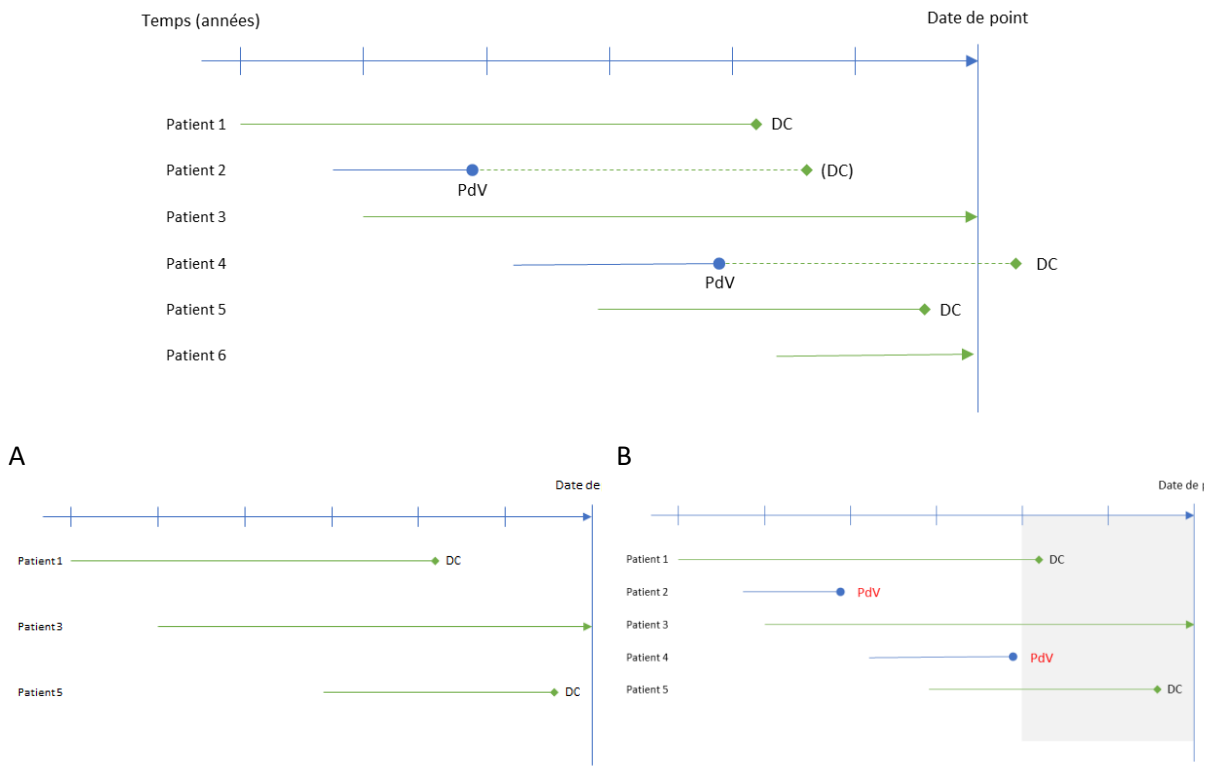


Figure 19 – Problèmes induits par une sélection des patients sur la durée de suivi et solution à adopter

La sous-figure A illustre une sélection des patients qui ont au moins 2 ans de suivi (durée considérée comme nécessaire pour voir apparaitre les événements que le traitement cherche à éviter). Dans le jeu de données (figure du haut) il y a deux patients qui sont perdus de vue moins de 2 ans après le début de traitement. Ils ne sont donc pas sélectionnés et l'étude ne contient aucun perdu de vue, faisant croire à tort qu'elle est exempte de ce problème.

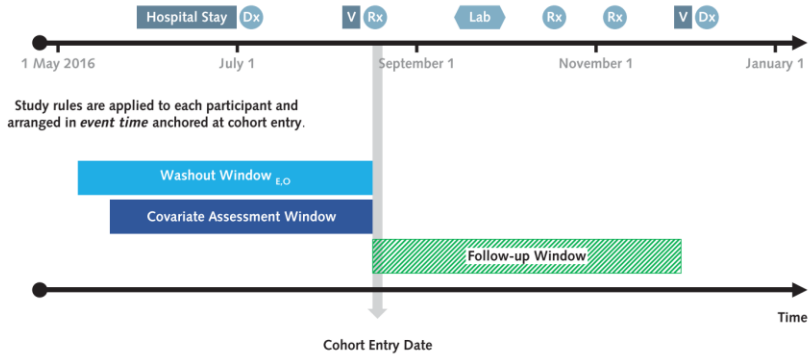
Dans la sous-figure B, les patients n'ont pas été sélectionnés sur la durée effective de leur suivi disponible dans la source de données. Seuls ont été exclus les patients dont le traitement avait débuté moins de 2 ans avant la date de point (de ce fait ces patients ne pouvaient pas avoir la durée minimale de suivi permettant de voir apparaitre les événements). En revanche, les patients perdus de vue ont tous débuté leur traitement plus de deux ans avant la date de point. Ils sont donc inclus. Il apparait alors qu'il y a deux perdus de vue pouvant entrainer un biais.

Une fois la date de point fixée, il est possible de trouver la date de fin de la fenêtre d'inclusion qui est cette date moins la durée minimale de suivi que l'on souhaite avoir (par exemple la même que dans le groupe traité). Ensuite il est possible de remonter dans le temps jusqu'à ce que le nombre de patients éligibles soit trouvé.

18.2 Représentation graphique du processus d'extraction des données de l'étude

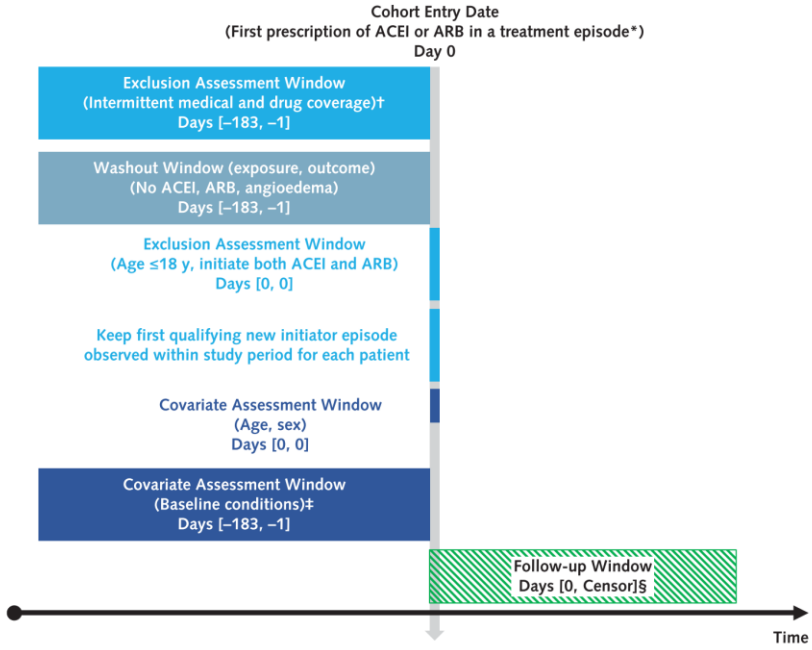
La reconstruction d'un suivi longitudinal de patients à partir de données historiques (bases de données) est un processus complexe. Une représentation graphique permettant de représenter cette complexité a été récemment proposée [204] et devient un standard dans les études de ce type. Bien que nous spécifiques des études de comparaison à un contrôle externe, cette représentation s'applique parfaitement à la constitution du groupe contrôle à partir de la (ou des) source de données choisies.

Figure 1. From transactional data to study implementation.



Individual patient data are documented as encounters from various sources and are arranged in calendar time. This work is licensed under CC BY, and the original versions can be found at www.repeatinitiative.org/projects.html. Dx = diagnosis; E = exposure; Lab = laboratory test; O = outcome; Rx = drug dispensing; V = visit.

Figure 3. Exposure-based cohort entry where the cohort entry date is selected after application of exclusion criteria.



This work is licensed under CC BY, and the original versions can be found at www.repeatinitiative.org/projects.html. ACEI = angiotensin-converting enzyme inhibitor; ARB = angiotensin-receptor blocker.
* Treatment episodes were defined by date of dispensing and days' supply with a stockpiling algorithm if a new dispensing occurred before the end of days' supply. Gaps of <30 d between end of days' supply and next dispensing were bridged. 30 d was added to the last dispensing days' supply in an exposure episode.
† Up to 45-d gaps in medical or pharmacy enrollment were allowed.
‡ Baseline conditions included allergic reactions, diabetes, heart failure, ischemic heart disease, and use of nonsteroidal anti-inflammatory drugs.
§ Earliest of outcome of interest (angioedema), switching or withdrawing study drugs, death, disenrollment, 365 d of follow-up, or end of study period.

19 Biais liés aux données

Les données peuvent être à l'origine de deux biais :

- Le biais de classification des expositions
- Le biais de classification/mesure des critères de jugement

Au premier abord, les données semblent aussi contribuer aux biais liés aux données manquantes, mais le mécanisme de ce biais est de nature différente par rapport aux deux précédents et conduit à traiter ce biais de façon séparée (section 17).

> *Dans les études observationnelles classiques, les erreurs de classification (ou de mesure) peuvent être symétriques (affectant de la même façon les 2 groupes, appelés aussi erreurs non différentielles) ou asymétriques (survenant de manière différente entre le groupe traité et le groupe contrôle, aussi appelée erreurs différentielles).*
> *Cette distinction est importante, car les erreurs symétriques ne peuvent pas créer un biais dans une conclusion de supériorité (différence entre les 2 groupes). Seule l'erreur asymétrique peut faire conclure à tort à la supériorité du nouveau traitement. En revanche, les erreurs symétriques peuvent biaiser les résultats vers l'absence de différence (*bias toward the null*) et faire conclure à tort à l'absence d'effet ou à une équivalence de traitement, conclusion à tort qui peut aussi provenir d'erreurs asymétriques.*
> *Les erreurs symétriques induisent aussi un bruit de fond aléatoire pénalisant pour la précision des estimations et la puissance des comparaisons.*

Contrairement aux études purement observationnelles où souvent les erreurs de classification affectent les 2 groupes, les erreurs de classification/mesure dans les comparaisons externes affectent surtout le groupe contrôle, de nature observationnelle, et peu ou pas le groupe traité, de nature expérimentale. Dans ce cas, les erreurs de classification purement aléatoire du groupe contrôle deviennent des erreurs asymétriques (contrairement à ce qui se passerait dans une étude observationnelle classique).

Cette différence avec les études observationnelles « classiques » est fondamentale et montre que les comparaisons à un groupe contrôle externe sont bien plus exposées aux biais de classification/mesure ce qui ajoute aux limites méthodologiques de ces études.

19.1.1 Biais de classification du critère de jugement

Le biais de classification du critère de jugement (biais de mesure) provient des erreurs de mesure sur le critère de jugement : erreur de mesure pour les critères continue ou erreur de classification pour les critères binaires types d'événements cliniques [205].

Pour les critères de jugement binaires, ces erreurs de classification sont soit du type « faux positif » : le patient est considéré à tort comme ayant présenté le critère de jugement ou soit du type « faux négatif » : le patient a réellement présenté l'événement, mais cela n'est pas enregistré dans les données utilisées.

Si le critère de jugement est une variable catégorielle à plus de deux modalités (et non pas une variable binaire) la situation est similaire, mais plus complexe.
Pour les critères de jugement continu il s'agit d'une erreur de mesure quantitative.

- Erreur de classification dans les études monobras

Les erreurs de mesure ou de classification du critère de jugement peuvent affecter aussi bien le groupe contrôle externe que la partie expérimentale. Dans une étude monobras, seule rentre en ligne de compte la valeur du critère de jugement.

Si ce critère a valeur de succès du traitement (comme la réponse objective en oncologie) il serait souhaitable, du point de vue du traitement, que ce critère survienne le plus fréquemment possible. Comme les études monobras ne sont pas en aveugle du traitement, on peut ainsi craindre que le processus de détermination de ce critère maximise la survenue de ce succès (sciemment ou subjectivement, au niveau de la définition ou de sa mise en application), introduisant des faux positifs. L'utilisation d'un comité central d'adjudication pour faire cette classification n'apporte pas de garantie supplémentaire, car le comité ne sera pas en aveugle (sauf si les cas d'une étude monobras sont adjudiqués en les mélangeant avec ceux d'autres études comparatives et si le comité est en aveugle de l'étude, ce qui est très rarement mis en œuvre). Cette situation conduira à une erreur de mesure asymétrique favorisant le groupe traité.

> Exemple Le dossier dévaluation par la FDA de l'eflornithine pour les neuroblastome pédiatriques à haut risque [195] analyse une comparassions externe réalisée à partir de l'essai monobras de l'eflornithine et un groupe contrôle externe historique. Une relecture en double aveugle de l'imagerie de l'étude monobras a été réalisée. La FDA regrette que cette relecture n'a pas été possible pour le groupe contrôle externe dans sa discussion des biais de la comparaison externe au niveau des critères de jugement (page 201 du document).

Si le critère de jugement est un critère d'échec, la situation serait l'inverse avec une tentation à minimiser le nombre d'échecs du traitement par des erreurs de classification de type faux négatifs.

- Différence de définition

L'erreur de mesure ou de classification peut aussi ne pas provenir d'une réelle erreur de mesure, mais être la conséquence d'une différence de définition entre la partie expérimentale et le groupe contrôle externe (situation assez fréquente). Dans ce cas les erreurs de classification seront forcément asymétriques entre les 2 groupes.

Dans beaucoup de cas, la situation peut mélanger plusieurs problématiques sur la mesure ou la classification des critères de jugement : différence de définition, de méthode de mesure ou de diagnostic, etc. et il est très difficile d'anticiper la direction du biais. C'est par exemple le cas en oncologie avec la progression tumorale qui n'est pas défini, mesuré et évalué de la même façon dans la vraie vie et dans les essais cliniques (cf. section 20.7).

- Prévention des biais

La prévention de ces biais nécessite que le processus de mesure ou de classification du critère de jugement du groupe contrôle soit le plus possible aligné avec celui de la partie expérimentale au niveau de la définition du critère de jugement, de la méthode de mesure ou de diagnostic, etc. Ce qui en pratique est parfois impossible à obtenir par exemple si le critère n'est pas employé en vraie vie (comme l'échelle ADAS-cog avec la maladie l'Alzheimer, ou la PFS en oncologie).

Dans les études observationnelles traditionnelles, l'impossibilité d'avoir avec les données de vraie vie les mêmes critères que ceux des essais cliniques est moins problématique et

peut être gérée avec l'utilisation de proxy, c'est-à-dire de critère qui ne reflète qu'approximativement et de façon indirecte le critère non disponible dans les données de vraie vie. Comme dans ces études, le même proxy est utilisé dans les deux groupes comparés, cette approximation peut ne pas être trop gênante (voir négligeable), car elle est identique dans les deux groupes comparés. Cette problématique est donc moins cruciale dans ces études classiques que dans les comparaisons externes où elles peuvent être insolubles.

La préplanification de la comparaison externe (cf. *externally controlled study* section 3) au moment de la conception de l'étude monobras offre plus de flexibilité pour prévenir ces biais que la conception a posteriori de cette comparaison externe. En effet, il est alors possible de choisir comme critère de jugement de l'étude monobras un critère qui sera émulable avec les données de vraie vie choisies, à condition que ce critère garde un sens par rapport au besoin d'évaluation (compatible avec les critères demandés par les agences dans la pathologie considérée).

La différence de critère de jugement entre la partie expérimentale et le groupe contrôle externe contribue à l'effet étude irréductible existant dans ces comparaisons comme l'identifie l'analyse en inférence causale de ces comparaisons (cf. section 13.3.1).

L'analyse quantitative de biais permettra a posteriori d'éprouver la robustesse des résultats obtenus vis-à-vis de différentes hypothèses/scénarios d'erreur de mesure ou de classification cf. section 23).

19.1.2 Le biais de classification de l'exposition

Le biais de classification de l'exposition provient des erreurs d'identification du traitement reçu par les patients. Dans les études observationnelles classiques, il s'agit d'une erreur dans la classification binaire : patients recevant le traitement étudié / patients recevant le traitement contrôle.

Au niveau des études de comparaisons externes, les erreurs de classification ne concernent que le groupe contrôle externe. Il n'est pas attendu d'erreurs au niveau de la partie expérimentale (monobras ou groupé traité d'un essai randomisé). Par conséquent les erreurs de classification de l'exposition dans les comparaisons externes sont asymétriques.

La classification (étiquetage) à tort de patients recevant le traitement étudié comme patients contrôles entrainera un biais conduisant à ne pas mettre en évidence la différence existante pourtant entre les 2 traitements. Ce cas de figure est peu probable, car si le traitement étudié est un nouveau traitement non commercialisé les patients ne peuvent pas le recevoir dans la vraie vie.

Un biais conduisant à conclure à tort à une différence entre les 2 traitements comparés surviendra si les patients du groupe contrôle ne reçoivent pas le traitement contrôle voulu et sont en réalité non traités. Cette situation conduira à conclure à la supériorité du traitement étudié par rapport au comparateur même si en réalité ces 2 traitements performent de la même façon.

Pour la safety d'un nouvel anticoagulant, si la comparaison est effectuée versus des patients non traité (non-utilisateurs d'un traitement), l'inclusion dans le groupe contrôle par erreur de patients recevant un traitement antiagrégant ou anticoagulant conduira à minimiser le risque d'hémorragies associé au nouveau produit (bof l'exemple).

Avec un vaccin, un biais de classification pourrait conduire à ne pas mettre en évidence l'efficacité de celui-ci par rapport à un groupe contrôle externe non vaccinée si celui-ci est construit à partir d'une

base de données de dispensation. En effet il est possible que les sujets de ce groupe aient pu bénéficier du vaccin en dehors d'une prescription médicale et d'une dispensation, par exemple, au niveau d'une institution ou d'une entreprise.

19.1.3 Erreur de mesure sur les covariables

Les erreurs de mesure affectant les variables autres que les critères de jugement peuvent aussi être problématiques dans les études observationnelles, en particulier les erreurs de mesure sur les facteurs de confusion [206]. En effet si les valeurs utilisées pour prendre en compte un facteur de confusion dans l'analyse sont peu fiables (inexactes) cela revient à ne pas prendre en compte ce facteur de confusion (l'ajustement se faisant sur une valeur plus ou moins aléatoire et non sur la vraie valeur du facteur de confusion).

Cette situation peut survenir lorsque des proxys sont utilisés pour ces covariables. Par exemple, le score ECOG est un facteur de confusion en oncologie. Il n'est pas mesuré en pratique médicale en dehors des études cliniques. Il est alors approximé à partir d'autre renseignant présent dans la base de données l'aide d'un algorithme [207] [208] [209]. Il est important que cet algorithme ait fait l'objet d'une validation et que ses performances diagnostiques soient documentées, satisfaisantes et transportables à l'étude considérée.

19.1.4 Précision des données afin d'assurer l'hypothèse de cohérence (*consistency*) (STUVA) de l'inférence causale

Le codage de la variable traitement doit être suffisamment précis pour éviter que deux modalités différentes de traitement soient codées de la même façon afin de satisfaire l'hypothèse fondamentale de cohérence de l'inférence causale qui est l'une des composantes de l'hypothèse STUVA (cf. section 13.1.2). Cette hypothèse implique que l'issue observée (l'outcome, valeur du critère de jugement) pour un individu sous le traitement qu'il reçoit correspond à son outcome potentiel avec ce traitement. Cela nécessite qu'il n'existe pas de multiples versions du traitement évalué associées chacune à un outcome potentiel différent. En effet, l'issue observée chez un patient, recevant apparemment le traitement A alors qu'en réalité il reçoit le traitement B, ne correspondra pas à son outcome potentiel avec le traitement A. Cela surviendrait si le codage de la variable traitement est insuffisamment précis pour distinguer ces versions différentes du traitement. Si la dose de la prescription est importante, les données doivent être suffisamment précises pour distinguer les doses prescrites.

Cela peut aussi survenir en cas d'erreur de classification de l'exposition dans le groupe contrôle. Par exemple dans un groupe contrôle non traité, si des patients en réalité traités sont inclus par erreur, leur issue observée ne correspondra pas à leur outcome potentiel sans traitement.

Avec un groupe contrôle traité, identifié à partir d'une base de données administrative de dispensation, des patients inclus dans ce groupe (considéré traité par le traitement contrôle) pourront en réalité être non traités (dispensation n'impliquant pas forcément la prise du traitement) et leur l'issue observée ne correspondra pas à leur outcome potentiel traité avec ce traitement.

La cible de l'inférence étant l'effet de l'assignement à un traitement (analyse en intention de traiter, cf. section 13.4) ces considérations sur les versions du traitement concernent uniquement le traitement initialement prescrit, et non pas les modifications de ce traitement (réduction de dose, arrêt prématuré, switch, etc.) survenant au cours du temps du fait des effets indésirables ou de l'évolution de l'état clinique du patient. L'outcome potentiel rattaché à un traitement intègre l'effet de ces adaptations inévitables et souhaitables, qui se reproduiront lors de l'utilisation en pratique du

traitement (outcome potentiel avec le traitement tel qu'assigné, indépendamment des événements intercurrents et de leurs conséquences sur le critère de jugement). Il s'agit alors de l'estimand « *policy treatment* » (cf. section 13.4.2).

20 La qualité des données

Une partie de la fiabilité des résultats produits par une comparaison à un groupe contrôle externe va directement dépendre de la qualité des données (cf. section 19), l'autre partie dépendra de la construction de l'étude, c'est-à-dire des biais induits par le design (cf. section 17), et de la possibilité de corriger le biais de confusion (cf. section 14).

La qualité des données est un terme générique qui recouvre plusieurs caractéristiques distinctes des données ayant chacune potentiellement des retentissements sur la fiabilité des résultats de l'étude.

Une partie des biais pouvant affecter une comparaison externe sont des biais inscrits dans les données elles-mêmes (contrairement à d'autres biais qui sont induits par des défauts de construction ou de réalisation de l'étude).

La fiabilité de ces études dépend aussi de la disponibilité dans les données de toutes les variables nécessaires à la bonne réalisation de l'étude : critère de jugement, facteurs de confusion potentiels, contrôle négatif, etc. Par exemple, si le jeu de données ne contient pas certains facteurs de confusions potentiels, il sera impossible de les prendre en compte dans l'analyse et un biais de confusion résiduel affectera les résultats produits.

Plusieurs guides ou documents de réflexion sur la qualité des données ont été produits par la FDA [51] [52], aussi EMA doc [55], MHRA [210], etc.

EMA	Data Quality Framework for EU medicines regulation: application to Real-World Data https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf
MHRA	MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions - GOV.UK https://www.gov.uk/government/publications/mhra-guidance-on-the-use-of-real-world-data-in-clinical-studies-to-support-regulatory-decisions
FDA	Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory
FDA	Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-registries-support-regulatory-decision-making-drug-and-biological-products

Ce document n'aborde pas les questions techniques concernant l'informatique, la gestion des données, l'extraction, les questions d'interopération, etc.

20.1 Exactitudes (accuracy)

20.1.1 Généralités

L'exactitude fait référence au degré avec lequel les valeurs enregistrées correspondent fidèlement aux vraies valeurs du patient. On peut aussi parler de fiabilité ou de taux d'erreur (cf. section 20.4).

Il s'agit par exemple de l'exactitude des valeurs enregistrées pour les diagnostics, les traitements reçus, des valeurs des caractéristiques démographiques, cliniques, biologiques, etc.

Le degré d'exactitude peut être apprécié par le taux d'erreur par variable, par patient ou global. Pour les variables binaires concernant les événements cliniques, les erreurs peuvent être soit des faux négatifs (l'événement qu'a présenté le patient n'est pas enregistré), soit des faux positifs (un évènement clinique est enregistré, mais le patient ne l'a pas présenté en réalité).

Un manque d'exactitude, de fiabilité des données va fausser les résultats de la comparaison externes de différentes manières.

Les erreurs sur les traitements reçus sont susceptibles d'introduire un biais de classification (cf. section 19.1.2), tandis que les erreurs sur les diagnostics, ou sur la survenue des événements cliniques conduiront à un biais de mesure (biais de classification des outcomes) (cf. section 19.1.1).

Un biais surviendra quand les erreurs de mesure sont asymétriques entre les deux groupes, mais aussi lorsque l'erreur est purement aléatoire (symétrique, identique entre les deux groupes) en biaisant les résultats vers l'absence de différence (*bias toward the null*), exposant au risque d'obtenir une étude non concluante ne montrant pas le bénéfice du traitement étudié ou conduisant à conclure à tort à une fausse bonne sécurité du traitement étudié.

Les erreurs concernant les facteurs de confusion, même purement aléatoire, rendront caduque la correction du biais de confusion par l'analyse statistique de l'étude (cf. section 19.1.3).

20.1.2 Origine des erreurs de classification

Ces erreurs peuvent être le fruit d'erreurs de saisie ou d'erreurs d'extraction par exemple lors d'une « *chart review* » ou lors de l'utilisation d'un système d'IA pour extraire de l'information structurée d'un dossier médical électronique. Les erreurs pures de saisie ou d'extraction sont aléatoires (ne dépendent ni de la valeur elle-même ni du traitement). Dans une étude observationnelle classique elles sont symétriques (affectant de la même façon les deux groupes) et non informatives. Elles ne peuvent pas induire un biais en faveur d'une fausse différence, cependant, elles peuvent induire un biais vers l'absence de différence. Mais dans une comparaison externe, ces erreurs aléatoires n'ont lieu que dans le groupe contrôle et non pas dans le groupe traité et conduisent à des erreurs asymétriques.

Ces erreurs peuvent aussi être une déformation volontaire de la réalité, principalement dans les bases de données de remboursement (*claims databases*, type SNDS ou autres). En effet la finalité de la saisie dans ces bases pour les organismes de soins est d'obtenir un payement suivant une nomenclature prenant ne compte les pathologies, les actes et les traitements réalisés. Il y a donc souvent une optimisation, par exemple, des diagnostics saisis en fonction de la valeur du remboursement prévu dans la nomenclature ce qui déforme les informations enregistrées par rapport à la réalité. Cette déformation peut aussi survenir sur les critères qui conditionnent pour le payeur le remboursement ou non d'un traitement.

> Il a été observé [211] qu'un changement de tarification Medicare, liant le remboursement des réparations de hernie à un seuil de 3 cm, a entraîné une chute soudaine des hernies codées comme « < 3 cm ». Comme les caractéristiques des patients n'ont pas changé, cette rupture suggère une adaptation du codage plutôt qu'un changement dans la présentation clinique de la pathologie. Les auteurs soulignent ainsi que des incitations financières pourraient influencer la manière dont les médecins mesurent ou déclarent certaines variables. Ce phénomène dégraderait la précision des bases administratives, qui deviennent moins fiables pour la recherche ou l'évaluation des pratiques.

Il faut aussi noter que les biais de classification ou de mesure (les biais d'information) ne proviennent pas que d'erreurs de valeur enregistrée. La valeur peut être exacte, mais elle ne reflète pas ce qu'elle est censée refléter ou la variable pour laquelle elle est utilisée.

> Une étude comparant la fréquence de survenue des hémorragies majeures entre un nouvel anticoagulant et la warfarine a été réalisée sur une base de données administrative. Pour le critère de jugement, les hospitalisations pour hémorragie ont été utilisées comme proxy des hémorragies majeures (information non disponible dans le thesaurus de codage des diagnostics). L'étude a montré à tort un surcroit d'hémorragie avec le nouvel anticoagulant du fait d'un biais de mesure, car les médecins traitants ont plus fréquemment hospitalisé les patients qui présentaient un saignement mineur dans le groupe du nouveau produit que dans le groupe contrôle, car ils ne disposaient pas d'un marqueur du « niveau d'anticoagulation » comme l'INR avec la warfarine et avaient moins d'expérience avec le nouveau produit. Le biais de mesure du critère de jugement est donc ancré dans la genèse de l'information « hospitalisation pour hémorragie » (qui, à gravité du saignement identique, est plus fréquent dans le groupe du nouvel anticoagulant) et ne provient pas d'une erreur de saisie ou de codage dans la base de données administrative.

20.2 Complétudes, exhaustivité

La complétude des données fait référence à la proportion de données manquantes.

Les données manquantes ont la potentialité de fausser les résultats de l'étude de plusieurs manières en introduisant un biais de sélection et/ou un biais d'attrition. Les techniques utilisées pour gérer les données manquantes entrainent soit une réduction de puissance/précision soit une sous-estimation des tailles d'effet quand ces données manquantes sont remplacées de manière conservatrice.

Les données manquantes rendent inutilisable pour l'analyse des patients (à partir du moment où il y a une donnée manquante sur une des variables nécessaires à cette analyse) ce qui conduit à les remplacer/imputer à l'aide d'une méthode statistique. L'imputation des données manquantes est un sujet très technique et les méthodes disponibles reposent sur des hypothèses fondamentales difficiles à vérifier. Ainsi même après remplacement les données manquantes sont susceptibles de fausser les résultats de l'étude.

La présence de données manquantes n'est vraiment identifiable que dans les sources de données structurées comme les registres, les cohortes où elles correspondent à une valeur d'une variable dont la saisie était prévue, mais qui n'a pas été renseignée. Ces valeurs manquantes font l'objet d'un code particulier. En revanche dans les sources non structurées comme les dossiers médicaux ou les bases administratives, le concept même de données manquantes n'existe pas en tant que-t-elle, étant donné qu'il n'y a pas de liste de variable à saisir. Dans ces sources, l'absence d'une information pourtant existante comme un diagnostic, un traitement ne sera pas identifiable. Sur ces variables d'occurrence

aléatoire, une information existante oubliée ou non mentionnée se traduira, in fine, par la modalité d'absence de la variable, donc par un faux négatif. Par exemple la non-mention dans le dossier médical de la survenue d'une hémorragie conduira à la notion d'absence d'hémorragie et non pas à la notion de données manquantes. Dans un recueil structuré, l'absence d'hémorragie sera renseignée quand cette absence aura été vérifiée explicitement et sera codée « données manquantes » quand cette vérification explicite n'aura pas été effectuée ou s'avère impossible à faire.

Les données manquantes sont fréquentes (voire très fréquente) dans beaucoup de sources de données type registre de maladie car la saisie des informations dans ces bases de données est un travail supplémentaire à réaliser lors des consultations ou des hospitalisations des patients sans qu'une aide spécifique soit disponible. Le data management est en général inexistant ou très limité ainsi que le monitorage. Au total, beaucoup de ces bases, attractives car étant théoriquement conçues pour faire de la recherche, s'avère inutilisable à cause des données manquantes.

20.3 Informativité, pertinence (*relevance*)

L'informativité des données fait référence à la présence de toutes les variables nécessaires à la réalisation de l'étude : critères de jugement, facteur de confusion, critère d'éligibilité de la population visée, contrôles négatifs et positifs, etc. L'absence des variables nécessaires à la réalisation d'une étude rend celle-ci impossible à réaliser avec la source donnée considérée.

20.3.1 Critères de jugement

Les critères de jugement nécessaire pour l'évaluation des nouveaux traitements sont parfois très spécifiques et non utilisés en pratique médicale courante. Dans ce cas aucune source de données de vraie vie ne contiendra les données nécessaires à la réalisation d'une comparaison externe.

> Dans la maladie d'Alzheimer, les essais cliniques utilisent des échelles spécifiques comme l'ADAS-cog, le CDR-SB, l'échelle NPI, ou dans les formes précoces l'ADAS-Cog + ADCS-ADL, le CDR-SB, etc. qui ne sont pas employées en pratique courante. Les registres ou autres données de vraie vie sur l'Alzheimer ne sont donc pas utilisables comme groupe contrôle externe d'une étude monobras ou d'un RCT.

Dans les études observationnelles classiques, cette difficulté est moindre, car il n'y a pas nécessité de se comparer à une étude ayant déjà utilisé un certain critère de jugement. Il est alors possible d'utiliser un autre critère en postulant qu'il s'agit d'un proxy et que l'effet relatif mesuré sur ce proxy est une bonne approximation de l'effet relatif sur le vrai critère de jugement, car il s'agira du même proxy dans les 2 groupes.

> En oncologie, la PFS n'est pas disponible dans les données de vraie vie (cf. section 20.7) mais il est possible de définir un autre critère la rwPFS à partir des progressions telles qu'identifiées en vraie vie. Dans une étude classique, les 2 groupes seront comparés avec ce même critère et le hazard ratio obtenu peut éventuellement être acceptable comme approximation du HR de PFS. Mais dans une comparaison externe cela conduirait à comparer la PFS dans le groupe traité avec la rwPFS dans le groupe contrôle ; sans aucune possibilité de savoir en quoi la rwPFS diffère de la vraie PFS, car il est impossible de mesurer la rwPFS dans les études mesurant la PFS. Il est possible en vraie vie que les progressions soient constatées plus tardivement que dans les études mettant en œuvre la fréquence régulière de l'imagerie nécessaire à l'application des critères RECIST de progression. Mais il est aussi possible qu'en vraie vie les traitements soient arrêtés plus précocement pour différentes raisons allant d'un souci de minimiser les effets

indésirables à la volonté d'accélérer la mise en œuvre d'un traitement de ligne ultérieure plus puissant[23]. Il est donc impossible d'anticiper la direction du biais qu'aurait une comparaison PFS versus rwPFS.

Mais pour une comparaison externe, l'utilisation d'un proxy ne solutionnera pas la question, car la comparaison impliquera toujours deux critères de jugement différents. Il faudrait que le proxy utilisé donne exactement les mêmes valeurs que le critère de jugement qu'il remplace ce qui est une hypothèse forte.

Ce point conduit à l'existence d'un « effet étude » souvent irréductible, mis en évidence par l'inférence causale (cf. section 13.3.1).

Les critères nécessaires pour apprécier la safety sont rarement disponibles dans les sources de données de vraie vie. Par exemple en oncologie les notions d'effet indésirable de grade 1, 2, 3, 4 sont impossible à retrouver dans les données de vraie vie ainsi que la notion d'arrêt de traitement liée à un effet indésirable ou la notion d'effet indésirable d'attribuable ou non au traitement. Or ces informations de safety sont indispensables pour apprécier correctement la balance bénéfice risque d'un nouveau traitement. Ce point peut être très pénalisant et empêcher l'utilisation de la voie de la comparaison externe pour baser un changement de pratique.

20.3.2 Critères d'éligibilité (de sélection des patients de la population visée)

Nombre d'études monobras évaluent des thérapeutiques ciblées sur une altération moléculaire particulière (par exemple l'étude monobras CodeBreak 200, NCT04303780, du sotorasib dans le cancer du poumon non à petites cellules a inclus des patients ayant une mutation KRASG12C). Il est donc nécessaire de sélectionner les patients du groupe contrôle externe sur la présence de la même altération moléculaire. Plusieurs difficultés peuvent survenir à ce niveau. La première est liée au fait que cette altération moléculaire n'est pas considérée en pratique, car elle n'était pas connue jusqu'à présent ou n'avait aucun intérêt avant l'apparition de la première thérapeutique la ciblant. Une autre difficulté peut provenir du fait que peu de sources de données enregistrent ce type d'information sur les altérations moléculaires ou les variantes génétiques. Par exemple en oncologie jusqu'à récemment la base la plus complète sur ces données était celle de Flatiron (flatiron.com).

Cette absence peut être rédhibitoire, car elle conduit à prendre un groupe contrôle de patients tous venant, présentant ou non cette altération. En absence de connaissance de la valeur pronostique de cette altération moléculaire, il devient impossible d'interpréter les résultats d'une comparaison impliquant des patients sélectionnés versus des patients non sélectionnés sur ce biomarqueur. En effet, si cette altération moléculaire est en fait un marqueur de bon pronostic, une comparaison à un groupe de patients tout-venant donnera l'impression d'un bénéfice du traitement étudié même si celui-ci n'est pas plus efficace que le traitement contrôle (utilisé chez les patients tout-venant).

[23] Dans beaucoup de cancer, les immunothérapies ont été d'abord développées en 2eme ligne où elles ont démontré des bénéfices notables en survie. Cependant tant que les essais de premières lignes ne sont pas disponibles, les prises en charge commencent avec les traitements validés de première qui peuvent être de simple chimiothérapie dont le bénéfice démontré est parfois assez ténu. On peut imaginer dans ce cas une certaine volonté de recourir à la 2ème ligne le plus rapidement possible, conduisant en vraie vie à noter une progression avant le moment où elle aurait identifié par l'imagerie programmée ou, voire, même avant le moment où elle remplirait les critères RECIST.

20.3.3 Facteurs de confusion

Un autre défaut d'informativité couramment rencontré est l'absence de certains facteurs de confusion potentiels. Cela peut provenir du fait que le recueil de données n'a pas été établi après une réflexion formalisée d'identification des facteurs de confusion d'une éventuelle comparaison de traitement.

Une autre raison peut être l'absence de mesure ou de détermination de ces facteurs dans la pratique médicale. Par exemple en oncologie un facteur pronostic important est le score ECOG. En pratique clinique courante, celui-ci n'est pas utilisé et il n'est donc pas enregistré dans les sources de données d'oncologie. Des tentatives de le récréer de façon algorithmique ont été proposées, mais il est alors nécessaire de disposer d'un algorithme validé pour la source de données envisagée [208] [207].

Toujours en oncologie, pour certains cancers à bon pronostic, les patients peuvent décéder de causes autres que leur cancer, par exemple, cardiologiques, neurologiques, etc. Les facteurs de risque de ces causes de décès seront donc des facteurs de confusion potentielle pour la survie globale (OS) dans des comparaisons externes et cela d'autant plus que l'un des traitements comparés à des effets indésirables ou des contre-indications dans la sphère cardiologique par exemple. Or les registres en oncologie sont souvent constitués dans une seule perspective oncologique et ces facteurs de risque de décès d'autres causes non recueillies.

20.3.4 Chainage

Un manque d'informativité d'une source de donnée peut être compensé en effectuant un chainage avec une autre base contenant les informations manquantes. Ce chainage constitue à identifier chaque patient de la première dans la seconde et de récupérer ainsi les données manquantes. Réaliser cela n'est absolument pas trivial en raison de l'anonymisation des bases et passe par des procédures de chainage probabiliste qui ont leur lot d'approximation. Cette limitation est levée dans les cas exceptionnels ou il est possible de faire un chainage exact des patients.

20.4 Origine des données

L'origine des données doit être transparente et validée. Avec la multiplication des jeux de données, il s'avère que des études peuvent se baser, volontairement ou involontairement, sur des données douteuses, parfois complètement artificielles, par exemple inventées ou entièrement simulées. [212].

Récemment il a été mis en évidence que 124 publications de modèles prédictifs, utilisés en pratique pour 3 d'entre eux et présente pour 86 dans des revues de la littérature, étaient basés sur deux jeux de données d'entrainement dont la provenance est intraçable (et qui sont certainement entièrement inventés) [212].

20.5 La validation des données

Pour juger de la qualité des données utilisées en général et de la possibilité de biais liés aux données en particulier, il est nécessaire de documenter numériquement le degré d'exactitude et de complétudes des données utilisées pour créer le groupe contrôle externe [87] [51].

Cette approche quantitative est complémentaire de la description transparente du processus de constitution de la source de données, des algorithmes utilisés pour créer les variables, de la gestion des données, de leur protection, de leur anonymisation et du respect de la législation, etc. Ces aspects

informatiques et de data management, fondamentaux eux aussi, ne sont pas abordés dans ce document, mais doivent être suivis et documentés [52] [213].

Plusieurs métriques sont utilisables pour mesurer l'exactitude (*accuracy*) des données comme les indices de performance diagnostique (sensibilité, spécificité, valeur prédictive positive, valeur prédictive négative) ou la différence des temps jusqu'à événements réels et rapportés dans les données pour les variables en « *time-to-events* » [214] [87].

Tous ces indices nécessitent de comparer le contenu des données avec un standard de référence (*reference standard, gold standard, ground truth*) constitué des vraies valeurs des patients. Cela implique la réalisation d'étude de validation des données ayant accès à ces vraies valeurs (par retour aux dossiers médicaux ou autres démarches) [215].

> *Ces études de validation pose la question du choix de métrique (Se,SP ou VPP,VPN, de la dépendance de VPP et VPN à la prévalence de la variable, du design de l'étude de validation et de la transportabilité des résultats entre études de validation [214].*

Dans le contexte des études de validation des données, la **valeur prédictive positive (VPP)** mesure le pourcentage de valeurs exactes pour un catégorie donnée. Par exemple une VPP de 80% sur la catégorie « fibrillation auriculaire » signifie que seulement 80% des patients qui seront étiquetés dans la source de données « fibrillation auriculaire » ont bien en réalité une fibrillation auriculaire.

Une valeur de VPP élevée assure que la classification des patients dans une catégorie est correcte.

VPP = Vrais Positifs / (Vrais Positifs + Faux Positifs)

Par exemple, si l'algorithme/l'extraction identifie 100 patients diabétiques dans une base de données et que la validation par rapport aux dossiers confirme que 90 d'entre eux sont réellement diabétiques, la VPP est de 90%.

La VPP dépend fortement de la prévalence de la catégorie dans la population. Une catégorie rare aura tendance à avoir une VPP plus faible même avec un bon algorithme. Ainsi les valeurs de VPP pour un algorithme donné rapportée dans la littérature ne correspondent pas forcément à la VPP de cet algorithme sur une nouvelle base de données [214]. Sensibilité et spécificité (qui rentre dans le calcul de la VPP) sont davantage transposable, car elles ne dépendent pas de la prévalence de la catégorie à détecter.

La **valeur prédictive négative VPN** mesure la proportion des cas classés comme négatifs par l'algorithme qui sont réellement des vrais négatifs lors de la vérification par l'étude de validation.

VPN = Vrais Négatifs / (Vrais Négatifs + Faux Négatifs)

Par exemple, si l'algorithme/l'extraction identifie 1000 patients comme "non-diabétiques" et que la validation confirme que 700 ne le sont effectivement pas, la VPN est de 70%.

Comme la VPP, la VPN dépend de la prévalence.

La VPN est importante quand il s'agit d'identifier les patients qui n'ont pas la maladie ou qui n'ont pas fait l'événement.

On trouve facilement des exemples d'études de validation de ce type dans la littérature [216] [217] [218] [219] [220] [221] [222], y compris des revues systématiques de ces études [223] [224].

En plus de documenter la qualité des données, ces indices quantitatifs permettront de paramétrer les analyses quantitatives du biais destinées à éprouver la robustesse des résultats produits (cf. section 16.2) [225].

Cette validation des données peut concerner toute la population analytique ou seulement un sous échantillon aléatoire [51].

Lorsque la validation par rapport à une référence standard n'est vraiment pas possible comme avec les données anonymisées, d'autres approches peuvent être mises en œuvre [226] :

- Comparaison de différente définitions, codes, etc.
- Comparaison des distributions des valeurs de l'échantillon avec la distribution de cette variable dans la population générale ou dans d'autres études
- Évaluation de la plausibilité des données
- Évaluation de la cohérence des données au niveau patient

Des outils sont aussi disponibles pour évaluer la fiabilité du processus d'extraction de sources non structurées (par exemple par IA, TAL) comme le framework VALID [227].

L'initiative PRINCIPLED des réseaux sentinels FDA détaille les aspects pratiques de la mise en œuvre de ces principes [87].

Une autre approche de validation des données est celle du benchmarking (cf. section 23) qui consiste à montrer qu'une source de données permet de retrouver les mêmes résultats d'efficacités et de safety (hazard ratio par exemple) que les essais randomisés de traitements précédents.

Bien que fondamental, l'exactitude et la complétude des données ne suffisent pas à elles-mêmes à garantir la fiabilité des résultats des comparaisons externes, encore faut-il que les autres problématiques méthodologiques aient pu être solutionnées (confusion, biais de sélection, limites de la démarche rétrospectives, défaut de conception de l'étude, etc.). En revanche, quelle que soit la méthodologie de l'étude, des données de qualité limitée interrogeront sur la fiabilité des résultats.

Dans les guides des agences, ces aspects de la validation des données sont systématiquement mentionnés :

- EMA Data Quality Framework for EU medicines regulation: application to Real-World Data [213]
- NICE NICE real-world evidence framework [64] (chapitre data validity)
- FDA Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products [51]
- FDA Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products [52]

20.6 Recommandations pour la constitution des sources de données

En pratique il s'avère très difficile de trouver des données adaptées à la réalisation des groupes contrôles externes, même au niveau des registres ou cohortes [228]. Les raisons sont principalement :

- L'absence des critères de jugements nécessaires
- La non-disponibilité de tous les facteurs de confusion identifiés
- Le taux très important de données manquantes à la fois sur les critères de jugement que sur les covariables
- L'absence de certaines variables nécessaires à la sélection des sujets (par exemple l'altération moléculaire cible du nouveau traitement)
- Le manque de variables pouvant servir à faire des contrôles négatifs

Ces limitations au niveau des données rendent les études, réalisées malgré tout, impropres à la décision, car insuffisamment robustes (ajustement insuffisant, impossibilité d'éprouver l'importance du biais de confusion résiduel) et/ou pertinentes (critères de jugement inadaptés ou différents de ceux attendus, sélection des patients inadaptée).

Ce retour d'expérience met en exergue la nécessité d'anticiper ces besoins au moment de la mise en place de la source de données (registres, cohortes, entrepôts de données de santé, etc.) et d'intégrer le recueil de ces variables fondamentales pour la réalisation de groupes contrôles externes.

En effet ces recueils d'information sont souvent conçus de manière générale, en intégrant les variables standards du domaine médical concerné, mais sans véritable anticipation des réels besoins des différents types d'études qui pourraient être réalisées dans le futur. Il est connu depuis longtemps, dans le domaine de l'épidémiologie, que ce genre de recueil conduit fréquemment à la constitution de cimetières de données, où il s'avère que des variables fondamentales manquent lorsque l'on envisage de les utiliser pour réaliser une étude.

Sur ce point il conviendrait donc que l'utilisation des données pour constituer des groupes contrôles externes soit envisagée lors de la mise en place d'une nouvelle source de données (registre par exemple). Il est en effet souvent possible d'anticiper les critères de jugements qui seront nécessaires (ils sont en général standard pour la ou les pathologies couvertes et ne dépendent pas vraiment des nouveaux traitements qui seront à évaluer). Une fois les critères de jugement connus, il est possible de faire la revue systématique de leurs facteurs pronostiques (qui seront les futurs facteurs de confusion des comparaisons externes) et de prévoir leur recueil. De même il est possible d'imaginer dès ce stade les contrôles négatifs.

Cette anticipation permettrait aussi d'identifier les chainages nécessaires (pour quelles variables, avec quelles autres sources de données) et de les mettre en place de manière prospective.

Cette analyse permet aussi d'identifier la liste des variables fondamentales de ce type d'étude, sur lesquelles il serait nécessaire d'assurer une qualité optimale du recueil, avec éventuellement un data management.

Entretemps, la difficulté de trouver des sources de données contenant d'emblée et de manière structurée toutes les données nécessaires peut être contournée en utilisant seulement la source de données comme moyen d'identification des patients potentiellement éligibles. Ensuite une *chart review* sera mise en place pour extraire les données nécessaires à l'étude directement dans les dossiers médicaux de ces patients.

20.7 La rwPFS en oncologie

Un critère de jugement habituel dans les essais cliniques en oncologie est la survie sans progression (*progression free survival, PFS*) qui mesure le temps jusqu'à la progression du cancer ou jusqu'au décès si celui survient sans progression préalable. La progression est mesurée à l'aide d'imageries répétées (comme un scanner tous les 3 mois, la périodicité dépendant de la pathologie étudiée) et avec l'utilisation de critère formel de progression, les critères RECIST.

Compte tenu de difficulté d'interprétation survenant parfois avec des clichés ambigus, ces mesures sont souvent effectuées, non pas par les investigateurs eux-mêmes, mais de façon centralisée par un comité d'adjudication, IRB.

Dans la pratique médicale courante, en dehors de la réalisation d'un essai clinique, les progressions sont diagnostiquées par les médecins traitants sans utiliser les critères RECIST, à partir d'imagerie réalisée en fonction des besoins cliniques et sans la stricte périodicité des protocoles des essais.

Il est donc impossible de mesurer la PFS dans des données de vraie vie.

Pour pallier cette impossibilité, il a été cherché des proxys évaluables avec les données disponibles dans les sources habituelles : dossiers médicaux, registres, etc. Ces proxys peuvent être le temps jusqu'à progression telle que notée dans les dossiers par les médecins traitants ou le temps jusqu'à l'arrêt du traitement ou à son changement pour un autre (le constat d'une progression du cancer conduit fréquemment à changer de ligne de traitement c'est-à-dire à utiliser un nouveau traitement). On rencontre ainsi à la place de la rwPFS le temps jusqu'au prochain traitement (rwTTNT *real world time to next treatment*). Se pose aussi la question des changements de traitement sans progression identifiable. Se superpose aussi la question des censures qui suivent des règles très strictes dans les essais randomisés sans équivalent en observationnel (la notion de scanner manqué n'existe pas en observationnelle par exemple).

Compte tenu de ces différences fondamentales, la mesure effectuée sur les données observationnelles est classiquement appelée real world PFS, rwPFS afin de bien expliciter qu'il ne s'agit pas de la PFS.

Ainsi le terme rwPFS peut couvrir des réalités différentes en fonction de la source de données utilisée ou de la situation (en dernière ligne la progression ne se traduira pas par un changement de traitement, toutes les ressources thérapeutiques ayant été épuisées par exemple). Il est aussi possible de définir la rwPFS de très nombreuses façons [229].

La rwPFS va différer de la PFS pour deux raisons (Figure 20) : les erreurs de classement liées à l'absence d'utilisation des critères RECIST et la non-périodicité des imageries [230] [231]. Plusieurs études ont montré que la rwPFS conduisait à une estimation biaisée de la PFS [231][231][26].

Les erreurs de classement conduisent à des faux positifs : à un temps donné, les données de RWD sont en faveur d'une progression (avis du clinicien ou du radiologue, changement de traitement, etc.) alors que si les critères RECIST avaient été appliqués la progression n'aurait pas (encore) été diagnostiquée ; mais aussi à des faux négatifs : dans la vraie vie, rien ne suggère une progression (pas de changement de traitement ou absence de mention de progression radiologique) alors que si les critères RECIST avaient été appliqués ils auraient conclu à une progression. Dans ces 2 cas, le temps jusqu'à événement mesuré avec la rwPFS sera différent de celui mesuré avec la PFS

L'irrégularité de la surveillance conduira aussi à des temps jusqu'à un événement différent avec la rwPFS et la PFS. Dans la vraie vie le scanner peut être effectué plus ou moins en fonction de la clinique, par exemple, quand celle-ci laisse fortement suspecter une progression (symptômes, altération de

l'état général, etc.). Dans ce cas la progression sera détectée plus tardivement qu'avec une imagerie régulière qui aurait détecté la progression radiologique avec les signes cliniques. Mais il est aussi possible que l'imagerie guidée par la clinique soit effectuée un peu plus tôt que ce qu'elle aurait dû être dans le schéma de surveillance d'un essai clinique (dans ce contexte malgré les signes d'appels, la date programmée du scanner est attendue).

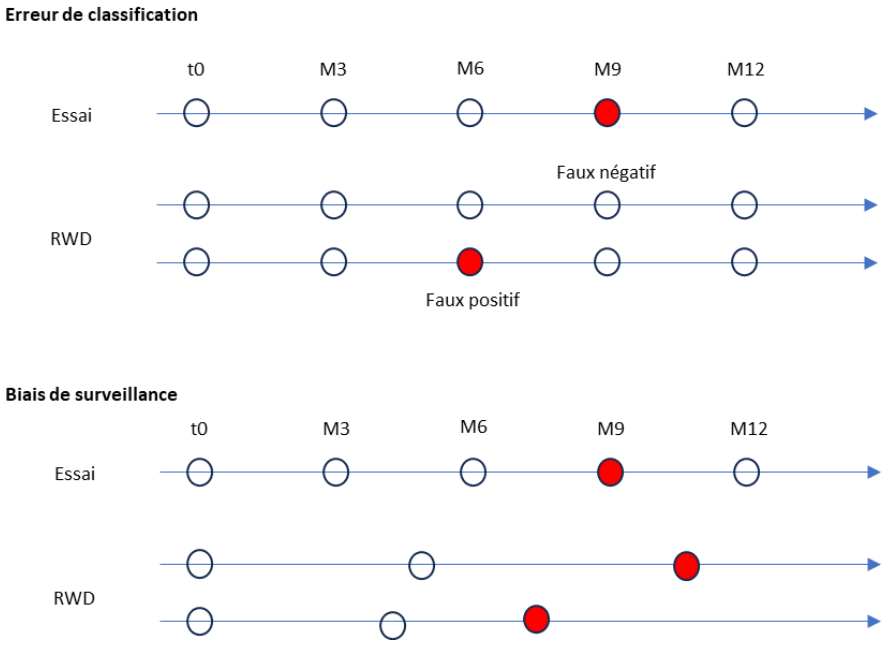


Figure 20 – Illustration des 2 biais affectant la rwPFS par rapport à la PFS

La ligne temporelle « essai » représente ce qui se passe avec l'évaluation de la PFS ; les 2 autres lignes illustrent ce qui peut se passer avec une évaluation de la rwPFS à partir de données de vraie vie. Dans tout les cas, l'analyse portera sur le temps jusqu'à événement (du t0 au point rouge) qui peut donc être notablement différent entre PFS et rwPFS.

Si, dans une certaine mesure, ces biais peuvent ne pas trop perturber la comparaison de deux groupes utilisant la même définition du critère de jugement, ils auront bien plus de conséquences dans les comparaisons à un groupe contrôle externe. En effet dans ce cas la comparaison se fera entre la PFS dans le groupe traité et la rwPFS dans le groupe contrôle, soit entre deux concepts différents. Cette problématique ne peut pas être solutionnée en cherchant à mesurer aussi la rwPFS dans le groupe expérimental, car cela est impossible.

Au total, l'évaluation du bénéfice d'un nouveau traitement sur la PFS par une comparaison externe à un groupe contrôle externe issu de données observationnelles s'avère impossible. L'utilisation d'un groupe contrôle issu des données d'un essai clinique précédent utilisant la PFS usuelle est cependant envisageable.

L'évaluation du bénéfice du nouveau traitement sur la PFS à l'aide d'un groupe contrôle externe basé sur des données de vraie vie est très problématique et doit être évitée

Des méthodes d'analyse statistique permettant de comparer sans biais une PFS à la rwPFS ont été récemment proposées [232] [233] mais cette méthode repose sur des hypothèses fortes, en général invérifiables et sa réelle aptitude à corriger des biais va être très difficilement à établir de manière empirique, car rwPFS et PFS ne peuvent être mesuré simultanément sur les mêmes données. D'autres propositions sont en cours de développement[24].

En oncohématologie, les critères de progression sont souvent biologiques et peut-être plus facilement identifiables dans les bases de données contenant la biologie des patients.

Les limites de la rwPFS dans les comparaisons externes font que le critère de jugement d'élection pour les comparaisons externes en oncologie tumeurs solides doit être la **survie globale** (*overall survival*, OS). L'OS pose moins de problèmes métrologiques et de plus c'est le critère cliniquement pertinent par excellence.

Il faut cependant noter que l'information sur les décès n'est pas forcément complète dans beaucoup de sources de données. Les dossiers médicaux hospitaliers peuvent ne pas rapporter le décès et la date de décès d'un patient quand celui-ci est survenu à l'extérieur de l'établissement (comme au domicile ou dans un autre établissement). Il en est de même pour les registres qui sont remplis au fur et à mesure des consultations ou hospitalisations. On parle d'ailleurs parfois de rwOS (real world Overall Survival) pour alerter sur ce point.

Un chainage avec un registre d'état civil peut solutionner le point [234] [235].

[24] https://www.ema.europa.eu/en/documents/presentation/presentation-quantifying-mitigating-measurement-bias-real-world-endpoints-when-constructing-external-control-arms-b-ackerman_en.pdf

21 Les outils d'évaluation du risque de biais

Il n'existe pas (encore) d'outils d'évaluation globale du risque de biais spécifique des études de comparaisons externes. Cependant compte tenu de leur forte similitude aux études observationnelles classiques, les outils dédiés à ces études sont parfaitement utilisables, car les comparaisons externes étant, par essence, des études observationnelles, elles sont soumises aux mêmes biais que les autres études comparatives observationnelles. De ce fait les outils comme le ROBINS-I [236], APPRAISE [237] sont utilisables pour la gradation formalisée du risque de biais des études de comparaisons externes.

De nombreuses outils d'évaluation des études observationnelles (NRS « *non randomized study* ») ont été proposés. Une récente revue en décompte 44 [238]. La plupart se focalisent sur la qualité de réalisation et non pas sur une évaluation du risque de biais. De toutes ces propositions, la plus fréquemment citée et/ou utilisée actuellement est ROBINS-I. APPRAISE est la proposition la plus récente (2025).

21.1 ROBINS-I

ROBINS-I [236] permet de grader le risque de chacun des biais affectant les études observationnelles évaluant une intervention de santé en 4 catégories : *low, moderate, serious, critical* et permet aussi de donner une évaluation globale du risque de biais. La version actuelle est la version 2. Les sept domaines de biais (types de biais) affectant les études observationnelles sont considérés :

1. Biais de confusion
2. Biais de sélection
3. Biais de classification
4. Biais de déviation
5. Biais du aux données manquantes
6. Biais de mesure
7. Biais de sélection des résultats rapportés

Le risque de biais existant au niveau de chaque domaine est coté en low, moderate, serious or critical. Des questions indicatrices guident le lecteur dans l'évaluation du niveau de risque de biais de ces domaines.

ROBBINS-I n'est pas spécifique des études à contrôle externe, mais s'applique à ces études qui sont des études observationnelles à part entière.

Compte tenu de l'enjeu de l'évaluation des nouveaux traitements, seules les études ayant un risque de biais faible pourront être prises en compte (en sachant que pour le biais de confusion le faible risque de biais s'entend indépendamment de la question des facteurs de confusion non pris en compte).

21.2 APPRAISE

L'outil APPRAISE (*APpraisal of Potential for bias in ReAl world evIdence StudiEs*) [237] est un outil d'aide à l'évaluation des biais des études de RWE[25]. Il guide la lecture et l'évaluation, mais ne propose pas de

[25] Outil disponible à https://osf.io/a4nhd/overview

système de cotation du risque de biais. Il propose d'explorer les biais potentiels à travers 3 domaines : utilisation d'un design ou d'une analyse inappropriée, erreur de classification de l'exposition ou de l'outcome, et contrôle du biais de confusion.

APPRAISE identifie neuf types de biais courants (comme le biais lié au temps, la causalité inverse, les erreurs de classification et la confusion résiduelle) grâce à des questions structurées. Contrairement aux simples listes de contrôle, APPRAISE relie clairement les choix de conception et d'analyse à la validité des résultats et au risque de décision.

L'outil APPRAISE concerne toutes les études observationnelles comparatives (sur l'efficacité ou la sécurité des traitements) et s'applique donc aux études à contrôle externe.

APPRAISE est une proposition récente (publication en 2025), propre au domaine du *Health Technology Assessment* (HTA) et non pas de l'épidémiologie.

Afin de faciliter la lecture critique de ces études différents guides et checklists de rédaction ont été élaborés pour assurer la pleine informativité de leurs publications (ou rapports) :

- *STROBE qui concerne les études observationnelles dans leur ensemble (https://www.strobe-statement.org/)*
- *RECORD qui est spécifique des études observationnelles basées sur des données de santé collectées en routine (https://www.record-statement.org/)*
- *RECORD-PE spécifique des études de pharmaco-épidémiologie (https://www.record-statement.org/)*

22 L'émulation d'un essai cible

L'émulation d'un essai cible est un cadre conceptuel proposé pour aider à la construction des études observationnelles d'évaluation des interventions de santé et éviter des défauts de design conduisant à l'introduction de biais dans les études [239] [240] [241] [185] [242]. L'idée est de construire et analyser une étude observationnelle de façon à reproduire (émuler) ce qui se passerait dans un essai randomisé hypothétique qui aurait le même objectif [239] [240] [242]

L'émulation d'un essai cible n'est pas une méthode ou un design d'étude, c'est seulement un cadre conceptuel d'aide à la construction d'études observationnelles

Il ne s'agit pas d'un design d'étude ou d'une méthode d'analyse particulière. L'émulation fait appel à la mise en œuvre des techniques habituelles des études observationnelles, mais guidées de telle façon à éviter quelques pièges invalidant les résultats de l'étude (en particulier en termes de biais liés au début de suivi ou de temps d'immortalité).

> *L'émulation d'un essai cible est parfois comprise ou présentée à tort comme étant une nouvelle méthodologie suppléant complètement l'essai randomisé. Il n'en est rien, même si cette approche améliore notablement la conception des études observationnelles, celles-ci restent exposées à leurs limites habituelles qui doivent être solutionnées pour elles-mêmes. Par exemple, l'émulation n'apporte rien au niveau de la correction du biais de confusion par l'analyse. Cette correction sera obtenue par l'identification de tous les facteurs de confusion et leur prise en compte dans l'analyse. Deux points sur lesquels n'intervient pas l'émulation.*

L'émulation d'un essai cible permet de clarifier et d'éviter des biais consécutifs à une mauvaise construction des études observationnelles. Elle ne solutionne pas les biais inhérents aux données qui sont le biais de confusion, les erreurs de classification et de mesure, ou le biais de sélection.

Bien qu'initialement proposé pour la construction d'études observationnelles classiques (où les deux groupes comparés sont issus de la même source de données), le concept d'émulation d'un essai cible se généralise parfaitement à la question des comparaisons externe [243].

En premier il convient donc d'élaborer le protocole (synopsis) d'un essai randomisé qui aurait le même but que l'étude de comparaison externe envisagée en définissant avec précision l'objectif, les critères d'éligibilité des patients (inclusion et exclusion), les traitements comparés, les patients visés, les hypothèses et le calcul du nombre de sujets (ou d'événements) nécessaires, etc.

Ces choix peuvent être présentés sous la forme tabulaire habituelle des synopsis de protocoles d'essais randomisés, auquel sera rajoutée une colonne pour décrire la façon où tous ces éléments seront émulés lors de l'étude observationnelle. Ce tableau est la base de l'approche d'émulation de l'essai cible et son analyse sera la clé de l'évaluation de la solidité de la comparaison externe.

Dans un premier temps, cette approche va de l'essai randomisé vers l'étude observationnelle, afin de construire celle-ci pour qu'elle émule point par point l'essai randomisé. Cependant, sur certains points, il peut apparaitre que l'émulation sera impossible avec les données disponibles ou envisagées. Dans ce cas il ne convient pas de rester sur ce statu quo, éventuellement en précisant qu'il y aura une approximation à ce niveau dans l'étude observationnel par rapport à l'essai randomisé, mais il convient de modifier en retour l'élément correspondant de l'essai pour rendre compte de ce qui se passera effectivement dans l'étude observationnelle. [240]. L'apport de ce retour est fondamental, car l'intérêt de l'émulation est tout autant de donner une représentation en termes d'essai randomisé ce qui sera évalué par l'étude observationnelle que d'aider à la construction de celle-ci.

> *Par exemple en oncologie la survie sans progression est un critère de jugement habituel. Sa détermination repose le plus souvent sur la réalisation de scanner périodique (tous les 3 mois par exemple) qui servent à suivre l'évolution de la masse tumorale à l'aide de critère de mesure précis, les critères RECIST et de déterminer s'il y a progression de cette masse tumorale. Avec les données de vraie vie ce critère est non-émulable car dans la vraie vie les scanners ne sont réalisés de manière strictement périodique (et sans la même périodicité) et leur interprétation ne se base pas sur les critères RECIST (cf. section 20.7). En pratique d'autres critères de jugement sont utilisés en vraie vie pour se substituer à la PFS comme le temps jusqu'au prochain traitement, ou le temps jusqu'à l'arrêt du traitement pour la dernière ligne. Ces critères sont souvent dénommés real world PFS (rwPFS), mais ils ne constituent pas une émulation de la PFS, tout au plus une approximation. De ce fait il convient donc de revenir à la spécification de l'essai cible et de changer le critère de jugement pour le temps jusqu'au prochain traitement décidé par l'investigateur à la place de la PFS initialement envisagée. Ainsi avec cette modification il apparaitra clairement à la lecture du synopsis de l'essai cible ce qu'évalue effectivement l'étude observationnelle et il sera alors possible de juger en connaissance de cause l'utilisabilité de ses résultats (par exemple s'il est démontré ou non empiriquement que cette rwPFS permet d'estimer le même effet traitement que la vraie PFS).*

22.1 Mise en œuvre

En pratique l'émulation d'un essai cible consiste à élaborer le protocole (ou tout du moins son synopsis) de l'essai cible, puis à émuler par l'analyse des données observationnelle chaque élément méthodologique de cet essai, c'est-à-dire à construire une étude observationnelle qui mime ce qui se passerait si l'essai cible avait été réalisé [244]. Cette façon de procéder permet d'éviter des défauts de conception dans l'analyse des données observationnelles en particulier au niveau du biais de sélection biais par temps d'immortalité ou lié à un objectif ne se traduisant pas par une question causale.

Les éléments constitutifs de l'essai cible qui sont à émuler sont les suivants :

- Critère d'éligibilité
- Les stratégies de traitement comparées
- L'assignement des traitements (la randomisation)
- La définition du temps de début de suivi des patients (t0)
- Les critères de jugement
- Le contraste causal
- Les éléments de l'analyse des données

Un point sensible dans l'émulation est la détermination du t0 de début de suivi qui conditionne la protection contre le biais de sélection des études [197]. Cette problématique est détaillée section 18.1.1.

En pratique cette démarche débouche sur un tableau présentant les éléments de l'essai cible et la façon dont ces éléments seront émulés dans l'étude observationnelle.

Table 1 | Target trial protocol for case example study evaluating the effect of sodium-glucose cotransporter-2 (SGLT-2) inhibitors on genital infections

Element	Specification	Emulation using real world data sources
Eligibility criteria	Patients with type 2 diabetes mellitus; aged ≥65 years; no use of study drug treatments before randomization; no history of end stage renal disease, HIV, or genital infections; continuous Medicare A, B, D enrolment for six months and recorded glycated hemoglobin (HbA_{1c}) test results in electronic health records in six months before treatment initiation	Same as target trial
Treatment strategies	Initiation of (1) SGLT-2 inhibitors (canagliflozin, dapagliflozin, empagliflozin); or (2) DPP-4 inhibitors (alogliptin, linagliptin, saxagliptin, sitagliptin). Under both strategies, use of antidiabetic treatment after initiation is left to physician and patients' discretion	Same as target trial
Treatment assignment	Randomized, non-blinded	Non-blinded and assumed to be randomized within levels of measured confounders*
Follow-up start (time 0)	At assignment	Same as target trial
Follow-up end	First of administrative end of follow-up (day 365), loss to follow-up, death, or outcome occurrence	Same as target trial
Primary outcome	Genital infections	Same as target trial
Causal contrast	Intention-to-treat effect (effect of being assigned to the treatment)	Observational analogue of intention-to-treat effect

SGLT-2=sodium-glucose cotransporter-2; DPP-4=dipeptidyl peptidase-4; HbA1c=glycated hemoglobin.
*Measured confounders include demographics (age, sex, race, socioeconomic status markers), diabetes severity related variables including microvascular and macrovascular complications, measures related to diabetes control such as HbA_{1c}, comorbid conditions, cotreatments, markers for healthy behavior, and healthcare use.

Figure 21 - Exemple de tableau décrivant l'émulation d'un essai cible (extrait de [87])

Un guide de rédaction EQUATOR[26] spécifique, dénommé TARGET, liste les informations indispensables à rapporter dans une publication (ou un rapport) d'une étude basée sur une émulation d'essais cible [245].

L'émulation de l'essai de cible est mentionnée dans plusieurs guidelines comme celui de la FDA sur les études observationnelles (non-interventionnelles) [48] ou le *position paper* de l'HAS [44].

Agence	Document	Page	Extrait
EMA	Reflection paper on use of real-world data in noninterventional studies to generate real-world evidence for regulatory purposes	7	The target trial emulation (TTE) framework should be considered as a strategy that uses existing tools and methods to formalise the design and analysis of NIS using RWD with causal objectives
MHRA	MHRA draft guideline on the use of external control arms based on real- world data to support regulatory decisions	NA	NA
MHRA	MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions	NA	NA
FDA	Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products[53]	5	Several conceptual approaches can be used to address concerns regarding causality when designing a noninterventional study,

[26] https://www.equator-network.org/reporting-guidelines/

			including, but not limited to, the emulation of a hypothetical clinical trial that addresses the research question of interest.
FDA	Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products [6]	NA	NA
HAS	Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health [44]		
NICE	NICE real-world evidence framework [64]		

22.2 Évaluation des performances de l'approche

La fiabilité des résultats obtenus avec cette approche a été évaluée de manière empirique en comparant les résultats de RCT avec ceux produits par leur émulation par des études observationnelles (RCT DUPLICATE [246]). Il faut remarquer que la validité des études observationnelle conçue par une approche d'émulation ne se réduit pas à l'émulation mais fait intervenir tous les autres éléments méthodologiques comme la correction du biais de confusion par analyse, la qualité des données, etc. Ces études de validation empirique ont donc évalué un ensemble et non pas l'apport isolé de l'approche d'émulation pour la conception de ces études. Il s'agit d'une évaluation de la fiabilité d'études observationnelles optimisées sur de nombreux points méthodologiques, conçues suivant une approche d'émulation et réalisées sur des données appropriées et de qualité. Ces résultats ne préjugent pas que toute étude conçue avec une approche d'émulation sera du même niveau de fiabilité, l'émulation n'intervenant en rien sur la qualité et la complétude des ajustements, sur la qualité des données (erreur de classification de l'exposition ou du critère de jugement, fréquence et nature des données manquantes, etc.).

22.3 Méta-épidémiologie

Il convient de noter que de nombreuses études se présentent comme des émulations, alors qu'une analyse critique des publications ne permet pas toujours de le confirmer [247] [200] [248] [249] [250]. Ce constat met en lumière une tendance à valoriser cette approche conformément aux recommandations, sans que la qualité des études soit nécessairement à la hauteur des annonces. Il est donc recommandé d'adopter une attitude prudente lors de l'évaluation critique de ces travaux.

23 Le benchmarking et les contrôles positifs

Compte tenu des nombreuses problématiques méthodologiques, il existe toujours un doute sur la réelle fiabilité de la démarche mise en œuvre (sources de données, validité du critère de jugement, qualités des données, choix du t0, etc.) pour constituer un groupe contrôle externe. Une possibilité pour évaluer cette fiabilité, et l'aptitude de la méthodologie employée à solutionner les différentes problématiques méthodologiques, est de recourir à des contrôles positifs, à un benchmarking [251] [252].

Cette approche consiste à montrer que la même méthodologie permet de retrouver un résultat connu. Ce résultat connu est appelé contrôle positif et cette démarche de vérification « benchmarking ». Il s'agit par exemple de retrouver les résultats d'un essai randomisé d'un autre traitement précédent à partir d'une émulation d'essai cible réalisée avec les données pressenties pour constituer le groupe contrôle externe. Reproduire le résultat de référence donne quelques réassurances indirectes sur la qualité des données, leur pertinence, les erreurs de classification des critères de jugement, etc. La validation de la capacité à pouvoir corriger du biais de confusion n'est que très partielle, car la structure de confusion affectant une comparaison de deux groupes issus des mêmes données n'est pas comparable à celle d'une comparaison externe où tous les facteurs pronostiques rentrent en ligne de compte (cf. section 14.1).

Cette approche de benchmarking est aussi proposée comme approche de recalibration dont le but est de corriger les résultats de l'étude du biais estimé par l'utilisation des contrôles de falsification (contrôles positifs, mais surtout contrôles négatifs) [253] [254] [255]. Ces approches reposent sur l'estimation empirique du biais et de son incertitude statistique (variance), permettant ensuite de corriger le résultat obtenu par l'étude ainsi que son intervalle de confiance (et la p value). La validité de la correction dépend bien entendu d'hypothèses comme l'échangeabilité du biais entre la comparaison contrôle et la comparaison d'intérêt. Une première validation empirique de cette proposition est planifiée dans le contexte de l'émulation d'essais et non pas de comparaison externe [256]. Ses résultats ne sont pas encore connus à la date de rédaction de ce document.

Le benchmarking peut aussi être utilisée comme point de repère pour ajuster la méthode d'analyse en fonction des écarts entre les résultats obtenus et les résultats attendus, par exemple, en choisissant la définition du critère de jugement (ou l'algorithme phénotypique) permettant le retrouver au mieux le résultat attendu, ou d'adapter la liste des covariables prises en compte, etc.

Cette démarche soulève deux problématiques. La modification itérative du plan d'analyse statistique qu'elle induit peut-être le point d'ancrage de p-hacking. Il est donc indispensable que cette adaptation soit prévue au protocole et au plan d'analyse statistique initial et qu'elle se fasse impérativement sans aucune analyse inférentielle. Elle devra être aussi rapportée de façon très transparente (historique détaillé des adaptations et des résultats produits).

L'autre problématique est le risque de surdétermination, appelée aussi *bias-variance tradeoff*. Cette problématique, bien connue dans le domaine de la construction des outils prédictifs et de l'apprentissage statistique, survient lorsqu'un modèle statistique, ou une analyse statistique, est optimisés afin d'expliquer le mieux possible des données d'étalonnage (ici le résultat servant au benchmark). Cette optimisation (réduction/suppression du biais sur les données d'étalonnage) va faire perdre en généralisabilité du modèle, entrainant une chute de ses performances sur de nouvelles données. Il s'agit de l'illustration en statistique du vieil adage « le mieux est l'ennemi du bien » ! L'optimisation conduit à un modèle (à une analyse) adapté au bruit particulier affectant la comparaison

d'étalonnage et ne devient plus du tout adapté à d'autres situations où le bruit sera différent par définition.

Bien entendu retrouver le résultat connu n'apporte pas une garantie absolue de l'exactitude du résultat de la future comparaison externe. En revanche l'impossibilité de retrouver le résultat connu doit conduire à de sérieuses réserves sur la possibilité d'exploiter le résultat de la comparaison externe d'intérêt.

Lorsque la comparaison externe est réalisée à partir d'un RCT dont le groupe contrôle n'est pas/plus approprié, un autre type de contrôle positif peut consister à rechercher le résultat produit par cet essai avec son groupe contrôle randomisé en émulant un groupe contrôle externe traité avec le même traitement. Par exemple, si l'essai randomisé compare un nouveau traitement N au traitement standard A, il est certainement possible de trouver des patients traiter par A dans la source de données envisagées pour faire le groupe externe d'intérêt (traité par le traitement standard actuel B qui a montré sa supériorité à A durant la réalisation de cet essai randomisé par exemple). À l'aide de ce groupe contrôle externe il est possible de faire la comparaison externe N versus A et de comparer le résultat à celui de l'essai randomisé lui-même (comparant aussi N versus A).

Pour les comparaisons externes partant d'une étude monobras, il n'est pas possible de suivre une telle démarche. Au mieux, le benchmarking pourra consister à retrouver le résultat concernant le traitement contrôle d'intérêt (c'est-à-dire le résultat de l'essai randomisé l'ayant évalué si un tel essai a lui lieu). Dans ce cas la source de données est utilisée pour émuler un tout autre essai que celui qui concerne la comparaison externe d'intérêt, mais un tel benchmarking permet dans une certaine mesure de valider la qualité des données, les critères de jugement utilisés et dans une certaine mesure les ajustements statistiques.

24 Analyses de sensibilité, analyses quantitatives du biais

Les études observationnelles et les modèles statistiques utilisés font de nombreuses hypothèses très variées concernant le design de l'étude, la spécification du modèle d'analyse, la façon dont les variables sont mesurées et définies, la manière dont sont échantillonnés les patients, etc. Ces hypothèses sont presque toutes intestables et la validité des résultats produits dépend du respect de ces hypothèses par les données de l'étude.

> *Par exemple, une hypothèse de base des études observationnelles est qu'il n'existe pas de facteur de confusion non pris en compte par l'analyse (hypothèse NUC,* no unmeasured confounder*).*

Même si ces hypothèses ne sont pas testables ou directement vérifiables, il est possible d'éprouver la sensibilité du résultat de l'étude à ces hypothèses par des analyses de sensibilité, ou mieux, des analyses quantitatives des biais.

Les analyses de sensibilité consistent à faire varier des choix de construction ou d'analyse afin d'éprouver la sensibilité des résultats à ces options.

Les analyses quantitatives de biais consistent à faire des hypothèses numériques sur l'importance de la déviation des données aux hypothèses structurelles de l'étude et de l'analyse et de voir si les résultats obtenus peuvent provenir en partie ou en totalités de ces déviations aux hypothèses. En d'autres termes ces analyses explorent s'il existe des explications alternatives aux résultats obtenus.

▪ Comparaison des logiques sous-jacentes aux 2 approches

Analyse de sensibilité pour justifier de l'absence de biais (ou, plus exactement, du respect des hypothèses faites par l'analyse)	Il est certain que l'analyse de sensibilité respecte l'hypothèse étudiée donc si elle donne un résultat similaire à l'analyse princeps cela amène à conclure que cette analyse princeps respecte aussi cette hypothèse
Analyse quantitative de biais	Compte tenu du contexte de l'étude et en cas d'absence d'effet du traitement, quelle serait la taille de l'écart à l'hypothèse sous-jacente, nécessaire pour expliquer le résultat obtenu. Un tel écart à l'hypothèse de l'analyse est-il possible ? S'il est raisonnablement possible de conclure qu'un tel écart à l'hypothèse faite par l'étude/analyse sur le biais étudié est peu plausible, cela pourra amener la conclusion que le résultat ne provient pas d'un tel biais

Par exemple, ces études font l'hypothèse de l'absence de biais de classification de l'outcome, donc l'hypothèse d'absence d'erreur de classification asymétrique du critère de jugement. Pour explorer la vérification de cette hypothèse dans le cadre de l'étude, la logique de l'analyse de sensibilité est de montrer qu'en utilisant une définition alternative du critère de jugement, pour laquelle il est certain qu'il n'y a pas d'erreur de classification, l'estimation de l'effet du traitement reste inchangée. Cette approche fait à son tour une hypothèse forte : il n'y a pas d'erreur de classification avec la nouvelle

définition du critère de jugement. Une autre limite réside dans la difficulté de faire une conclusion objective à l'absence de changement de résultat.

Pour le même objectif, l'analyse quantitative de biais pourra procéder en faisant l'hypothèse qu'il existe bien une erreur de classification asymétrique et en calculant pour différente valeur de cette erreur (en termes de sensibilité et de spécificité ou simplement en termes de taux de faux positifs et faux négatif) la taille de l'effet traitement que cela pourrait produire sous l'hypothèse nulle d'absence d'effet du traitement. La principale difficulté réside au niveau de l'appréciation de la plausibilité de la taille de l'erreur de classification qui invalide les résultats obtenus.

24.1 Analyses de sensibilité

Les analyses de sensibilités ont pour but de montrer la robustesse des résultats « primaires » vis-à-vis des choix méthodologiques ou de stratégie d'analyse. Elles ont pour objectifs de montrer que le résultat obtenu, dans sa taille ou sa direction, ne dépend pas de la méthode choisie et qu'il est donc robuste.

Derrière cette définition se cachent de nombreuses difficultés :

- Quel degré de variation dans les résultats des analyses de sensibilité récuse la robustesse du résultat ?
- Quelles alternatives d'analyse (méthodes, stratégie) tester ? À quelle condition un changement de méthode d'analyse est en mesure de vraiment montrer la fragilité d'un résultat ?

Les analyses de sensibilité ne peuvent que réfuter la robustesse du résultat et ne doivent pas être utilisées pour choisir le résultat qui sera mis en avant.

Plusieurs options sont possibles pour conclure à la non-confirmation de la robustesse du résultat par une analyse de sensibilité : une simple différence numérique dans les estimations de taille d'effet, la perte de la signification statistique ou la mise en évidence d'une différence statistiquement significative entre les deux estimations. En général rien n'est formalisé et l'interprétation s'effectue dans la nuance ouvrant la voie à des conclusions arbitraires.

Les analyses de sensibilité doivent être construites pour tester des aspects très précis. Par exemple pour explorer la sensibilité des résultats aux erreurs de classification, les analyses de sensibilités peuvent être effectuées en faisant varier les définitions des expositions et des critères de jugement.

D'autres analyses de sensibilités peuvent faire varier la définition des patients et/ou explorer d'autres sources de données. Les options de modélisations peuvent être explorées en faisant varier les modèles et méthodes d'analyses.

Indirectement les analyses de sensibilité peuvent éclairer la discussion de l'absence de HARKing et de p hacking. En effet cette question va devenir centrale s'il s'avère que d'autres stratégies d'analyse, que celle utilisée pour produire les résultats primaires mis en avant, donnent des résultats et/ou des conclusions différentes. Encore faut-il que de telles analyses de sensibilité soient rapportées. Cette discussion renforce la nécessité que ces études apportent une garantie tangible d'absence ce de HARKing et de p hacking (cf. section 10).

Bien que très classiques, les analyses de sensibilité présentent des limites qui résident dans le fait que les différentes variantes d'analyses testées peuvent être toutes sujettes au même biais (par exemple

les différentes définitions des critères de jugement sont sujettes aux mêmes erreurs de classification). L'analyse quantitative des biais permet de dépasser cette limite en introduisant explicitement une erreur dans les données ou l'analyse et en étudiant la sensibilité du résultat à cette erreur.

Ces analyses de sensibilité ne sont très souvent pas ou mal décrites dans les plans d'analyses ce qui ne permet pas de vérifier leur pré-spécification.

24.2 Analyse quantitative du biais

L'analyse quantitative des biais vise à estimer numériquement l'impact potentiel des biais inhérents aux études observationnelles, afin d'évaluer la robustesse des résultats et la plausibilité d'explications alternatives à l'effet observé du médicament [257] [170] [258]. Cette approche s'applique aussi aux comparaisons externes de facto [169] [172].

Dans quelle mesure les biais (confusion, sélection, classification, etc.) pourraient-ils expliquer, à eux seuls, l'effet observé ?

L'idée clé de l'analyse quantitative des biais est de quantifier l'impact potentiel d'un biais sur le résultat de l'étude : "Si ce biais existe, de quelle ampleur faudrait-il qu'il soit pour expliquer entièrement (ou partiellement) l'effet observé ?". Une analyse de point de bascule (*tipping point analysis*) peut-être réalisée en faisant varier numériquement l'hypothèse de biais et en regardant comment l'estimation de l'effet change et, à quel point, la conclusion qualitative change.

Les analyses quantitatives du biais reviennent à explorer si des explications alternatives aux résultats prenant la forme de biais sont plausibles.

Si des hypothèses de biais faibles ou réalistes suffisent à annuler l'effet, les résultats de l'étude apparaitront fragiles. Par contre si seules les hypothèses très fortes et donc peu plausibles modifient la conclusion, les résultats sont robustes et leur crédibilité augmente.

Cette approche a d'abord été appliquée au biais de confusion (cf. section 16.2), entre autres avec la E value (cf. section 16.2.2). Elle est généralisable à d'autres biais potentiels comme les biais de classification des expositions ou des critères de jugement.

Ainsi, l'analyse quantitative des biais permet d'aller au-delà du simple "discussion des biais". Elle permet de passer d'une simple analyse qualitative en une estimation chiffrée de la sensibilité du résulta aux biais. Elle permet numérique de répondre à la question : "À quel point la conclusion de l'étude dépendent-elles d'hypothèses non vérifiables ?".

Les limites de l'analyse quantitative des biais résident dans la difficulté de prendre en compte simultanément toutes les sources potentielles de biais pouvant affecter l'étude et dans l'arbitraire affectant l'évaluation de la plausibilité des tailles de biais remettant en cause le résultat (difficultés d'interprétation des *tipping point analyses* en général).

La réalisation de ces analyses (sensibilité, analyse quantitative de biais) dans un cadre bayésien est en cours d'élaboration [259]. Cette approche permet de prendre en compte l'idée a priori que l'on peut avoir sur l'importance du non-respect des hypothèses de validité de l'inférence effectuée par l'étude.

Principe général de l'analyse quantitative de biais, application au biais de confusion

Soit un facteur pronostic du critère de jugement qui multiplie le risque par RR_c. Soit r_0 le risque d'événements chez les patients qui n'ont pas le facteur pronostic C (patients C-) et $r_0 RR_C$ le risque chez les patients qui présente le facteur pronostique (patients C+). Si dans un échantillon la proportion de patients C+ est p, la proportion de patients C- est donc 1-p, la fréquence de l'événement dans l'échantillon est :

$$\mathrm{r}_0 * (1 - \mathrm{p}) + \mathrm{r}_0 * \mathrm{RR_c} * \mathrm{p} = \mathrm{r0} * \big(1 + \mathrm{p} * (\mathrm{RR_c} - 1)\big)$$

Si maintenant dans le groupe traité (monobras) la proportion de patients C+ est p_1 et celle dans le groupe contrôle externe est p_0 , la fréquence des événements observée dans le groupe traité sera $\mathrm{r}_0 * \big(1 + \mathrm{p}_1 * (\mathrm{RR_c} - 1)\big)$ et $\mathrm{r}_0 * \big(1 + \mathrm{p}_0 * (\mathrm{RR_c} - 1)\big)$ dans le groupe contrôle.

Le risque r_0 des C- est identique dans ces deux groupes du fait de l'ajustement qui a porté sur tous les facteurs de confusion sauf le facteur C.

Sous l'hypothèse nulle d'absence de tout effet du traitement, le risque d'événement est donc le même dans les deux groupes après ajustement et le risque relatif ajusté vrai est $RR_{vrai} = r_0/r_0 = 1$.

En revanche, en raison de la non-prise en compte du facteur C, le risque relatif observé malgré l'ajustement est :

$$\mathrm{RR_{obs}} = \frac{\mathrm{r}_0 * \big(1 + \mathrm{p}_1 * (\mathrm{RR_c} - 1)\big)}{\mathrm{r}_0 * \big(1 + \mathrm{p}_0 * (\mathrm{RR_c} - 1)\big)} = \frac{1 + \mathrm{p}_1 * (\mathrm{RR_c} - 1)}{1 + \mathrm{p}_0 * (\mathrm{RR_c} - 1)}$$

Ce rapport est donc la taille du biais induit par la différence entre les deux groupes de la proportion de patients porteurs d'un facteur pronostique non pris en compte et qui augmentent le risque de l'évènement critère de jugement par un risque relatif RR_c .

Par exemple en prenant RR_c=2, $p_1 = 0.20$ et $p_0 = 0.40$, sous l'hypothèse nulle l'étude donnera un risque relatif de 0.86. Si le risque relatif effectivement obtenu est supérieur à cette valeur, par exemple 0.90, il peut être entièrement expliqué par un tel biais de confusion et ne pourra pas être considéré comme robuste, à l'abri d'un biais. En revanche si le résultat est un risque relatif de 0.50, ce résultat ne peut pas être entièrement dû au biais modélisé. À lui seul cette analyse quantitative de biais ne permet pas d'écarter tout risque de biais, car elle n'a porté que sur un seul des biais qui peuvent affecter une telle étude, mais elle contribue à documenter en partie la robustesse du résultat.

25 Calcul d'effectif

Comme dans un essai clinique, la puissance statistique de la comparaison à un groupe contrôle externe doit être assurée pour limiter le risque d'obtenir une étude non concluante alors que le traitement étudié est réellement supérieur à son contrôle. Il est donc nécessaire qu'un nombre suffisant de patients (conduisant à suffisamment d'événements) soit inclus dans l'étude pour garantir cette puissance. Cela est d'autant plus important que l'étude de comparaison externe est une étude de confirmation.

> *La nécessité d'un calcul d'effectif apparait dans le guide FDA sur les essais non randomisés [48] et sur les études comparatives à contrôle externe [6] (cf. Tableau 14).*

Ce nombre se détermine de la même façon que dans l'essai clinique à partir d'hypothèses sur la taille de l'effet du traitement et sur le risque de base du critère de jugement pour faire simple. Cependant plusieurs points seront spécifiques.

- ## Conséquences des analyses ajustées

Le calcul d'effectif ne peut être qu'approximatif, car il est impossible de prendre en compte les conséquences de l'analyse ajustée en termes de puissance de la comparaison statistique. L'ajustement impacte la puissance en fonction des relations complexes existant entre les covariables, le traitement et le critère de jugement. Ces relations et leurs conséquences en termes de puissance sont impossibles à anticiper. En pratique, il est indispensable de majorer le nombre de sujets donnés par le calcul, de façon arbitraire, pour compenser ce phénomène.

> *La situation la plus simple pour illustrer cette problématique est la pondération par IPW qui conduit à un effective sample size (ESS) réduit par rapport au nombre initial de sujets. L'ESS peut être assimilé aux nombres de sujets contribuant effectivement à la comparaison. Le fait que l'ESS soit inférieur à l'effectif initial montre la réduction de puissance induite par l'ajustement. Mais la valeur de l'ESS n'est pas anticipable avant de réaliser l'analyse et il est donc impossible de le prendre en compte dans le calcul d'effectif.*

- ## Déséquilibres d'effectifs

En général, le groupe contrôle externe est constitué après la réalisation de l'étude monobras (ou du RCT qu'il doit complémenter). L'effectif du groupe traité est donc déjà défini et non modifiable. Seule se posera la question du nombre de sujets du groupe contrôle. En effet, la puissance est conditionnée par le nombre total de patients (ou d'événements) et à effectif fixé du groupe traité, la seule possibilité pour avoir le nombre total de patients (ou d'événements) garantissant la puissance est de jouer sur l'effectif du groupe contrôle.

Cela conduit très fréquemment à un déséquilibre des effectifs, car la taille de l'étude monobras n'a pas été fixée dans la perspective de garantir une certaine puissance à une éventuelle comparaison externe subséquente, mais sur une tout autre base (par exemple obtenir une certaine précision sur l'intervalle de confiance du critère de jugement principal). Le déséquilibre d'effectif n'est pas un problème au niveau statistique, mais un grand déséquilibre (>1:5, 1 :10) pose cependant la question d'un résultat dans lequel la contribution des patients traités sera très minoritaire.

Un déséquilibre d'effectif, nécessaire pour obtenir la puissance voulue, entraîne aussi des répercussions sur la méthode de prise en compte des facteurs de confusion : par exemple un matching devra s'aligner sur ce déséquilibre.

- Rareté des patients contrôles indépendamment de la fréquence de la maladie

Apparait aussi la question de pouvoir constituer un groupe contrôle externe de taille conséquente. Lorsque l'étude concerne une pathologie peu fréquente, il peut s'avérer difficile de trouver une source de données pouvant fournir un grand nombre de patients contrôles. Une solution est d'envisager de constituer le groupe contrôle à partir de plusieurs sources de données (cf. ci-dessous).

- Importance du calcul d'effectifs

Ces difficultés mettent en évidence le fait que la question des effectifs devrait être anticipée dès la conception de l'étude monobras. L'effectif de celle-ci devant être déterminé, non pas seulement pour ses propres besoins, mais surtout dans la perspective de la comparaison externe qu'il faudra à coup sûr réaliser. Cet aspect renforce l'idée que les études monobras ne doivent pas être envisagées comme telles, mais d'emblée comme des études comparatives à contrôle externe. C'est le concept des « *externally controlled study* » proposé par ICH et pour lesquelles la FDA a produit un document spécifique [6]. Cette anticipation permettra de limiter un éventuel déséquilibre d'effectif entre les 2 groupes comparés, diminuera l'effectif nécessaire pour el groupe contrôle le rendant ainsi plus facile à constituer.

Les comparaisons externes ayant pour finalité de démontrer le bénéfice clinique d'un nouveau traitement, il est indispensable de garantir leur puissance statistique pour éviter d'obtenir une étude non concluante aux conséquences fâcheuses du fait d'un manque de puissance

Pour certains auteurs le calcul d'effectifs n'a pas lieu d'être dans les études observationnelles [260]. Cette discussion concerne principalement les études observationnelles classiques réalisées à l'aide de grande base de données nationale dans le cadre de la pharmacoépidémiologie. Dans ces études, l'objectif n'est pas de « détecter » l'effet du traitement, mais de l'estimer le mieux qu'il est possible de faire. De ce fait la question de la précision/puissance statistique a priori ne se pose pas comme dans une étude dont l'objectif est de « détecter » l'effet comme dans les études expérimentales.

Les comparaisons externes sont des approches plus proches des études expérimentales que des études de pharmacoépidémiologie et leur calibration pour garantir une certaine puissance reste un élément indispensable. Si cette comparaison a pour objectif de démontrer le bénéfice d'un nouveau produit, une étude non concluante, simplement par manque de puissance, entraînera des répercussions fâcheuses. En effet, dans le cadre d'une étude de confirmation, un résultat non statistiquement significatif ne permet pas de conclure au bénéfice du nouveau traitement et ne fournit donc pas de preuve pour son adoption dans la stratégie thérapeutique, même si les résultats montrent une tendance en faveur du nouveau traitement. L'étude sera déclarée négative et le nouveau traitement, qui avait besoin de cette étude pour rentrer dans la stratégie thérapeutique, ne le pourra pas. Le calcul d'effectif a priori permet de limiter ce risque.

- Calcul d'effectif et choix de la source de données

Le calcul d'effectif est un élément fondamental qui permettra de choisir la source de données. En effet il est inutile, voire dangereux (cf. supra), de partir d'emblée sur une source qui ne permettra pas de

constituer un groupe contrôle de taille suffisante pour garantir la puissance statistique de la comparaison. Même si le choix de la source de données est un compromis entre la qualité des données, leur informativité (critères d'éligibilité, critères de jugement, facteurs de confusion, contrôles négatifs) et la taille, la question de la puissance de la comparaison ne peut être négligée. Il est donc nécessaire, lors de la phase de faisabilité ou de qualification des données, de tester le nombre potentiel de patients éligible au groupe contrôle externe que peuvent fournir les différentes sources de données envisagées. Cet exercice nécessite parfois d'avoir accès au critère de jugement si le calcul de calibration a été fait en nombre d'événements (ce qui est fortement recommandé pour les critères binaires). Dans ce cas, toutes les précautions doivent être prises pour éviter de compromettre la garantie d'absence de HARKing et de p hacking en documentant de façon convaincante qu'il n'y a pas eu de comparaisons des groupes (analyse inférentielle).

- Que faire en cas de nombre de sujets disponibles insuffisants

Lorsqu'il s'avère qu'aucune source de données ne peut donner le nombre de patients/événements nécessaire, deux solutions peuvent être envisagées.

La première solution consiste à construire le groupe contrôle externe non pas à partir d'une seule source de données, mais de plusieurs (par exemple le registre français de la maladie associée à d'autres registres européens, nord-américains, etc.) [92]. Cette approche augmente la complexité de l'étude et soulève à son tour de nouvelles problématiques pratiques : homogénéisation des variables et des définitions, disponibilité dans toutes sources de données utilisées des variables nécessaires, etc. Se posent aussi des questions en termes d'homogénéité à travers les populations de patients réunies des facteurs de confusion, des déterminants des critères de jugements et d'éventuels modificateurs d'effet.

Se pose aussi la question de réaliser une seule comparaison à un groupe contrôle externe unique, construite par regroupement des différentes sources, ou de réaliser une comparaison pour chaque source et de regrouper ces différentes estimations de l'effet du traitement par méta-analyse. Cette dernière option présente l'avantage supplémentaire de documenter la stabilité des estimations quelle que soit la source utilisée, et, le cas échéant, de renforcer la confiance accordée au résultat.

L'autre solution consiste à différer l'analyse afin d'attendre l'intégration dans la source de données du nombre suffisant de patients et/ou d'événements (de la même façon de ce qui se passe dans un recrutement prospectif). Outre l'inconvénient de ne pouvoir apporter disposer des résultats de la comparaison externe immédiatement, cette option amène aussi la question de la réalisation d'un groupe contrôle comptemporain (cf. section 9.2).

Tableau 14 – Mention du calcul d'effectif dans les guides des agences

Agence	Document	Page	Extrait
EMA	Reflection paper on use of real-world data in noninterventional studies to generate real-world evidence for regulatory purposes	NA	
EMA	Draft Concept Paper on the Development of a Reflection Paper on the Use of External Controls for Evidence Generation in Regulatory Decision-Making	NA	
MHRA	MHRA draft guideline on the use of external control arms based on real- world data to support regulatory decisions	NA	

MHRA	MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions	§3	it is recommended that statistical power calculations are used to assess whether the potential number of patients would enable a clinically important treatment effect to be detected
FDA	Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products[53]	7	Assessment of feasibility, including sample size calculation and anticipated operating characteristics (e.g., statistical power)
FDA	Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products [6]	13	Before conducting an externally controlled trial, sponsors should develop a statistical analysis plan that prespecifies analyses of interest, such as analyses of primary and secondary endpoints, calculations of statistical power and sample size,
HAS	Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health [44]	NA	
NICE	NICE real-world evidence framework [64]	NA	

26 Contrôle du risque alpha global

Au niveau statistique, la production de preuves au-delà de tout doute raisonnable nécessite un contrôle strict du risque alpha global, c'est-à-dire du risque de conclure à tort à un quelconque bénéfice du traitement uniquement du fait des fluctuations aléatoires d'échantillonnage. Tout comme les essais randomisés pivots de phase 3, ces études de comparaisons externes doivent prévoir une gestion de la multiplicité des comparaisons statistiques afin de contrôler le risque alpha global et utiliser les mêmes techniques que celles mises en œuvre habituellement dans les essais cliniques : critère de jugement principal unique ou hiérarchisation des critères de jugement ou répartition du risque alpha avec ou sans réallocation, etc.

Si l'étude est strictement rétrospective, la question des analyses intermédiaires ne se pose pas. Cependant, dans certains cas, ces études peuvent comporter une partie prospective, par exemple lorsque le nombre initial de patients (ou d'événements) dans la base de données (registre, cohorte) au moment de la mise en place de l'étude est insuffisant. Il convient alors d'attendre le nombre nécessaire avant de réaliser l'analyse. Dans ces situations un schéma d'analyses séquentielles peut être prévu en utilisant les méthodes habituelles de contrôle du risque alpha global lors des analyses intermédiaires. Sans cette protection, tout résultat d'analyse « intermédiaire » qui serait réalisé s'apparenterait directement à du p-hacking et rendrait caduque la possibilité de produire des résultats démontrés avec cette étude.

La p value n'est pas un indicateur de la solidité du résultat, en particulier vis-à-vis des biais. La p valeur n'est qu'un instrument pour contrôler (au sens de limiter) le risque de fausse conclusion à un intérêt du traitement, uniquement du fait du hasard (cf. livre blanc Dossier 3 – Le contrôle du risque alpha global [27]).

Une p-value très petite (<0.001 par exemple) ne signifie pas que le résultat est robuste vis-à-vis des biais, elle ne signifie pas que les limites habituelles des comparaisons observationnelles ont été dépassées. Cette robustesse vis-à-vis des biais est en partie documentée par l'analyse quantitative des biais et la E-value (cf. section 16.2) mais pas par la p value.

De même, la p-value n'est pas un indicateur de pertinence clinique de la taille de l'effet. Une p-value petite ne montre pas que la taille de l'effet est importante. Ce sont deux notions différentes. Un petit effet, non cliniquement pertinent, peut-être significatif, voire très significatif, si la puissance de l'étude est importante. Or avec les données médico-adminsitratives, il est possible d'avoir des effectifs considérables (en particulier avec les données populationnelles comme les bases administratives type SNDS) et toutes différences obtenues s'avèrent très significatives. Cela ne donne aucune garantie que la taille de l'effet est importante ou que le risque de biais est supprimé.

[27] https://sfpt-fr.org/livreblancmethodo/part6/file_0.htm

27 Pertinence clinique

Les études de comparaisons externes doivent être en mesure d'apporter des résultats de même pertinence clinique que les essais randomisés pivots de phase 3, étant donné que leurs résultats seront utilisés pour les mêmes usages de décision et de construction de la stratégie thérapeutique.

Ces études doivent donc démontrer le bénéfice clinique du traitement étudié sur des critères de jugement cliniquement pertinents, par rapport à un traitement comparateur loyal et cliniquement pertinent et assurer que la balance bénéfice risque est favorable.

Sur ces points ces études doivent être strictement similaires à ce qu'aurait été un essai randomisé visant à faire la même démonstration du bénéfice clinique du traitement évalué.

27.1 Critère de jugement

Les études de comparaison externes doivent démontrer le bénéfice du traitement sur les critères de jugement nécessaire pour positionner le traitement évalué dans la stratégie thérapeutique. Les critères de jugement doivent donc être des critères cliniques pertinents et non des critères intermédiaires.

Une problématique fréquente est que les critères nécessaires pour l'évaluation du traitement ne sont pas disponibles (documenté) dans les sources de données, car non mesuré en routine clinique et non reconstituable à partir des informations disponibles.

L'exemple typique est celui de la survie sans progression en oncologie (cf. section 20.7) qui nécessite une imagerie itérative régulière et l'utilisation des critères RECIST jamais utilisés en pratique.

L'utilisation de proxy pour remplacer les critères de jugement inaccessibles (comme le temps jusqu'à l'arrêt du traitement pour remplacer la PFS en oncologie) pose alors la question de la démonstration qu'il s'agit bien de véritables surrogates des critères qu'ils veulent remplacer. Démonstration impossible à obtenir compte tenu de l'impossibilité d'avoir les critères dont ils veulent être les surrogates.

27.2 Taille de l'effet

La pertinence clinique de la taille de l'effet dans une étude de comparaison externe se pose de la même façon que dans un essai randomisé de phase 3.

La taille de l'effet est aussi fréquemment regardée sur le plan de la robustesse aux biais du résultat, en partant de l'idée qu'un effet de grande taille est moins susceptible d'être dû au biais qu'un effet plus modeste. La solidité de ce raisonnement peut se discuter en particulier, car il a été montré que des résultats de très grande taille observés dans des études préliminaires étaient rarement retrouvés dans les études de confirmation subséquentes [261]. La question de la solidité méthodologique d'une étude de comparaison externe doit être appréhendée par solidité de la méthodologie de l'étude elle-même et non pas à travers le prisme des résultats produits.

27.3 Traitement comparateur

Tout comme dans les essais randomisés pivots de phase 3, le traitement de comparaison doit être cliniquement pertinent, c'est-à-dire être le traitement connu comme étant le plus intéressant au moment de l'évaluation des résultats de la comparaison externe.

Dans les domaines où les réelles innovations thérapeutiques se succèdent rapidement (comme en oncologie) il est parfois difficile de trouver un contingent suffisant de patients recevant ce traitement dans les sources de données, du fait du peu de recul d'utilisation de ce traitement et du délai d'intégration des données. Cette situation peut conduire à différer l'analyse de l'étude de comparaison externe par rapport à sa conception afin d'attendre l'apparition d'un nombre suffisant de patients dans la source de données. Cette situation nécessite donc de retarder l'utilisation du traitement évalué jusqu'au moment où la comparaison externe est possible. Dans ces situations l'avantage du recrutement prospectif du groupe contrôle en même temps que le groupe traité dans l'essai randomisé devient évident.

27.4 Réalisation, suivi

Dans un essai d'oncologie, la survie globale est souvent influencée par les traitements que reçoivent les patients après la progression (traitements de ligne ultérieure). Le groupe contrôle ne sera pas un comparateur approprié s'il reçoit moins de traitement postprogression ou des traitements moins efficaces que le groupe traité. Une différence de survie pourrait être notée, même si le nouveau traitement n'est pas supérieur sur ce critère de jugement, simplement car il sera favorisé par une prise en charge postprogression plus performante.

Cette situation est d'autant plus possible que le groupe contrôle provient d'une source de données historique, avec des patients traités dans un contexte plus ancien où les traitements postprogression actuels (dont ont bénéficié les patients du groupe traité) n'existaient pas. A l'inverse, cette situation peut aussi être due à un groupe expérimental réalisé dans un contexte de soin plus intensiviste que la pratique médicale habituelle conduisant au même biais de réalisation.

Cette problématique ne concerne pas que l'oncologie et la survie globale, et se généralise à tous les domaines que ce soit en termes de traitement concomitant, de traitement de secours, de traitement subséquent.

Cette problématique met en évidence que les groupes contrôles externes ne doivent pas être trop anciens par rapport au contexte de soins dans lequel les données concernant le nouveau traitement ont été acquises. Il est impératif que le groupe contrôle ait pu bénéficier de la même prise en charge en dehors des traitements comparés en termes de traitements efficaces sur le critère de jugement considéré, de contexte de soin (nursing, mesure d'accompagnement, soins palliatifs, etc.).

> *Suissa et al. évoquent la possibilité d'évolution séculaire dans les groupes contrôles historiques couvrant une période très large [186]. Dans le cas la valeur du critère de jugement dépend de l'année. Cette évolution séculaire peut en théorie être modélisée, mais en faisant des hypothèses souvent irréalistes.*

Ainsi des contrôles historiques recueillis avant les dernières avancées thérapeutiques concernant ces patients ne seront pas adaptés pour ces comparaisons.

En lecture critique, pour évaluer cette limite méthodologique, il va être nécessaire de positionner dans le temps la période de traitement des patients contrôles par rapport à la chronologie d'apparitions des dernières avancées thérapeutiques du domaine.

27.5 Balance bénéfice risque

Démontrer le bénéfice clinique d'un traitement nécessite d'apporter la preuve que la balance bénéfice risque est favorable, c'est-à-dire que l'efficacité du traitement n'est pas susceptible d'être contrebalancée de manière qualitative ou quantitative par les effets indésirables du traitement.

Tout comme les essais randomisés pivots de phase 3, les études de comparaisons externes doivent non seulement démontrer l'efficacité du traitement évalué, mais aussi documenter correctement les effets indésirables afin de permettre une évaluation fiable de la balance bénéfice risque.

En pratique il s'avère difficile et souvent impossible de documenter les effets indésirables dans le groupe contrôle en raison de l'impossibilité de trouver cette information dans les sources de données disponibles. Ces situations conduisent à devoir apprécier la balance bénéfice risque de manière très indirecte à partir des données de safety de l'étude du traitement testé et de données publiées pour le traitement contrôle, avec toutes les limites inhérentes à cet exercice : différences de définition, mesure, appréciation des données de safety, différences de patients, de temporalités, etc.

28 Méta-épidémiologie et étude de cas

28.1 Méta-épidémiologie

Plusieurs travaux de méta-épidémiologie de description des études de comparaisons externes sont disponibles fin 2025 [262] [67] [263] [69] [264] [265] [266] [13] [267] [268].

L'étude par Liu et al. couvre la période allant de 2010 à 2023 [262]. Leurs résultats montrent que la méthodologie, la réalisation et l'analyse de ces études sont sous-optimales limitant la fiabilité et la crédibilité de ces études.

Farah et coll. [264] analysent les comparaisons externes publiées jusqu'en 2022 en oncologie. Parmi les 23 études identifiées et analysées, seules 52% ont utilisé une méthode pour aligner le groupe contrôle avec les caractéristiques du groupe traité.

Une revue systématique des études de comparaison externe en oncohématologie publiée en aout 2024 [265] trouve 32 études et mets en évidence des limitations méthodologiques sur tous les plans qui remettent en question les conclusion de ces études.

Presque toutes les autres études de méta-épidémiologies qui se sont intéressées à la qualité méthodologique des études de comparaisons externes publiées jusqu'à présent [264] [265] [266] [262] mettent en évidence une faible qualité méthodologie de ces travaux limitant fortement leur conclusion.

D'autres études ont revu les évaluations méthodologiques de ce types d'études par des agences de régulations et de HTA [69] [46][269]. Les conclusions sur la qualité de ces études sont similaires et ces travaux sont très rarement pris en considération, avec plusieurs agences ne considérant jamais ces travaux dans leur évaluation.

Confirmant la méta-épidémiologie, à la date de janvier 2026, ils ne nous ont pas été possible d'identifier un travail publié qui pourrait servir d'exemple satisfaisant d'étude de comparaison externe. Toutes les publications passées en revue présentaient de gros manques au niveau de la méthodologie et de l'informativité de la publication.

28.2 Validation empirique

Au-delà de la compréhension au niveau théorique des conditions nécessaires à la fiabilité des résultats des comparaisons externes, leur adoption comme approche possible d'évaluation des nouveaux traitements nécessite de vérifier leur fiabilité par des études de validité empirique.

Ces études consistent à reproduire par des comparaisons externes des résultats d'essais randomisés connus. Si pour la même comparaison de traitement chez les mêmes patients les comparaisons externes permettent d'obtenir les mêmes estimations des bénéfices cliniques que les essais randomisés, ces études apportent des éléments importants de validation de l'approche, même si au niveau théorique persistent des limites méthodologiques non solutionnables ou non solutionnées. Une limite importante de cette approche provient de son aspect rétrospectif. Les résultats à confirmer des essais randomisés sont connus avant de réaliser les comparaisons externes. Ces études doivent donc apporter toutes les garanties d'absence de p-hacking.

Une étude a été réalisée avec les données de la base Flatiron par ses promoteurs [270]. Il s'agit d'un travail rétrospectif réalisé alors que les résultats à reproduire à l'aide d'une comparaison à un groupe contrôle externe étaient connus. Bien que les résultats de ce travail soient en faveur d'une reproductibilité assez bonne des résultats des RCT ils présentent plusieurs limites méthodologiques. [271] empêchant de conclure définitivement.

Depuis ce premier travail de validation empirique, plusieurs autres ont été publiés [270, 271] [272] [273][274] [275] [276] [277] [278] [279] [280] [281] [282], présentant tous les mêmes limites liées à leur réalisation alors que les résultats à reproduire sont connus.

28.3 Études de cas

Plusieurs histoires passées illustrent le manque de fiabilité des comparaisons externe telle que réalisée jusqu'à présent et les dangers de les exploiter pour la prise de décision.

28.3.1 Viltolarsen

Le viltolarsen dans la maladie de Duchenne a été enregistré par la FDA en aout 2020 à partir d'une comparaison de 16 enfants traités par viltolarsen à un groupe contrôle externe de 69 enfants issus de la cohorte d'histoire naturelle CINRG DNHS [283]. Cette comparaison externe montrant une différence en termes de temps ou de vitesse pour parcourir 10m ainsi que sur d'autres critères fonctionnels.

Cet enregistrement était un enregistrement accéléré qui nécessitait la réalisation d'un essai randomisé de confirmation. Un communiqué de presse[28] de la firme en mai 2024 a indiqué que cet essai (NCT04060199, non publié) n'avait pas confirmé le bénéfice du produit chez ces enfants.

28.3.2 Sodium phenylbutyrate et taurursodiol dans la SLA

L'association sodium phenylbutyrate et taurursodiol a été évalué versus placebo dans le traitement de la SLA dans l'essai randomisé CENTAUR. Une comparaison externe du groupe traité de cet essai versus un groupe contrôle externe issu de la cohorte PRO-ACT (Pooled Ressource Open Access ALS Clinical trials) a été réalisé pour évaluer l'effet du traitement sur la survie [12]. Cette comparaison externe met en évidence une amélioration de la survie statistiquement significative.

Un essai randomisé de confirmation (PHOENIX, NCT05021536), entrepris ultérieurement, a échoué à mettre en évidence un tel bénéfice du traitement conduisant la firme à retirer le produit du marché[29].

[28] https://www.nippon-shinyaku.co.jp/file/download.php?file_id=7613
[29] https://ansm.sante.fr/tableau-acces-derogatoire/relyvrio

29 Synopsis - les critères d'acceptabilité des études de comparaisons externes pour la modification des stratégies thérapeutiques

Cette section propose une liste de vérification pour examiner si une étude particulière de comparaison à un groupe contrôle externe satisfait les attentes méthodologiques nécessaires pour exploiter ses résultats pour la construction de la stratégie thérapeutique.

Elle a été établie en tenant compte de la nécessité de disposer de preuves solides pour la modification des stratégies thérapeutiques, des problématiques méthodologiques posées par les comparaisons externes et des solutions actuellement disponibles dans l'état de l'art.

Un résultat issu d'une comparaison externe sera acceptable pour baser un changement de stratégie thérapeutique aux conditions suivantes.

▪ L'étude est une étude de confirmation

Respect de la démarche hypothético-déductive comme dans un essai de phase 3 pivot (essais de confirmation). La satisfaction de ces critères d'acceptabilité est à juger sur la base des éléments suivants :

- Spécification d'un objectif/hypothèses explicite

▪ Il est possible d'écarter la présence de HARKing

(Principalement pour les études rétrospectives)

La satisfaction de ce critère d'acceptabilité est à juger sur la base des éléments attestant que l'objectif/hypothèse a bien été formulé indépendamment des données (avant toute analyse inférentielle) :

- Justification de l'objectif/hypothèse par des études exploratoires précédentes ou des connaissances fondamentales (cohérence des dates).
- Attestation explicite des investigateurs dans le protocole, le rapport, la publication.
- Cohérence des dates entre dates du protocole, date d'accès aux données, date d'analyses.
- Enregistrement/publication du protocole.
- etc.

▪ Il est possible d'écarter la présence de p hacking

Ce critère d'acceptabilité est satisfait si l'analyse statistique n'a pas pu être adaptée en fonction des résultats obtenus (pas d'analyses cachées), à juger sur les éléments suivants :

- Existence d'un plan d'analyse statistique (éventuellement inclus dans le protocole).

- Élément attestant que le *statistical analysis plan* (SAP) a été établi a priori avant toute analyse inférentielle :
    - Attestation explicite des investigateurs dans le protocole, le SAP, le rapport, la publication,
    - Cohérence des dates entre date du protocole, date d'accès aux données, date d'analyse,
    - Enregistrement/publication du protocole.

Les hypothèses de l'inférence causale sont vérifiées

Ce critère d'acceptabilité est satisfait si l'estimand statistique utilisé permet l'identification de l'estimand causal avec les données utilisées, ce qui nécessite que les hypothèses fondamentales de l'inférence causale sont vérifiées par les données et la méthodologie de l'étude :

- Hypothèse de positivité vérifiée.
- Hypothèse de cohérence vérifiée.
- Hypothèse de non-interférence vérifiée.
- Hypothèse d'échangeabilité conditionnelle (appelée aussi hypothèse NUC, no *uncontrolled confounding*) vérifiée (cf. biais de confusion)

(Une simple comparaison avant-après n'est pas une comparaison contrefactuelle acceptable)

L'étude fait une émulation satisfaisante d'un essai cible et l'essai cible émulé par l'étude est satisfaisant

La satisfaction de ce critère d'acceptabilité est à juger sur la base des éléments suivants :

- Protocole ou synopsis de l'essai émulé disponible.
- Protocole de l'essai cible satisfaisant et correspond bien à la question causale de l'étude observationnelle.
- Description de la façon dont ce protocole a été émulé (ou pas) point par point.
- Rapport au format TARGET

Le bénéfice du traitement est appréhendé 1) en termes d'ATT (average treatment effet among treated) et 2) en termes d'effet de l'assignement à un traitement

La satisfaction de ce critère d'acceptabilité est à juger sur la base des éléments suivants :

- Utilisation d'une méthode statistique permettant d'identifier l'ATT comme une pondération inverse IPTW avec des poids appropriés
- Analyse en intention de traiter (« *as started* ») par un estimand de type « *policy treatment* » (et non pas analyse en traitement reçu)

Correction du biais de confusion

L'hypothèse NUC (*no uncontrolled confounders*) est vérifiée. À juger sur la base des éléments suivants :

- Identification de tous les facteurs pronostiques des critères de jugement utilisés par une revue systématique satisfaisante, complété par une recherche satisfaisante des modificateurs d'effets
- Prise en compte de tous les facteurs de confusion (FdC) affectant l'étude pour chaque critère de jugement (ou les covariables prises en compte permettent de bloquer tous les chemins de confusion)
- Détermination formalisée des facteurs de confusion par un graphique de causalité (diagramme acyclique orienté – DAGs - par exemple).
    - Le réseau de casualité a été a bien été le support de la détermination des facteurs de confusion (par opposition à un DAG établi après avoir établi la liste des covariables et qui dans ce cas devient une simple représentation graphique et non pas un outil d'analyse),
    - Pas de sur-ajustement sur des collisionneurs ou des médiateurs,
- Tous les facteurs de confusion identifiés étaient disponibles).
- Les facteurs de confusion ont été mesurés sans erreur.
- La méthode de prise en compte des FdC a atteint son but (par exemple comparabilité des distributions des FdC entre les 2 groupes) attesté par les « *standardized mean difference* » (SMD), représentation des distributions, etc.
- Conformité de l'analyse au SAP.

Le biais de confusion résiduel est négligeable

Justification de l'absence de biais de confusion résiduel (ou que le biais de confusion résiduel n'explique pas la totalité du résultat) à juger sur la base des éléments suivants :

- Contrôles négatifs (ou positif en fonction du sens du résultat) appropriés et montrant l'absence de biais de confusion.
- Analyse quantitative des biais (E value pe) montrant l'absence de biais de confusion.

(attention la recalibration est difficilement acceptable)

L'étude de comparaison externe a été construite en évitant d'introduire un biais de sélection

En particulier

- Absence de sélection des patients sur des variables post début de suivi
- Définition correcte du t0 des débuts de suivi
- Synchronisation des t0 du groupe contrôle avec le groupe traité
- Courbe de survie non évocatrices d'un bais de temps d'immortalité

- Absence de censure informative (liées à des facteurs de risque du endpoint) → vraie analyse en ITT (« *as started* »)

Le risque de biais est faible ou modéré (mesuré par l'outil ROBINS-I ou APPRAISE)

En particulier :

- Pour le biais de sélection
    - Absence de temps d'immortalité, assurée par le design ou l'analyse, confirmée par les résultats (courbes de survie),
    - Synchronisation correcte des t0 dans les deux groupes (hormis études autocontrôlées),
- Pour les biais de mesure et de classification (qualité des données)
    - Validation des données par une vérification cas par cas ou par sondage assurant l'exactitude des données,
    - Validation à l'aide de contrôle positif (ou d'un « benchmarking » préalable de la source de données),
    - Absence d'erreur de mesure ou de classification symétrique pour les résultats concluants à l'absence de différence (« *safety* » par exemple).

Le risque alpha global est strictement contrôlé

Contrôle strict du risque alpha global pour les résultats décisionnels du fait de l'utilisation d'une méthode appropriée comme :

- Répartition (*co-primary endpoints*, Bonferroni, etc.).
- Hiérarchisation.
- Réallocation (Holmes, Hochberg, méthode graphique de Bretz, etc.).

Les résultats sont cliniquement pertinents

La satisfaction de ce critère d'acceptabilité est à juger en fonction de :

- Pertinence des critères de jugement.
- Pertinence des tailles des effets.
- Pertinence du comparateur.
- Pertinence de la balance bénéfice risque.

Il est possibilité d'écarter un biais de publication ou de selective reporting

La satisfaction de ce critère d'acceptabilité est à juger sur la base des éléments suivants :

- Attestation de l'unicité de l'étude réalisée (vérifiée à l'aide des registres et des publications) ou liste de toutes les études similaires réalisées (avec leurs résultats et la méta-analyse).
- Utilisation de plusieurs groupes contrôle externe d'emblée.

Références

1 European Medicines Agency (EMA). ICH E10 Choice of control group in clinical trials - Scientific guideline | European Medicines Agency (EMA) 2001. Available at: https://www.ema.europa.eu/en/ich-e10-choice-control-group-clinical-trials-scientific-guideline Accessed December 28, 2025.

2 Rippin G, Largent J, Hoogendoorn WE, et al. External Comparator Cohort studies - clarification of terminology. *Front. Drug Saf. Regul.* 2024;3 doi:10.3389/fdsfr.2023.1321894;

3 Pasculli G, Virgolin M, Myles P, et al. Synthetic Data in Healthcare and Drug Development: Definitions, Regulatory Frameworks, Issues. *CPT Pharmacometrics Syst Pharmacol* 2025;14:840–52 doi:10.1002/psp4.70021; PMID:40193292;

4 Concato J, Corrigan-Curay J. Real-World Evidence — Where Are We Now? *New England Journal of Medicine* 2022;386:1680–82 doi:10.1056/NEJMp2200089; PMID:35485775;

5 Thorlund K, Dron L, Park JJH, et al. Synthetic and External Controls in Clinical Trials - A Primer for Researchers. *Clin Epidemiol* 2020;12:457–67 doi:10.2147/CLEP.S242097; PMID:32440224;

6 FDA/CDER. Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products ;

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

8 Ou S-HI, Lin HM, Hong J-L, et al. Comparative effectiveness of mobocertinib and standard of care in patients with NSCLC with EGFR exon 20 insertion mutations: An indirect comparison. *Lung Cancer* 2023;179:107186 doi:10.1016/j.lungcan.2023.107186; PMID:37075617;

9 Ludwig H, Terpos E, Boccadoro M, et al. Plitidepsin in combination with dexamethasone (ADMYRE trial) versus an external control arm of pomalidomide plus dexamethasone in patients with relapsed/refractory multiple myeloma. *Ann Hematol* 2026;105:26 doi:10.1007/s00277-026-06811-w; PMID:41545603;

10 Coelho T, Marques W, Dasgupta NR, et al. Eplontersen for Hereditary Transthyretin Amyloidosis With Polyneuropathy. *JAMA* 2023;330:1448–58 doi:10.1001/jama.2023.18688; PMID:37768671;

11 Paganoni S, Macklin EA, Hendrix S, et al. Trial of Sodium Phenylbutyrate-Taurursodiol for Amyotrophic Lateral Sclerosis. *N Engl J Med* 2020;383:919–30 doi:10.1056/NEJMoa1916945; PMID:32877582;

12 Paganoni S, Quintana M, Sherman AV, et al. Analysis of sodium phenylbutyrate and taurursodiol survival effect in ALS using external controls. *Ann Clin Transl Neurol* 2023;10:2297–304 doi:10.1002/acn3.51915; PMID:37807839;

13 Burcu M, Dreyer NA, Franklin JM, et al. Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms. *Pharmacoepidemiol Drug Saf* 2020;29:1228–35 doi:10.1002/pds.4975; PMID:32162381;

14 EMA. Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation.

15 Collignon O, Schritz A, Senn SJ, et al. Clustered allocation as a way of understanding historical controls: Components of variation and regulatory considerations. *Stat Methods Med Res* 2020;29:1960–71 doi:10.1177/0962280219880213; PMID:31599194;

16 Pocock SJ. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases* 1976;29:175–88 doi:10.1016/0021-9681(76)90044-8; PMID:770493;

17 Foucher Y. Boosting the performances of clinical trials from external data or related algorithms/models: conditions to respect ;

18 J. De Keizer, S. Chevret, A. Fernandes, et al. Hybrid randomized clinical trials incorporating external controls: assumptions and related recommendations.

19 Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. *The Lancet* 2002;359:57–61 doi:10.1016/S0140-6736(02)07283-5; PMID:11809203;

20 Cucherat M, Demarcq O, Chassany O, et al. Methodological expectations for demonstration of health product effectiveness by observational studies. *Therapie* 2025;80:47–59 doi:10.1016/j.therap.2024.10.062; PMID:39694790;

21 Abbasi AB, Curtis LH, Califf RM. The Promise of Real-World Data for Research - What Are We Missing? *N Engl J Med* 2025;393:318–21 doi:10.1056/NEJMp2416479; PMID:40689459;

22 Dahly DL, Wilkinson J. Nonrandomized studies of interventions - complementary or just convenient? *Fertility and Sterility* 2025;0 doi:10.1016/j.fertnstert.2025.07.019; PMID:40685106;

23 Fonarow GC. Randomization-There Is No Substitute. *JAMA Cardiol* 2016;1:633–35 doi:10.1001/jamacardio.2016.1792; PMID:27439153;

24 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;

25 Dahly DL, Wilkinson J. Nonrandomized studies of interventions - complementary or just convenient? *Fertility and sterility* 2025;124:657–58 doi:10.1016/j.fertnstert.2025.07.019; PMID:40685106;

26 Hernán MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. *Am J Public Health* 2018;108:616–19 doi:10.2105/AJPH.2018.304337; PMID:29565659;

27 Grodstein F, Stampfer MJ, Manson JE, et al. Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. *N Engl J Med* 1996;335:453–61 doi:10.1056/NEJM199608153350701; PMID:8672166;

28 Manson JE, Hsia J, Johnson KC, et al. Estrogen plus progestin and the risk of coronary heart disease. *N Engl J Med* 2003;349:523–34 doi:10.1056/NEJMoa030808; PMID:12904517;

29 Elm E von, Egger M. The scandal of poor epidemiological research. *BMJ* 2004;329:868–69 doi:10.1136/bmj.329.7471.868; PMID:15485939;

30 Cummings JL, Atri A, Feldman HH, et al. evoke and evoke+: design of two large-scale, double-blind, placebo-controlled, phase 3 studies evaluating efficacy, safety, and tolerability of semaglutide in early-stage symptomatic Alzheimer's disease. *Alzheimers Res Ther* 2025;17:14 doi:10.1186/s13195-024-01666-7; PMID:39780249;

31 Wang W, Wang Q, Qi X, et al. Associations of semaglutide with first-time diagnosis of Alzheimer's disease in patients with type 2 diabetes: Target trial emulation using nationwide real-world data in the US. *Alzheimers Dement* 2024;20:8661–72 doi:10.1002/alz.14313; PMID:39445596;

32 Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ* 2016;352:i493 doi:10.1136/bmj.i493; PMID:26858277;

33 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;

34 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;

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

36 Woolacott N, Corbett M, Jones-Diette J, et al. Methodological challenges for the evaluation of clinical effectiveness in the context of accelerated regulatory approval: an overview. *Journal of Clinical Epidemiology* 2017;90:108–18 doi:10.1016/j.jclinepi.2017.07.002; PMID:28709997;

37 Heyard R, Held L, Schneeweiss S, et al. Design differences and variation in results between randomised trials and non-randomised emulations: meta-analysis of RCT-DUPLICATE data. *BMJ Medicine* 2024;3:e000709 doi:10.1136/bmjmed-2023-000709; PMID:38348308;

38 CHMP. Abecma; INN-idecabtagene vicleucel ;

39 CHMP. Abecma; INN-idecabtagene vicleucel ;

40 Vickers AJ, Assel M, Dunn RL, et al. Guidelines for Reporting Observational Research in Urology: The Importance of Clear Reference to Causality. *Eur Urol* 2023 doi:10.1016/j.eururo.2023.04.027; PMID:37286459;

41 Dahabreh IJ, Bibbins-Domingo K. Causal Inference About the Effects of Interventions From Observational Studies in Medical Journals. *JAMA* 2024;331:1845–53 doi:10.1001/jama.2024.7741; PMID:38722735;

42 Wieseler B, Neyt M, Kaiser T, et al. Replacing RCTs with real world data for regulatory decision making: a self-fulfilling prophecy? *BMJ* 2023;380:e073100 doi:10.1136/bmj-2022-073100; PMID:36863730;

43 FDA Eliminates Major Barrier to Using Real-World Evidence in Drug and Device Application Reviews: FDA Mon, 12/15/2025 - 12:55.

44 Vanier A, Fernandez J, Kelley S, et al. Rapid access to innovative medicinal products while ensuring relevant health technology assessment. Position of the French National Authority for Health. *BMJ Evidence-Based Medicine* 2024;29:1–5 doi:10.1136/bmjebm-2022-112091; PMID:36788020;

45 Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Oncology Center of Excellence (OCE). Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products.

46 Monnereau M, Delord J-P, Michiels S, et al. Acceptance of external control arms by HTA agencies: a review of oncology submissions in France, England, Germany and Norway from 2021 to 2023. *British journal of cancer* 2025 doi:10.1038/s41416-025-03155-6; PMID:40940536;

47 McCord M. FDA Broad Agency Announcement (BAA) for Advanced Research and Development of Regulatory Science.

48 Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Oncology Center of Excellence (OCE). Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products Guidance for Industry ;

49 (Mon, 09/22/2025 - 19:08). FDA use of Real-World Evidence in Regulatory Decision Making. *FDA,* Mon, 09/22/2025 - 19:08. Available at: https://www.fda.gov/science-research/real-world-evidence/fda-use-real-world-evidence-regulatory-decision-making Accessed September 26, 2025.

50 CBER C. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices - Guidance for Industry and Food and Drug Administration Staff ;

51 FDA/ CDER CBER OCE. Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products ;

52 Stewart A. Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products ;

53 Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Oncology Center of Excellence (OCE). Real-World Evidence: Considerations Regarding Non-Interventional Studies for Drug and Biological Products Guidance for Industry ;

54 draft-reflection-paper-establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence-marketing-authorisation_en ;

55 Abellan Andres Juan Jose, RWE tDG. Reflection paper on use of real-world data in noninterventional studies to generate real-world evidence for regulatory purposes ;

56 European Medicines Agency. Concept paper on the revision of the guideline on the evaluation of anticancer medicinal products and appendices.

57 European Medicines Agency. Draft Concept Paper on the Development of a Reflection Paper on the Use of External Controls for Evidence Generation in Regulatory Decision-Making ;

58 uk Cg. MHRA draft guideline on the use of external control arms based on real-world data to support regulatory decisions ;

59 Weiss M. ICH_M14_Step4_Final_Guideline_2025_0905 ;

60 Berger ML, Dreyer N, Anderson F, et al. Prospective observational studies to assess comparative effectiveness: the ISPOR good research practices task force report. *Value Health* 2012;15:217–30 doi:10.1016/j.jval.2011.12.010; PMID:22433752;

61 Berger ML, Mamdani M, Atkins D, et al. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part I. *Value in Health* 2009;12:1044–52 doi:10.1111/j.1524-4733.2009.00600.x; PMID:19793072;

62 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. *Pharmacoepidemiol Drug Saf* 2017;26:1033–39 doi:10.1002/pds.4297; PMID:28913966;

63 Cox E, Martin BC, van Staa T, et al. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report--Part II. *Value Health* 2009;12:1053–61 doi:10.1111/j.1524-4733.2009.00601.x; PMID:19744292;

64 National Institute for Health and Care Excellence (NICE). NICE real-world evidence framework ;

65 JCA. Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons ;

66 JCA. Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons ;

67 Patel D, Grimson F, Mihaylova E, et al. Use of External Comparators for Health Technology Assessment Submissions Based on Single-Arm Trials. *Value Health* 2021;24:1118–25 doi:10.1016/j.jval.2021.01.015; PMID:34372977;

68 Jaksa A, Louder A, Maksymiuk C, et al. A Comparison of Seven Oncology External Control Arm Case Studies: Critiques From Regulatory and Health Technology Assessment Agencies. *Value in Health* 2022;25:1967–76 doi:10.1016/j.jval.2022.05.016; PMID:35760714;

69 Wang X, Dormont F, Lorenzato C, et al. Current perspectives for external control arms in oncology clinical trials: Analysis of EMA approvals 2016-2021. *Journal of Cancer Policy* 2023;35:100403 doi:10.1016/j.jcpo.2023.100403; PMID:36646208;

70 Mangla KK, Kolovos S, Lisica A, et al. Acceptability of external control-arm use in nononcology health technology assessment submissions. *J Comp Eff Res* 2026;15:e250073 doi:10.57264/cer-2025-0073; PMID:41384576;

71 Hwang TJ, Carpenter D, Lauffenburger JC, et al. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Intern Med* 2016;176:1826–33 doi:10.1001/jamainternmed.2016.6008; PMID:27723879;

72 Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics* 2019;20:273–86 doi:10.1093/biostatistics/kxx069; PMID:29394327;

73 Lefeuvre C, Antonio M de, Bouhour F, et al. Characteristics of Patients With Late-Onset Pompe Disease in France: Insights From the French Pompe Registry in 2022. *Neurology* 2023;101:e966-e977 doi:10.1212/WNL.0000000000207547; PMID:37419682;

74 Broussais F, Bay JO, Boissel N, et al. DESCAR-T, le registre national des patients traités par CAR-T Cells. *Bull Cancer* 2021;108:S143-S154 doi:10.1016/j.bulcan.2021.07.002; PMID:34920797;

75 Bai YG, Xu L, Duan XN, et al. The Breast Cancer Cohort Study in Chinese Women: research design and preliminary results of clinical multi-center cohort. *Zhonghua Liu Xing Bing Xue Za Zhi* 2020;41:2046–52 doi:10.3760/cma.j.cn112338-20200507-00694; PMID:33378815;

76 Rahman R, Ventz S, Redd R, et al. Accessible Data Collections for Improved Decision Making in Neuro-Oncology Clinical Trials. *Clin Cancer Res* 2023;29:2194–98 doi:10.1158/1078-0432.CCR-22-3524; PMID:36939557;

77 Tennant PWG, Arnold KF, Ellison GTH, et al. Analyses of 'change scores' do not estimate causal effects in observational data. *Int J Epidemiol* 2022;51:1604–15 doi:10.1093/ije/dyab050; PMID:34100077;

78 Jahanshahi M, Gregg K, Davis G, et al. The Use of External Controls in FDA Regulatory Decision Making. *Ther Innov Regul Sci* 2021;55:1019–35 doi:10.1007/s43441-021-00302-y; PMID:34014439;

79 Wagner AK, Soumerai SB, Zhang F, et al. Segmented regression analysis of interrupted time series studies in medication use research. *J Clin Pharm Ther* 2002;27:299–309 doi:10.1046/j.1365-2710.2002.00430.x; PMID:12174032;

80 Berger KI, Chien Y-H, Dubrovsky A, et al. Changes in forced vital capacity over ≤ 13 years among patients with late-onset Pompe disease treated with alglucosidase alfa: new modeling of real-world data from the Pompe Registry. *J Neurol* 2024 doi:10.1007/s00415-024-12489-9; PMID:38896264;

81 Gault N, Castañeda-Sanabria J, Rycke Y de, et al. Self-controlled designs in pharmacoepidemiology involving electronic healthcare databases: a systematic review. *BMC Med Res Methodol* 2017;17:25 doi:10.1186/s12874-016-0278-0; PMID:28178924;

82 Kawala CR, Ma X, Sykes J, et al. Real-world use of ivacaftor in Canada: A retrospective analysis using the Canadian Cystic Fibrosis Registry. *J Cyst Fibros* 2021;20:1040–45 doi:10.1016/j.jcf.2021.03.008; PMID:33810992;

83 San Sebastián M, Mosquera PA, Gustafsson PE. Do cardiovascular disease prevention programs in northern Sweden impact on population health? An interrupted time series analysis. *BMC Public Health* 2019;19:202 doi:10.1186/s12889-019-6514-x; PMID:30770750;

84 Wanner C, Feldt-Rasmussen U, Jovanovic A, et al. Cardiomyopathy and kidney function in agalsidase beta-treated female Fabry patients: a pre-treatment vs. post-treatment analysis. *ESC Heart Fail* 2020;7:825–34 doi:10.1002/ehf2.12647; PMID:32100468;

85 Tchetgen Tchetgen EJ, Park C, Richardson DB. Universal Difference-in-Differences for Causal Inference in Epidemiology. *Epidemiology (Cambridge, Mass.)* 2024;35:16–22 doi:10.1097/EDE.0000000000001676; PMID:38032801;

86 Petersen I, Douglas I, Whitaker H. Self controlled case series methods: an alternative to standard epidemiological study designs. *BMJ* 2016;354:i4515 doi:10.1136/bmj.i4515; PMID:27618829;

87 Desai RJ, Wang SV, Sreedhara SK, et al. Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center. *BMJ* 2024;384:e076460 doi:10.1136/bmj-2023-076460; PMID:38346815;

88 Kok JWTM de, van Bussel BCT, Schnabel R, et al. Table 0; documenting the steps to go from clinical database to research dataset. *Journal of Clinical Epidemiology* 2024;170:111342 doi:10.1016/j.jclinepi.2024.111342; PMID:38574979;

89 Gatto NM, Campbell UB, Rubinstein E, et al. The Structured Process to Identify Fit-For-Purpose Data: A Data Feasibility Assessment Framework. *Clin Pharmacol Ther* 2022;111:122–34 doi:10.1002/cpt.2466; PMID:34716990;

90 Gatto NM, Vititoe SE, Rubinstein E, et al. A Structured Process to Identify Fit-for-Purpose Study Design and Data to Generate Valid and Transparent Real-World Evidence for Regulatory Uses. *Clinical Pharmacology & Therapeutics* 2023;113:1235–39 doi:10.1002/cpt.2883; PMID:36871138;

91 Hall GC, Sauer B, Bourke A, et al. Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf* 2012;21:1–10 doi:10.1002/pds.2229; PMID:22069180;

92 Du Ogier Terrail J, Klopfenstein Q, Li H, et al. FedECA: federated external control arms for causal inference with time-to-event data in distributed settings. *Nat Commun* 2025;16:7496 doi:10.1038/s41467-025-62525-z; PMID:40804048;

93 Arora A, Wagner SK, Carpenter R, et al. The urgent need to accelerate synthetic data privacy frameworks for medical research. *The Lancet Digital Health* 2025;7:e157-e160 doi:10.1016/S2589-7500(24)00196-1; PMID:39603900;

94 Azizi Z, Zheng C, Mosquera L, et al. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 2021;11:e043497 doi:10.1136/bmjopen-2020-043497; PMID:33863713;

95 Akiya I, Ishihara T, Yamamoto K. Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study. *JMIR Medical Informatics* 2024;12:e55118 doi:10.2196/55118; PMID:38889082;

96 Elvatun S, Knoors D, Brant S, et al. Synthetic data as external control arms in scarce single-arm clinical trials. *PLOS Digit Health* 2025;4:e0000581 doi:10.1371/journal.pdig.0000581; PMID:39847598;

97 Koul A, Duran D, Hernandez-Boussard T. Synthetic data, synthetic trust: navigating data challenges in the digital revolution. *The Lancet Digital Health* 2025;7:100924 doi:10.1016/j.landig.2025.100924; PMID:41330822;

98 Elm E von, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLOS Medicine* 2007;4:e296 doi:10.1371/journal.pmed.0040296; PMID:17941714;

99 FDA. CRL_NDA210862_20251104.

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

101 Huebner M, Vach W, Le Cessie S, et al. Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. *BMC Med Res Methodol* 2020;20:61 doi:10.1186/s12874-020-00942-y; PMID:32169053;

102 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;

103 Hiemstra B, Keus F, Wetterslev J, et al. DEBATE-statistical analysis plans for observational studies. *BMC Med Res Methodol* 2019;19:233 doi:10.1186/s12874-019-0879-5; PMID:31818263;

104 Keele L, Grieve R. So Many Choices: A Guide to Selecting Among Methods to Adjust for Observed Confounders. *Stat Med* 2025;44:e10336 doi:10.1002/sim.10336; PMID:39947224;

105 Wang SV, Pottegård A, Crown W, et al. HARmonized Protocol Template to Enhance Reproducibility of Hypothesis Evaluating Real-World Evidence Studies on Treatment Effects: A Good Practices Report of a Joint ISPE/ISPOR Task Force. *Value in Health* 2022;25:1663–72 doi:10.1016/j.jval.2022.09.001; PMID:36241338;

106 Castelo-Branco L, Pellat A, Martins-Branco D, et al. ESMO Guidance for Reporting Oncology real-World evidence (GROW). *Annals of Oncology* 2023;34:1097–112 doi:10.1016/j.annonc.2023.10.001; PMID:37848160;

107 Berlin JA, Fihn SD. Encouraging the Registration of Observational Studies. *JAMA Netw Open* 2025;8:e2524181 doi:10.1001/jamanetworkopen.2025.24181; PMID:40711795;

108 Loder E, Groves T, MacAuley D. Registration of observational studies. *BMJ* 2010;340:c950 doi:10.1136/bmj.c950; PMID:20167643;

109 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;

110 Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* 2021;372:m4856 doi:10.1136/bmj.m4856; PMID:33436424;

111 Wang SV, Schneeweiss S. Data Checks Before Registering Study Protocols for Health Care Database Analyses. *JAMA* 2024;331:1445–46 doi:10.1001/jama.2024.2988; PMID:38587830;

112 Hernán MA. Causal Inference. What if.

113 Pearl J. Causal Inference in Statistics.

114 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;

115 Goetghebeur E, Le Cessie S, Stavola B de, et al. Formulating causal questions and principled statistical answers. *Stat Med* 2020;39:4922–48 doi:10.1002/sim.8741; PMID:32964526;

116 Humphreys ABC, Matthews AA, Young JC, et al. The definition of treatment assignment in observational emulations of target trials - an empirical examination in the Swedish Primary Care Cardiovascular Database. *Ann Epidemiol* 2025;108:56–62 doi:10.1016/j.annepidem.2025.06.003; PMID:40506003;

117 Fang Y, Zhong S. The Targeted Virtual Control Approach for Single-Arm Clinical Trials with External Controls. *Statistics in Biopharmaceutical Research* 2023;15:802–11 doi:10.1080/19466315.2022.2154260;

118 David M. Phillippo, A. E. Ades, Sofia Dias, Stephen Palmer, Keith R. Abrams, Nicky J. Welton. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE 2016.

119 Weckstein AR, Wang SV, Wyss R, et al. Scalable confounding adjustment in real-world evidence: benchmarking data-adaptive and investigator-specified strategies in a large-scale trial emulation study. *J Am Med Inform Assoc* 2025 doi:10.1093/jamia/ocaf204; PMID:41338229;

120 Shrier I, Platt RW. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol* 2008;8:70 doi:10.1186/1471-2288-8-70; PMID:18973665;

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

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

123 Lipsky AM, Greenland S. Causal Directed Acyclic Graphs. *JAMA* 2022;327:1083–84 doi:10.1001/jama.2022.1816; PMID:35226050;

124 Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology* 2022;142:264–67 doi:10.1016/j.jclinepi.2021.08.001; PMID:34371103;

125 Williamson EJ, Aitken Z, Lawrie J, et al. Introduction to causal diagrams for confounder selection. *Respirology* 2014;19:303–11 doi:10.1111/resp.12238; PMID:24447391;

126 Biostatistics in Biopharmaceutical Research and Development: Springer, Cham 2024.

127 Eriksson JW, Bodegard J, Nathanson D, et al. Sulphonylurea compared to DPP-4 inhibitors in combination with metformin carries increased risk of severe hypoglycemia, cardiovascular events, and all-cause mortality. *Diabetes research and clinical practice* 2016;117:39–47 doi:10.1016/j.diabres.2016.04.055; PMID:27329021;

128 Tools | Cochrane Prognosis 2025. Available at: https://methods.cochrane.org/prognosis/tools Accessed December 23, 2025.

129 Damen JAA, Moons KGM, van Smeden M, et al. How to conduct a systematic review and meta-analysis of prognostic model studies. *Clin Microbiol Infect* 2023;29:434–40 doi:10.1016/j.cmi.2022.07.019; PMID:35934199;

130 Riley RD, Moons KGM, Snell KIE, et al. A guide to systematic review and meta-analysis of prognostic factor studies. *BMJ* 2019;364:k4597 doi:10.1136/bmj.k4597; PMID:30700442;

131 Hayden JA, Côté P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. *Ann. Intern. Med.* 2006;144:427–37 doi:10.7326/0003-4819-144-6-200603210-00010; PMID:16549855;

132 Henry ML, O'Connell NE, Riley RD, et al. AMSTAR-PF: a critical appraisal tool for systematic reviews of prognostic factor studies. *BMJ* 2025;391:e085718 doi:10.1136/bmj-2025-085718; PMID:41429459;

133 Pufulete M, Mahadevan K, Johnson TW, et al. Confounders and co-interventions identified in non-randomized studies of interventions. *Journal of Clinical Epidemiology* 2022;148:115–23 doi:10.1016/j.jclinepi.2022.03.018; PMID:35346782;

134 Hogervorst MA, Soman KV, Gardarsdottir H, et al. Analytical Methods for Comparing Uncontrolled Trials With External Controls From Real-World Data: A Systematic Literature Review and Comparison With European Regulatory and Health Technology Assessment Practice. *Value in Health* 2025;28:161–74 doi:10.1016/j.jval.2024.08.002; PMID:39241824;

135 Loiseau N, Trichelair P, He M, et al. External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning. *BMC Med Res Methodol* 2022;22:335 doi:10.1186/s12874-022-01799-z; PMID:36577946;

136 Holt M, Kelly RJ, Fermont JM, et al. Effectiveness of Iptacopan Versus C5 Inhibitors in Complement Inhibitor-Naive Patients With Paroxysmal Nocturnal Haemoglobinuria. *EJHaem* 2025;6:e270055 doi:10.1002/jha2.70055; PMID:40395624;

137 Privitera S, Sedghamiz H, Hartenstein A, et al. An evolutionary algorithm for the direct optimization of covariate balance between nonrandomized populations. *Pharm Stat* 2024;23:288–307 doi:10.1002/pst.2352; PMID:38111126;

138 Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983;70:41 doi:10.2307/2335942;

139 Shiba K, Kawahara T. Using Propensity Scores for Causal Inference: Pitfalls and Tips. *Journal of epidemiology* 2021;31:457–63 doi:10.2188/jea.JE20210145; PMID:34121051;

140 Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011;46:399–424 doi:10.1080/00273171.2011.568786; PMID:21818162;

141 Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149–56 doi:10.1093/aje/kwj149; PMID:16624967;

142 Andrew BY, Alan Brookhart M, Pearse R, et al. Propensity score methods in observational research: brief review and guide for authors. *Br J Anaesth* 2023;131:805–09 doi:10.1016/j.bja.2023.06.054; PMID:37481434;

143 Simoneau G, Pellegrini F, Debray TP, et al. Recommendations for the use of propensity score methods in multiple sclerosis research. *Mult Scler* 2022;28:1467–80 doi:10.1177/13524585221085733; PMID:35387508;

144 Rizk JG. When and why to use overlap weighting: clarifying its role, assumptions, and estimand in real-world studies. *Journal of Clinical Epidemiology* 2025;187:111942 doi:10.1016/j.jclinepi.2025.111942; PMID:40850393;

145 Liu Y, Wang Y, Gao Y, et al. A tutorial for propensity score weighting methods under violations of the positivity assumption 2025.

146 Lee J, Lee H, Yoon D, et al. Lazertinib versus Platinum-Based Chemotherapy with Epidermal Growth Factor Receptor (EGFR)-Positive Non-Small-Cell Lung Cancer after Failing EGFR-Tyrosine Kinase Inhibitor: A Real-World External Comparator Study. *Cancers (Basel)* 2024;16 doi:10.3390/cancers16122169; PMID:38927875;

147 Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009;28:3083–107 doi:10.1002/sim.3697; PMID:19757444;

148 Wang SV, Schneeweiss S, Rassen JA. Optimal matching ratios in drug safety surveillance. *Epidemiology* 2014;25:772–73 doi:10.1097/EDE.0000000000000148; PMID:25076153;

149 Rassen JA, Shelat AA, Myers J, et al. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol Drug Saf* 2012;21 Suppl 2:69–80 doi:10.1002/pds.3263; PMID:22552982;

150 Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American journal of epidemiology* 2010;172:1092–97 doi:10.1093/aje/kwq224; PMID:20802241;

151 Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ* 2019;367:l5657 doi:10.1136/bmj.l5657; PMID:31645336;

152 Brookhart MA, Wyss R, Layton JB, et al. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes* 2013;6:604–11 doi:10.1161/CIRCOUTCOMES.113.000359; PMID:24021692;

153 Källberg D, Waernbaum I. Large Sample Properties of Entropy Balancing Estimators of Average Causal Effects. *Econometrics and Statistics* 2023 doi:10.1016/j.ecosta.2023.11.004;

154 Phillippo DM, Dias S, Ades AE, et al. Equivalence of entropy balancing and the method of moments for matching-adjusted indirect comparison. *Research synthesis methods* 2020;11:568–72 doi:10.1002/jrsm.1416; PMID:32395870;

155 Hainmueller J. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Polit. anal.* 2012;20:25–46 doi:10.1093/pan/mpr025;

156 Rolfo C, Hess LM, Jen M-H, et al. External control cohorts for the single-arm LIBRETTO-001 trial of selpercatinib in RET+ non-small-cell lung cancer. *ESMO Open* 2022;7:100551 doi:10.1016/j.esmoop.2022.100551; PMID:35930972;

157 Nguyen T-L, Collins GS, Spence J, et al. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC Med Res Methodol* 2017;17:78 doi:10.1186/s12874-017-0338-0; PMID:28454568;

158 Shinozaki T, Nojima M. Misuse of Regression Adjustment for Additional Confounders Following Insufficient Propensity Score Balancing. *Epidemiology* 2019;30:541–48 doi:10.1097/EDE.0000000000001023; PMID:31166216;

159 Moccia C, Moirano G, Popovic M, et al. Machine learning in causal inference for epidemiology. *Eur J Epidemiol* 2024;39:1097–108 doi:10.1007/s10654-024-01173-x; PMID:39535572;

160 Bi Q, Goodman KE, Kaminsky J, et al. What is Machine Learning? A Primer for the Epidemiologist. *American journal of epidemiology* 2019;188:2222–39 doi:10.1093/aje/kwz189; PMID:31509183;

161 Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *American journal of epidemiology* 2017;185:65–73 doi:10.1093/aje/kww165; PMID:27941068;

162 Talbot D, Diop A, Mésidor M, et al. Guidelines and Best Practices for the Use of Targeted Maximum Likelihood and Machine Learning When Estimating Causal Effects of Exposures on Time-To-Event Outcomes. *Stat Med* 2025;44:e70034 doi:10.1002/sim.70034; PMID:40079648;

163 van der Laan MJ. Targeted maximum likelihood based causal inference: Part I. *Int J Biostat* 2010;6:Article 2 doi:10.2202/1557-4679.1211; PMID:21969992;

164 Groenwold RHH, Hak E, Hoes AW. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. *Journal of Clinical Epidemiology* 2009;62:22–28 doi:10.1016/j.jclinepi.2008.02.011; PMID:18619797;

165 Groenwold RHH. Falsification end points for observational studies. *JAMA* 2013;309:1769–70 doi:10.1001/jama.2013.3089;

166 Lipsitch M, Tchetgen ET, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology* 2010;21:383–88 doi:10.1097/EDE.0b013e3181d61eeb;

167 Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? *JAMA* 2013;309:241–42 doi:10.1001/jama.2012.96867;

168 Gray C, Ralphs E, Fox MP, et al. Use of quantitative bias analysis to evaluate single-arm trials with real-world data external controls. *Pharmacoepidemiol Drug Saf* 2024;33:e5796 doi:10.1002/pds.5796; PMID:38680093;

169 Gupta A, Hsu G, Kent S, et al. Quantitative Bias Analysis for Single-Arm Trials With External Control Arms. *JAMA Netw Open* 2025;8:e252152 doi:10.1001/jamanetworkopen.2025.2152; PMID:40136297;

170 Lash TL, Fox MP, Cooney D, et al. Quantitative Bias Analysis in Regulatory Settings. *Am J Public Health* 2016;106:1227–30 doi:10.2105/AJPH.2016.303199; PMID:27196652;

171 Leahy TP, Durand-Zaleski I, Sampietro-Colom L, et al. The role of quantitative bias analysis for nonrandomized comparisons in health technology assessment: recommendations from an expert workshop. *Int J Technol Assess Health Care* 2023;39:e68 doi:10.1017/S0266462323002702; PMID:37981828;

172 Thorlund K, Duffield S, Popat S, et al. Quantitative bias analysis for external control arms using real-world data in clinical trials: a primer for clinical researchers. *J Comp Eff Res* 2024:e230147 doi:10.57264/cer-2023-0147; PMID:38205741;

173 Yin X, Stuart E, Burcu M, et al. Assessing the impact of unmeasured confounding in external control arms via tipping point analyses. *J. Clin. Oncol.* 2024;42:e23065-e23065 doi:10.1200/JCO.2024.42.16_suppl.e23065;

174 Risk of bias tools - ROBINS-I V2 tool 2025. Available at: https://www.riskofbias.info/welcome/robins-i-v2 Accessed November 02, 2025.

175 Popat S, Liu SV, Scheuer N, et al. Addressing challenges with real-world synthetic control arms to demonstrate the comparative effectiveness of Pralsetinib in non-small cell lung cancer. *Nat Commun* 2022;13:3500 doi:10.1038/s41467-022-30908-1; PMID:35715405;

176 Rippin G, Sanz H, Hoogendoorn WE, et al. Examining the Effect of Missing Data and Unmeasured Confounding on External Comparator Studies: Case Studies and Simulations. *Drug Saf* 2024;47:1245–63 doi:10.1007/s40264-024-01467-9; PMID:39102176;

177 Soutar S, Macdougall A, Wallis J, et al. Flexible quantitative bias analysis for unmeasured confounding in subject-level indirect treatment comparisons with proportional hazards violation. *BMC Med Res Methodol* 2025;25:131 doi:10.1186/s12874-025-02551-z; PMID:40348970;

178 Gaster T, Eggertsen CM, Støvring H, et al. Quantifying the impact of unmeasured confounding in observational studies with the E value. *bmjmed* 2023;2:e000366 doi:10.1136/bmjmed-2022-000366; PMID:37159620;

179 VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann. Intern. Med.* 2017;167:268–74 doi:10.7326/M16-2607; PMID:28693043;

180 Faillie J-L, Suissa S. Le biais de temps immortel dans les études pharmacoépidémiologiques définition, solutions et exemples. *Therapie* 2015;70:259–63 doi:10.2515/therapie/2014207; PMID:25487848;

181 Yadav K, Lewis RJ. Immortal Time Bias in Observational Studies. *JAMA* 2021;325:686–87 doi:10.1001/jama.2020.9151; PMID:33591334;

182 Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol* 2015;11:437–41 doi:10.1038/nrrheum.2015.30; PMID:25800216;

183 Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Current epidemiology reports* 2015;2:221–28 doi:10.1007/s40471-015-0053-5; PMID:26954351;

184 Stürmer T, Wang T. Active Comparator New User Cohort Studies and Matching. *JAMA Intern Med* 2026;186:122 doi:10.1001/jamainternmed.2025.5792;

185 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;

186 Suissa S. Single-arm Trials with Historical Controls: Study Designs to Avoid Time-related Biases. *Epidemiology* 2021;32:94–100 doi:10.1097/EDE.0000000000001267; PMID:33009252;

187 Hernán MA. How to estimate the effect of treatment duration on survival outcomes using observational data. *BMJ* 2018;360:k182 doi:10.1136/bmj.k182; PMID:29419381;

188 Antunes L, Rippin G, Ralphs E, et al. Choosing an Index Date for Untreated Patients in External Comparator Studies. *Drug Saf* 2025:1–12 doi:10.1007/s40264-025-01613-x; PMID:41021206;

189 Backenroth D. How to choose a time zero for patients in external control arms. *Pharm Stat* 2021;20:783–92 doi:10.1002/pst.2107; PMID:33655598;

190 Cui Z, Khanal M, Chen Y, et al. MSR105 Proportional Randomization Method to Identify Index Line of Therapy in Externally Controlled Trials. *Value in Health* 2024;27:S280 doi:10.1016/j.jval.2024.03.2420;

191 Hatswell AJ, Deighton K, Snider JT, et al. Approaches to Selecting "Time Zero" in External Control Arms with Multiple Potential Entry Points: A Simulation Study of 8 Approaches. *Med Decis Making* 2022;42:893–905 doi:10.1177/0272989X221096070; PMID:35514320;

192 van Le H, Benedetti M de, Yue L, et al. Effect of designations of index date in externally controlled trials: an empirical example. *Epidemiologic Methods* 2024;13 doi:10.1515/em-2023-0041;

193 Orbach D, Carton M, Khadir SK, et al. Therapeutic benefit of larotrectinib over the historical standard of care in patients with locally advanced or metastatic infantile fibrosarcoma (EPI VITRAKVI study). *ESMO Open* 2024;9:103006 doi:10.1016/j.esmoop.2024.103006; PMID:38657345;

194 van der Sluis, Inge M, Lorenzo P de, Kotecha RS, et al. Blinatumomab Added to Chemotherapy in Infant Lymphoblastic Leukemia. *N Engl J Med* 2023;388:1572–81 doi:10.1056/NEJMoa2214171; PMID:37099340;

195 FDA. eflornithine (brand name Iwilfin) for pediatric high-risk neuroblastoma.

196 Willems S, Schat A, van Noorden MS, et al. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. *Stat Methods Med Res* 2018;27:323–35 doi:10.1177/0962280216628900; PMID:26988930;

197 Fu EL, Harhay MO, Schneeweiss S, et al. Starting right: aligning eligibility and treatment assignment at time zero when emulating a target trial. *BMJ* 2026;392:e084909 doi:10.1136/bmj-2025-084909; PMID:41526041;

198 Danaei G, Rodríguez LAG, Cantero OF, et al. Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. *Stat Methods Med Res* 2013;22:70–96 doi:10.1177/0962280211403603; PMID:22016461;

199 Dickerman BA, García-Albéniz X, Logan RW, et al. Avoidable flaws in observational analyses: an application to statins and cancer. *Nature medicine* 2019;25:1601–06 doi:10.1038/s41591-019-0597-x; PMID:31591592;

200 Ren Y, Jia Y, Liu L, et al. Design and Implementation of Observational Studies Emulating a Target Trial. *JAMA Netw Open* 2026;9:e2558262 doi:10.1001/jamanetworkopen.2025.58262; PMID:41712213;

201 Zhou Z, Rahme E, Abrahamowicz M, et al. Survival bias associated with time-to-treatment initiation in drug effectiveness evaluation: a comparison of methods. *Am J Epidemiol* 2005;162:1016–23 doi:10.1093/aje/kwi307; PMID:16192344;

202 García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol* 2017;32:495–500 doi:10.1007/s10654-017-0287-2; PMID:28748498;

203 Christiaens A, Simon-Tillaux N, Thompson W, et al. Impact of deintensifying hypoglycaemic drugs in older adults with type 2 diabetes: protocol for an emulation of a target trial. *BMJ Open* 2023;13:e073081 doi:10.1136/bmjopen-2023-073081; PMID:37984943;

204 Schneeweiss S, Rassen JA, Brown JS, et al. Graphical Depiction of Longitudinal Study Designs in Health Care Databases. *Ann. Intern. Med.* 2019;170:398–406 doi:10.7326/M18-3079; PMID:30856654;

205 Nourredine M, Gavoille A, Lepage C, et al. Accounting for Misclassification of Binary Outcomes in External Control Arm Studies for Unanchored Indirect Comparisons: Simulations and Applied Example. *Stat Med* 2025;44:e70236 doi:10.1002/sim.70236; PMID:40930536;

206 Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980;112:564–69 doi:10.1093/oxfordjournals.aje.a113025; PMID:7424903;

207 Sheffield KM, Bowman L, Smith DM, et al. Development and validation of a claims-based approach to proxy ECOG performance status across ten tumor groups. *J Comp Eff Res* 2018;7:193–208 doi:10.2217/cer-2017-0040; PMID:29533694;

208 Graham S, Graham J, O'Rourke J, et al. Machine Learning Approach to Estimating ECOG PS for a Multiple-Myeloma Cohort from Real World Data. *Blood* 2023;142:4700 doi:10.1182/blood-2023-182252;

209 Salloum RG, Smith TJ, Jensen GA, et al. Using claims-based measures to predict performance status score in patients with lung cancer. *Cancer* 2011;117:1038–48 doi:10.1002/cncr.25677; PMID:20957722;

210 MHRA guidance on the use of real-world data in clinical studies to support regulatory decisions - GOV.UK ;

211 Hallway A, Isenberg E, Howard R, et al. Medicare Coding Changes and Reported Hernia Size. *The Journal of the American Medical Association* 2025 doi:10.1001/jama.2024.26829;

212 Gibson AD, White NM, Collins GS, et al. Evidence of Unreliable Data and Poor Data Provenance in Clinical Prediction Model Research and Clinical Practice 2026.

213 European Medicines Agency. Data Quality Framework for EU medicines regulation: application to Real-World Data ;

214 Fox MP, Lash TL, Bodnar LM. Common misconceptions about validation studies. *Int J Epidemiol* 2020;49:1392–96 doi:10.1093/ije/dyaa090; PMID:32617564;

215 Schelde AB, Kornholt J. Validation studies in epidemiologic research: estimation of the positive predictive value. *Journal of Clinical Epidemiology* 2021;137:262–64 doi:10.1016/j.jclinepi.2021.05.009; PMID:34022395;

216 Wang Y-W, Liu C-C, Chen H-C, et al. Assessing the Validity of Claims-Based Diagnostic Codes for Psychotic and Affective Disorders and the Influence of the Coding Transition from the ICD-9 to the ICD-10 in Taiwan's National Health Insurance Research Database. *Clin Epidemiol* 2025;17:635–45 doi:10.2147/CLEP.S522618; PMID:40661787;

217 Ando T, Ooba N, Mochizuki M, et al. Positive predictive value of ICD-10 codes for acute myocardial infarction in Japan: a validation study at a single center. *BMC Health Serv Res* 2018;18:895 doi:10.1186/s12913-018-3727-0; PMID:30477501;

218 Fujihara K, Yamada-Harada M, Matsubayashi Y, et al. Accuracy of Japanese claims data in identifying diabetes-related complications. *Pharmacoepidemiol Drug Saf* 2021;30:594–601 doi:10.1002/pds.5213; PMID:33629363;

219 Lee H, Sparks JA, Lee SB, et al. Validation of serostatus of rheumatoid arthritis using ICD-10 codes in administrative claims data. *Pharmacoepidemiol Drug Saf* 2023;32:586–91 doi:10.1002/pds.5597; PMID:36728737;

220 Paik JM, Patorno E, Zhuo M, et al. Accuracy of identifying diagnosis of moderate to severe chronic kidney disease in administrative claims data. *Pharmacoepidemiol Drug Saf* 2022;31:467–75 doi:10.1002/pds.5398; PMID:34908211;

221 Roy L, Zappitelli M, White-Guay B, et al. Agreement Between Administrative Database and Medical Chart Review for the Prediction of Chronic Kidney Disease G category. *Can J Kidney Health Dis* 2020;7:2054358120959908 doi:10.1177/2054358120959908; PMID:33101698;

222 Thurin NH, Bosco-Levy P, Blin P, et al. Intra-database validation of case-identifying algorithms using reconstituted electronic health records from healthcare claims data. *BMC Med Res Methodol* 2021;21:95 doi:10.1186/s12874-021-01285-y; PMID:33933001;

223 Lee CD, Carnahan RM, McPheeters ML. A systematic review of validated methods for identifying Bell's palsy using administrative or claims data. *Vaccine* 2013;31 Suppl 10:K7-11 doi:10.1016/j.vaccine.2013.04.040; PMID:24331076;

224 Abraha I, Montedori A, Serraino D, et al. Accuracy of administrative databases in detecting primary breast cancer diagnoses: a systematic review. *BMJ Open* 2018;8:e019264 doi:10.1136/bmjopen-2017-019264; PMID:30037859;

225 Lanes S, Beachler DC. Validation to correct for outcome misclassification bias. *Pharmacoepidemiol Drug Saf* 2023;32:700–03 doi:10.1002/pds.5601; PMID:36751117;

226 European Medicines Agency. Data Quality Framework for EU medicines regulation: application to Real-World Data ;

227 Estevez M, Singh N, Dyson L, et al. Ensuring Reliability of Curated EHR-Derived Data: The Validation of Accuracy for LLM/ML-Extracted Information and Data (VALID) Framework 2025.

228 Velummailum RR, McKibbon C, Brenner DR, et al. Data Challenges for Externally Controlled Trials: Viewpoint. *J Med Internet Res* 2023;25:e43484 doi:10.2196/43484; PMID:37018021;

229 Mhatre SK, Machado RJM, Ton TGN, et al. Real-World Progression-Free Survival as an Endpoint in Lung Cancer: Replicating Atezolizumab and Docetaxel Arms of the OAK Trial Using Real-World Data. *Clin Pharmacol Ther* 2023;114:1313–22 doi:10.1002/cpt.3045; PMID:37696652;

230 Ackerman B, Gan RW, Meyer CS, et al. Measurement error and bias in real-world oncology endpoints when constructing external control arms. *Front. Drug Saf. Regul.* 2024;4 doi:10.3389/fdsfr.2024.1423493;

231 Zeng L, Cook RJ, Wen L, et al. Bias in progression-free survival analysis due to intermittent assessment of progression. *Stat Med* 2015;34:3181–93 doi:10.1002/sim.6529; PMID:26011411;

232 Zhu J, Tang RS. A proper statistical inference framework to compare clinical trial and real-world progression-free survival data. *Stat Med* 2022;41:5738–52 doi:10.1002/sim.9590; PMID:36199170;

233 Edwards JK, Cole SR, Zivich PN, et al. Risk functions with outcome measurement error. *Biostatistics* 2026;27 doi:10.1093/biostatistics/kxaf052; PMID:41555577;

234 Cren P-Y, Leguillette C, Craynest F, et al. Estimating overall survival by combining administrative and hospital death data: a methodological challenge. *Eur J Epidemiol* 2025:1–11 doi:10.1007/s10654-025-01278-x; PMID:41118097;

235 Hsu W-C, Crowley A, Parzynski CS. The impact of different censoring methods for analyzing survival using real-world data with linked mortality information: a simulation study. *BMC Med Res Methodol* 2024;24:203 doi:10.1186/s12874-024-02313-3; PMID:39272007;

236 Sterne JA, 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; PMID:27733354;

237 Bykov K, Jaksa A, Lund JL, et al. APPRAISE: A Tool for Appraising Potential for Bias in Real-world Evidence Studies on Medication Effectiveness or Safety. *Value Health* 2025 doi:10.1016/j.jval.2025.07.024; PMID:40774597;

238 D'Andrea E, Vinals L, Patorno E, et al. How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools. *BMJ Open* 2021;11:e043961 doi:10.1136/bmjopen-2020-043961; PMID:33762237;

239 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;

240 Hernán MA, Dahabreh IJ, Dickerman BA, et al. The Target Trial Framework for Causal Inference From Observational Data: Why and When Is It Helpful? *Ann Intern Med* 2025;178:402–07 doi:10.7326/ANNALS-24-01871; PMID:39961105;

241 Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American journal of epidemiology* 2016;183:758–64 doi:10.1093/aje/kwv254; PMID:26994063;

242 Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA* 2022;328:2446–47 doi:10.1001/jama.2022.21383; PMID:36508210;

243 Arnold K, Antunes L, Coles B, et al. Application of the target trial emulation framework to external comparator studies. *Front. Drug Saf. Regul.* 2024;4 doi:10.3389/fdsfr.2024.1380568;

244 Tran V-T, Porcher R, Perrodeau E, et al. Practical elements to consider when emulating a target trial. *Journal of Clinical Epidemiology* 2026;0:112205 doi:10.1016/j.jclinepi.2026.112205;

245 Cashin AG, Hansford HJ, Hernán MA, et al. Transparent Reporting of Observational Studies Emulating a Target Trial-The TARGET Statement. *JAMA* 2025 doi:10.1001/jama.2025.13350; PMID:40899949;

246 Wang SV, Schneeweiss S, Franklin JM, et al. Emulation of Randomized Clinical Trials With Nonrandomized Database Analyses: Results of 32 Clinical Trials. *JAMA* 2023;329:1376–85 doi:10.1001/jama.2023.4221; PMID:37097356;

247 Simon-Tillaux N, Martin GL, Hajage D, et al. Conducting observational analyses with the target trial emulation approach: a methodological systematic review. *BMJ Open* 2024;14:e086595 doi:10.1136/bmjopen-2024-086595; PMID:39532374;

248 Zhao SS, Lyu H, Solomon DH, et al. Improving rheumatoid arthritis comparative effectiveness research through causal inference principles: systematic review using a target trial emulation framework. *Ann Rheum Dis* 2020;79:883–90 doi:10.1136/annrheumdis-2020-217200; PMID:32381560;

249 Scola G, Chis Ster A, Bean D, et al. Implementation of the trial emulation approach in medical research: a scoping review. *BMC Med Res Methodol* 2023;23:186 doi:10.1186/s12874-023-02000-9; PMID:37587484;

250 Zuo H, Yu L, Campbell SM, et al. The implementation of target trial emulation for causal inference: a scoping review. *Journal of Clinical Epidemiology* 2023;162:29–37 doi:10.1016/j.jclinepi.2023.08.003; PMID:37562726;

251 Merola D, Campbell U, Lenis D, et al. Calibrating Observational Health Record Data Against a Randomized Trial. *JAMA Netw Open* 2024;7:e2436535 doi:10.1001/jamanetworkopen.2024.36535; PMID:39348118;

252 Dahabreh IJ, Robins JM, Hernán MA. Benchmarking Observational Methods by Comparing Randomized Trials and Their Emulations. *Epidemiology* 2020;31:614–19 doi:10.1097/EDE.0000000000001231; PMID:32740470;

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

254 Schuemie MJ, Hripcsak G, Ryan PB, et al. Robust empirical calibration of p-values using observational data. *Stat Med* 2016;35:3883–88 doi:10.1002/sim.6977; PMID:27592566;

255 Schuemie MJ, Hripcsak G, Ryan PB, et al. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci U S A* 2018;115:2571–77 doi:10.1073/pnas.1708282114; PMID:29531023;

256 Wang SV, Russo M, Glynn RJ, et al. A Benchmark, Expand, and Calibration (BenchExCal) Trial Emulation Approach for Using Real-World Evidence to Support Indication Expansions: Design and Process for a Planned Empirical Evaluation. *Clin Pharmacol Ther* 2025;117:1820–28 doi:10.1002/cpt.3621; PMID:40067205;

257 Haine D. Quantitative Bias Analysis for Epidemiologic Data 2025. Available at: https://cran.r-project.org/web/packages/episensr/vignettes/episensr.html Accessed January 25, 2026.

258 Brown JP, Hunnicutt JN, Ali MS, et al. Quantifying possible bias in clinical and epidemiological studies with quantitative bias analysis: common approaches and limitations. *BMJ* 2024;385:e076365 doi:10.1136/bmj-2023-076365; PMID:38565248;

259 Oganisian A. Stress-Testing Assumptions: A Guide to Bayesian Sensitivity Analyses in Causal Inference 2026.

260 Hernán MA. Causal analyses of existing databases: no power calculations required. *Journal of Clinical Epidemiology* 2022;144:203–05 doi:10.1016/j.jclinepi.2021.08.028; PMID:34461211;

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

262 Liu J, Yao M, Wang M, et al. Design, Conduct, and Analysis of Externally Controlled Trials. *JAMA Netw Open* 2025;8:e2530277 doi:10.1001/jamanetworkopen.2025.30277; PMID:40906478;

263 Silva P, Janjan N, Ramos KS, et al. External control arms: COVID-19 reveals the merits of using real world evidence in real-time for clinical and public health investigations. *Frontiers in Medicine* 2023;10:1198088 doi:10.3389/fmed.2023.1198088; PMID:37484840;

264 Farah E, Kenney M, Warkentin MT, et al. Examining external control arms in oncology: A scoping review of applications to date. *Cancer Medicine* 2024;13:e7447 doi:10.1002/cam4.7447; PMID:38984669;

265 Hermans SJF, van der Maas NG, van Norden Y, et al. Externally Controlled Studies Using Real-World Data in Patients With Hematological Cancers: A Systematic Review. *JAMA Oncol* 2024;10:1426–36 doi:10.1001/jamaoncol.2024.3466; PMID:39207765;

266 Hogervorst MA, Soman KV, Gardarsdottir H, et al. Analytical Methods for Comparing Uncontrolled Trials With External Controls From Real-World Data: A Systematic Literature Review and Comparison With European Regulatory and Health Technology Assessment Practice. *Value Health* 2025;28:161–74 doi:10.1016/j.jval.2024.08.002; PMID:39241824;

267 Vaghela S, Tanni KA, Banerjee G, et al. A systematic review of real-world evidence (RWE) supportive of new drug and biologic license application approvals in rare diseases. *Orphanet J Rare Dis* 2024;19:117 doi:10.1186/s13023-024-03111-2; PMID:38475874;

268 Zayadi A, Edge R, Parker CE, et al. Use of external control arms in immune-mediated inflammatory diseases: a systematic review. *BMJ Open* 2023;13:e076677 doi:10.1136/bmjopen-2023-076677; PMID:38070932;

269 Appiah K, Rizzo M, Sarri G, et al. Justifying the source of external comparators in single-arm oncology health technology submissions: a review of NICE and PBAC assessments. *J Comp Eff Res* 2024;13:e230140 doi:10.57264/cer-2023-0140; PMID:38174576;

270 Carrigan G, Whipple S, Capra WB, et al. Using Electronic Health Records to Derive Control Arms for Early Phase Single-Arm Lung Cancer Trials: Proof-of-Concept in Randomized Controlled Trials. *Clin Pharmacol Ther* 2020;107:369–77 doi:10.1002/cpt.1586; PMID:31350853;

271 Larrouquere L, Giai J, Cracowski J-L, et al. Externally Controlled Trials: Are We There Yet? *Clin Pharmacol Ther* 2020;108:918–19 doi:10.1002/cpt.1881; PMID:32542679;

272 Ventz S, Khozin S, Louv B, et al. The design and evaluation of hybrid controlled trials that leverage external data and randomization. *Nat Commun* 2022;13:5783 doi:10.1038/s41467-022-33192-1; PMID:36184621;

273 Ventz S, Lai A, Cloughesy TF, et al. Design and Evaluation of an External Control Arm Using Prior Clinical Trials and Real-World Data. *Clin Cancer Res* 2019;25:4993–5001 doi:10.1158/1078-0432.CCR-19-0820; PMID:31175098;

274 Ganame S, Walter T, Durand A, et al. Proof of concept and design of an externally controlled trial for patients with gastro-enteropancreatic neuroendocrine carcinomas based on the randomized phase II BEVANEC trial. *European journal of cancer (Oxford, England 1990)* 2025;225:115450 doi:10.1016/j.ejca.2025.115450; PMID:40340189;

275 Signorovitch J, Moshyk A, Zhao J, et al. Overall survival in the real-world and clinical trials: a case study validating external controls in advanced melanoma. *Future Oncol* 2022;18:1321–31 doi:10.2217/fon-2021-1054; PMID:35048743;

276 Jemielita T, Widman L, Fox C, et al. Replication of Oncology Randomized Trial Results using Swedish Registry Real World-Data: A Feasibility Study. *Clinical Pharmacology & Therapeutics* 2021;110:1613–21 doi:10.1002/cpt.2424; PMID:34549809;

277 Bahmane S, Harbron C, Incerti D, et al. Meta-Analysis of Bias in Non-Small Cell Lung Cancer External Control Arms That Use Real-World Progression-Free Survival as the End Point. *JCO Clin Cancer Inform* 2025;9:e2500198 doi:10.1200/CCI-25-00198; PMID:41270248;

278 Swaminathan AC, Snyder LD, Hong H, et al. External Control Arms in Idiopathic Pulmonary Fibrosis Using Clinical Trial and Real-World Data Sources. *American journal of respiratory and critical care medicine* 2023;208:579–88 doi:10.1164/rccm.202210-1947OC; PMID:37384378;

279 Schröder C, Lawrance M, Li C, et al. Building External Control Arms From Patient-Level Electronic Health Record Data to Replicate the Randomized IMblaze370 Control Arm in Metastatic Colorectal Cancer. *JCO Clin Cancer Inform* 2021;5:450–58 doi:10.1200/CCI.20.00149; PMID:33891473;

280 Ton TGN, Pal N, Trinh H, et al. Replication of Overall Survival, Progression-Free Survival, and Overall Response in Chemotherapy Arms of Non-Small Cell Lung Cancer Trials Using Real-World Data. *Clin Cancer Res* 2022;28:2844–53 doi:10.1158/1078-0432.CCR-22-0471; PMID:35511917;

281 Tan K, Bryan J, Segal B, et al. Emulating Control Arms for Cancer Clinical Trials Using External Cohorts Created From Electronic Health Record-Derived Real-World Data. *Clin Pharmacol Ther* 2022;111:168–78 doi:10.1002/cpt.2351; PMID:34197637;

282 Abrahami D, Pradhan R, Yin H, et al. Use of Real-World Data to Emulate a Clinical Trial and Support Regulatory Decision Making: Assessing the Impact of Temporality, Comparator Choice, and Method of Adjustment. *Clin Pharmacol Ther* 2021;109:452–61 doi:10.1002/cpt.2012; PMID:32767673;

283 Clemens PR, Rao VK, Connolly AM, et al. Safety, Tolerability, and Efficacy of Viltolarsen in Boys With Duchenne Muscular Dystrophy Amenable to Exon 53 Skipping: A Phase 2 Randomized Clinical Trial. *JAMA Neurol* 2020;77:982–91 doi:10.1001/jamaneurol.2020.1264; PMID:32453377;

284 Mackenzie CF, Dubé GP, Pitman AN. Re-analysis of the PolyHeme Phase III trial dataset: Lessons for future haemoglobin-based oxygen carrier trauma trials. *Injury* 2023;54:110712 doi:10.1016/j.injury.2023.03.040; PMID:37100694;

30 Annexes

Tableau 15 - Exemples de nouveaux traitements (ou nouvelles indications) qui se sont avérés augmenter la mortalité dans leur essais pivots sans que cela soit suspecté avec les études préliminaires

Umbralisib	Lymphome folliculaire L4+	UNITY-CLL NCT02612311
Idelalisib, copanlisib, duvelisib		Analyse FDA[30]
Gemtuzumab ozogamicin	CD33+ AML	SWOG S0106 NCT00085709
Olaparib	Cancer de l'ovaire avancé BRCA mute L4+	SOLO 3
Melphalan flufenamide (Pepaxto)	Myélome multiple L5+	OCEAN
Magrolimab (anticorps anti-CD47) (https://www.gilead.com/-/media/files/pdfs/other/magrolimab-trials-summary.pdf)	patients with untreated higher-risk myelodysplastic syndrome	ENHANCE
	patients with untreated acute myeloid leukemia with mutated TP53	ENHANCE 2
	patients with untreated acute myeloid leukemia unfit for intensive therapy	ENHANCE 3
Encainide et flecainide (anti-arythmiques, essai)	Post infarctus du myocarde	CAST
Torcetrapib	Prevention cardiovascualire	ILLUMINATE
PolyHeme (substitut de sang) [284]	treatment of hemorrhagic shock	NCT00076648
V710 Vaccine (Staphylococcus aureus)	Prevention of Staphylococcus aureus Infections After Cardiothoracic Surgery	NCT00518687

[30] U.S. Food and Drug Administration. Food And Drug Administration. Center For Drug Evaluation and Research. Oncologic Drugs Advisory Committee (ODAC) meeting. Virtual meeting. April 21, 2022. https://www.fda.gov/media/159920/download

Tableau 16 – Molécules commercialisées à tort dans le cadre d'enregistrement accéléré FDA et retirées en raison de la non-confirmation du bénéfice lorsque celui-ci était recherché ultérieurement par un essai de confirmation

Molécule	Pathologie	Essai de confirmation
Atezolizumab	Cancer du sein avancé triple négatif et PD-L1 positif	IMpassion-131 NCT03125902 OS 1.11 [0.76; 1.64]
Pembrolizumab	Cancer du poumon à petite cellule métastatique L2+	KEYNOTE-604 NCT03066778 OS 0.80 [0.64; 0.98] NS
Nivolumab	Cancer du poumon à petite cellule métastatique L2+	CheckMate 451 NCT02538666 OS 0.84 [0.69; 1.02]
Pembrolizumab	Adénocarcinome gastrique ou gastro-œsophagien avancé PD-L1 positif L3+	KEYNOTE-061 NCT02370498 OS 0.94 [0.79; 1.12]
Nivolumab	Carcinome hépatocellulaire L2	CheckMate 459 NCT02576509 OS 0.85 [0.72; 1.02]
Durvalumab	Carcinome urothélial avancé L2	DANUBE NCT02516241 OS 0.99 [0.83; 1.17]
Atezolizumab	Carcinome urothélial avancé PD-L1 positif	IMvigor-130 NCT02807636 OS 1.02 [0.83; 1.24]
Gefitinib	Cancer du poumon non à petite cellule avancé L2	ISEL OS 0.89 [0.77; 1.02]
Umbralisib	Lymphome folliculaire L4+ marginal zone lymphoma (MZL) L2	UNITY-CLL NCT02612311 increasing imbalance in survival in favor of the control arm
Belantamab mafodotin-blmf	Myeloma multiple L5+	DREAMM-3 NCT04162210 OS 1.14 [0.77; 1.68]
Duvelisib	Lymphome folliculaire L3+	DUO NCT02004522 OS 0.99 [0.65; 1.50]
Ibrutinib	Marginal Zone Lymphoma (MZL) L2+	SELENE NCT01974440 did not meet its primary endpoint of progression-free survival.
Ibrutinib	Lymphome cellule du manteau	SHINE NCT01776840 OS 1.07 [0.81; 1.40]
Panobinostat	Myélome multiple L3+	PANORAMA 1 NCT01023308 Abandon de l'essai de confirmation
Idelalisib	follicular B-cell Non-Hodgkin Lymphoma (FL) L3+	Non présentation de l'essai de confirmation demandé par la FDA
Romidepsin	Lymphoma à cellules T périphériques L2	Ro-CHOP NCT01796002 OS 0.90 [0.68; 1.20]
Gemtuzumab ozogamicin	CD33+ AML	SWOG S0106 NCT00085709 Augmentation des décès toxique
Olaparib	Cancer de l'ovaire avancé BRCA mute L4+	SOLO 3 Augmentation de la mortalité
Melphalan flufenamide (Pepaxto)	Myélome multiple L5+	OCEAN Tendance augmentation DC 1·10 [95% CI 0·85-1·44]

Tableau 17 – Exemples de molécules initialement adoptées sur la base d'une étude non comparative et dont le bénéfice clinique n'a pas été confirmé par un essai randomisé

Traitement	Étude mono-bras	Essai randomisé non concluant
Pembrolizumab plus docetaxel, mCRPC cancer de la prostate métastatique résistant à la castration	KEYNOTE-365 (NCT02861573)	Keynote-921 (NCT03834506)
Sotorasib, cancer du poumon NAPC	CodeBreak 100 (NCT03600883)	CodeBreak 200 (NCT04303780)
Epacadostat par-dessus le pembrolizumab dans le mélanome avancé	ECHO-202/KEYNOTE-037 (Taux de réponse 55%)	ECHO-301/KEYNOTE-252 (taux de réponse groupe traité 34%)
Pembrolizumab monothérapie dans le cancer de la vessie	KEYNOTE-05231	KEYNOTE-361 arrêté prématurément pour surmortalité
Pembrolizumab dans le carcinome hépatocellulaire	KEYNOTE-224	KEYNOTE-240
L'atezolizumab en première ligne du cancer de la vessie métastatique	IMvigor 210 (NCT02108652)	IMvigor 211 et Imvigor130 (NCT02807636)
Atezolizumab en association avec le nab-paclitaxel) dans le cancer du sein triple négatif métastatique PD-L1-positif	IMpassion130	IMpassion131
Belantamab mafodotin dans le myélome multiple RRMM	DREAMM-2 (NCT03525678)	DREAMM-3 (NCT04162210)
Pembrolizumab pour le cancer du poumon métastatique à petites cellules en 2ème ligne	KEYNOTE-158 (NCT02628067) KEYNOTE-028 (NCT02054806)	KEYNOTE-604 (NCT03066778)
Nivolumab pour le cancer du poumon métastatique à petites cellules en 2ème ligne	CheckMate-032 (NCT01928394)	CheckMate-451 (NCT02538666) CheckMate-331 (NCT02481830)
Nivolumab dans l'hépatocarcinome	CHECKMATE-040 (NCT 01658878)	CheckMate-459 (NCT02576509)
ibrutinib in MZL (marginal zone lymphoma)	PCYC-1121 (NCT01980628)	SELENE (NCT01974440)
ibrutinib in MCL (mantle cell lymphoma)	PCYC-1104 (NCT01236391)	SHINE (NCT01776840)
umbralisib in CLL (chronic lymphocytic leukemia)	UNITY-NHL (NCT02793583)	UNITY-CLL (NCT02612311)
mobocertinib in non-small cell lung cancer with EGFR exon 20 insertion mutations	Exclaim (NCT02716116)	Exclaim-2 (NCT04129502) arrêté pour futilité
Olaparib cancer de l'ovaire	NCT01078662	SOLO 3 (NCT02282020) toujours pas publié, OS HR=1 et >1 chez les L3+

31 10.1016/S1470-2045(17)30616-2

Index

A

AIPW, 82, 105
algorithme phénotypique, 48
analyse de faisabilité, 56
analyse en intention de traiter, 137
analyse per protocole, 80
analyses intermédiaires, 167
APPRAISE, 151
ATC, 71
ATE, 74
ATT, 71, 78
Augmented Inverse Probability Weighting, 105
average treatment effect, 74
average treatment effect among control, 71
average treatment effect among the treated, 78
average treatment effect among treated, 71
avis d'experts, 45

C

causal machine learning, 107
chainage, 147
change from baseline, 52
change score, 52
chart review, 47, 147
claims databases, 47
clinicaltrials.gov, 62
collider, 94
collisionneur, 94
comparaison indirecte, 20
critère de jugement, 168

D

data management, 147
Démarche hypothético déductive, 67
déplétion des susceptibles, 114
données artificielles, 21
double robust method, 105

E

échangeabilité conditionnelle, 73, 82
effective sample size, 103
effet traitement moyen, 74
émulation d'un essai cible, 153
émulation de l'essai cible, 80
émulation séquentielle, 128
entrepôt de données de santé, 147
entrepôt de données de santé de soin courant, 47
ESMO, 62
ESS, 103
essai clinique randomisé, 45
essai pivot, 25
étude monobras, 22
external control arm (ECA), 21
externally controlled study, 21, 164
externally controlled trial, 21, 40, 64

F

fluctuations aléatoires d'échantillonnage, 167

G

g-computation, 82
groupe contrôle externe
    définition, 19
groupe contrôle synthétique, 20
guide de rédaction, 152

H

HARKing, 56, 57
HARPER, 65
HAS, 62
hypothèse d'échangeabilité conditionnelle, 73
hypothèse de cohérence, 71, 137
Hypothèse de positivité, 70
hypothèse NUC, 73
hypothèse NUC no unobserved confusion, 85
hypothèse STUVA, 137
Hypothèse SUTVA, 71

I

IA, 21, 92, 107
inception cohort, 116
inférence causale, 137
in-silico (étude), 21
instrument, 94
intention de traiter, 79
intention de traiter (analyse), 137
inverse probability of censuring weighting, 124
IPCW, 124
ISPOR, 62, 65
ITT, 79

J

jumeau numérique, 21

M

machine learning, 92, 107

mécanisme d'action, 45
médiateur, 94
méthode doublement robuste, 105
méthode double-robuste, 82
multiplicité, 167

N

NCT (numéro), 62
NICE, 62
no unmeasured confounder, 73
non randomized interventional study, 40
non-infériorité, 67
non-interférence, 71
nouvelles méthodologies, 45

O

OS, 150
outcome potentiel, 74
overall survival, 150

P

p value, 167
patient incidents, 115
patients prévalents, 114
pertinence clinique, 168
PFS, 148
p-hacking, 56, 59
phase 2, 25
phase 3, 25
phases 2, 45
phases 3, 45
plan d'analyse statistique, 56, 64
plan d'analyse statistique (SAP), 61
positivity assumption, 70
potential outcome, 74
progression free survival, 148
propensity score, 93
protocole, 56, 61, 64

R

real world Overall Survival, 150
real world PFS, 148
RECORD, 152
RECORD-PE, 152
registered report, 58
registre, 147
règle du 0/100%, 23
rétrospective (analyse), 56
rétrospective (étude), 56
risque alpha global, 167
ROBINS-I, 151
rwOS, 150
rwPFS, 148

S

safety, 143
SAP (plan d'analyse statistique), 61
score de propension, 93
sensibilité, 145
série temporelle interrompue, 52
SMD, 100
sources de données, 47
spécificité, 145
Stable Unit Treatment Values Assumption, 71
standardized mean difference, 100
STROBE, 152
STUVA, 137
supériorité, 67
survie sans progression, 148

T

t0, 80
taille de l'effet, 168
TARGET, 155
tipping point analysis, 161
TMLE, 82, 92, 107

V

valeur prédictive négative, 145
valeur prédictive positive, 145
validation des données, 144
variable instrumentale, 94