Société Française de Pharmacologie et de Thérapeutique

Groupe de Travail Méthodologie

Livre blanc SFPT

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

Document compagnon

Acceptabilité des « nouvelles méthodologies » pour l'évaluation des médicaments

14 février 2022

Groupe de rédaction / relecture

Par ordre alphabétique

- Pr Theodora Angoulvant, Tours
- Pr Laurent Bertoletti, Saint Etienne
- Pr Jean-Luc Cracowski, Grenoble
- Pr Michel Cucherat, Lyon
- Pr Dominique Deplanque, Lille
- Dr Guillaume Grenet, Lyon
- Pr François Gueyffier, Lyon
- Dr Silvy Laporte, Saint Etienne
- Pr Bruno Laviolle, Rennes
- Pr Jean-Christophe Lega, Lyon
- Dr Clara Locher, Rennes
- Pr Florian Naudet, Rennes
- Pr Antoine Pariente, Bordeaux
- Pr Matthieu Roustit, Grenoble
- Pr Tabassome Simon, Paris

Table des matières

1 Objectifs, démarche mise en œuvre 9
2 Fondamentaux et principes de base considérés 10
2.1 Principes de base 10
2.2 Apports et intérêt de la méthodologie classique 11
2.3 Limites des raisonnements mécanicistes, *medical reversals* 11
2.4 Éthique, intégrité scientifique et open science 11
2.5 Vérificationnisme 12
2.6 Tableau de synthèse 13
3 Définition et classification des nouvelles « méthodologies » 16
4 L'acceptabilité des méthodologies « moins-disantes » 20
5 Retour des premières utilisations de nouvelles méthodologies 21
6 Évaluation des revendications de bénéfice clinique d'un nouveau traitement 23
7 Les *real world evidences* (RWE) 24
8 Les études observationnelles 25
8.1 Problématiques méthodologiques spécifiques et solutions possibles 27
8.1.1 Confusion 27
8.1.2 Autres biais 28
8.1.3 Autres éléments de méthode 29
8.1.4 Analyse en intention de traiter 32
8.1.5 Inférence causale (causal inference) 32
8.2 Synthèses des problématiques et de leurs solutions 32
8.3 Études de cas, retour sur expérience 34
8.3.1 Comparaison de la chlorthalidone et de l'hydrochlorothiazide pour le traitement de l'hypertension 35
8.3.2 Sécurité cardiovasculaire de l'insuline 36
8.4 Meta-recherche 37
8.5 Avis de la SFPT 38
9 L'approche d'émulation d'un essai cible 40
9.1 Étude de cas 40
9.2 Méta-recherche 40
9.3 Avis de la SFPT 41
10 Les registres 42
11 Les essais pragmatiques 43
Avis de la SFPT 43

12 Les essais plateformes 45
12.1 Problématiques méthodologiques 46
12.2 Etude de cas 47
12.3 Méta-recherche 48
12.4 Avis de la SFPT 48
13 Les essais bayésiens 50
13.1 Principes des essais bayésiens 50
13.1.1 Résultats des essais bayésiens 51
13.1.2 Risque alpha et multiplicité 53
13.1.3 Dépendance des résultats à l'apriori 54
13.1.4 Études de cas - exemples de présentation de résultats bayésiens 56
13.2 Problématiques méthodologiques spécifiques des essais bayésiens 57
13.3 Méta-recherche 58
13.4 Avis de la SFPT 58
14 Les essais adaptatifs 60
14.1 Problématiques méthodologiques 61
14.2 Etudes de cas 62
14.3 Méta-recherche 63
14.4 Avis de la SFPT 63
15 Les essais combinés (« sans couture », *seamless*) 65
15.1 Problématiques méthodologiques 65
15.2 Étude de cas 65
15.3 Méta-recherche 67
15.4 Avis de la SFPT 67
16 Études mono-bras (non comparative) 68
16.1 Problématiques méthodologiques 68
16.2 Études de cas 69
16.3 Méta-recherche 69
16.4 Avis de la SFPT 69
17 Études à contrôle externe (groupes contrôles synthétiques) 70
17.1 Problématiques méthodologiques spécifiques et solutions possibles 70
17.1.1 Comparaison post hoc 70
17.1.2 Biais de confusion 71
17.1.3 Autres biais 72
17.1.4 Pertinence clinique 72
17.2 Étude de cas 72

17.3 Solutions possibles 73
17.3.1 La comparaison externe doit être formalisée 73
17.3.2 La comparaison externe doit être clairement explicitée 74
17.3.3 Il doit être possible d'écarter un choix arbitraire de la référence de comparaison, destiné à favoriser le traitement évalué 74
17.3.4 Les ajustements effectués doivent permettre d'écarter un biais de confusion. 75
17.3.5 La référence de comparaison doit être cliniquement pertinente et loyale 76
17.3.6 Les revues systématiques doivent être de bonne qualité 76
17.3.7 L'exposition potentielle aux biais de l'étude mono-bras et des études de références doit être acceptable 76
17.3.8 Le résultat suggéré par la comparaison externe doit être cliniquement pertinent 78
17.4 Synthèses des problématiques et de leurs solutions 78
17.5 Méta-recherche 79
17.6 Avis de la SFPT 80
18 L'emprunt d'information 83
18.1 Problématiques méthodologiques spécifiques 84
18.2 Études de cas 85
18.3 Avis de la SFPT 86
19 Les *surrogates* (critères de substitution) 87
19.1 Problématiques méthodologiques 87
19.2 Solution 87
19.3 Études de cas 89
19.3.1 Le LDL cholestérol, un contre-exemple 89
19.3.2 PFS et OS, un autre contre-exemple 90
19.3.3 Metastasis-Free Survival dans le cancer de la prostate 91
19.4 Méta-recherche 92
19.5 Avis de la SFPT 94
20 Les essais basket 95
20.1 Problématiques méthodologiques 95
20.2 Étude de cas 95
20.3 Avis de la SFPT 96
21 Les analyses poolées, les méta-analyses 97
21.1 Problématiques méthodologiques 97
21.2 Étude de cas 98
21.3 Avis de la SFPT 99
22 Les comparaisons indirectes en remplacement d'études « head to head » manquantes 100

22.1 Problématiques méthodologiques 100
22.2 Méta-recherche 101
22.3 Avis de la SFPT 101
23 Les maladies rares 103

1 Objectifs, démarche mise en œuvre

Régulièrement, de nouvelles méthodologies sont proposées pour l'évaluation des traitements, principalement avec l'idée de rendre cette évaluation moins complexe et plus rapide, par exemple pour permettre des enregistrements accélérés.

Ce document propose une analyse de ces « nouvelles méthodologies » d'évaluation des traitements dont le but est :

- D'identifier et décrire les problématiques méthodologiques qui pourraient conduire à des résultats ne reflétant pas le réel bénéfice clinique du traitement évalué et exposant ainsi au risque d'enregistrer, recommander et rembourser à tort le nouveau traitement.
- De compléter cette évaluation sur les principes, par des données empiriques en provenance d'études méta-recherche (méta-épidémiologie) si elles existent (pouvant éventuellement montrer qu'un problème anticipé au niveau des principes théoriques n'apparait finalement pas en pratique de l'évaluation).
- D'identifier si une solution spécifique à ces problématiques peut être envisagée et quelles sont les conditions ou les démonstrations que ces solutions devront apporter pour garantir le degré de certitude des résultats produits. S'ils existent, des retours d'expérience de la mise en œuvre de ces solutions seront analysés ainsi que des études de méta-recherche documentant l'efficience en pratique de ces solutions.

La finalité est d'**identifier, pour chaque nouvelle méthodologie, les conditions nécessaires pour produire des résultats présentant le même degré de certitude que ceux apportés par la méthodologie « habituelle »** (si cela s'avère possible).

Il ne s'agit pas directement de recommandation de réalisation des essais/études. Il s'agit de préciser les garanties méthodologiques qui seront attendues pour recommander et positionner un nouveau traitement dans la stratégie thérapeutique à partir d'études utilisant ces nouvelles méthodologies.

Seules les « nouvelles méthodologies » rencontrées actuellement (novembre 2021) dans des essais pivots sont considérées dans ce document.

Une grande partie de ce document est directement issu du livre blanc de la SFPT sur la place de la méthodologie dans l'évaluation des médicaments ou de ces dossiers compagnons et découle de la même réflexion. Ces éléments ont été mis en forme et complétés pour en faire un document autonome spécifique de la question des « nouvelles méthodologies ».

2 Fondamentaux et principes de base considérés

L'analyse des « nouvelles méthodologies[1] » réalisée dans ce document a été effectuée en considérant les principes de base, les attendus méthodologiques et les fondamentaux médicaux, déontologiques, éthiques et scientifiques suivants.

2.1 Principes de base

L'adoption d'un nouveau traitement et son intégration à la stratégie thérapeutique nécessitent de disposer de **preuves fiables** de son **bénéfice clinique[2]** pour des raisons médicales, éthiques et déontologiques[3]. Pour cela, la méthodologie des études doit garantir la « qualité de la démonstration et la validité des résultats obtenus » et leur pertinence clinique.

Pour apporter une preuve fiable, les résultats produits par l'étude doivent refléter la réalité du bénéfice clinique apporté par le traitement évalué avec un haut degré de certitude[4]. En particulier, l'étude ne doit pas pouvoir être « positive » en l'absence de bénéfice du nouveau traitement. Or il existe de nombreuses circonstances qui peuvent faire produire à une étude des résultats positifs à tort : les biais, les erreurs aléatoires, etc. (cf. Tableau 1). La méthodologie « actuelle » regroupe un **ensemble de principes** apportant une solution à toutes ces problématiques et empêchent, ou limitent à un niveau voulu, le risque de résultats faussement positifs et donc de conclusion à tort à l'intérêt du traitement. Les « nouvelles méthodologies » doivent apporter, aussi, mais par d'autres moyens, un contrôle identique de ces différentes problématiques (cf. Figure 1).

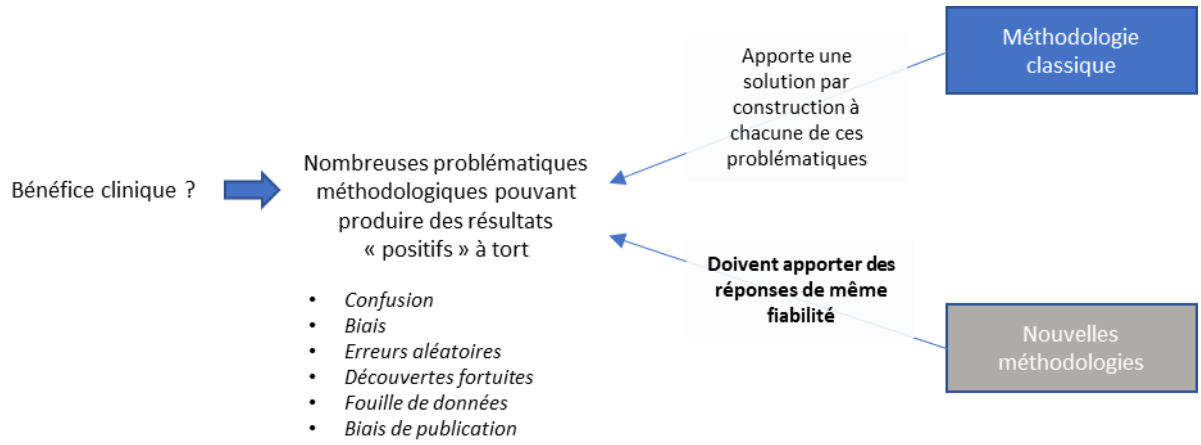


Figure 1 – Représentation schématique de l'analyse des nouvelles méthodologies réalisée dans ce document

[1] D'évaluation de l'efficacité et de la sécurité des nouveaux médicaments

[2] Le bénéfice clinique correspond à une balance bénéfice risque favorable. Un traitement ne bénéficie aux patients que si les risques ne contrebalancent pas l'efficacité (comme il n'y a de bénéfice financier que lorsque les recettes sont supérieures aux dépenses)

[3] Article 22 du code de déontologie en France par exemple

[4] Nous utilisons dans ce document la notion de degré de certitude des résultats proposé par GRADE

2.2 Apports et intérêt de la méthodologie classique

La **méthodologie standard n'est pas basée sur des principes gratuits** ou des considérations arbitraires, dogmatiques ou idéologiques. Elle a été construite progressivement au cours du temps, de manière pragmatique, pour garantir la fiabilité des résultats produits « au-delà de tout doute raisonnable » [1]. Cette méthodologie a été construite pour apporter des solutions simples et efficaces à tous les problèmes qui ont été progressivement découverts avec la pratique de l'évaluation des traitements (cf. Tableau 1), principalement à l'occasion de retours d'expériences où il s'est avéré *a posteriori* que les premiers résultats produits étaient faux (faux positifs). Les causes de ces faux positifs ont alors été identifiées et des solutions techniques ont été imaginées et intégrées à la méthodologie pour empêcher à l'avenir la reproduction de ces erreurs (cf. Tableau 1). La méthodologie actuelle est simple, non arbitraire, parfaitement bien codifiée et connue par toutes les parties prenantes, réalisables et couramment réalisées (Pubmed recense 27 260 publications d'essais randomisés en 2020). Elle est un juge impartial et strict, permettant de révéler l'apport réel des traitements (ce qui peut être perçu néanmoins comme une contrainte du point de vue de l'intérêt du traitement).

2.3 Limites des raisonnements mécanicistes, *medical reversals*

Les éléments précliniques et cliniques précoces (phase 2) ne permettent pas de prédire le bénéfice clinique et l'utilité médicale des traitements : 50% des phases 3 sont ainsi des échecs [2], signifiant que la moitié des traitements environ arrivant à ce stade ne déboucheront pas sur un enregistrement. Ces résultats montrent aussi que l'expertise fondamentale ou clinique ne permet pas de prédire le bénéfice clinique et qu'il n'est pas possible de s'affranchir de la démarche de vérification (vérificationnisme) que représente la réalisation des essais pivots (vérifier par les faits que le mécanisme par lequel on pense qu'un nouveau médicament pourrait apporter un bénéfice aux patients débouche réellement sur ce bénéfice attendu, et le quantifier).

Les nouvelles méthodologies doivent donner les mêmes garanties de fiabilité (tout en procédant autrement) car des méthodologies sous performantes exposent au risque d'enregistrer, de recommander et de rembourser des médicaments qui n'apportent pas le bénéfice clinique escompté. Ainsi il existe de nombreux exemples de « *medical reversals* » [3, 4, 5, 6, 7] où il s'est avéré *a posteriori* que des technologies médicales (médicaments ou autres) initialement acceptées sur la base de résultats fragiles ne répondaient pas aux attentes médicales et des patients.

2.4 Éthique, intégrité scientifique et open science

Produire des résultats fiables est bien évidement un impératif éthique pour répondre aux attentes légitimes des médecins (code de déontologie), des patients, des autorités de santé, et de la société. L'acceptation des résultats d'études de faible méthodologie soulève de véritables problèmes éthiques car pouvant conduire à utiliser à tort, sur une période de temps parfois prolongée et chez de nombreux patients, des traitements n'apportant pas le bénéfice escompté [8, 9, 10].

La solidité méthodologique des études est donc aussi un impératif dans le cadre de l'intégrité scientifique. « La fiabilité dans la conception, la méthodologie, l'analyse et l'utilisation des ressources » est la première valeur de l'intégrité scientifique mise en avant[5].

L'approche de l'Open Science formalise aussi des éléments importants pour la fiabilité des études scientifiques, y compris ceux conduisant à la construction des stratégies thérapeutiques : « règles du jeu » établies *a priori*, avec les critères de réussite définis explicitement ; utilisation de critères de jugement issus de « *core outcome set* » [11] ; et transparence sur tout le processus, de l'élaboration du protocole au partage des données. Le partage des données peut être vu aussi comme un rempart à la fraude.

2.5 Vérificationnisme

Compte-tenu des limites des extrapolations théoriques à partir des mécanismes d'action et des études exploratoires ou explicatives, la méthodologie « classique » s'inscrit pleinement dans une approche de vérificationnisme par les faits et uniquement les faits. L'hypothèse spéculative qu'un nouveau traitement pourrait avoir un intérêt clinique en raison de son mécanisme d'action est vérifiée dans la réalité à l'aide de données expérimentales observées. La conclusion définitive sera objective, faite à partir de résultats issus de données réelles, effectivement observés, s'affranchissant ainsi de tout caractère subjectif (spéculation, espoirs ou croyance) ou théorique.

Une partie des nouvelles propositions méthodologiques repose sur des techniques d'emprunt d'information, comme l'inférence bayésienne par exemple. Dans ces approches les résultats proviendront à la fois des données objectives observées, mais aussi du résultat supposé par l'investigateur qui est injecté dans le calcul du résultat. Cette insertion d'information *a priori* peut être purement arbitraire (ce que pré suppose l'investigateur sur l'effet du traitement) ou être basée sur de premiers résultats (et dans ce cas s'apparente à une démarche de méta-analyse). Si de l'information arbitraire est utilisée, ces nouvelles « méthodologies » dérogent au principe du vérificationnisme et déguisent en résultats apparemment factuels de simples valeurs arbitraires.

[5] The European code of conduct for research integrity. ALLEA, 2017
http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020-ethics_code-of-conduct_en.pdf

2.6 Tableau de synthèse

Tableau 1 – Liste des problématiques rencontrées dans la recherche de preuves du bénéfice clinique d'un nouveau traitement et les solutions apportées par la méthodologie habituelle des essais cliniques et leurs règles d'interprétation

Problématique méthodologique (au sens large) pouvant conduire à des résultats positifs à tort		Solution apportée par la méthodologie classique et les règles d'interprétation (essai contrôlé randomisé type essai pivot de phase 3)
Problématiques méthodologiques		
Raisonnement contrefactuel	Nécessité d'un raisonnement contrefactuel pour identifier l'effet propre d'un traitement et mesurer son importance en raison de la variabilité du vivant (inter- et intra-sujet)	Groupe contrôle contemporain apportant le contrefait
Biais de sélection	(au sens du biais de sélection des études observationnelles)[6]	Démarche expérimentale, étude prospective, début du suivi identique dans les 2 groupes, absence d'attrition ou gestion conservatrice de l'attrition
Biais de confusion	(appelé biais de sélection dans le domaine de l'essai randomisé)	Randomisation imprévisible (qui avec les autres principes permet de montrer la causalité)
Biais de réalisation, de suivi		Double insu
Biais de mesure, d'évaluation		Double insu
Biais d'attrition		Analyse en intention de traiter (ITT) avec remplacement des données manquantes sur le critère de jugement de manière conservatrice (pour l'essai de supériorité)
Estimation de l'effet traitement correspond à ce que la recommandation future du traitement produirait comme changement dans le devenir des patients (compte tenu de tout le reste de la stratégie thérapeutique)[7]		Estimation d'un effet traitement en ITT (*treatment policy estimand*)
Inflation du risque alpha	Risque de conclure à tort à l'intérêt du traitement du fait de l'erreur statistique alpha (de premier type)	Plan de contrôle du risque alpha global Définition des comparaisons inférentielles (comparaisons qui peuvent conduire à la conclusion à

[6] Au sens épidémiologique moderne du terme (différent du biais de sélection dont on parle dans l'essai randomisé), cf. ROBINS-I

[7] Et non pas une mesure d'un effet théorique idéal (analyse en per protocole), non représentatif du bénéfice à attendre a priori, pour un patient donné, pour lequel on ne sait pas, au moment où le traitement est initié, s'il tolèrera le traitement, aura une bonne observance et n'aura pas d'effet indésirable.

		l'intérêt du traitement et donc à la recommandation de son utilisation)
Multiplicité des comparaisons	pouvant amener à conclure à l'intérêt du traitement ; multiplicité induisant une inflation du risque alpha global	Plan de contrôle du risque alpha global (non prise en considération de la signification nominale, non-présentation des p values non inférentielles pour éviter les surinterprétations des résultats exploratoires sans contrôle du risque alpha global)
Non-respect de la démarche hypothético-déductive		Formulation des hypothèses a priori garantie par le caractère prospectif de l'essai (validation prospective de l'hypothèse)
Les résultats produits ne dépendent que des données observées et non pas des convictions personnelles		Les résultats produits ne proviennent que des données observées et la méthode ne fait pas d'hypothèse fondamentale forte non vérifiée ou non vérifiable8
Découverte fortuite, fouille de données	data dredging data milking	Réalisation d'essais de confirmation respectant pleinement la démarche hypothético déductive avec un objectif fixé a priori attesté par le plan de développement du médicament. Exclusion des résultats post hoc, ou exploratoire de la prise de décision
P-hacking	Modification de l'analyse statistique jusqu'à l'obtention des résultats voulus	Définition d'un protocole fixant les critères de jugement et les grandes lignes de l'analyse. Définition d'un plan d'analyse statistique (SAP) précis avant que les données (même parcellaires) soient disponibles et réalisation de l'analyse statistique en stricte conformité avec ce SAP (anciennement triple aveugle du statisticien !)
Fiabilité des données	Les résultats produits reposent sur des données de bonne qualité et aussi complètes que possible	Monitorage des données, traçabilité, audit interne et externe
Pertinence clinique		
Absence de pertinence clinique	(effet non nul, mais sans plus-value médicale, ne représentant pas un progrès thérapeutique, car trop petit en taille, montré sur un critère peu pertinent, ou par rapport à un comparateur non optimal, etc.)	Définition dans l'objectif de l'essai : • D'un critère de jugement clinique (ou d'un *surrogate* validé)

[8] Comme avec par exemple faire l'hypothèse d'un bénéfice clinique à partir d'un effet sur un critère intermédiaire ou introduire une croyance arbitraire sur l'effet du traitement dans une inférence bayésienne

		• D'un comparateur loyal et représentant le traitement standard du moment où l'essai est analysé pour intégrer le nouveau traitement dans la stratégie thérapeutique • D'une population cible recherchée correspondant à la totalité des patients relevant du traitement évalué (essai pragmatique)
Bénéfice complètement compensé par des effets délétères de manière quantitative ou qualitative		Évaluation de la sécurité Prise de décision basée sur la balance bénéfice risque = bénéfice clinique (et non pas seulement sur l'efficacité ou sur la sécurité de façon isolée et séparée)
Problématiques liés à la transparence, *reporting*, spin, intégrité		
Selective reporting	Présentation ou publication des seuls résultats positifs de l'étude, permettant d'aboutir à la conclusion recherchée	Protocole, enregistrement dans un registre, vérification des résultats mis en avant par rapport au protocole
Biais de publication		Plan de développement, enregistrement dans un registre, déclaration administrative (et aussi la lourdeur des études qui fait qu'il est difficile de les répéter et de ne retenir que les positives)
Spin de conclusion	Conclusion positive en faveur de l'intérêt du traitement dans une étude en réalité non concluante	Ne pas lire les conclusions, ne regarder que les résultats et la méthode
Manque de transparence des rapports ou des publications, ne permettant pas de voir les limites des résultats		Rapport standard exhaustif (peut-être un peu trop !) dans le cadre de la standardisation CDISC Guide EQUATOR (CONSORT) garantissant l'informativité des publications pour en faire l'analyse critique
Fraude des investigateurs		Monitorage de terrain, bonnes pratiques cliniques, recherche systématique de la fraude lors de l'analyse statistique, exclusion des centres en cas de suspicion de fraude
Fraude au niveau de l'investigateur principal (sponsor, centre de coordination)		Principes méthodologiques et système d'assurance qualité (procédure opératoire standard), traçabilité, audit interne et externe

3 Définition et classification des nouvelles « méthodologies »

Il n'existe pas de définition de ce qu'est une « nouvelle méthodologie ».

La méthodologie utilisée actuellement (que l'on peut qualifier de « classique », « standard » ou « habituelle ») a fait l'objet d'amélioration continue depuis son origine et continue à être améliorée régulièrement. Les derniers perfectionnements ont mis l'accent sur le contrôle du risque alpha global, la notion d'*estimand*, les essais randomisés plateformes, les essais randomisés pragmatiques sur registre, etc.

Les principes de la méthodologie classique, utilisés pour garantir un haut degré de certitude aux résultats, constituent cependant des contraintes qui rendent la réalisation des essais parfois lourde[9], complexe, longue, nécessitant de nombreux patients et in fine couteuse (même si cette augmentation des coûts est aussi liée à l'évolution réglementaire en générale). Ces aspects conduisent régulièrement à des contrepropositions cherchant à rendre l'évaluation de l'efficacité et de la sécurité des nouveaux médicaments plus simple, plus rapide et moins couteuse. Ce sont ces propositions qui sont communément dénommées « nouvelles méthodologies ».

Un des leviers actionnés par ces nouvelles propositions méthodologiques pour rendre l'évaluation des nouveaux traitements moins contraignante est d'abandonner certains principes méthodologiques en argumentant que les problématiques qu'ils solutionnent n'existent plus. Par exemple abandonner le principe du vérificationnisme[10] en faisant l'hypothèse que les connaissances physiopathologiques et pharmacologiques sont suffisantes pour être sûr qu'un effet pharmacologique ou qu'un effet sur un critère intermédiaire se traduit bien en bénéfice clinique. C'est l'argumentaire principal de la revendication à l'accès précoce des médicaments [12, 13] (cf. section 5).

Une autre voie utilisée par les « nouvelles méthodologies » est de chercher à faire aussi bien que la méthodologie classique, mais autrement. Par exemple, remplacer la randomisation par les méthodes de correction du biais de confusion ; supprimer le groupe contrôle en argumentant que le raisonnement contrefactuel peut se faire aussi bien à l'aide de contrôles externes (historiques par exemple), etc. Ces techniques alternatives reposent alors sur des hypothèses fondamentales conditionnant leur validité (par exemple l'hypothèse de transitivité dans les comparaisons indirectes utilisées en remplacement d'essais de comparaison directe « *head to head* »).

Une troisième approche consiste à faire reposer la production de la preuve du bénéfice clinique, non plus sur une étude suffisante en elle-même, mais sur un ensemble d'informations issues de différentes sources. L'idée est de faire des « études augmentées » en intégrant dans le processus de production des résultats de l'information externe qui renforcera l'information[11] produite par l'étude elle-même. On parle d'emprunt d'information. [14, 15]. Du fait que ce type d'étude va emprunter de l'information, elle n'a plus besoin d'apporter toute l'information nécessaire par elle-même. Elle peut donc être plus petite (en taille) et/ou moins longue que si elle avait la charge, à elle seule, d'apporter la totalité de l'information nécessaire. Cette approche repose la plupart du temps sur l'approche bayésienne en

[9] Indépendamment des contraintes administratives, légales et règlementaires. Uniquement sur le plan méthodologique, les autres contraintes administratives et règlementaires, dont l'intérêt est autre (protéger les patients principalement), se rajoutent à ces contraintes méthodologiques (garantissant la solidité du résultat).

[10] Démontrer par les faits que l'effet pharmacologique induit bien un bénéfice clinique

[11] Le terme information est utilisé dans son acception statistique. Dans la comparaison de deux groupes à la recherche de l'effet d'un traitement, l'information est apportée par les évènements et les sujets inclus.

utilisant la possibilité d'introduire de l'information *a priori* dans la production du résultat de l'étude. Ces méthodes reposent ainsi sur l'hypothèse que l'information empruntée est correcte pour documenter l'effet recherché.

Ces trois approches conduisent donc à des méthodologies moins-disant, reposant entièrement sur des **hypothèses simplificatrices fortes** (Tableau 2). Elles sont donc incapables de garantir par elles-mêmes un haut degré de crédibilité des résultats. Elles ne pourront produire des résultats fiables que si les hypothèses sur lesquelles elles se basent sont effectivement vérifiées (par exemple l'effet sur le critère intermédiaire prédit avec certitude l'effet sur le critère clinique, c'est-à-dire que ce critère intermédiaire est un véritable *surrogate* ; l'information empruntée reflète bien le réel effet du traitement et n'est pas une estimation abusivement optimiste ; etc.) (cf. Tableau 2). En pratique, pour que les résultats obtenus soient recevables comme preuves du bénéfice clinique, il est nécessaire que soit aussi démontré que les hypothèses fondamentales sous-jacentes à la nouvelle méthodologie sont effectivement vérifiées dans le cadre considéré. Ces approches doivent donc apporter des études (ou arguments) complémentaires démontrant que les résultats produits sont effectivement à l'abri des « biais » contre lesquels la nouvelle méthodologie ne protège pas par principe. Ces démonstrations complémentaires font appel à des méthodologies spécifiques (par exemple méthodologie de validation d'un *surrogate*).

Tableau 2 – Exemples d'hypothèses simplificatrices que font les nouvelles propositions méthodologiques pour simplifier la production des résultats.
La crédibilité des résultats produits dépend directement de la vérification de la plausibilité de l'hypothèse qui peut être, dans certains cas, invérifiable et laisser un caractère spéculatif au résultat produit

Nouvelles propositions méthodologiques	Hypothèses simplificatrice	Validation de l'hypothèse nécessaire pour rendre acceptables les résultats
Méthodologie classique	**Aucune hypothèse** : le principe est de vérifier directement que le traitement apporte le bénéfice clinique escompté *(hormis des hypothèses sur la qualité de la réalisation assurées par le système d'assurance qualité et vérifiées par le monitorage)*	
Surrogate	L'effet sur le *surrogate* prédit l'effet sur le critère clinique	Validation de cette hypothèse par une étude de corrélation des effets observés dans des essais précédents
Étude observationnelle	Les ajustements ont permis de supprimer complètement le biais de confusion L'association observée peut être interprétée de manière causale + nombreuse autres hypothèses	Démonstration que tous les facteurs ont été pris en considération dans les ajustements Démonstration que tous les autres biais sont contrôlés Démonstration de l'absence de *data dredging*, *p haking*, etc.
Emprunt d'information (essais bayésiens, codata, etc.)	L'information empruntée correspond bien au vrai effet du traitement que l'on cherche à estimer (dans la population de l'étude)	Démonstration que l'information empruntée correspond bien à l'effet qui est recherché
Étude mono-bras	Le contrôle externe constitue un contrefait correct	Démonstration que tous les facteurs ont été pris en considération dans les ajustements

	OU le changement avant/après conduit à un raisonnement contrefactuel correct	Démonstration que tous les autres biais sont contrôlés Démonstration de l'absence de *data dredging*, *p haking*, etc.
Design adaptatif	Adaptations effectuées ne dépendent pas des résultats qu'elles produisent	Structure des méthodes utilisées
Essais pragmatiques randomisés en vie réelle	Pas de défaut de réalisation	Invérifiable, car pas de monitorage

D'autres propositions sont simplement des améliorations ou optimisations de la démarche habituelle, comme les **« masters protocols »**, qui reposent sur une infrastructure, un design d'essai et un protocole uniques pour évaluer un ou plusieurs médicaments dans une ou plusieurs maladies [16]. Parmi eux, on distingue différentes approches :

- Les **essais « baskets** » évaluent un même traitement, mais dans plusieurs maladies ou sous-types de maladies
- Les **essais « umbrella »** étudient au contraire différents traitements dans une même maladie
- Enfin les **essais « plateformes »** permettent de comparer différents traitements dans une même maladie, mais de manière continue, les différents traitements étant amenés à entrer ou sortir de la plateforme sur la base d'un algorithme de décision.

Ces approches n'abandonnent aucun principe méthodologique, mais rendent la production de preuves plus fluides, rapides et moins lourdes. Il existe bien des conditions de validité spécifiques (utilisation de patients contrôles contemporains pour les essais plateformes par exemple), mais aucun abandon des principes de la méthodologie. Aucune hypothèse n'est nécessaire pour assurer la crédibilité des résultats (qui ne dépend que de la méthodologie de l'étude).

Le Tableau 3 tente de récapituler ces éléments de différenciation entre méthodologie classique et les nouvelles propositions méthodologiques.

Tableau 3 – Éléments de différenciation entre méthodologie classique et les nouvelles propositions méthodologiques

Méthodologie classique	La méthodologie de l'étude est autosuffisante pour garantir un haut degré de crédibilité aux résultats qui ne dépendent que de données observées (confrontation de la théorie à la réalité)
Méthodologie moins-disante	Fait l'hypothèse que certaines problématiques méthodologiques n'affectent pas le domaine de l'étude ce qui permet de simplifier la méthodologie en abandonnant les principes utilisés dans la méthodologie classique pour parer à ces problèmes
Emprunt d'information	Le degré de crédibilité du résultat va dépendre de la crédibilité de l'information empruntée (non arbitraire, non subjective, et applicable (représentative) à la situation de l'évaluation)
Optimisation de la méthodologie classique	Aucune hypothèse simplificatrice présupposée, le degré de crédibilité des résultats ne dépend que de la méthodologie de l'étude

Même si la motivation des propositions moins-disantes est clairement d'alléger l'évaluation des traitements de contraintes qui peuvent être inutiles (ou de la rendre réalisable en faisant des concessions sur la crédibilité des résultats), on peut néanmoins craindre que dans certaines réalisations, une partie de la motivation de ce choix réside dans la recherche de méthodes plus flexibles, permettant un plus grand contrôle des résultats obtenus.

4 L'acceptabilité des méthodologies « moins-disantes »

Les nouvelles méthodologies représentent souvent des approches produisant des résultats de plus faible niveau de crédibilité que l'approche classique (cf. section 3), la question de leur acceptabilité découle entièrement de la question du niveau des exigences demandées pour introduire un nouveau traitement dans la stratégie thérapeutique.

Jusqu'à présent un haut degré de certitude était exigé (« preuves au-delà de tout doute raisonnable ») et la méthodologie classique a été construite pour produire des résultats répondant à cette exigence. Pour prétendre atteindre le même niveau d'exigence, les nouvelles méthodologies basées sur des hypothèses simplificatrices ou des emprunts d'informations doivent alors apporter une démonstration complémentaire qui est celle de la validité de leurs hypothèses ou informations empruntées afin de lever les réserves méthodologiques sous-jacentes.

Si désormais, il était accepté collectivement que ce niveau d'exigence est excessif, contreproductif, car handicapant trop l'accès aux patients pour les nouvelles propositions thérapeutiques, et qu'il soit possible d'admettre un risque plus important d'accepter des traitements n'apportant pas les bénéfices escomptés ou un risque d'effets secondaires non maitrisé, les nouvelles méthodologies seraient alors susceptibles de produire des résultats d'un niveau de crédibilité compatible avec ce niveau d'exigence plus faible que dans la méthode classique.

Finalement, l'acceptabilité de méthodologies moins-disantes dépend directement de ce qui est attendu comme niveau de degré de certitude. La méthodologie est une ressource technique qui s'adapte au cahier des charges qu'on lui demande. La question de l'acceptabilité des nouvelles méthodologies basées sur des hypothèses simplificatrices ou des emprunts d'information revient donc en fait à la question du niveau de certitude attendue pour prendre la décision d'utiliser un nouveau traitement, autrement dit, du risque consenti d'effectuer cette décision à tort. Ce n'est pas une question méthodologique, mais plutôt une question politique.

Cependant une difficulté apparait, celle de savoir **quel est le niveau de risque pris avec une méthodologie non optimale**. Cette détermination n'est possible que s'il est faisable de comparer les résultats produits avec les nouvelles méthodologies et ceux obtenus avec l'approche habituelle. Peu de données de ce type sont actuellement disponibles.

Des données indirectes peuvent être trouvées dans l'étude des molécules en oncologie ayant eux un enregistrement accéléré par la FDA. Ces enregistrements ont été obtenus conditionnellement à la réalisation d'études de phase 3 classiques. Compte tenu maintenant du recul qui a permis d'obtenir les études de confirmation et du nombre substantiel d'indications accordées par cette voie, il est donc possible de comparer les résultats des 2 approches (ce qui ne sera plus le cas si cette obligation de réaliser les études de confirmation classiques disparait).

Aucune étude n'a cherché à estimer la proportion d'enregistrements accélérés accordées à tort. Cependant une liste de seize indications pour lesquelles l'étude de confirmation a été négative est disponible [17] montrant que ce cas de figure n'est pas rare. Pour l'instant il n'y a donc pas de validation empirique que les nouvelles approches assurent le même de niveau de fiabilité des décisions que l'approche classique. Cette constatation associée à d'autres éléments fait que cette approche d'enregistrement accéléré fait maintenant l'objet de nombreuses critiques qui seront discutées dans le chapitre suivant (cf. section 5).

5 Retour des premières utilisations de nouvelles méthodologies

Les nouvelles propositions « méthodologiques » sont d'ores et déjà utilisées, par exemple dans le cadre des demandes à la FDA d'enregistrement accéléré (*accelerated approval*), en particulier en oncologie. Dans ce cadre, les autorisations de commercialiser sont accordées sur la base de résultats préliminaires (effet sur des *surrogates*, études mono-bras [18, 19, 20, 21]) laissant présupposer un éventuel bénéfice clinique du traitement. Cet accord est conditionnel à la confirmation du bénéfice après commercialisation par une étude de phase 3. Pour l'instant il est plus ou moins exigé que cette phase 3 soit une étude randomisée, mais il existe aussi des demandes pour que cette étude repose sur des nouvelles « méthodologies ».

Les « nouvelles méthodologies » utilisées dans ces enregistrements accélérés sont principalement des études mono-bras ou l'utilisation d'un *surrogate* [22].

Il apparait aussi que les études de confirmation demandées ne sont pas toujours réalisées ou terminées [23], montrant que le système actuel présente des failles qui peuvent être graves de conséquence. Cela montre aussi la fragilité de l'idée que l'évaluation puisse se faire après commercialisation, les investigateurs pouvant considérer qu'il n'est plus acceptable de réaliser un essai comparatif étant donné que la molécule est disponible en standard, alors que ne pas faire l'essai expose à une méconnaissance du rapport bénéfice/risque du médicament ce qui est éthiquement inacceptable [24, 25, 26, 27].

La tendance actuelle est plutôt de s'orienter vers des études observationnelles (registre de patients tous traités par la nouvelle molécule, posant la problématique des études monobras, ou études de « *comparative effectiveness* » si la totalité des patients ne la reçoit pas) pour chercher cette démonstration du bénéfice à la place de la réalisation d'essais comparatifs randomisés *post-approval*.

Actuellement cette pratique des enregistrements accélérés fait l'objet de nombreuses critiques, principalement en oncologie, car il s'avère qu'il a conduit à « abaisser la barre » des exigences de preuve du bénéfice clinique de manière substantielle [28, 29, 30, 31, 32]. Au-delà de l'oncologie, ces enregistrement accéléré (*accelerated approvals* ou *breakthrough approval*) ont été utilisés dans des indications comme la dépression résistante ou la dépression du *post partum*, arguant d'un certain caractère « compassionnel » alors que ces situations sont bien différentes des patients ayant un cancer sans aucune perspective thérapeutique (citons les exemples de l'esketamine ou de la brexanolone).

L'étude des molécules enregistrée en oncologie à la FDA entre 2000 et 2016 met en évidence la fragilité des données disponibles au moment de la mise sur le marché et la petitesse des bénéfices avec une médiane des bénéfices absolus sur la survie de 2.4 mois [21].

Il a été estimé qu'en 2019 aux USA les médicaments enregistrés par une procédure accélérée représentait en valeur 9.1% des dépenses de médicaments payés par Medicaid sans qu'il soit possible de connaitre le réel bénéfice de ces traitements, l'enregistrement accéléré de ces molécules reposant sur des critères intermédiaires pour la majorité d'entre elles [33]. Une autre étude portant sur le même sujet arrive au même constat en trouvant que 2/3 des essais présentent des limites empêchant de savoir ce qu'apporte réellement le traitement au patient [18]. Au niveau européen, la situation s'avère identique avec la moitié des essais sur lesquels sont basés les enregistrements par l'EMA en oncologie qui sont à haut risque de biais [20].

Outre la difficulté d'obtenir des preuves fiables avec ce type d'études, ces enregistrements accélérés reviennent aussi à faire financer la recherche de la preuve par les payeurs et non plus par l'industriel qui percevra ensuite le bénéfice de cette démonstration. Même en cas d'échec, ces études auront permis à l'industriel d'engranger du chiffre d'affaires avec une molécule n'apportant pas de bénéfice au patient et donc ainsi de limiter la charge financière qu'il devrait subir du fait de l'échec de la molécule. De ce fait les payeurs deviennent des co-investisseurs à perte dans le développement industriel des nouvelles molécules.

Sur un autre plan, à l'issu de ces enregistrements accélérés, les médecins prescripteurs et les patients traités participent ainsi à l'évaluation de nouveaux médicaments sans que cette situation leur soit clairement perceptible. Sans le recours aux nouvelles méthodologies, cette obtention, ou non, de la preuve du bénéfice clinique aurait été réalisée par des investigateurs (et non des prescripteurs) dans le cadre éthique et réglementaire d'une étude de phase 3 où le consentement des patients aurait été recueilli.

6 Évaluation des revendications de bénéfice clinique d'un nouveau traitement

D'une façon générale, l'évaluation de tout résultat utilisé pour revendiquer le bénéfice clinique d'un nouveau traitement consiste à répondre aux questions suivantes, et ce quel que soit le type de méthodologie, classique ou nouvelle :

1. Le résultat est-il potentiellement une découverte fortuite (résultat purement exploratoire, issu d'une fouille de données, absence de démarche hypothético déductive, etc.) ?
2. Le résultat obtenu peut-il provenir d'autre chose que le traitement étudié (problématique de la confusion dans les études non randomisées « en vie réelle », contrôle externe, etc.) ?
3. Le résultat obtenu peut-il conduire à conclure à tort à l'intérêt du traitement du fait du hasard, à cause d'un risque non contrôlé (risque alpha **global** non contrôlé) ?
4. Le résultat obtenu peut-il provenir de biais : biais entrainant une différence à tort lorsque les résultats sont en faveur d'une différence, biais « *toward the null* » quand le résultat est en faveur d'une absence de différence (pour la sécurité par exemple) ?
5. Le résultat est-il cliniquement pertinent : pathologie, absence de *disease mongering* (façonnage de maladie), comparateur loyal et pertinent, critère de jugement pertinent, taille du bénéfice suffisante ?
6. La balance-bénéfice risque est-elle démontrée comme favorable qualitativement (absence d'effet indésirable de gravité disproportionnée par rapport au bénéfice recherché) et quantitativement (bénéfice clinique net favorable en cas d'augmentation certaine des effets indésirables avec le nouveau traitement par rapport au comparateur) ?

Lorsque la méthodologie utilisée repose sur des **hypothèses simplificatrices** (« nouvelles méthodologies », cf. section 3) il est important de faire attention à la plausibilité de ces hypothèses. L'étude doit apporter la **preuve** complémentaire **que ces hypothèses étaient vérifiées** et que de ce fait l'approche utilisée garantit, malgré cette simplification, un haut degré de certitude.

Quand il y a **emprunt d'information** (cf. section 3), la part du résultat due à l'information empruntée doit être quantifiée (et comparée à celle due aux données originales amassées par l'étude). De plus, l'hypothèse fondamentale, à savoir que l'information empruntée documente correctement l'effet du traitement, doit être aussi vérifiée. Ces aspects sont très techniques et nécessitent pour les juger une bonne expertise à la fois statistique et clinique du domaine. De plus cette technicité peut être faussement rassurante, mais la question fondamentale est une question de sens commun : l'information injectée dans le processus de production du résultat est-elle appropriée, non-arbitraire et consensuelle, et ne distord-elle pas la réalité ?

7 Les *real world evidences* (RWE)

Le terme *real world evidence* désigne une preuve (*evidence*) obtenue, non plus à partir de données spécifiques recueillies dans le cadre expérimental d'un essai clinique randomisé, mais à partir de données d'observation de la pratique médicale dans la « vraie vie » (études observationnelles sur des *real world data*, RWD).

Le terme *evidence* marque bien qu'il s'agit de produire des résultats de haut degré de certitude, recevables au niveau règlementaire et pour modifier les pratiques (comme ce que permettent actuellement les résultats des essais randomisés pivots).

Les *real world evidences* peuvent être générées à partir des *real world data* par différent type d'études comme des essais randomisés nichés dans des registres ou des bases de soins, des essais randomisés pragmatiques « classique » sans sur-sélection des patients (cf. section 11) ou des études observationnelles (s'inscrivant dans une approche d'émulation d'un essai cible, cf. section 9).

Dans ce document les essais pragmatiques sont abordés section 11 et les études observationnelles section 8.

8 Les études observationnelles

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

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

Les études observationnelles peuvent être utilisées avec d'autres objectifs (décrire les traitements utilisés, recherche du mésusage, etc.), mais ces applications sont hors du champ de ce document. Les études observationnelles considérées ici s'inscrivent dans une tentative d'obtenir des résultats aussi solides que ceux produits par l'approche habituelle, c'est-à-dire obtenir des preuves (*evidences*) (cf. section 7). L'idée est alors de produire des *real world evidence* (RWE), similaires aux *evidences* produite par les études expérimentales randomisés[12,13].

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

L'analyse statistique est complexe, car elle cherche à corriger les résultats des biais inhérents à l'approche observationnelle, principalement le biais de confusion. Elle cherche aussi à s'inscrire dans le cadre de l'inférence causale (voire section 8.1.5) afin de tenter d'établir un lien de causalité entre le traitement étudié et les bénéfices mis en évidence, comme ce qui est obtenu avec l'essai randomisé. Enfin, ces études doivent aussi s'inscrire dans une approche d'émulation d'un essai cible [34] afin de conforter leurs résultats en se rapprochant le plus possible de la méthodologie de l'essai clinique.

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

Cette approche est inadaptée à l'évaluation des traitements avant leur commercialisation (ou leur mise à disposition), car l'approche de vraie vie sous-entend que les traitements étudiés sont déjà utilisés dans la pratique médicale courante. Leur utilisation est donc limitée à la confirmation de l'efficacité des traitements dans la vraie vie en comparant les résultats d'études observationnelles à ceux des essais randomisés (avec une problématique importante de p hacking, car l'analyse est réalisée en connaissant le résultat à obtenir) ou à étudier ce qui se passe sur des durées de suivi plus longues.

---

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

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

Les études observationnelles peuvent aussi être exploitée pour rechercher des hypothèse de repositionnement d'ancienne molécule dans d'autre pathologie que leur usage initial (« *repurposing* ») [35].

Les études observationnelles sont parfois aussi envisagées pour apporter la confirmation du bénéfice après un enregistrement précoce, à la place d'un essai randomisé. L'accès précoce est accordé sans démonstration du bénéfice clinique, avec des études de faible méthodologie, comme des études mono-bras ou des essais sur critères intermédiaires (n'ayant pas valeur de surrogate), à condition qu'une étude de confirmation soit entreprise. Si cette étude ne confirme pas le bénéfice, le traitement devrait être retiré. Cet affaiblissement de la qualité des essais initiaux engendrera une perte considérable de niveau de preuve que les meilleures études observationnelles ne pourront totalement compenser. Le degré de certitude du bénéfice clinique des produits serait bien inférieur à ce qu'apporte la méthodologie classique ou même ce qui est exigé actuellement pour les accès précoces, une confirmation par un essai randomisé (même s'il est vrai que ces études ne sont pas toujours produites. [17, 23, 36, 37, 38], cf. section 5).

Depuis des années, les études observationnelles à promotion industrielle et s'intéressant à l'efficacité des médicaments sont, à de rares exceptions, réalisées dans le cadre de phase 4 sans finalité d'enregistrement. La robustesse méthodologique de ce type d'études est en général faible. [39, 40] et leurs résultats sont principalement exploités à but de communication promotionnelle (ou d'abondement aux scores bibliométriques des auteurs). Pour parer à l'éventualité de résultats opposés à ceux souhaités, ces études aménagent souvent une ambigüité quant à leur objectif : décrire et non pas comparer (voir faire une comparaison descriptive !). Dans ce contexte, ces études ne cherchent pas, en général, à apporter une solution à toutes les problématiques méthodologiques que posent ces études (cf. infra) et ne correspondent absolument pas à ce qui actuellement envisagée comme « *real world evidences* ». S'il existe une possibilité de produire des preuves solides à partir des données observationnelles, de degré de certitude identique à celui apporté par la méthodologique classique, cela passe par une tout autre approche, bien plus complexe et sophistiquée, et il ne faut pas considérer que ces études indigentes méthodologiquement correspondent à ce qui est envisagé pour produire des RWE.

> Par exemple, dans la récente crise de la COVID-19, de nombreuses études « observationnelles » ont été produites pour justifier *a posteriori* des usages compassionnels ou des repositionnements. Ces études reposaient, à de rares exceptions près[14], sur des méthodologies simplistes, comme un ajustement arbitraire non raisonné. Les conséquences ont été particulièrement délétères, car non seulement leurs résultats ne sont pas pertinents, mais plus grave encore, elles ont entravé lourdement la réalisation des essais randomisés bien faits, les médecins préférant prescrire de manière "compassionnel" ces traitements non évalués correctement plutôt que d'inclure les patients dans un essai correct. Ce qui retarde et empêche d'avoir la réponse d'une démonstration d'efficacité ou au contraire de l'échec du médicament à évaluer.

[14] Voir par exemple l'étude de l'ivermectine dans la COVID de Soto-Becerra et al. [41] qui met en œuvre une approche d'inférence causale et d'émulation d'un essai cible et qui ne permet pas de conclure au bénéfice de celle-ci.

8.1 Problématiques méthodologiques spécifiques et solutions possibles

8.1.1 Confusion

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

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

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

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

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

- L'identification de tous les facteurs de confusion potentiels par une revue systématique de tous les caractéristiques (prédicteurs ou déterminants) conditionnant le critère de jugement considéré
- L'élaboration d'un réseau de causalité (causal graph) [42] sous la forme d'un graphe orienté acyclique (Directed acyclic graph, DAG) pour identifier les facteurs d'ajustement (et ceux sur lesquels il ne faut pas ajuster, comme les colliders)
- La prise en compte de tous les facteurs de confusion identifié ce qui suppose La disponibilité des données (de bonne qualité) pour chacune de ces variables.

Juger de l'optimalité d'un ajustement *a posteriori* n'est pas aisé même si l'ensemble des points précédents ont été suivis par l'étude. En effet, juger de l'identification complète de tous les facteurs de confusion pour chaque critère demande une expertise thématique et méthodologique poussée. De plus, dans la réalité, il est rare que l'analyse puisse prendre en compte tous les facteurs de confusion

potentiels (non-disponibilité par exemple avec les données secondaires). In fine se pose donc la question d'un **biais de confusion résiduel**.

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

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

Exemples de contrôle négatif.

Dans les études des conséquences de l'exposition in utero à un médicament, la prise du médicament longtemps avant ou après la grossesse sont des contrôles négatifs potentiels. Biologiquement il ne peut pas y avoir d'association avec les malformations par exemple, mais cette relation est susceptible d'être affectée par les mêmes facteurs de confusion que la relation d'intérêt (exposition in utero et malformation). Si l'ajustement est insuffisant, une relation apparait sur ce contrôle négatif entrainant la réfutation de l'absence de biais de confusion résiduelle.

Un évènement clinique type effet indésirable médicamenteux, que l'on sait parfaitement exclu avec le médicament d'intérêt, mais dont certains facteurs de risques sont communs avec les facteurs de confusion potentielle (facteurs de fragilité des patients par exemple) peut servir de contrôle négatif.

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

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

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

8.1.2 Autres biais

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

Des outils comme l'échelle ROBINS-I ont été développés pour estimer le risque de biais de manière standardisée. ROBINS-I [43] est maintenant largement adopté (méta-analyses intégrant des études observationnelles, recommandations GRADE [44], etc.). Il permet une évaluation du risque de biais approfondi portant sur toutes les dimensions de biais existants dans une étude observationnelle. Il a été conçu de façon à unifier les biais affectant les études contrôlées randomisées et les études observationnelles. Ainsi son niveau de « low risk of bias » correspond au degré de certitude apporté par un essai randomisé correctement conçu et réalisé [43].

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

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

L'approche d'émulation d'un essai cible (cf. section 9) a été proposée pour permettre, par une approche systématisée, d'anticiper la survenue de biais lors de la conception puis lors de l'analyse des études observationnelles , en particulier concernant la sélection des patients, la définition de la date index de suivi des patients, le recueil des évènements critères de jugement, et la définition de la population et des modalités d'analyse [34, 45]

8.1.3 Autres éléments de méthode

8.1.3.1 Étude exploratoire, étude de confirmation, découverte fortuite

Les études observationnelles, surtout celles réalisées a posteriori à partir de données secondaires, constituent des outils très intéressant pour la génération d'hypothèses / pour es approches exploratoires sans objectif spécifiquement défini a priori. C'est en particulier le cas pour la détection de risques potentiels associés à l'utilisation des médicaments. L'approche exploratoire, conduite par les données (data-driven) et non par des hypothèses préalables, va conduire à une fouille des données sans objectif défini a priori. Cette fouille large expose à un risque important de découverte fortuite, qui peut être contenu, pour une part, par des approches statistiques reposant sur l'utilisation des approches de *false discovery rate*. La possibilité de trouver une explication *a posteriori* au résultat trouvé ne permet pas de renforcer sa robustesse compte tenu du nombre d'explications possibles compte tenu de la complexité des phénomènes biologiques ([46], chapitre 1, inductivism, page 7).

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

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

Dans le contexte de construction des stratégies thérapeutiques, les études exploratoires sont donc non recevables. Les RWE pouvant être considérées pour ces définitions de stratégies thérapeutiques

doivent impérativement être issues d'études de confirmation ayant clairement un objectif en phase avec la revendication. Les objectifs compatibles avec l'acceptabilité des résultats pour confirmer la place du traitement d'intérêt dans la stratégie sont alors identiques à ceux des essais randomisés : démontrer la supériorité de N par rapport au traitement de référence sur un critère cliniquement pertinent et dans la population cible du traitement[15].

8.1.3.2 Études multibases

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

L'utilisation dans la même étude de plusieurs sources de données secondaires (bases de données) permet de limiter les conséquences de cette problématique [47]. L'association d'intérêt est recherchée simultanément avec la même méthode dans plusieurs bases et une conclusion générale ne sera effectuée que s'il y a homogénéité des résultats à travers les bases (après avoir exclu la possibilité d'une hétérogénéité explicable par les différences de populations et des modificateurs d'effet). Le résultat global de l'étude sera la méta-analyse des résultats obtenus sur chaque base. Cette approche multibases permet d'augmenter la reproductibilité des résultats et est maintenant fréquemment utilisé.

8.1.3.3 P hacking

Les termes *p hacking* ou *data dredging* désignent l'adaptation de l'analyse statistique en cours de réalisation, en fonction des résultats qu'elle produit. Ces adaptations peuvent concerner aussi bien la méthode statistique (choix de la méthode, transformation de variables, choix des covariables d'ajustement, etc.) que le jeu de données (exclusion de patients, gestion des évènements intercurrents, restriction de l'analyse à une sous population, etc.). Ces adaptations sont d'autant plus faciles à effectuer que l'étude nécessite une analyse statistique complexe, comme avec les études observationnelles par exemple.

Avec cette pratique, il est ainsi possible d'orienter les résultats dans la direction souhaitée, tout du moins en termes de signification statistique (d'où le nom de *p hacking*) [48, 49].

Il a ainsi été montré qu'avec un même jeu de données, confié à des équipes scientifiques différentes ayant des conceptions théoriques antithétiques, il était possible d'obtenir des résultats très différents et même opposés [50, 51]. L'étude perd ainsi sa valeur scientifique (assurée par le fait que la réponse à la question posée est fournie uniquement par les données) pour devenir une simple opération à produire les résultats escomptés. Il ne s'agit plus d'un test loyal d'une hypothèse thérapeutique où seule la réalité pourra la réfuter ou la confirmer, mais d'une démarche de recherche active de la façon d'analyser des données afin d'obtenir un résultat le plus proche de la réponse voulue ! Un *p-hacking reverse* a aussi été mis en évidence où l'analyse statistique est construite pour ne pas donner de différence significative [52].

Cette potentialité peut être aussi illustrée par le concept de vibration des effets [49]. Il s'agit de visualiser l'ampleur suivant laquelle « vibrent » les différents résultats (taille d'effet et p value) obtenus par toutes les possibilités d'analyse d'une même recherche d'association. Ces vibrations peuvent

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

déboucher dans certains cas sur des effets Janus où des résultats opposés sont obtenus à partir du même jeu de données.

Dans la littérature ces aspects sont souvent introduits par l'aphorisme dû à Ronald Coase : « if you torture the data long enough, it will confess to anything »[16]. On parle aussi de *data-dredging* ou partie de pêche [53, 54].

La solution réside dans la conception *a priori* de l'analyse statistique, complètement indépendante des données et des résultats produits. Cela est obtenu par l'élaboration d'un plan d'analyse statistique (*statistical analysis plan*, SAP) en amont de la disponibilité des données. Ainsi aucune adaptation de la stratégie d'analyse ne peut s'effectuer au moment de sa réalisation (sans que cela soit détectable en comparant le plan d'analyse statistique et l'analyse effectivement réalisée).

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

Cependant, pour les études réalisées a posteriori (on parle aussi d'études historiques) sur données secondaires, le SAP sera par définition élaboré alors que les données sont déjà disponibles. Pour donner la garantie de l'absence de tout p hacking (choix post hoc des variables d'ajustements, de la population d'analyse, des définitions des expositions et des critères de jugement), de publication sélective en fonction des résultats, de HARKing ou autre opération de data dredging, il est impératif que soit explicitement mentionné dans le protocole et le rapport de l'étude que l'analyse a été conçue indépendamment des données et des résultats produits [55].

Pour lever ces réserves, ces études doivent donner la garantie qu'elles ont bien procédé à une validation prospective a priori sur des données historiques d'une hypothèse formulée a priori. L'enregistrement des protocoles et des plans d'analyses statistiques, l'utilisation d'algorithme standard de phénotype, la transparence et l'attestation explicite de l'absence de ces pratiques par les investigateurs sont des éléments permettant de lever ces réserves [55, 56, 57].

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

8.1.3.4 Biais de publication et selective reporting

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

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

La solution à cette problématique n'est pas simple. L'enregistrement des protocoles est la première mesure possible, mais se heurte au fait qu'avec les études réalisées a posteriori, il est possible

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

d'enregistrer les protocoles alors que l'analyse a déjà été réalisée. L'enregistrement doit donc est associé à un engagement explicite des investigateurs que le protocole, le SAP et l'enregistrement ont été effectués avant toutes analyses et production de résultats.

8.1.4 Analyse en intention de traiter

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

Lorsqu'une étude observationnelle est réalisée dans un objectif d'évaluation d'efficacité et, éventuellement, de recommandation de l'utilisation d'un traitement, le schéma employé doit être choisi différemment, pour permettre la mesure de l'effet de l'initiation d'un traitement (et non pas celui de recevoir un traitement) [58]. Pour documenter cette décision d'instaurer un nouveau traitement et mesurer ce que cette décision induira comme amélioration dans le devenir des patients, une approche d'analyse en intention de traire est nécessaire (ou un estimé de type « treatment policy »).

8.1.5 Inférence causale (causal inference)

L'inférence causale est une approche récente, encore en plein développement, basée sur une théorie et des hypothèses, des designs et des techniques d'analyse qui permettent de tirer des conclusions de causalité à partir de données observationnelles [34, 59, 60, 61, 62]. Cette approche basée sur une mathématisation de la causalité permet de construire des stratégies et des modèles d'analyses des données permettant de conclure à la causalité. Cette approche est donc naturellement la plus appropriée pour des études qui essayent de se passer de l'apport de la randomisation en termes de causalité.

8.2 Synthèses des problématiques et de leurs solutions

Tableau 4 – Fiche de synthèse récapitulative des problématiques méthodologiques et des solutions attendues afin d'accepter un résultat d'étude observationnelle pour la construction des stratégies thérapeutiques

Problématique méthodologique ***et particularité spécifique à la nouvelle méthodologie.***	**Solution spécifique à apporter avec cette nouvelle méthodologie** ***pour garantir l'obtention du même degré de certitude qu'avec la méthodologie classique***
Nécessité d'un raisonnement contrefactuel pour identifier l'effet propre du traitement et mesurer son importance en raison de la variabilité du vivant (inter et intra sujet)	Étude observationnelle analytique (comparative) type étude de cohorte ou étude cas-témoins intégrant un groupe contrôle contemporain apportant le contrefait, s'inscrivant dans une approche d'émulation d'un essai cible
Biais de sélection Nombreuse possibilité de biais de sélection dans les études observationnelles (temps d'immortalité, ajustement ou restriction sur un collider, biais protopathique, etc.)	Utilisation d'un design d'émulation d'un essai cible, synchronisation des débuts de suivi entre les 2 groupes Ajustement après modélisation du réseau de causalité (DAGs) pour éviter l'ajustement sur les colliders Cotation par ROBINS -I en low risk of bias sur cette dimension

Biais de confusion majeur lié à la nature observationnelle (biais par indication, channeling biais)	Prise en compte de tous les facteurs de confusion dans l'analyse (quel que soit la méthode). Cela nécessite 1) d'identifier tous les facteurs déterminant le critère de jugement considéré pour établir le graphique de causalité (DAGs) et déterminer les réels facteur de confusion et 2) pouvoir prendre en compte toutes ces covariables identifiées Démontrer l'absence de biais de confusion résiduelle (contrôle négatif ou positif, analyse quantitative de biais) L'absence de randomisation ne permettant d'obtenir par design une estimation causale de l'effet traitement, doit être compensée par une approche d'inférence causale.
Biais de réalisation Aucun contrôle par design de ce biais dans les études observationnelles	Cotation par ROBINS -I en low risk of bias sur cette dimension
Biais de suivi Aucun contrôle par design de ce biais dans les études observationnelles	Cotation par ROBINS -I en low risk of bias sur cette dimension
Biais d'attrition Aucun contrôle par design de ce biais dans les études observationnelles	Cotation par ROBINS -I en low risk of bias sur cette dimension
Estimation de l'effet traitement correspond à ce que la recommandation future du traitement produirait comme changement dans le devenir des patients (compte tenu de tout le reste de la stratégie thérapeutique)	Analyse en intention de traiter
Risque de conclure à tort à l'intérêt du traitement du fait de l'erreur statistique alpha (de premier type)	Plan de contrôle du risque alpha global Définition des comparaisons inférentielles (tests qui peuvent conduire à la conclusion à l'intérêt du traitement et donc à la recommandation de son utilisation)
Multiplicité des comparaisons pouvant amener à conclure à l'intérêt du traitement ; multiplicité induisant une inflation du risque alpha global	Plan de contrôle du risque alpha global (non prise en considération de la signification nominale, non-présentation des p values non inférentielles pour éviter les surinterprétations des résultats exploratoires sans contrôle du risque alpha global)
Fraude des investigateurs	Pour les études prospectives : Monitorage de terrain, bonnes pratiques cliniques, recherche systématique de la fraude lors de l'analyse statistique, exclusion des centres en cas de suspicion de fraude Pour les études sur bases : non applicable
Fraude au niveau de l'investigateur principal (sponsor, réalisation de l'étude) possible (cf. Surgisphere)	Système d'assurance qualité (procédure opératoire standard), traçabilité, audit interne et externe (disponibilité des données)
Découverte fortuite, fouille de données (data dredging, data milking)	Réalisation d'étude observationnelle de confirmation respectant pleinement la démarche hypothético déductive avec un objectif fixé *a priori* certifié par les investigateurs. Exclusion des résultats post hoc, ou exploratoire de la prise de décision
Respect de la démarche hypothético déductive	Formulation des hypothèses *a priori*, garantie soit 1) par une démarche prospective ou 2) certifiée par les investigateurs dans le protocole
P hacking (modification de l'analyse statistique jusqu'à l'obtention des résultats voulus) fréquent dans les études observationnelles	Définition d'un protocole fixant les critères de jugement et les grandes lignes de l'analyse Définition d'un plan d'analyse statistique (SAP) précis avant que les données (même parcellaires) soient disponibles et réalisation de l'analyse statistique en stricte conformité avec ce SAP (démarche certifiée par les investigateurs)

Selective reporting (présentation, publication au niveau de l'étude que des résultats positifs permettant de faire la conclusion recherchée)	Protocole établi a priori, enregistrement dans un registre Vérification des résultats mis en avant par rapport au protocole
Biais de publication Risque important avec les études observationnelles rétrospectives compte tenu de la relative facilité de les multiplier	Difficile à exclure. Il faudrait avoir connaissance de toutes les études rétrospectives entreprises sur la même question Garantie apportée par le dossier que toutes les études observationnelles entreprises sont rapportées (solutionnable que par des obligations réglementaires) **Point majeur de dégradation du degré de certitude apporté par les études observationnelles**
Pertinence clinique	Utilisation : • D'un critère de jugement clinique (ou d'un surrogate démontré) • D'un comparateur loyal et représentant le traitement standard du moment où l'étude est analysée pour intégrer le nouveau traitement dans la stratégie thérapeutique • D'une population cible recherchée correspondant à la totalité des patients relevant du traitement évalué
Bénéfice complètement compensé par des effets délétères de manière quantitative ou qualitative	Évaluation de la safety avec la même précision et robustesse que l'efficacité Prise de décision basée sur la balance bénéfice risque = bénéfice clinique (et non pas seulement sur l'efficacité ou sur la sécurité de façon isolée et séparée)
Spin de conclusion (conclusion positive en faveur de l'intérêt du traitement dans une étude en réalité non concluante)	Ne pas lire les conclusions, ne regarder que les résultats et la méthode
Manque de transparence des rapports ou des publications ne permettant pas de voir les limites des résultats	Rapport exhaustif (pas de standardisation actuellement) Guide EQUATOR (STROBE) garantissant l'informativité des publications pour disposer de tous les éléments nécessaires à la l'évaluation critique de l'étude

8.3 Études de cas, retour sur expérience

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

Lorsque ces études observationnelles sont réalisées après les essais cliniques pour confirmer en vraie vie leur résultat, un constat complètement différent est fait avec très peu d'échecs de confirmation. Cette situation s'explique parfaitement par la problématique connue du p hacking dans les études observationnelles [63]. Quand l'étude observationnelle est réalisée en premier, avant les essais cliniques, les retours d'expériences attirent l'attention sur une faible aptitude à estimer correctement le réel bénéfice des traitements. Tandis que lorsque ces études sont réalisées après les essais alors que le résultat à obtenir est connu, elle réussit presque toujours à retrouver le résultat attendu. Le phénomène connu de p hacking provenant de la possibilité d'adapter les choix d'analyse (covariable d'ajustement ou population d'analyse) en fonction des résultats produits pourrait expliquer la bonne

performance des études observationnelles lorsqu'elles sont réalisées alors que le résultat à produire est connu. Un biais de publication n'est pas à exclure aussi dans cette situation.

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

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

L'étude par Hripcsak et al. donne un exemple complet de la méthodologie sophistiquée pour produire des RW evidence [64]. L'évaluation complète de la méthodologie de cette étude

L'étude est clairement une étude de confirmation (cf. section 8.1.3.1) dont l'objectif global est de comparer l'efficacité et la sécurité relative de la chlorthalidone et de l'hydrochlorothiazide.

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

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

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

L'étude est multibase (cf. section 8.1.3.2) afin de limiter le risque de découverte fortuite et augmenter la reproductibilité des résultats.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

8.3.2 Sécurité cardiovasculaire de l'insuline

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

Cependant il existe de nombreux exemples d'études observationnelles montrant à tort un effet indésirable. Par exemple chez les diabétiques un surcroit de mortalité et d'évènements cardiovasculaire a été observé avec l'insuline en deuxième ligne par rapport à un inhibiteur de la DDP4

dans une étude observationnelle [65]. Ce résultat n'a pas été retrouvé dans un l'essai randomisé de grande taille ORIGIN dédié à l'évaluation de l'insuline basale [66].

La possibilité de faux positif sur la sécurité des médicaments avec les études observationnelles purement exploratoire est certainement majorée par deux éléments. Le raisonnement en sécurité se base sur le principe de précaution et un simple doute suffit à faire prendre des décisions. Ainsi, les résultats de ces études sont souvent considérés malgré leurs fragilités méthodologiques reconnues. Le 2ème facteur de risque de faux positif est consécutif à la multiplicité de comparaisons présentent dans ces études exploratoires souvent réalisées pour « évaluer la sécurité », sans plus hypothèse construite. Ces limites sont levées par la réalisation de réelles études de confirmation (cf. section 8.1.3.1 et section 8.1.3.3).

8.4 Meta-recherche

La réflexion sur la fiabilité des études observationnelles pour l'évaluation de l'efficacité et de la sécurité des traitements à débuter au début des années 2000 [67, 68, 69].

Plusieurs études de méta-recherche ont étudié la concordance des résultats obtenus par les données observationnelles par rapport à ceux produits par les essais randomisés comparant les mêmes traitements dans la même situation pathologique. Aucune de ces études ne permet de conclure que les études observationnelles donnent systématiquement le même résultat que les essais randomisés [67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84]. Il faut cependant noter qu'il s'agit ici d'une des études telles que réalisées (avec leurs limites et leurs défauts, aussi bien pour les études observationnelles et les essais randomisés) et ce n'est pas une comparaison, stricto sensu, de la science observationnelle à l'approche expérimentale des essais randomisés.

Trois études par Tanen [85], Dahabreh [86] et Lonjon [87] ont été regroupée dans une analyse conjointe [72]. Une méta-analyse de la Cochrane est aussi disponible [88]

Les résultats les plus récents montrent qu'en moyenne les études observationnelles pourraient donner les mêmes résultats que les RCT, mais avec une forte variabilité [88, 89, 90]. Cependant il convient de remarquer que la mesure de concordance en moyenne sur des tailles d'effet n'est pas appropriée, car les surestimations compensent en moyenne les sous-estimations ce qui peut conduire à une moyenne des différences des estimations égale à zéro alors que dans aucun cas les 2 estimations n'ont été identiques. L'écart quadratique (ou la valeur absolue des différences) est un meilleur paramètre, mais rarement utiliser dans ces études de concordances.

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

En oncologie, Kumar et al. [74] ont utilisé les données du *National Cancer Database (NCDB)*, un registre de cancérologie, pour reproduire les résultats de survie de 141 essais randomisés. En utilisant des analyses réalisées par score de propension, une concordance des hazard ratio n'a été trouvée que dans 64% des cas et celle des p value dans 45% des cas. La conclusion proposée est que l'approche de *"comparative effectiveness"* à l'aide des données de registre de cancer produit souvent des résultats discordants avec ceux des essais randomisés. Soni et al. [91] évaluent la concordance des hazard ratio de survie entre des études observationnelles publiées et les essais randomisés correspondants. Aucune corrélation n'est retrouvée entre les hazard ratio des deux types d'études et le taux de concordances des hazard ratio n'est pas trouvé supérieur à celui attendu du fait du hasard. Aucune

caractéristique des études observationnelles améliorant cette concordance n'a été retrouvée. Dans 9% des cas, des résultats statistiquement significatifs diamétralement opposés sont observés (*Janus effect*) (5% des cas dans le papier par Kumar). Une revue de la littérature portant sur ces résultats en oncologie [70] conclue que l'essai randomisé doit donc rester le standard pour l'évaluation des médicaments en oncologie.

Durant la crise de la COVID-19, une profusion d'étude observationnelle a été publiée (dans des revues ou en préprint). Pour toutes les molécules qui ont échoué ultérieurement à monter un réel intérêt dans des essais randomisés, de nombreuses études observationnelles en faveur de leur efficacité sont disponibles [89], comme par exemple avec les plasmas de patients convalescents [92, 93]. Il faut cependant noter que dans la plupart de ces études observationnelles la méthodologie était catastrophique.

8.5 Avis de la SFPT

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

- Une étude observationnelle de confirmation, hypothético déductive, dont l'objectif défini *a priori* était clairement la comparaison de l'efficacité et de la sécurité du traitement d'intérêt à un comparateur pertinent (comparative effectiveness and safety)
- Une approche d'inférence causale aboutie
- Un design d'émulation d'un essai cible correctement conçu et réalisé
- Une analyse en intention de traiter
- La démonstration de l'absence de biais de confusion résiduelle
    - en montrant que les ajustements ont porté sur la totalité des facteurs de confusion identifiés par une approche formelle (revue systématique des facteurs de risques ou pronostiques des critères de jugement, modélisation des relations de causalité, DAGs), y compris les facteurs de confusion dépendant du temps
    - complété par une démonstration de l'absence de biais de confusion résiduel :
        - contrôles négatifs ou positifs en fonction de la nature du résultat (raisonnablement convainquant de l'absence de biais résiduel en nombre, captation des facteurs de confusion et en résultats), recalibration éventuelle,
        - analyses de biais quantitatif raisonnablement convaincantes (par exemple à l'aide, entre autres, d'approches de type « rule-out approach » ou « array approach » qui consistent à chercher le déséquilibre de prévalence de facteur de confusion hypothétique et la force d'association de celui-ci à l'événement qui seraient nécessaires pour remettre en cause les résultats. Si ces déséquilibre ou forces d'association sont irréalistes, alors la confusion résiduelle ne peut être un argument invoqué pour expliquer les résultats
- Un niveau de risque de biais coté à « low risk of bias » par ROBINS-I

- Une approche permettant d'écarter avec certitude une sélection post hoc des résultats, un p hacking, une approche exploratoire, un HARKing et les limites des approches rétrospectives : démonstration que les résultats n'étaient pas connus (même partiellement) avant l'analyse des données (certifier par le protocole de manière explicite, enregistrement du protocole, SAP daté, etc.)
- Éventuellement, une approche multibase prévue d'emblée et démontrant la reproductibilité des résultats et permettant d'exclure une découverte fortuite
- Comme pour un essai clinique, il faut aussi exiger
    - Des résultats cliniquement pertinents en termes de critères de jugement, comparaison effectuée, contexte de soins contemporain, taille d'effet
    - Une documentation satisfaisante d'une balance bénéfice risque favorable

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

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

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

9 L'approche d'émulation d'un essai cible

L'approche d'émulation d'un essai cible [34, 45, 58] représente à la date de rédaction de ce document (novembre 2021) le cadre formel le plus abouti pour tenter de produire, avec une étude observationnelle d'efficacité relative, des résultats fiables, similaires à ceux qui auraient été obtenus si un essai randomisé avait été réalisé.

Cette approche est affectée par toutes les problématiques discutées précédemment pour les études observationnelles (section 8) et doit apporter les démonstrations permettant de lever ces limites méthodologiques, en particulier, mais pas uniquement, la problématique du biais de confusion [34].

L'idée générale est de calquer le processus de conception et d'analyse de données observationnelles sur ce qui se serait passé si la question avait été abordée par un essai randomisé et d'obtenir une estimation causale. Ainsi doivent être précisément définis et appliquer sur la base de données : les critères d'éligibilité, la définition des interventions comparées, la durée de suivi, le ou les critères de jugement et la procédure d'allocation au traitement. Cette approche repose aussi sur l'utilisation d'un estimand causal pour déterminer l'effet du traitement soit en effet en intention de traiter soit un effet en per protocole.

Un autre apport de l'émulation d'un essai cible est de définir de manière explicite le début de suivi au moment où les sujets vérifient les critères d'éligibilité et débute le traitement. Il est alors possible de synchroniser le suivi des 2 groupes afin de réduire une partie des biais de sélection comme ceux liés à un temps d'immortalité.

Il reste cependant impossible d'émuler le double aveugle (comme avec toutes approches observationnelles ou essai pragmatique).

9.1 Étude de cas

Relativement peu d'exemples d'application de cette approche existent [94, 95, 96, 97].

9.2 Méta-recherche

Plusieurs travaux de méta-recherche ont consisté à utiliser l'approche d'émulation de l'essai cible pour reproduire les résultats d'essais randomisés (benchmark) [98, 99].

Dans RCT DUPLICATE [98], dix réplications d'essais randomisés à l'aide de données observationnelles basées sur l'émulation d'un essai cible ont été effectuées (cf. Figure 2). Une concordance des estimations des bénéfices est retrouvée en général avec des résultats conduisant à la même décision d'enregistrement dans 6 cas sur 10. Dans 8 cas sur 10, l'estimation de l'émulation de l'essai cible tombe dans l'intervalle de confiance du RCT.

Figure 2 – Résultats des comparaisons des résultats des essais randomisés et des émulations d'un essai cible sur des données observationnelles dans RCT DUPLICATE [98].

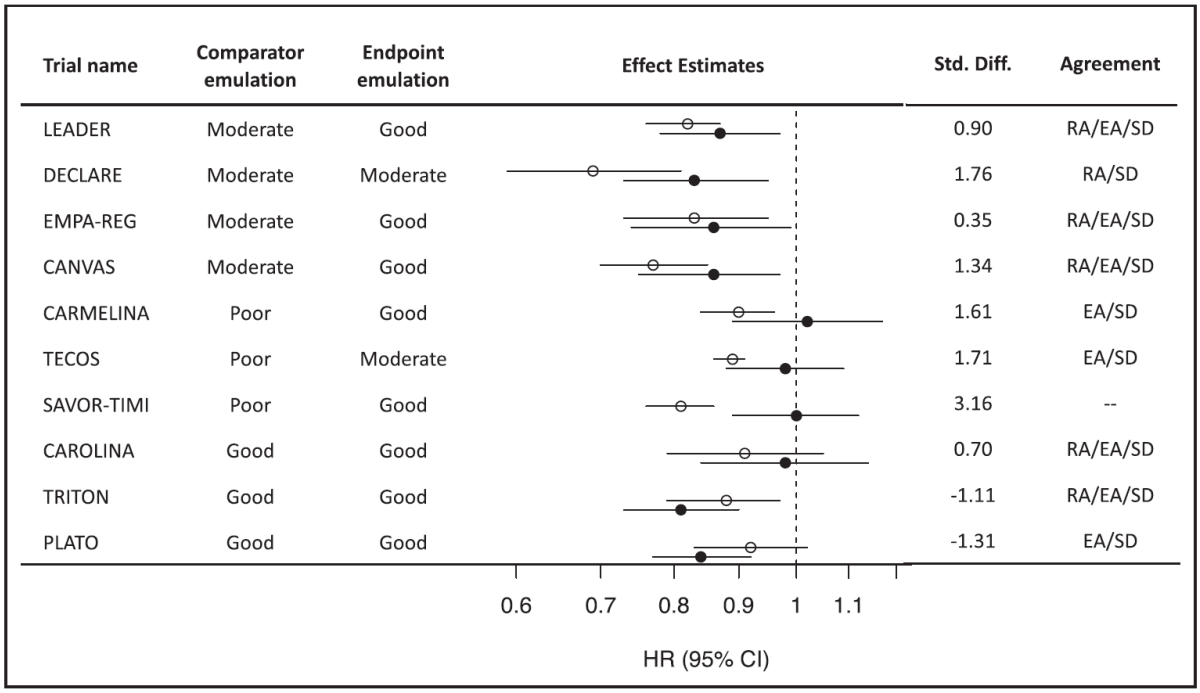


Figure 2. Agreement between randomized, controlled trial (RCT) findings and their prespecified real-world evidence (RWE) emulations.
Open circles represent the estimated hazard ratio (HR) from RWE, and filled circles represent the estimated HR from the corresponding RCT. Under the null hypothesis of no bias in the RWE, we would expect ≈5% of emulations to have a standardized difference >2. CANVAS indicates Canagliflozin Cardiovascular Assessment Study; CARMELINA, Cardiovascular and Renal Microvascular Outcome Study With Linagliptin in Patients With Type 2 Diabetes Mellitus; CAROLINA, Cardiovascular Outcome Study of Linagliptin Versus Glimepiride in Patients With Type 2 Diabetes; DECLARE, Multicenter Trial to Evaluate the Effect of Dapagliflozin on the Incidence of Cardiovascular Events; EA, estimate agreement reached; EMPA-REG, BI 10773 (Empagliflozin) Cardiovascular Outcome Event Trial in Type 2 Diabetes Mellitus Patients; HR, hazard ratio; LEADER, Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results; PLATO, Platelet Inhibition and Patient Outcomes; RA, regulatory agreement reached; SAVOR-TIMI, Does Saxagliptin Reduce the Risk of Cardiovascular Events When Used Alone or Added to Other Diabetes Medications; SD, standardized difference <1.96; TECOS, Sitagliptin Cardiovascular Outcomes Study (MK-0431-082); and TRITON, Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition With Prasugrel–Thrombolysis in Myocardial Infarction 38.

9.3 Avis de la SFPT

L'émulation d'un essai cible apparait être la formalisation la plus avancée à l'heure actuelle pour la réalisation d'études observationnelles sur l'efficacité des médicaments. Elle intègre aussi les principes de l'inférence causale.

Cette approche semble être à privilégier pour l'obtention de résultats de haut de degré de crédibilité à partir des données de la vraie vie (RWD). Ces études doivent apportées toutes les garanties de fiabilité des résultats nécessaire pour interpréter les résultats des études observationnelles (cf. section 8.5).

10 Les registres

Les registres (*registry*) sont des collections de données [100] sur des patients présentant une certaine pathologie ou traité avec un traitement particulier (étude post-enregistrement par exemple). Il peut s'agir aussi de registres populationnels (*population registries*). Dans d'autres contextes, le terme cohorte (à ne pas confondre avec les études de cohortes) est utilisé dans un sens similaire comme le terme de base de données.

Pour les questions concernant l'efficacité des traitements, les **registres définis par le traitement** sont des **études mono-bras** (cf. section 16). Ces registres peuvent aussi fournir des groupes contrôles externes pour des études mono-bras de nouveau traitement.

Dans les **registres définis par la pathologie**, les patients sont traités de différentes manières, ce qui permet la réalisation **d'études observationnelles d'efficacité relative** (cf. section 8).

Le terme registre désigne l'infrastructure de recueil et de conservation des données. Avec ces données des études observationnelles (*registry based study*) peuvent ensuite être réalisées avec des objectifs divers et variés (description, estimation d'incidence ou de prévalence, et éventuellement efficacité ou sécurité des traitements) et différents types d'analyses (études de cohortes, cas témoin, transversales, avant-après, etc.) [101].

Il est aussi possible de réaliser des **essais randomisés** se basant sur la structure du registre pour le recueil des données (*registry based randomised controlled trials*) [102, 103, 104]. Ces essais sont aussi classés parmi les **essais pragmatiques** (cf. section 11).

Actuellement, l'accent est particulièrement mis sur la qualité des données (comme par exemple avec le document de guidance européen [100]). La qualité des données est un point important, particulièrement lorsque les résultats de l'étude sont en faveur de l'absence d'effet, mais le problème structurel des données observationnelles reste le biais de confusion. Des données exactes et complètes sont nécessaires, mais n'évince pas la problématique des ajustements, du biais de confusion résiduel et du biais de sélection.

11 Les essais pragmatiques

Le terme « essai pragmatique » peut désigner plusieurs concepts en l'absence de stabilisation terminologique [105], parmi lesquels on retrouve :

- Des essais randomisés qui cherchent à évaluer une pratique médicale basée sur le traitement, plus que le traitement lui-même (en utilisant, par exemple, des critères de sélection de patients les plus larges possibles) [106].
- Des essais randomisés dont le recueil des données a été simplifié, en s'appuyant sur un ou des systèmes d'information recueillant déjà des données cliniques en routine sur les patients inclus [107]. Le but est de simplifier la réalisation de l'essai afin de réduire les coûts et d'accélérer leur réalisation.
- Une randomisation en « vraie vie », où la totalité du recueil de l'information s'appuie sur des bases de données administratives. La FDA intègre ces essais dans la classe des « *real world evidence* », mais avec un niveau de preuve bien supérieur à celui des RWE obtenues par des études observationnelles [17].

Le terme pragmatique peut aussi être utilisé pour faire la distinction avec les essais explicative [108].

Ces essais pragmatiques présentent comme avantage, outre un coût optimisé et une charge de travail réduite pour les investigateurs, une représentativité et une pertinence clinique satisfaisantes. Et au niveau méthodologique, ils ont la valeur des essais randomisés.

Les essais pragmatiques sont aussi la réponse à une critique souvent faite aux essais randomisés qui est de ne pas inclure des patients représentatifs de la vraie vie. Ce point n'est pas une limite intrinsèque des essais randomisés en eux-mêmes, mais provient des pratiques qui tendent à réaliser les essais dans des conditions le plus favorables aux traitements étudiés. Rien n'empêche d'inclure dans un essai randomisé une population large, la plus représentative possible des patients. Ceci est régulièrement le cas, par exemple avec les méga-essais incluant plusieurs milliers de patients, et qui n'effectuent pas de sélection inutile. En réponse à cette critique, il est souvent avancé que les études observationnelles seraient la réponse pour évaluer le bénéfice du traitement dans une population large, mais cette approche expose au risque d'obtenir une réponse biaisée (si des moyens très importants n'ont pas été mis en œuvre dans la conception de l'étude et son analyse pour éviter ces biais).

Avis de la SFPT

Pour positionner ou confirmer la position d'un nouveau traitement dans la stratégie thérapeutique à partir d'un essai pragmatique, il faut que tous les critères nécessaires pour les essais randomisés en général soient remplis auxquels s'ajoutent les critères suivants :

La représentativité de la population : il est attendu des essais thérapeutiques, quel que soit leur type, qu'ils incluent des patients représentatifs de la population cible envisagée. Sur ce point l'approche pragmatique (définie par la recherche de la représentativité) apporte la garantie que cette attente est remplie. Cependant beaucoup de nouveaux traitements, notamment en oncologie, sont ciblés sur des biomarqueurs (comme des sous-types moléculaires ou génétiques

17 https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

de la tumeur). Dans ce cas il n'est pas attendu une représentativité des patients inclus par rapport à la population générale des patients, mais bien par rapport à la sous-population des patients présentant le biomarqueur considéré. En particulier, il est attendu que les patients ne soient pas sursélectionnés, par exemple sur leur état général (si une option à visée curatrice est envisagée en pratique chez les patients avec un état général dégradé) ou leurs comorbidités.

La qualité des données malgré la simplification de l'étude : pour les considérations liées à la simplification du recueil de l'information, il est important que ces essais apportent la garantie de la qualité des données (tout particulièrement pour les résultats de sécurité ou, de façon plus générale, tous les résultats qui ne montrent pas de différence entre les groupes comparés). Les sources de données utilisées contiennent souvent beaucoup de données manquantes (de façon implicite ou explicite) et parfois des données faussées comme dans les bases administratives pour optimiser les aspects financiers. Il doit être démontré que ce problème n'affecte pas le ou les critères de jugement et qu'un biais d'attrition (dû à des données manquantes informatives) ou un biais « *toward the null* » est exclu.

12 Les essais plateformes

Les essais plateformes, également appelés **essais multi-bras et multi-étapes** (*MAMS: multi-arm, multi-stage trials*), peuvent s'avérer utiles lorsque plusieurs nouveaux traitements sont disponibles pour une indication donnée. En effet, les essais plateformes ont pour objectif **d'évaluer plusieurs traitements expérimentaux à un contrôle unique**.

Le principe consiste à mettre en place et à maintenir une logistique d'inclusion de patients présentant une même maladie, permettant de comparer plusieurs bras, simultanément ou les uns après les autres [109, 110, 111], à un groupe contrôle unique. L'essai est régi par un « master protocol » auquel est annexé, pour chaque nouveau traitement à l'étude, un protocole spécifique.

L'essai est par essence adaptatif au sens où certains traitements testés seront abandonnés et d'autres ajoutés au cours du temps. Le groupe contrôle peut lui aussi évoluer au cours du temps, si le nouveau standard de soin (*standard-of-care*) a changé depuis la mise en place de l'essai.

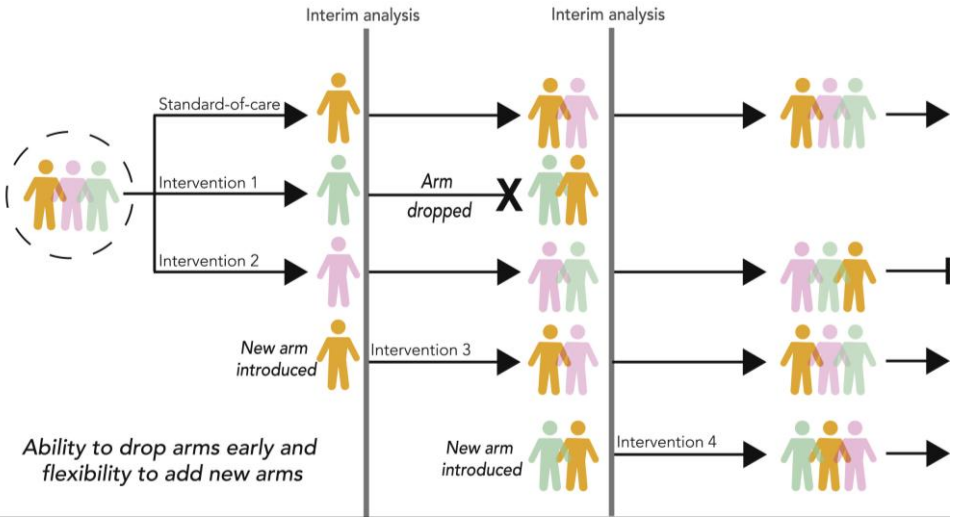


Figure 3 – Design d'un essai plateforme
(reproduit avec autorisation de [112])

L'intérêt de cette approche est d'**optimiser la durée** d'évaluation de nouveaux traitements en évitant de mettre en place une logistique spécifique pour chaque médicament (recrutement de centre, formation, mise en place des procédures). Du fait d'un groupe contrôle partagé entre tous les traitements expérimentaux, les essais plateformes permettent également de **réduire le nombre de sujet inclus** par rapport à la réalisation d'essais séparés.

Les essais plateformes sont d'abord apparus pour la réalisation d'essais de phases 2 [113] et sont majoritairement utilisés dans des essais de phase 3 actuellement [114]. Dans certains cas, la transition de la phase 2 à la phase 3 se fait au sein du même essai, en s'appuyant sur un design « sans couture » (*seamless*) : parmi les différents traitements évalués en phase 2, les plus prometteurs passent en phase 3. Le suivi des patients initialement inclus dans la phase 2 peut alors se poursuivre sur une plus longue période, et le critère de jugement principal peut être différent entre les 2 phases. Par exemple, l'évaluation de la phase 2 peut reposer sur un *surrogate* tel que la survie sans progression (PFS: *progression-free survival*) alors que la phase 3 est évaluée sur un critère clinique tel que la mortalité toute cause (OS: *overall survival*). Ce design particulièrement flexible a d'abord été proposé dans le domaine de l'oncologie [113, 115], et a récemment été utilisé à plusieurs reprises pour l'évaluation rigoureuse des traitements dans la COVID-19 [116][117, 118].

12.1 Problématiques méthodologiques

Au niveau méthodologique les essais plateformes apportent la plupart des garanties souhaitées lorsqu'ils s'appuient sur un design adéquat. Le terme de « plateforme » ne préjuge en rien des caractéristiques méthodologiques de l'étude sous-jacente. Toutefois, certains aspects méthodologiques inhérents à cette approche méritent une attention particulière.

Un point important concerne le **respect de la contemporanéité du groupe contrôle**. Dans les essais plateformes, la randomisation de nouveaux patients dans le groupe contrôle est continue, sur des périodes pouvant être prolongées, et antérieurement à l'ajout d'un nouveau traitement expérimental dans l'essai. Ainsi, il est possible de constituer un groupe contrôle au début de la mise en place de l'essai plateforme, avant l'introduction du nouveau traitement, ce qui reviendrait à faire des comparaisons à une cohorte historique et ne donnerait aucune garantie dans le contrôle de la confusion. En effet, même sur de courtes périodes, le standard de soin peut évoluer rapidement et avoir un impact sur la comparabilité entre les groupes. Ce fut par exemple le cas avec l'introduction des corticoïdes dans la prise en charge des patients atteints de formes graves de COVID-19 [119].

Le **ratio d'allocation** entre les groupes expérimentaux et le groupe contrôle est également un élément à prendre en compte. Il est généralement recommandé d'allouer plus de patients au groupe contrôle que dans les bras expérimentaux afin d'optimiser la puissance globale de l'essai, sauf dans quelques cas très particuliers où pour des raisons éthiques il peut être préférable de privilégier un traitement expérimental [120]. Le fait d'avoir un contrôle partagé d'effectif trop faible augmente aussi le risque de conclure à tort (dans un sens ou dans l'autre) si par malchance ce groupe à un risque basal supérieur ou inférieur aux autres, due aux fluctuations d'échantillonnage [121].

Le **double-aveugle**, bien qu'il soit théoriquement possible, s'avère complexe à mettre en œuvre. En pratique, la grande majorité des essais plateformes est donc réalisée en ouvert [114]. L'utilisation du double-aveugle nécessite l'utilisation de *multiple dummy*, d'autant plus difficile qu'il y a de traitements expérimentaux aux formes galéniques différentes évalués simultanément. De plus, cela implique de connaître à l'avance les différents traitements sui seront étudiés, ce qui limite la flexibilité qui est l'un des atouts majeurs des essais plateformes [121].

Une autre problématique méthodologique importante des essais plateformes est la **gestion de la multiplicité des comparaisons** (dans le temps du fait d'analyses intermédiaires, et entre les différents bras de traitement), qui peut s'avérer assez complexe. Ces essais incluent en effet des analyses intermédiaires régulières, sur lesquelles sont basées les décisions de maintien ou d'exclusion, pour futilité le plus souvent, des traitements expérimentaux dans l'essai. Par ailleurs, si plusieurs traitements sont comparés à un même groupe contrôle, ils ne sont pas nécessairement comparés entre eux. Ainsi, deux situations différentes sont envisagées [122] :

- Les comparaisons au groupe contrôle des différents traitements testés le sont indépendamment les uns des autres (comme autant d'essais à 2 bras successifs dans la même pathologie). La mesure de l'erreur de type I est qualifiée de *pairwise type I error rate* (PWER) : elle correspond à la probabilité de rejeter à tort l'hypothèse nulle pour le résultat principal d'un bras expérimental particulier à la fin de l'essai, indépendamment des autres bras.
- Des comparaisons sont également effectuées entre les différents bras afin de trouver le nouveau traitement (ou la posologie) le plus efficace dans cette pathologie. On s'intéresse alors au risque alpha global, qualifié de *familywise type I error rate* (FWER) : c'est la probabilité de rejeter à tort l'hypothèse nulle pour le résultat principal pour au moins un des bras expérimentaux d'un ensemble de comparaisons dans un essai à plusieurs bras.

Le choix de l'une ou l'autre de ces deux approches dépend étroitement de l'objectif global de l'essai.

12.2 Etude de cas

Un des premiers exemples d'essai plateforme est l'essai STAMPEDE dans le cancer de la prostate [123, 124, 125]. Cet essai emblématique démarré en 2005 a permis de comparer jusqu'à huit traitements expérimentaux simultanément. La figure suivante illustre les différents bras de l'étude, avec une modification du standard de soin (initialement traitement expérimental C) en 2015.

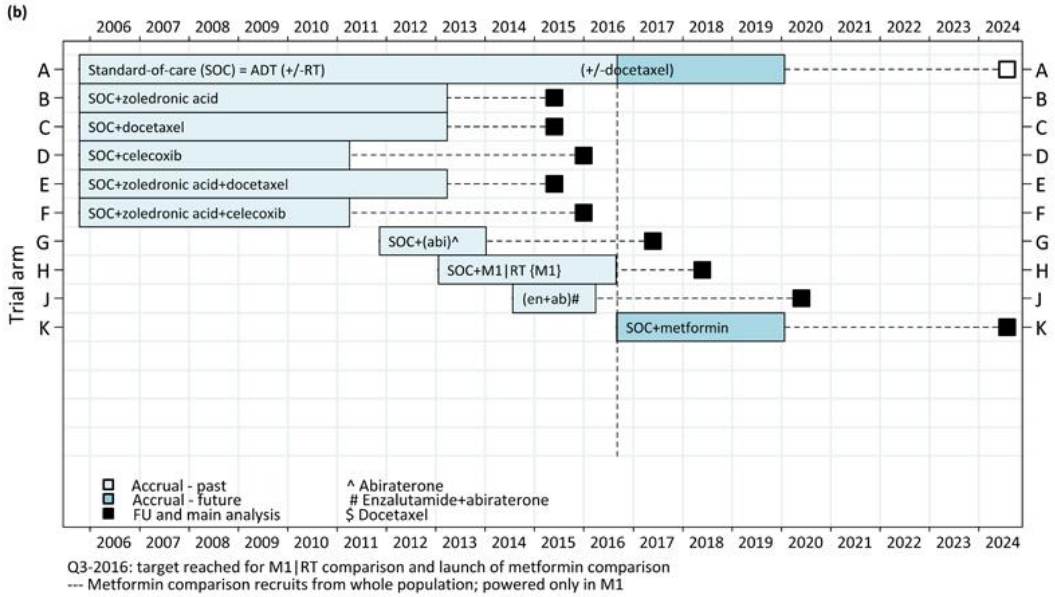


Figure 4 – ***Recrutement dans l'essai STAMPEDE [126]***

Cet essai a utilisé l'approche *seamless* pour comparer les traitements expérimentaux au groupe contrôle :

- dans un premier temps sur un **critère intermédiaire**, la survie sans progression (FFS: *failure free survival*),
- puis dans un second temps, sur un **critère définitif** pour l'analyse finale, la survie globale (OS: *overall survival*).

Le tableau suivant résume les différentes analyses intermédiaires réalisées, le critère de jugement (OM) et le seuil de risque alpha retenu pour chacune d'elle. Pour les analyses intermédiaires, des seuils élevés ont été retenus, alors que le seuil de risque final était conventionnel (0.025 en unilatéral, selon l'approche PWER).

Stage	Type	OM	HR	Power	Sig. α_{jk}	Critical HR	Control Events
1	Activity	FFS	0.75	95%	0.500	1.00	114
2	Activity	FFS	0.75	95%	0.250	0.92	215
3	Activity	FFS	0.75	95%	0.100	0.89	334
4	Efficacy	OS	0.75	90%	0.025	-	403

Source: MRC Clinical Trials Unit at UCL

Plus récemment, l'essai RECOVERY dans la COVID-19 illustre bien l'intérêt et l'efficience des essais plateformes [118]. Cet essai a permis d'inclure près de 45000 patients et de statuer sur l'intérêt clinique de neuf traitements dans la COVID-19 (parmi lesquels dexaméthasone, lopinavir/ritonavir, hydroxychloroquine, tocilizumab, plasmathérapie, etc.) [127, 128, 129] en un temps record et avec une précision/puissance adaptée. Notons que ce succès n'est pas uniquement lié au design type plateforme, mais également au côté pragmatique de l'essai, qui a simplifié au maximum le recueil d'information afin de faciliter sa mise en œuvre. L'essai s'est ainsi démarqué d'autres essais plateformes plus lourds, qui n'ont pas eu la même efficience.

12.3 Méta-recherche

Une seule étude de méta-recherche est disponible. Publiée en novembre 2019 [114], donc avant les nombreux essais plateformes mis en place lors de la pandémie de COVID-19, elle recensait 16 essais plateformes publiés jusqu'en juillet 2019. Les principales caractéristiques de ces 16 essais plateformes sont : (i) essais randomisés (15/16), (ii) essais de phases 3 (7/15) – dont 4 essais combinés (*seamless*) de phase II/III –, (iii) aucun en double aveugle, du fait de la complexité évoquée plus haut.

Bien que cette étude ne se penche pas sur la fiabilité des résultats de cette approche, la question ne se pose pas car elle repose sur la méthodologie classique et ne change que des aspects liés à la réalisation pratique de cette méthodologie. Il conviendra cependant de surveiller que ces études mettent bien en places tous les principes méthodologiques classiques et qu'aucune régression vers moins de rigueur ne survienne cachée derrière la nouveauté et la complexité de ces études.

12.4 Avis de la SFPT

Afin de garantir la comparabilité entre les groupes, il doit être explicite que les patients contrôles ont bien été contemporains des patients traités [119].

Le ratio d'allocation doit généralement être déterminé afin d'augmenter la taille du groupe contrôle, d'un coefficient multiplicateur proche de $\sqrt{t}$, t étant le nombre de traitement expérimentaux testés simultanément et comparés au groupe contrôle. Ainsi dans l'essai STAMPEDE, qui comportait initialement 5 bras expérimentaux, le ratio retenu était 2:1:1:1:1:1 [126].

Le double-aveugle, bien que théoriquement possible, n'est jamais mis en œuvre car difficile à réaliser. Cela fragilise le maintien de la comparabilité entre les groupes lors du suivi ou de l'évaluation des critères de jugement.

Compte-tenu de la multiplicité des comparaisons, dans le temps et parfois entre les groupes, la stratégie de contrôle du risque alpha doit être clairement détaillée et définie *a priori*. Pour les analyses intermédiaires, comme dans les essais classiques, les décisions d'arrêt (le plus souvent pour futilité) reposent sur des règles usuelles, telles que Haybittle-Peto ou DeMets-Lan. Selon l'objectif de l'étude, la méthode de O'Brien-Flemming pourrait être jugée trop stricte.

Un point important concerne le choix du contrôle du PWER ou du FWER. Le contrôle du PWER est recommandé lorsque les bras expérimentaux sont très différents les uns des autres. Cette situation a l'avantage de ne pas poser de problème de multiplicité lors de l'ajout d'un nouveau traitement

expérimental dans l'essai [122]. En revanche, l'emphase doit être mise sur le contrôle du FWER lorsque les différents traitements expérimentaux sont proches (par exemple différentes doses, durées, ou schémas d'administration d'un même traitement, etc.). L'ajout d'un nouveau bras expérimental impactera alors le contrôle du risque alpha global ; des méthodes ont récemment été proposées pour prendre en compte ce risque, mais leur mise en œuvre est complexe [122].

Ces essais apportent une véritable optimisation de l'évaluation concomitante de différents traitements. Ils facilitent l'accès à des évaluations de haut standard et apportent de façon plus rapide et plus efficiente des réponses fiables et cliniquement pertinentes (pour les essais correctement conçus et réalisés). Un autre avantage des essais plateformes est d'être centrés sur la pathologie à traiter. Ces essais sont moins sujets à une optimisation centrée sur l'intérêt du traitement que les essais ad-hoc traditionnel, mis en place uniquement pour tester le nouveau traitement [130].

Compte-tenu de ces avantages évidents, cette approche devrait être généralisée. Par exemple, les groupes collaboratifs d'investigateurs qui réalisent des essais indépendants, au coup par coup, pourraient mutualiser les ressources et offrir une plateforme continue d'inclusion des patients basée sur un « *master protocol* ».

Toutefois, le design, la mise en œuvre et l'analyse de tels essais est complexe. En outre, il existe aujourd'hui des points de blocages d'ordres administratif et financier. Cette approche étant encore inhabituelle, les cadres actuels d'accompagnement à la promotion (pour les essais académiques notamment) et de financement doivent évoluer pour s'adapter à cette approche.

13 Les essais bayésiens

Les essais bayésiens sont des essais de méthodologie classique (randomisation, etc.), mais dont l'exploitation quantitative des données s'effectue dans un cadre bayésien et non plus fréquentiste.

Le réel intérêt de l'inférence bayésienne dans la recherche de preuve de haut degré de certitude du bénéfice clinique des traitements est de produire des résultats basés sur des concepts qui sont directement intelligibles (probabilité *a posteriori* d'efficacité, intervalle de crédibilité) et qui correspondent directement à la question du chercheur (quelle est la probabilité que le traitement « marche » ?). Les résultats familiers bien connus comme la p value ou l'intervalle de confiance, qui sont spécifique de l'approche fréquentiste, ne sont pas estimés avec cette méthode.

Par contre, pour l'essai thérapeutique, l'intérêt du bayésien n'est pas dans la possibilité, qui est souvent mise en avant comme avantage de l'approche, de prendre en considération l'idée *a priori* que peut avoir le chercheur sur le résultat de l'étude. À l'inverse, c'est cet aspect de l'inférence bayésienne qui est exploitée dans l'emprunt d'information (cf. section 18).

13.1 Principes des essais bayésiens

L'approche bayésienne repose sur la distribution de probabilité du paramètre d'intérêt, par exemple le risque ratio. Les résultats peuvent être donnés sous forme graphique en représentant cette distribution (par un histogramme parfois) ou sous forme résumée par la médiane (ou la moyenne) et les 2.5^{e} et 97.5^{e} percentiles (qui constituent l'intervalle de crédibilité à 95%, car 95% de la distribution est contenu entre ces 2 percentiles).

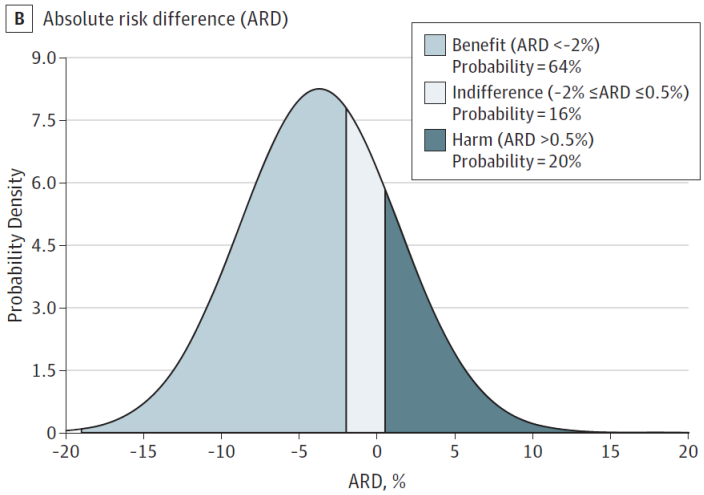


Figure 5 – Exemple de présentation de la distribution de l'effet traitement produite par une approche d'inférence bayésienne
Ici l'effet traitement est mesuré par la différence des risques (ARD). L'absence d'effet correspond à la valeur zéro. Les valeurs négatives correspondent à un bénéfice et les valeurs positives à un effet délétère.

Ce résultat est produit à partir des données fournies par l'essai combinées avec une idée *a priori* de cette distribution de l'effet du traitement, appelé couramment « l'apriori » (*prior* en anglais) pour produire une distribution « *a posteriori* », le résultat de l'étude.

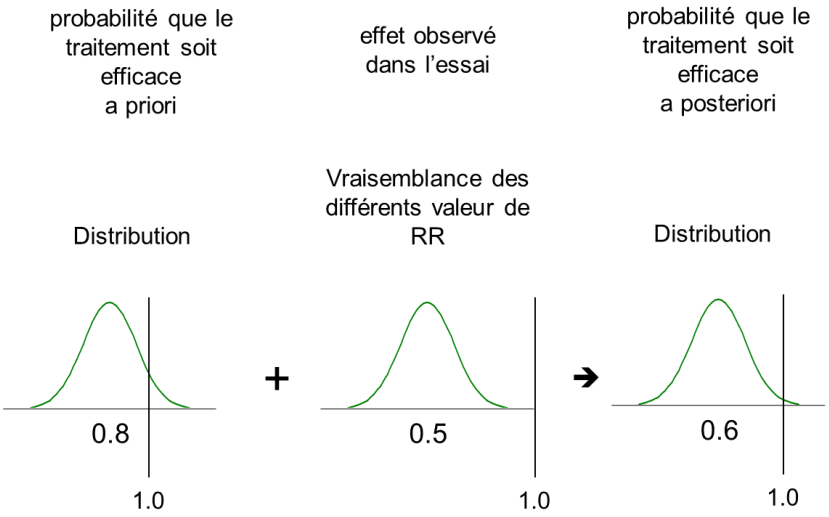


Figure 6 – Illustration du processus de production du résultat *a posteriori* dans l'inférence bayésienne
Le résultat est la combinaison de l'information apportée par l'essai (résultat de l'essai) avec une idée ***a priori*** de la distribution de l'effet du traitement (souvent arbitraire)

13.1.1 Résultats des essais bayésiens

À partir de cette distribution « *a posteriori* » de l'effet traitement, plusieurs mesures sont dérivées, principalement une estimation ponctuelle (médian ou moyenne), l'intervalle de crédibilité à 95% et la probabilité *a posteriori* que le traitement soit efficace.

La distribution de l'effet du traitement représente l'incertitude avec laquelle on connait l'effet du traitement. Plus la distribution est étalée, plus l'incertitude est grande. Sur cette distribution (cf. Figure 7) on peut calculer la probabilité que l'effet du traitement soit dans la zone du bénéfice (c'est la zone en dessous de l'absence d'effet). Cette probabilité est l'aire sous la courbe de cette zone (en dessous de 1 pour une mesure relative, risque ratio, odds ratio, hazard ratio ; ou en dessous de zéro pour une différence de moyenne, de risque).

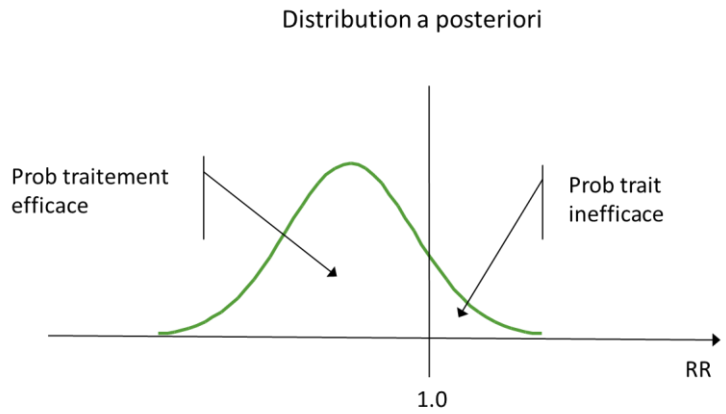


Figure 7 – Illustration du calcul de la probabilité *a posteriori* que le traitement soit efficace
Il s'agit de l'aire sous la courbe en dessous de la valeur de l'absence d'effet. Ici inférieure à 1 comme il s'agit d'une distribution de risque ratio (RR). La distribution représente l'incertitude sur l'estimation du risque ratio. Il s'avère qu'il y a une possibilité que celui-ci soit 1 ou supérieur à 1 (effet délétère), mais il est plus probable qu'il soit inférieur à 1 (la plus grande partie de la distribution est en dessous de 1). La probabilité que ce risque ratio soit inférieur à 1 est calculée par l'aire sous la courbe.

Pour pouvoir conclure à l'intérêt du traitement à partir de la probabilité *a posteriori*, il est nécessaire de fixer un seuil de décision, sinon la conclusion serait arbitraire, décidée au cas par cas et très influencée par l'intérêt de conclure au bénéfice du traitement. Ce seuil correspond au degré de certitude que l'on souhaite avoir pour adopter le nouveau traitement. Par analogie avec le niveau de risque alpha utilisé dans l'approche classique fréquentiste, ce seuil est souvent fixé à 97.5% même s'il ne s'agit pas des mêmes concepts. Comme la notion de risque alpha existe aussi dans l'essai bayésien (cf. section suivante), ce seuil est maintenant fixé pour garantir un risque alpha de 97.5%. Cependant compte tenu de la nouveauté de cette approche, aucun standard n'a été établi et certains essais peuvent revendiquer la démonstration de la supériorité du traitement avec un seuil nettement inférieur à ces valeurs.

Exemple de choix effectué dans un essai

The criterion for declaring a most or least effective treatment was a probability greater than 0.975. The threshold of 0.975 was chosen by convention (analogous to an alpha of 0.025 in a one-sided comparison) and because a simulation study showed that with this threshold and trial design, the type I error rate was controlled.

Il est aussi possible de calculer l'intervalle qui englobe 95% des valeurs les plus probables. C'est l'intervalle de crédibilité à 95%. On peut remarquer que la définition de l'intervalle de crédibilité bayésien correspond à l'interprétation erronée de l'intervalle de confiance à 95% fréquentiste qui est souvent faite[18], illustrant au passage que les concepts bayésiens sont plus intuitifs et plus faciles à interpréter correctement. Ce dernier point représente l'avantage de l'approche bayésienne dans le cadre des essais de phase 3.

[18] L'intervalle de confiance à 95% est l'intervalle qui a une probabilité de 95% de contenir la vraie valeur de l'effet du traitement (vraie valeur qui est considérée comme fixe). En d'autre termes, si l'on réplique (hypothétiquement) un grand nombre de fois le même essai, 95% des IC ainsi générés contiendront la vraie valeur.

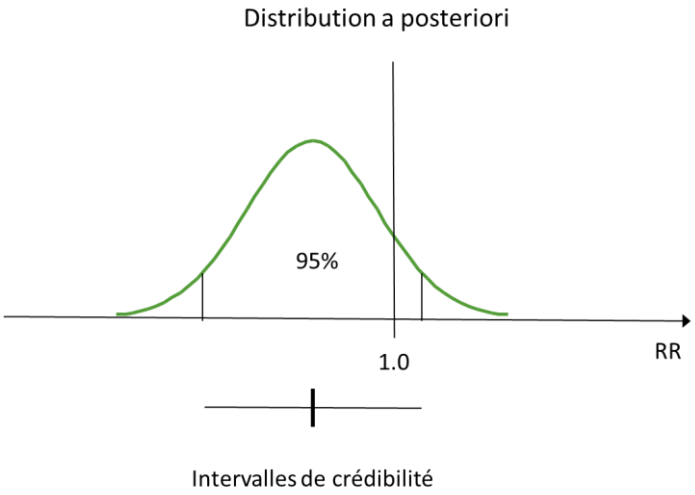


Figure 8 – Illustration de la détermination de l'intervalle de crédibilité

13.1.2 Risque alpha et multiplicité

Dans l'inférence bayésienne, la notion d'erreur statistique dans la conclusion perdure, car cette problématique concerne la décision prise à partir des résultats et non pas l'approche d'estimation utilisée pour produire les résultats. Le risque alpha trouve son essence dans les fluctuations aléatoires d'échantillonnage qui, sous l'hypothèse nulle, peuvent quand même produire, de manière aléatoirement, une structure de données en faveur du traitement étudié. Dans ce cas, quelle que soit l'approche d'estimation utilisée, l'appréciation de l'effet traitement sera erronée vu que ce sont les données elles-mêmes qui sont, à tort, en faveur de l'effet du traitement. Pour gérer cela, les essais calculent le seuil de probabilité *a posteriori* de l'efficacité autorisant de conclure à la démonstration de l'effet de telle façon qu'il garantisse un risque alpha au niveau habituel (2.5%, car la décision est par essence unilatérale).

Exemple de fixation du seuil en fonction du risque alpha global

The criterion for declaring a most or least effective treatment was a probability greater than 0.975. The threshold of 0.975 was chosen by convention (analogous to an alpha of 0.025 in a one-sided comparison) and because **a simulation study showed that with this threshold and trial design, the type I error rate was controlled**. [131]

De même la multiplicité des comparaisons pouvant conduire à conclure à l'intérêt du traitement entraine une inflation du risque alpha global. Les méthodes de gestion de la multiplicité utilisées habituellement sont utilisées dans l'essai bayésien.

The first coprimary outcome, time to first recovery, was analysed using a Bayesian piecewise exponential model regressed on treatment and stratification covariates (age and comorbidity), and included parameters for time interval (0–7 days, 8–14 days, 15–21 days, and >21 days from random allocation). The second coprimary outcome, hospitalisation or death, was analysed using a Bayesian logistic regression model regressed on treatment and stratification covariates (age and comorbidity). We included these

stratification covariates in the primary analysis as response adaptive randomisation increases the risk of imbalance on these variables. **The coprimary outcomes were evaluated using a so-called gate-keeping strategy**. For a given treatment, the hypothesis for the time to first recovery endpoint was evaluated first and, if the recovery null hypothesis was rejected, the hypothesis for the second coprimary endpoint of hospitalisation or death was evaluated. **This gate-keeping strategy preserves the overall type I error** of the primary endpoints without additional adjustments for multiple hypotheses. In the context of multiple interim analyses, the master protocol specified each null hypothesis to be rejected if the Bayesian posterior probability of superiority exceeded 0·99 for the time to recovery endpoint and 0·975 (via gate-keeping) for the hospitalisation or death endpoint [132].

	Azithromycin plus usual care	Usual care alone	Estimated treatment effect (95% Bayseian credible interval)	Probability of meaningful effect	Probability of superiority
Primary outcomes (primary analysis population)					
First reported recovery	402/500 (80%)	631/823 (77%)	..	..	..
Time to first reported recovery (days)	7 (3 to 17)	8 (2 to 23)	1·08 (0·95 to 1·23)*	0·23*	0·89*
Hospitalisation or death at 28 days	16/500 (3%)	28/823 (3%)	0·3% (−1·7 to 2·2)†	0·042†	0·64†
Primary outcomes (SARS-CoV-2-positive analysis population)					
First reported recovery	136/186 (73%)	163/236 (69%)	..	..	..
Time to first reported recovery (days)	9 (4 to not reached)	13 (5 to not reached)	1·12 (0·91–1·38)*	0·47*	0·86*
Hospitalisation or death at 28 days	11/186 (6%)	17/236 (7%)	1·6% (−3·1 to 6·2)†	0·43†	0·76†

Data are n/N (%) or median (IQR). HR=hazard ratio. *Estimated HR derived from a Bayesian piecewise exponential model adjusted for age and comorbidity at baseline, with 95% Bayesian credible interval. HR >1 favours azithromycin. †Estimated absolute benefit in percentage of hospitalisation or death derived from a Bayesian logistic regression model adjusted for age and comorbidity at baseline, with 95% Bayesian credible interval. A positive value favours azithromycin.

Table 2: **Primary outcomes**

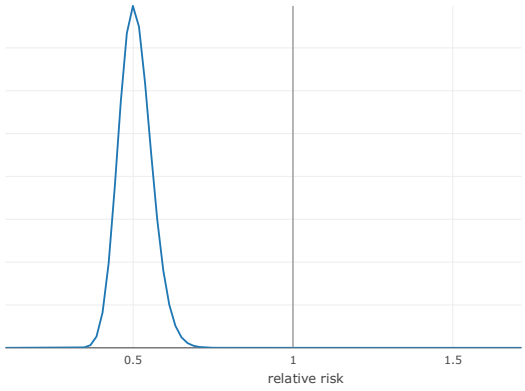
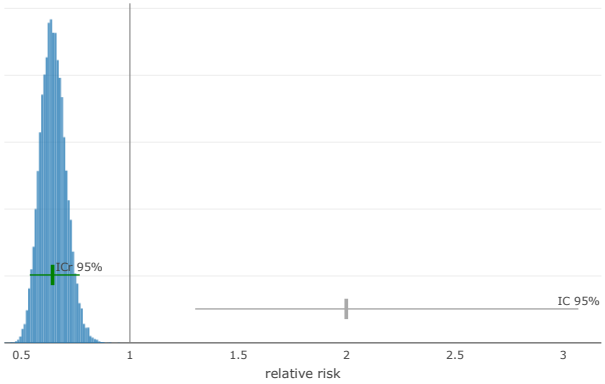
13.1.3 Dépendance des résultats à l'apriori

La principale limite de l'inférence bayésienne pour l'évaluation des nouveaux traitements (et qui a été longtemps un frein à l'adoption de cette approche) est le fait que l'apriori arbitraire peut conditionner presque en totalité le résultat (*a posteriori*) indépendamment de ce qui a été observé dans l'essai, cette limite a été longtemps un frein à l'adoption de cette approche.

> *Plus l'apriori est non informatif, plus le résultat a posteriori est conditionné par les données observées. Plus l'apriori est informatif (de façon objective du fait de données déjà connues, ou subjective du fait de croyance pure), moins le résultat a posteriori est conditionné par les données observées.*

Ce risque est illustré par l'exemple présenté Figure 9.

Figure 9 – Illustration de la dépendance des résultats bayésiens à l'apriori

Distribution *a priori*	**Distribution *a posteriori*** (et résultat fréquentiste IC 95%)
	
Distribution *a priori* choisie, cette distribution est en faveur d'une réduction de 50% du critère de jugement (RR=0.5) avec une certitude élevée (distribution étroite)	Les données de l'essai sont en faveur d'un doublement de la fréquence du critère de jugement (RR=2, cf. l'intervalle de confiance IC 95% fréquentiste à droite en gris). Malgré cela, en raison de l'apriori choisi, la distribution *a posteriori* (en bleu) des risques relatifs est entièrement faveur de la réduction du critère de jugement, avec une moyenne de 0.6 (cf. intervalle de crédibilité ICr, à gauche, en vert) et une probabilité que RR<1.0 égale à 1.

Cette possibilité est inacceptable dans le contexte de la confirmation par les faits du bénéfice des traitements. Il est cependant possible de faire de l'inférence bayésienne avec un apriori complètement non informatif, c'est-à-dire qui ne privilégie *a priori* aucune valeur de l'effet traitement (comme dans le cadre fréquentiste où aucune hypothèse n'est faite sur l'effet du traitement). Dans cette situation le résultat *a posteriori* est entièrement conditionné par les données de l'étude.

Exemple d'une documentation du choix de l'apriori

The prior probability of outcome for each treatment group was assumed to follow a noninformative beta distribution, which yielded a beta distribution for the posterior probability when a binomial likelihood was assumed for the outcome [133]

Dans certaines situations, un apriori informatif sceptique (pessimiste, *skeptical* en anglais) peut être utilisé comme analyses de robustesse. Il s'agit d'un apriori qui « croit », qu' *a priori*, le traitement n'est pas efficace avec une assez forte certitude (distribution étroite). Si malgré cet *a priori*, un bénéfice est mis en évidence cela témoigne de données fortement en faveur de l'efficacité. Cette approche est surtout utilisée pour interpréter à but **exploratoire** des résultats de **découverte fortuite** (cf. par exemple [134]).

Il est aussi possible d'utiliser comme « *a priori* » les résultats d'une étude précédente. Cependant une étude fréquentiste ne permet pas d'estimer une distribution de l'effet traitement. Il est alors nécessaire de réanalyser cette étude initiale en Bayésien avec un apriori non informatif pour obtenir un résultat sous une forme utilisable comme « *a priori* » de la nouvelle étude. Finalement cette

opération revient à faire la méta-analyse des 2 études et reconnecte avec la problématique de l'acceptabilité des méta-analyses comme preuve de l'efficacité (cf. section 21)

13.1.4 Études de cas - exemples de présentation de résultats bayésiens

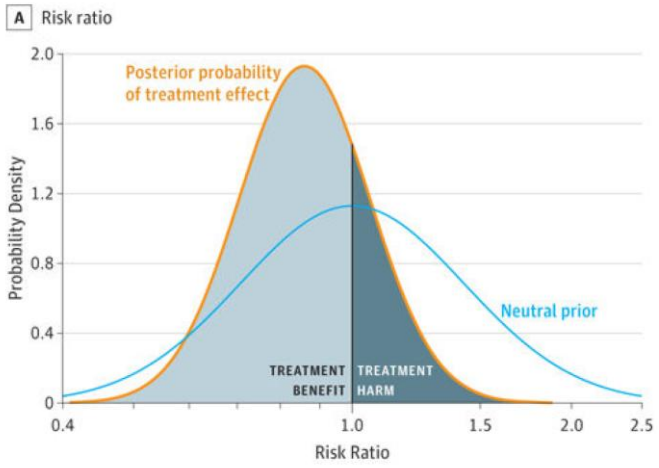
13.1.4.1 Exemple 1

Disease progression occurred in 77 patients (30.0%) in the convalescent-plasma group and in 81 patients (31.9%) in the placebo group (risk difference, 1.9 percentage points; 95% credible interval, −6.0 to 9.8; posterior probability of superiority of convalescent plasma, 0.68).

C3PO [133]

Outcome	Intention-to-Treat Population (N=511)			
	Convalescent Plasma (N=257)	Placebo (N=254)	Risk Difference (95% Credible Interval)‡ *percentage points*	Posterior Probability of Superiority of Convalescent Plasma
Patients with a disease-progression event — no. (%)	77 (30.0)	81 (31.9)	1.9 (−6.0 to 9.8)	0.68

13.1.4.2 Exemple 2 [135]



Outcome	No. (%)		Enthusiastic Prior (RR, 0.72)		Neutral Prior (RR, 1.0)		Skeptical Prior (RR, 1.10)	
	Hypothermia (n = 78)	Noncooled (n = 79)	aRR (95% Credible Interval)	P-TB, %	aRR (95% Credible Interval)	P-TB, %	aRR (95% Credible Interval)	P-TB, %
Primary Outcome								
Death or moderate-severe disability	19 (24.4)	22 (27.9)	0.78 (0.52-1.15)	90	0.86 (0.58-1.29)	76	0.89 (0.60-1.32)	73

Abbreviations: aRR, adjusted risk ratio; P-TB, posterior probability of treatment benefit (risk ratio <1.0); RR, risk ratio

In Bayesian analyses, the probability of treatment effect (posterior probability) is estimated after the trial and incorporates the prior probability estimated from the best data from previous studies (clinical trials or pilot trials). Judgment of the prior probability may vary and be neutral, enthusiastic, or skeptical. Therefore, analyses were performed using 3 different prior probabilities: (1) a neutral prior, assuming no treatment effect (RR, 1.0); (2) an enthusiastic prior, assuming a 28% reduction in the risk of death or disability as in the earlier Neonatal Research Network trial (RR, 0.72); and (3) a skeptical prior, assuming a 10% increase in the risk of death or disability (RR, 1.10). Whether neutral, enthusiastic, or skeptical, assessments of prior probability involve uncertainty about the minimum and maximum likely treatment effects. To reflect this uncertainty in each analysis, a probability distribution for the treatment effect with the 95% credible intervals that ranged from half to twice the assumed RR (SD, 0.35 in the log scale) was used. For example, the probability distribution for the neutral prior was centered at an RR of 1.0 (mean of 0 in the log scale) with a 50% prior probability of a better outcome, a 50% prior probability of a worse outcome, and a 95% credible interval for the RR of 0.5 to 2.0

13.2 Problématiques méthodologiques spécifiques des essais bayésiens

Problématique méthodologique spécifique (exposant à un risque de production de résultat favorable à tort au traitement étudié)	**Démonstration que doivent apporter les solutions à ces problématiques** (pour garantir la disparition du risque de conclure à tort)
Conditionnement du résultat par l'apriori utilisé (pouvant conduire à des résultats à l'opposé de l'observation)	Utilisation d'un apriori réellement non informatif (même si cela réduit l'attrait de ces études qui est potentiellement de pouvoir conclure avec moins de patients si utilisation d'un apriori informatif, cf. section 18)
Utilisation des résultats d'essais précédents comme apriori	Revient à décider sur une méta-analyse, idem à la situation où la méta-analyse est la seule preuve du bénéfice avec aucun essai concluant par lui-même Difficulté d'exprimer des résultats fréquentiste en distribution d'effet *a priori*
Choix arbitraire du seuil de probabilité *a posteriori* pour définir de « positivité » de l'essai	Sans utilisation d'un seuil standard, l'interprétation de la probabilité *a posteriori* est arbitraire et variera d'un essai à l'autre car la décision de conclure à l'intérêt du traitement s'effectuera alors de manière post hoc, en connaissant le résultat de l'étude. Aucun standard n'existe pour le moment, mais les pratiques (cf. exemples) utilisent 97.5% par analogie avec le risque alpha contrôlé en fréquentiste (même si le risque alpha n'a aucune relation directe avec la probabilité a posteriori)
Risque alpha et multiplicité non pris en compte	Définition du seuil de « positivité » de la probabilité a postériorité afin de contrôler le risque alpha global au niveau habituel (2.5% unilatéral)
Multiplicité (AI, critères, etc.)	Utilisation d'une méthode habituelle de gestion de la multiplicité des comparaisons inférentielles (pouvant conduire à conclure à l'intérêt du traitement) comme la hiérarchisation ou ajustement du seuil de « positivité » de la probabilité a postériorité afin de contrôlé le risque alpha global

AI : analyse intermédiaires

13.3 Méta-recherche

Aucune étude de méta-recherche n'a été trouvée. Le nombre d'essais bayésiens publiés jusqu'à présent (novembre 2021) semble restreint (recherche non exhaustive) [132, 133, 135, 136, 137, 138, 139, 140, 141, 142, 143]. Cette approche a été utilisée à plusieurs reprises pour des traitements de la COVID-19. Antérieurement la majorité des essais concernait des dispositifs médicaux.

La plupart de ces essais a utilisé des « a priori » non informatif et des seuils de probabilité *a posteriori* d'efficacité d'au moins 97.5%, avec fréquemment un ajustement pour la multiplicité :

Essai [ref]	**Choix du prior (dans matériel et méthodes)**
Early Convalescent Plasma for High-Risk Outpatients with COVID-19 [133]	The prior probability of outcome for each treatment group was assumed to follow a noninformative beta distribution, which yielded a beta distribution for the posterior probability when a binomial likelihood was assumed for the outcome.
Effect of Tocilizumab vs Usual Care in Adults Hospitalized With COVID-19 and Moderate or Severe Pneumonia [142]	For the day 4 outcome, we used a beta prior distribution with parameters 1 and 1 for the proportion in each arm (eFigure 1 in Supplement 2). For the day 14 outcome, we used a Gaussian prior distribution with a mean of 0 and variance of 106 for the log hazard ratio (HR) For the primary analyses, a non-informative flat prior distribution for the log HR was used, as a Gaussian distribution with mean 0 and variance 106
Effect of anakinra versus usual care in adults in hospital with COVID-19 and mild-to-moderate pneumonia (CORIMUNO-ANA-1) [138]	For the day 4 outcome, we used a β prior distribution with parameters 1 and 1 for the proportion in each treatment group. For the day 14 outcome, we used a Gaussian prior distribution with a mean log hazard ratio (HR) of 0 and variance of 1×10^6 for the log HR.
Interleukin-6 Receptor Antagonists in Critically Ill Patients with COVID-19 [141]	Prior distributions for individual treatment effects were neutral Pas plus de précision dans l'article
REMAP-CAP protocol [117]	REMAP-CAP launches with no prior assumptions regarding which interventions are superior, akin to a typical RCT design.
Azithromycin for community treatment of suspected COVID-19 in people at increased risk of an adverse clinical course in the UK (PRINCIPLE): a randomised, controlled, open-label, adaptive platform trial [132]	The log hazard ratio for treatment has the weak informative prior j N(0; 0:32); and is assumed to be constant over time. The weak informative prior for the log hazard ratio places the prior mass of the HR between 0.5 and 2.0, which in line with clinical expectations for potential therapies, and also will be quickly overwhelmed with accruing data.
Therapeutic Anticoagulation with Heparin in Noncritically Ill Patients with COVID-19 The ATTACC, ACTIV-4a, and REMAP-CAP Investigators [137]	The primary model incorporated weakly informative Dirichlet prior distributions for the number of days without organ support

13.4 Avis de la SFPT

Pour être recevable comme preuve du bénéfice clinique, il est attendu spécifiquement pour les essais bayésiens :

L'utilisation d'un « *a priori* » strictement non informatif (ou septique)

L'utilisation d'un seuil pour la probabilité *a posteriori* d'au moins 97.5%, idéalement recalibrée pour garantir le risque alpha global de la décision compte tenu de la multiplicité

Une gestion de la multiplicité au niveau des analyses intermédiaires, des critères de jugements, des doses, etc.

L’utilisation de l’inférence bayésienne dans les essais randomisés pivots permet de produire des résultats statistiques directement compréhensibles comme la probabilité que le traitement soit efficace.

Cette approche a longtemps été écartée pour les essais pivots en raison de la possibilité de conditionner le résultat par l’information arbitraire introduite *a priori*. L’utilisation d’un a apriori non informatif est donc nécessaire (cette approche peut aussi être utilisée pour faire de l’emprunt d’information, cf. section 18).

La conclusion à la démonstration du bénéfice clinique repose sur la comparaison de la probabilité *a posteriori* d'efficacité par rapport à un seuil qui doit avoir été prédéfini pour garantir un contrôle du risque alpha global (et qui ne peut pas être inférieur à 97.5%).

14 Les essais adaptatifs

Le terme de design adaptatif (ou essai adaptatif) recouvre un ensemble hétérogène de schémas expérimentaux ayant en commun, selon la définition de la FDA, la possibilité d'apporter des modifications planifiées de manière prospective à un ou plusieurs aspects de l'étude, sur la base de données accumulées au cours de l'essai [144]. Le caractère adaptatif d'un essai peut ainsi concerner des objectifs variés [144, 145, 146], parmi lesquels :

- Adapter le nombre total de patients à inclure dans l'essai : une vérification des hypothèses sur lesquelles se basait le calcul d'effectif initial après avoir inclus une partie des patients permet de réévaluer (en aveugle ou non) l'effectif en cours d'étude ou évaluer sa puissance conditionnelle, qui est la probabilité que le résultat final de l'étude soit statistiquement significatif, compte-tenu des données déjà observées [147].
- Prendre une décision sur la poursuite ou non de l'essai : les essais séquentiels en groupes (*group sequential designs*) prévoient différentes analyses intermédiaires ainsi que des critères d'arrêt définis *a priori*. Ces règles d'arrêt offrent la possibilité de conclure précocement à l'efficacité ou à la futilité (c'est-à-dire une faible probabilité que le traitement soit efficace à la fin de l'essai) du traitement.
- Adapter la population de l'essai : les stratégies dites « d'enrichissement » (*adaptive enrichment*) permettent de définir des sous-groupes de patients répondant le plus au traitement, sur la base d'une ou plusieurs analyses intermédiaires, et de focaliser sur la poursuite de l'essai sur ces sous-groupes.

Le terme « d'essai adaptatif » recouvre également d'autres approches, que nous ne détaillerons pas dans ce chapitre :

- Adapter le traitement ou la dose : la modification de la dose est de longue date réalisée dans les essais de phases précoces, selon une démarche algorithmique (par exemple 3+3). Une approche adaptative consiste à modéliser la probabilité de survenue d'un événement. C'est par exemple le cas des méthodes CRM (*Continual Reassessment Methods*) qui utilisent les données accumulées dans l'essai pour déterminer la dose administrée au patient suivant, selon une approche bayésienne.

- Introduire ou arrêter de nouveaux bras de traitement : dans certains cas, plusieurs doses sont testées en parallèles, et certaines peuvent être abandonnées en cours d'essais à la suite d'analyses intermédiaires, selon un design séquentiel en groupes. Les essais plateformes (cf chapitre 12) représentent un autre exemple d'essais adaptatifs : plusieurs traitements sont comparés à un même contrôle, qui représente le standard de prise en charge ; certains bras peuvent être introduits ou arrêtés, selon des règles prédéfinies.

- Adapter l'allocation des patients : c'est le cas par exemple des méthodes de minimisation, qui visent à équilibrer les groupes sur certaines caractéristiques lors de la randomisation, en se basant sur les caractéristiques des patients déjà inclus. D'autres approches, plus controversées, consistent à adapter la randomisation en fonction des résultats intermédiaires accumulés au cours de l'étude [148].

- Adapter la dose ou le choix du traitement qui sera évalué dans une phase ultérieure du développement, dans le cadre d'essais combinés (également appelés « sans couture », *seamless*). Ces essais peuvent regrouper des études de phase 1 et de phase 2, ou de phase 2 et de phase 3 (cf chapitre 15).

Les designs séquentiels en groupes et la réévaluation du nombre de sujets nécessaires en cours d'essai sont actuellement les formes les plus courantes d'essais adaptatifs. Ils présentent de nombreux avantages [144], que nous détaillerons dans ce chapitre :

- Sur le plan statistique, ils maximisent les chances de démontrer l'efficacité du traitement (donc la puissance), ou à puissance équivalente minimisent le nombre de patients inclus.
- Sur le plan éthique, ils permettent d'arrêter précocement un essai qui aura trop peu de chances de démontrer une supériorité du traitement (futilité), évitant ainsi d'exposer inutilement des patients à un risque d'effet indésirable. A l'inverse, un arrêt précoce pour supériorité permet de proposer à tous les patients le traitement le plus efficace.
- Leur flexibilité est un atout pour les promoteurs et les financeurs des essais, les arrêts précoces étant synonymes d'essais moins longs et moins couteux.

14.1 Problématiques méthodologiques

La flexibilité des essais adaptatifs ne doit pas compromettre la validité et l'intégrité des résultats. En effet, les adaptations réalisées en cours d'étude ne peuvent pas être déterminées arbitrairement, mais en fonction de règles et de méthodes préétablies. Par définition, **un essai adaptatif prévoit la possibilité et les modalités d'adaptation *a priori*** ; il ne s'agit pas d'une modification substantielle du protocole apportée en cours d'étude. Il est donc essentiel de fournir suffisamment de détails sur le type d'adaptation et les méthodes utilisées (types d'analyses intermédiaires, définition des règles d'arrêt, méthodes statistiques utilisées) avant le début de l'étude.

Les essais adaptatifs s'appuyant sur l'utilisation de données accumulées au cours de l'étude elle-même, ils impliquent fréquemment des analyses intermédiaires. L'une des principales problématiques méthodologiques qui en découle est la **gestion de la multiplicité des comparaisons**, avec l'augmentation du risque de conclure à tort à l'intérêt du traitement (erreur de type I). Dans les essais séquentiels en groupes, qui comportent une ou plusieurs analyses intermédiaires, si le seuil de risque de 5% classiquement retenu est utilisé pour chaque analyse, le risque global dépassera ce seuil. Le seuil de significativité retenu pour chaque analyse doit ainsi être ajusté pour préserver un risque global acceptable. Différentes approches ont été proposées, plus ou moins conservatrices sur les probabilités d'arrêt précoce du traitement :

- La méthode de Pocock utilise un seuil constant à chaque analyse. Elle expose à un risque élevé d'arrêt prématuré.
- La méthode de O'Brien et Flemming est couramment utilisée car elle a un impact faible sur l'effectif global. En revanche la probabilité d'arrêt précoce est plus faible, ce qui dans certaines situations (comme les essais plateformes, où plusieurs traitements sont en concurrence), peut être jugé trop conservateur.
- La méthode de Lan et DeMets généralise cette approche en utilisant une fonction de consommation du risque alpha au cours du temps. Elle est plus flexible et permet des analyses « flottantes », basée sur une fraction d'information non définie *a priori*. Ainsi, ni le nombre d'analyses intermédiaires ni le moment de leur réalisation nécessitent d'être fixés à l'avance. Elle s'affranchit ainsi des contraintes logistiques liées au gel partiel de la base de données à un instant précis.

Notons que l'utilisation de ces seuils peut avoir pour finalité d'évaluer la supériorité et/ou la futilité. Le gain en termes de flexibilité (en l'occurrence la possibilité d'arrêter l'essai plus précocement) se fait

au détriment d'une légère augmentation du nombre total de patients à inclure, qui s'explique par la multiplicité des comparaisons. Dans le cas des règles d'arrêt pour futilité, on distingue deux situations : en l'absence d'engagement sur l'arrêt de l'essai (*nonbinding stopping rules*) le comité de pilotage de l'essai peut décider de poursuivre l'étude même si les critères d'arrêt pour futilité sont satisfaits. Cette approche est plus flexible que les règles avec engagement d'arrêt (*binding*), au détriment là encore d'une légère augmentation de l'effectif total.

Pour les designs d'enrichissement, qui peuvent se baser sur l'utilisation d'un biomarqueur, il est impératif que celui-ci soit déterminé a priori, ou si c'est en cours d'essai que son utilisation ne soit pas motivée par les données de l'essai [146].

Notons que certaines méthodes adaptatives sont non comparatives, elles n'auront donc pas d'impact sur le risque d'erreur de type I. Il s'agit par exemple de la ré-estimation du calcul du nombre de sujets basée sur la variance du critère de jugement principal recueilli dans l'étude, sans levée d'aveugle, ou encore la valeur pronostique d'un biomarqueur pour les designs d'enrichissement.

Un autre aspect essentiel des essais adaptatifs est de **maintenir l'intégrité de l'essai, alors que l'accès à une partie de l'information est nécessaire** pour effectuer les adaptations prévues. Aux problématiques méthodologiques énoncées ci-dessus s'ajoutent donc des contraintes logistiques visant à garantir :

- L'obtention de données de bonne qualité pour la réalisation des analyses intermédiaires, au moment où celles-ci sont prévues.
- Une restriction de l'accès aux résultats intermédiaires comparatifs (particulièrement aux personnes directement impliquées dans la conduite de l'essai).

Les moyens mis en œuvre peuvent parfois être complexes et ajouter des coûts à l'essai.

14.2 Etudes de cas

L'essai dit PARADIGM HF [149] est une étude contrôlée, randomisée, en double-aveugle, comparant l'association sacubitril + valsartan à l'enalapril chez des patients ayant une insuffisance cardiaque à fraction d'éjection diminuée. Le critère de jugement principale était un composite de décès de cause cardiovasculaire ou d'hospitalisations pour insuffisance cardiaque. Le protocole prévoyait initialement deux analyses intermédiaires :

> The O'Brien-Fleming type of boundary with Lan-DeMets alpha spending function will be used for the interim efficacy analyses to assess superiority. As currently planned, two interim efficacy analyses are to be expected approximately at 1/3 and 2/3 of information time (i.e. approximately 803 and 1607 patients, respectively, with a primary events of CV mortality or HF hospitalization). The interim efficacy analysis with the boundary will spend approximately an alpha of 0.0001 (one-sided) at the first interim analysis and 0.00605 (one-sided) at the second interim analysis. The actual alpha to be spent for the interim efficacy analyses will be precisely determined based on the Lan-DeMets alpha spending function using the actual number of patients who have experienced a primary events at the interim efficacy analyses.

Finalement une troisième analyse intermédiaire a été ajoutée à la demande du DMSB, en adaptant les règles d'arrêt (ce qui est permis en utilisant la méthode flexible de Lan et DeMets). Une p-value nominale <0.001 ayant été obtenue lors de la 3ème analyse intermédiaire, l'essai a été arrêté

précocement pour supériorité, bien que le recrutement ait été terminé au moment de cette analyse [150].

Mi-2020 un essai contrôlé, randomisé, en double-aveugle, contre placebo, a évalué l'efficacité de l'hydroxychloroquine par rapport à un placebo dans la prévention de la COVID-19, chez 821 adultes asymptomatiques ayant été en contact avec une personne contaminée. Le protocole prévoyait au total trois analyses intermédiaires. L'incidence d'une maladie symptomatique suite à une exposition étant encore peu connue au début de l'essai, une analyse intermédiaire ayant pour objectif un re-calcul d'effectif était prévue au protocole [151] :

> Because the estimates for both incident symptomatic COVID-19 after an exposure and loss to follow-up were relatively unknown in early March 2020, the protocol prespecified a sample-size reestimation at the second interim analysis. This reestimation, which used the incidence of new infections in the placebo group and the observed percentage of participants lost to follow-up, was aimed at maintaining the ability to detect an effect size of a 50% relative reduction in new symptomatic infections.

Cette analyse intermédiaire a permis de réduite l'effectif total de l'étude, compte-tenu d'une survenue d'infections symptomatiques dans le groupe contrôle plus importante qu'initialement prévue.

Par ailleurs, la puissance conditionnelle, qui est la probabilité que le résultat final de l'étude soit statistiquement significatif, compte tenu des données observées jusqu'alors [147], était estimée à chacune des analyses intermédiaires. Cette approche a permis l'arrêt prématuré de l'essai pour futilité à la troisième analyse, la puissance conditionnelle étant alors un inférieure à 1% [151].

14.3 Méta-recherche

Une étude de méta-recherche conduite dans le domaine de la sclérose latérale amyotrophique montre que parmi l'ensemble des essais contrôlés randomisés, peu adoptent un schéma séquentiel en groupes. Pourtant, cette approche permet fréquemment de réduire la durée des études [152].

En revanche, une étude de coûts assez récente confirme que les différents designs adaptatifs sont plus consommateurs de ressources que les designs fixes traditionnels [153].

14.4 Avis de la SFPT

Par définition, un essai adaptatif prévoit *a priori*, dans le protocole, la possibilité de modifier certains aspects de l'essai selon les données accumulées. La méthodologie est ainsi adaptée pour prendre en compte les éventuels risques associés à ces modifications. Les essais adaptatifs regroupent un ensemble hétérogène de designs. Les designs séquentiels en groupes et la réévaluation du nombre de sujets nécessaires en cours d'essai sont des approches robustes, désormais courantes, **bien acceptées par les autorités et les agences**. Toutefois nous notons trois points essentiels :

Le type d'adaptation et les méthodes utilisées doivent être définis avant le début de l'étude, ou si s'ils sont ajoutés en cours d'étude, que ce soit indépendamment des résultats obtenus.

Sur le plan méthodologique, les analyses multiples augmentent le risque de conclure à tort à l'efficacité d'un nouveau traitement ; elles nécessitent donc la **mise en œuvre de méthodes adaptées pour prendre en compte cette inflation du risque d'erreur de type I**.

L’accès aux données accumulées lors de l’essai pour réaliser ces analyses ajoute des contraintes logistiques, avec une augmentation des coûts associés. Par ailleurs, pour respecter l’intégrité et la validité de l’essai, **un DSMB avec des membres expérimentés sur la gestion de ces designs est indispensable** [146].

15 Les essais combinés (« sans couture », *seamless*)

Les essais combinés (*seamless*, « sans couture ») sont un type de schéma expérimental adaptatif (cf chapitre 14) où le même « essai » combine un essai de phase 1 et un essai de phase 2, ou encore une phase 2 et une phase 3. La partie phase 2 servira par exemple à déterminer la dose optimale qui sera ensuite utilisée dans la phase 3 de l'étude destinée à montrer l'efficacité et la sécurité du médicament. Les premiers patients inclus pour la partie phase 2 contribueront aussi à la partie phase 3 s'ils ont reçu la dose retenue. Si un problème de sécurité survient à une dose donnée lors de la phase 2, celle-ci sera arrêtée. Dans certains cas (cf exemples ci-dessous), les approches *seamless* associent les trois phases du développement clinique.

Ces designs évitent ainsi les temps morts présents dans l'approche traditionnelle entre les différentes phases, et permettent de limiter le nombre total de patients.

15.1 Problématiques méthodologiques

Les problématiques soulevées par les essais combinés se recoupent largement avec celles des essais adaptatifs en général (cf chapitre 14) : pré-spécification des objectifs, des méthodes et du plan d'analyse ; gestion de la multiplicité des comparaisons ; et maintien de l'intégrité de l'essai tout au long du programme de recherche.

Le premier de ces trois points fondamentaux a notamment été rappelé par un groupe de travail du *National Cancer Institute*, à la suite de certaines dérives. En effet, dans le domaine de l'oncologie, les approches *seamless* phases 1-2 se sont largement développées dans les années 2010, en ajoutant notamment des cohortes d'expansion à des essais de phase 1 de type *first in human*. Pour que ces approches restent valides, la définition précise du design de l'étude, les caractéristiques des différentes cohortes d'expansion et les méthodes statistiques mises en œuvre doivent être décrites **avant le début de l'essai** [154].

15.2 Étude de cas

L'essai STAMPEDE est l'un des premiers exemples d'essai « plateforme » (cf chapitre 12), encore appelés multi-bras multi-étapes, car ils combinent en effet différentes phases dans une approche *seamless*. La figure ci-dessous détaille les 5 étapes de l'essai : l'étape pilote a pour objectif principal la sécurité ; elle est suivie de trois étapes d'évaluation de l'efficacité, séparées par des analyses intermédiaires, avec pour critère de jugement principal la survie sans progression. Puis la dernière étape est la phase confirmatoire, qui a pour critère principal la survie globale. Ainsi la survie globale des patients inclus lors des premières phases de l'essai sera utilisée pour la phase confirmatoire, sous réserve que ces bras de traitement n'aient pas été interrompus pour des raisons de sécurité ou de manque d'efficacité (futilité) suite aux différentes analyses intermédiaires.

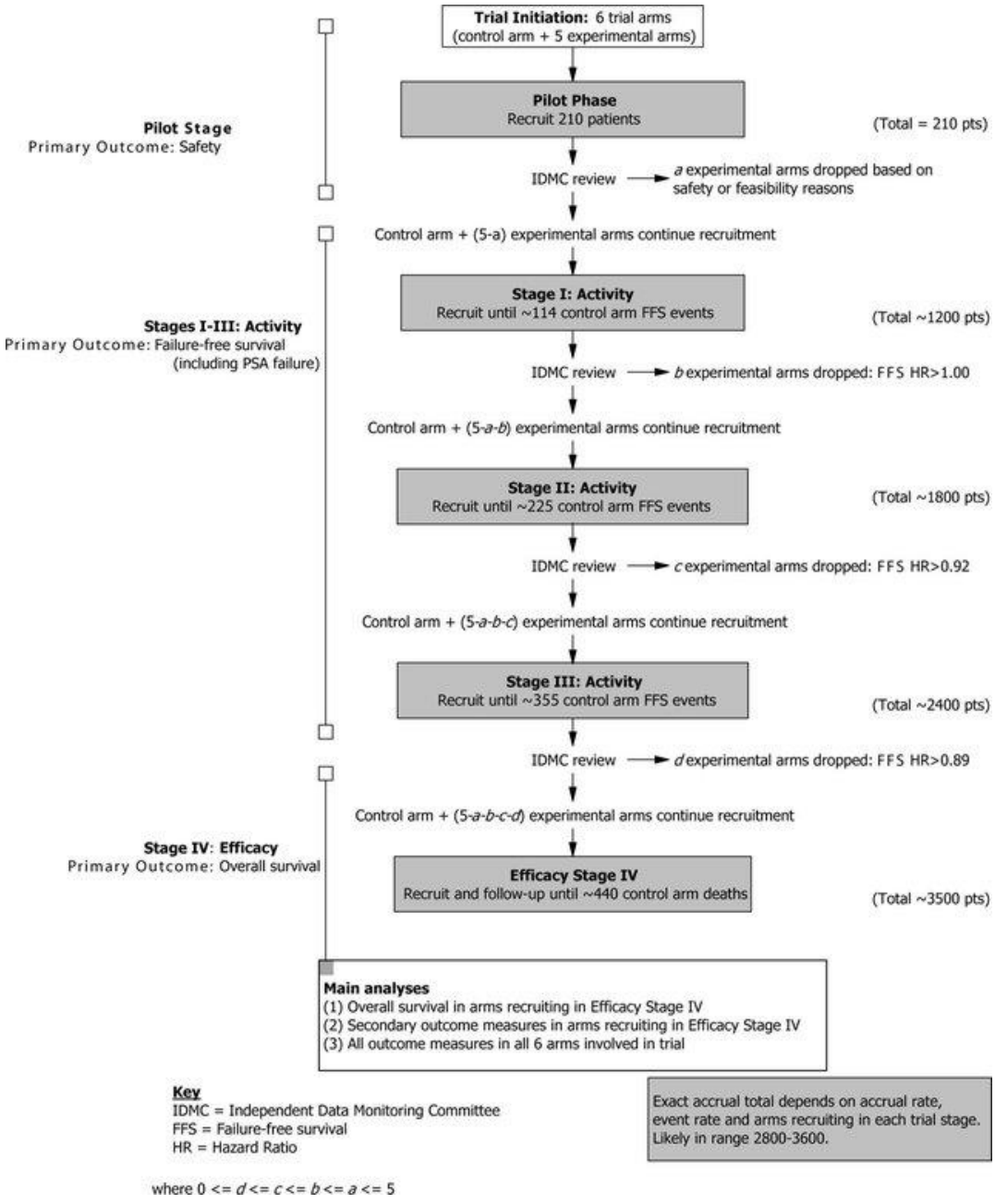


Schéma des différentes étapes de l'essai STAMPEDE, évaluant différents traitements dans le cancer de la prostate [126]

Plus récemment, l'intérêt de cette approche a été parfaitement illustré par l'essai vaccinal du BNT162b2 (Comirnaty®) dans la COVID-19. Devant l'urgence, un seul « essai » a été mis en place pour servir de phase 1, 2 et 3 (NCT04368728). Les premiers patients inclus ont servi de phase 1 [155] puis les résultats de phase 2 sur la réactogénicité et la tolérance ont été publiés [156] avant ceux concernant l'efficacité clinique sur la prévention des maladies symptomatiques et des formes sévères [157, 158].

15.3 Méta-recherche

Une étude a recensé tous les abstracts présentés au congrès annuel de l'*American Society of Clinical Oncology* entre 2010 et 2017, et concernant des essais de phase précoce. Parmi 1786 essais identifiés, seulement 3% étaient des essais combinés, mais ils incluaient environ 15% des patients. Le nombre médian de cohortes d'expansion était de 3 [159].

Les récents essais combinés mis en œuvre lors de la pandémie de COVID-19 ont montré à bien des égards leur efficience, en accélérant le développement de et en rationalisant le nombre de patients inclus.

15.4 Avis de la SFPT

Au niveau méthodologique les essais combinés apportent toutes les garanties nécessaires lorsqu'ils répondent à un certain nombre de critères, communs à tous les essais adaptatifs (cf chapitre 14) :

- Définition *a priori* des objectifs, du design de l'essai, des méthodes d'analyse

- Contrôle approprié du risque d'erreur de type I

- Maintien de l'intégrité de l'essai en cours d'étude

Lorsque ces conditions sont remplies, cette approche devrait être encouragée compte-tenu de ses avantages, y compris pour les essais académiques. Dans ce cas dernier, ils évitent que ces essais ne s'arrêtent à la phase 2 en raison de difficulté à trouver les financements complémentaires, et entrainant des prises de décision sur la base de ces données préliminaires donnant des résultats de faible degré de certitude. Cette approche permettrait d'apporter un élément de solution au gaspillage des ressources de recherche [160, 161, 162, 163, 164, 165, 166].

16 Études mono-bras (non comparative)

Les études mono-bras (single-arm study) consistent en une série prospective de patients recevant tous le traitement évalué. Elles donnent uniquement la valeur du critère de jugement sous traitement, « dans l'absolu » par exemple la valeur d'un taux de succès, de réponse, de survie, ou la valeur d'un paramètre biologique (« 42% de décès à 12 mois » par exemple). Ces études ne permettent donc pas d'apporter directement la démonstration du bénéfice d'un nouveau traitement. Pourtant, de plus en plus d'études mono-bras sont proposées comme seule étude « pivot » pour l'enregistrement, le remboursement ou la modification des pratiques. [36] [167].

16.1 Problématiques méthodologiques

Les études mono-bras n'apportent, en fait, que la moitié de l'information nécessaire pour déterminer l'effet d'un traitement : une valeur dans « l'absolu » du critère de jugement chez les patients traités. Suivant les termes de l'épidémiologie moderne [168], il manque le contrefait (« counterfact ») qui permettrait de déterminer l'effet du traitement. La valeur sous traitement obtenue dans l'étude mono-bras est le « fait ». Pour vérifier s'il y a bien un effet du traitement, il faut le « contrefait ». Dans un essai randomisé, le « fait » est apporté par le groupe actif expérimental et le « contrefait » est apporté par le groupe contrôle expérimental (contrôle interne).

Pour déduire si cette valeur représente un bénéfice du traitement, il est nécessaire de procéder à une comparaison pour montrer que cette valeur est différente de celle qui aurait été obtenue chez ces patients, sans traitement (avec le traitement standard ou sous placebo). Dans le cas des essais mono-bras, cette comparaison ne peut s'effectuer qu'avec des données externes à l'étude (contrôle historique par exemple). Cette comparaison externe est l'élément essentiel, à tel point que la littérature dans le domaine, comme ICH[19], n'aborde pas cette problématique sous l'angle des études mono-bras, mais bien sous celui de la comparaison externe (*externally controlled trial*) [169].

Problématique méthodologique spécifique (Exposant à un risque de production de résultat favorable à tort au traitement étudié)	**Démonstration que doivent apporter les solutions à ces problématiques (pour garantir la disparition du risque de conclure à tort)**
Impossibilité de raisonnement contrefactuel, impossibilité de déterminer l'effet du traitement	Utilisation d'un groupe contrôle externe avec une comparaison indirecte non ancrée (pour prendre en compte les biais de confusion et autres). Réalisation d'une étude comparative contrôle externe (cf. section suivante)

Les études mono-bras ne permettent de conclure à l'intérêt du traitement évalué à elles seules que dans des situations exceptionnelles dites 0%/100% correspondant à l'observation confirmée d'évolutions favorables avec le nouveau traitement dans une pathologie où l'évolution serait péjorative dans 100% des cas (comme par exemple le traitement insulinique dans le diabète insulino-dépendant). Ce type de situation est exceptionnel[20] [170]. Même dans le cas de l'infection par le virus

[19] ICH E10, section 2.5 page 24.

[20] Par exemple même dans le cas de l'infection par le virus Ebola le pronostic, bien que très péjoratif, n'est pas suffisamment sombre avec l'existence de patient qui survive spontanément pour rentrer dans la règle des 100%/0%. Un essai randomisé a donc été réalisé pour évaluer l'efficacité clinique du ZMapp (7. Group PIW, Multi-National PIIST, Davey

Ebola par exemple, le pronostic, bien que très péjoratif, n'est pas suffisamment sombre pour rentrer dans la règle des 0%/100% avec l'existence de patients qui survivent spontanément. Un essai randomisé a donc été réalisé pour évaluer l'efficacité clinique du ZMapp [5].

16.2 Études de cas

De nombreux retours d'expérience donnent des exemples de traitements considérés comme apportant un bénéfice avec une comparaison non formalisée des résultats d'une étude mono-bras par rapport à une référence subjective ou implicite, mais pour lesquels ce bénéfice n'a pas été mis en évidence ultérieurement lorsque d'un ECR a été réalisé :

Traitement	Étude mono-bras	Phase 3 non concluante
Epacadostat par-dessus pembrolizumab dans le mélanome avancé	ECHO-202/KEYNOTE-037 (Taux de réponse 55%)	PFS et OS négatives ECHO-301/KEYNOTE-252 (taux de réponse groupe traité 34%)
Pembrolizumab monothérapie dans le cancer de la vessie	KEYNOTE-052[21]	KEYNOTE-361 arrêté prématurément pour surmortalité[22]
Pembrolizumab dans le carcinome hépatocellulaire	KEYNOTE-224	KEYNOTE-240 non conclusive[23]
L'atezolizumab en première ligne du cancer de la vessie métastatique	IMvigor 210	IMvigor 211 non concluante
Remdesivir dans la COVID-19	SIMPLE [171]	RECOVERY, DISCOVERY

16.3 Méta-recherche

La réalisation d'étude mono-bras est souvent justifiée par l'impossibilité de réaliser un essai comparatif randomisé en raison d'un nombre trop faible de patients. En oncologie, Rittberg et al. a montré que pour 31 enregistrements récents basés sur une étude mono-bras, un essai randomisé aurait pu réalisable dans plus de 80% des cas [172].

16.4 Avis de la SFPT

Les études mono-bras ne permettent pas d'évaluer le bénéfice clinique en raison de l'absence de raisonnement contrefactuel.

Dans les rares situations où ces études seraient proposées, elles doivent obligatoirement prendre la forme d'études comparatives à contrôles externes

RT, Jr., Dodd L, Proschan MA, Neaton J, et al. A Randomized, Controlled Trial of ZMapp for Ebola Virus Infection. N Engl J Med. 2016;375(15):1448-56.).

21 10.1016/S1470-2045(17)30616-2

22 https://www.esmo.org/Oncology-News/Patients-with-mUC-and-Low-PD-L1-Expression-May-Have-Decreased-Survival-When-Treated-with-Pembrolizumab-or-Atezolizumab

23 https://www.mrknewsroom.com/news-release/oncology/merck-provides-update-keynote-240-phase-3-study-keytruda-pembrolizumab-previou

17 Études à contrôle externe (groupes contrôles synthétiques)

Les études à contrôle externe sont des études qui formalisent la comparaison des résultats d'une étude mono-bras avec celle d'un contrôle externe (contrôle historique par exemple). Ces études peuvent être prévues *a priori*, au moment de la planification de l'étude mono-bras, mais le plus souvent, elles sont réalisées a posteriori après la disponibilité des résultats de l'étude mono-bras.

> *Dans les études mono-bras (cf. section 16), en l'absence de point de comparaison apporté par un groupe contrôle, il est impossible de savoir si la valeur du critère de jugement observée sous traitement est meilleure que celle qui aurait été obtenue sans ce traitement dans la même étude (mêmes patients, même mesure du critère de jugement, même suivi, même contexte de soins). L'exploitation des études mono-bras pour démontrer le bénéfice d'un nouveau traitement nécessite donc une comparaison implicite qui fera obligatoirement appel à une référence de comparaison externe (comme une comparaison historique par exemple).*

Plusieurs types de comparateurs externes sont envisageables (cf. Tableau 5).

Tableau 5 – Types de comparateurs externes possibles

- Norme, valeur de référence qui peut être soit fixée par une « exigence règlementaire », soit issue de la littérature (étude de cohorte par exemple) ou d'une conférence de consensus
- Cohorte disponible sous la forme d'une publication (données individuelles inaccessibles) ou publication d'un essai randomisé sont un des groupes peut servir de comparaison externe qui pourra servir de base à une MAIC ou d'une autre méthode de ce type
- Cohorte avec accès aux données individuelles qui pourra servir à une approche type étude observationnelle ou à la constitution d'un groupe contrôle synthétique
- Données individuelles d'un essai randomisé qui pourra servir à une approche type étude observationnelle
- Évolution des mêmes patients avant le traitement (design autocontrôlé de type avant/ après)
- Emprunt de données (voire section 18)

Les comparaisons externes sont souvent appelées historiques car les groupes contrôles servant à la comparaison sont non contemporains de l'étude monobras.

17.1 Problématiques méthodologiques spécifiques et solutions possibles

Les comparaisons externes présentent plusieurs limites méthodologiques fortes.

17.1.1 Comparaison post hoc

La première est que, le plus souvent, le choix de la référence de comparaison s'effectue de manière post hoc (non prévue au protocole, souvent après que les résultats de l'étude monobras soient connues) et que rien ne permet de garantir que ce choix n'ait pas été fait pour favoriser le nouveau traitement. La solution à cette limite est simple. Ces études ne doivent plus être conçues comme des essais mono-bras descriptif, mais comme de véritables essais comparatifs dont le groupe contrôle n'est

pas interne et contemporain, issu d'une randomisation, mais externe (Tableau 6). De ce fait l'objectif de l'étude est bien celui de montrer l'intérêt clinique du nouveau traitement (et non pas de décrire l'évolution des patients avec le traitement). Le choix de la référence de comparaison est fixé au protocole ainsi que la méthode d'analyse, empêchant tous choix post hoc. Cette approche est avalisée dans la nouvelle version de ICH E10 qui propose le terme d'essai à contrôle externe (*externally controlled trial*) [173].

La détermination *a priori* de la référence de comparaison au moment de la construction du protocole de l'étude permet aussi une meilleure prévention des biais en utilisant, par exemple, le même critère de jugement et la méthode de sa mesure (cf. infra).

Tableau 6 – Apport d'une véritable étude comparative à contrôle externe par rapport aux études mono-bras telles que réalisées actuellement.

	Étude mono-bras (telle que réalisée actuellement)	**Étude comparative à contrôle externe**
Objectif	Décrire le devenir des patients avec le nouveau traitement	Montrer la supériorité du nouveau traitement
Détermination du bénéfice du traitement	En post hoc, par une analyse complémentaire souvent mise en place après l'obtention des résultats	Objectif de l'essai défini a priori dans le protocole
Choix de la référence de comparaison	Post hoc, potentiellement influencée par les résultats observée de l'étude	Fixée *a priori*, prévue par le protocole

17.1.2 Biais de confusion

L'autre limite majeure de la comparaison à un contrôle externe est le biais de confusion : rien ne garantit que les groupes ou les résultats comparés sont réellement comparable.

Plusieurs méthodes de formalisation des comparaisons externes (aussi appelée parfois comparaison indirecte, *unanchored indirect comparison*) ont été proposées afin de tenter de fiabiliser les conclusions tirées des études à comparateur externe. Ces approches ont pour objectif de suppléer l'absence de contrôle par design (plan expérimental) des facteurs de confusion par leur prise en compte dans l'analyse statistique (ajustement). Les méthodes permettent en théorie de supprimer le biais de confusion à condition que tous les facteurs de confusion puissent être pris en compte (facteurs conditionnant le critère de jugement et ayant une distribution différente entre la groupe traitée et le comparateur externe, y compris les modificateurs de l'effet du traitement). Pour cela les facteurs de confusion potentiels de chaque critère de jugement doivent être préalablement identifiés par une revue systématique des études pronostiques. L'ajustement statistique réalisé (quelle que soit la méthode) devra prendre en compte tous ces facteurs. Pour confirmer que ces ajustements ont permis de supprimer le biais de confusion, l'importance du biais de confusion résiduel doit être appréciée par l'utilisation de contrôles négatifs comme variables de falsification et par une analyse quantitative du biais (« *bias analysis* ») [8]. Compte tenu des difficultés rencontrées pour apporter ces garanties, les résultats ne seront convaincants qu'en cas de résultats montrant un large effet, non explicables par les biais et la confusion résiduelle.

Le choix *a priori* de la référence de comparaison au moment de la construction du protocole contribuera aussi à la minimisation du biais de confusion en permettant d'aligner les critères d'éligibilité avec ceux utilisés pour la construction du contrôle externe (approche par restriction).

Cette limite du biais de confusion fait qu'au niveau règlementaire, ICH E10 n'envisage le recours aux essais à contrôles externes que dans des situations très particulières où l'effet du traitement est extrêmement important et le cours de la maladie hautement prévisible (« *Use of the external control design is restricted to situations in which the effect of treatment is dramatic and the usual course of the disease highly predictable* ») [173].

17.1.3 Autres biais

D'autres limites existent en termes de biais de mesure, de sélection et d'attrition. Le Tableau 2 résume ces limites et les solutions envisageables.

17.1.4 Pertinence clinique

L'étude à comparateur externe doit aussi assurer la pertinence clinique de la comparaison effectuée en termes de : critères de jugement principaux pertinents, comparateur loyal représentant le traitement recommandé « standard of care » en vigueur au moment de la prise de décision, durée de suivi, évaluation de la sécurité. Sur ces points, l'étude doit être comparable à ce qu'aurait été un essai randomisé pivot.

17.2 Étude de cas

L'atezolizumab a bénéficié d'un enregistrement accéléré dans le traitement de première ligne du cancer de la vessie métastatique de patients non éligibles à un traitement par platine sur la base de l'étude mono-bras IMvigor 210 [174] qui montrait un taux de réponse de 23%, statistiquement significatif par rapport à une valeur de référence de 10% fixée au protocole. La publication ne donne pas de justification factuelle de la norme, qui est présentée comme étant le taux de réponse attendue chez ces patients compte tenu des traitements utilisés.

	Patients	Complete response	Partial response	Objective response, n (% [95% CI])*	Median duration of response (95% CI)
	119	11	16	27 (23% [16–31])	NE (14·1–NE)
IC2/3	32	4	5	9 (28% [14–47])	NE (11·1–NE)
IC1/2/3	80	8	11	19 (24% [15–35])	NE (NE–NE)
IC1	48	4	6	10 (21% [10–35])	NE (NE–NE)
IC0	39	3	5	8 (21% [9–36])	NE (12·8–NE)

Data cutoff was July 4, 2016. PD-L1=programmed death-ligand 1. IC=tumour-infiltrating immune cell. NE=not estimable. *Includes objective response rate per Response Evaluation Criteria in Solid Tumors version 1.1 (independent review facility).

Table 2: **Objective response by PD-L1 status on tumour-infiltrating immune cells**

Cet enregistrement accéléré était conditionné à la réalisation de la phase 3 comparative versus chimiothérapie (IMvigor 211) [175]. Cet essai a été négatif sur son critère de jugement principal, ne montrant pas de bénéfice sur la survie globale (HR 0.87 [0.63 ; 1.21]).

De plus, le taux de réponse dans le groupe chimiothérapie a été de 21.6%, très proche des 23% du critère de jugement principal de l'étude mono-bras IMvigor210, montrant bien, a posteriori, le caractère arbitraire et subjectif de fixer le taux de réponse attendue sous chimiothérapie à 10%.

Cet exemple illustre la difficulté de déterminer la valeur contrefactuelle (« *conterfactual value* ») du critère de jugement, c'est-à-dire ce qui aurait été observé, chez ces patients, sans le nouveau traitement (et qui aurait donné par le groupe contrôle d'un essai comparatif). Lorsqu'une valeur est précisée (comme ici 10%), cette valeur peut être arbitraire, basée sur un avis d'expert qui est une évaluation subjective de cette valeur contrefactuelle. Compte des enjeux de ces études la crainte est qu'une valeur particulière favorable au traitement étudié soit choisie (comme ce fût le cas dans cette étude). Le but des essais à contrôle externe de bonne méthodologie est d'éviter de tels choix arbitraires.

On peut aussi noter que dans cette étude le critère de jugement était de faible pertinence clinique et ne correspondait pas au critère attendu, l'OS, (qui a bien était pris en comme critère de jugement principal dans l'essai subséquent de phase 3). Ces études conduisent donc à ce que la prise de décision d'utiliser le nouveau traitement s'effectue non seulement en se basant sur une méthodologie très dégradée, mais aussi sur des données sans réelle pertinence clinique. Sans formalisation par une vraie comparaison externe l'utilisation isolée de ces études mono-bras s'apparente à un déni complet d'évaluation.

17.3 Solutions possibles

Les solutions envisageables à ces problématiques passent par la mise en œuvre des éléments suivants.

17.3.1 La comparaison externe doit être formalisée

La comparaison externe doit avoir été planifiée *a priori*, en même temps que l'étude mono-bras. L'étude devait avoir pour objectif de déterminer le bénéfice du traitement par comparaison externe et non pas simplement de décrire le devenir de quelques patients tous traités.

La méthode de comparaison externe doit être clairement définie dans les méthodes : comparaison à une norme, comparaison « ajustée » par rapport à une cohorte de référence (*Matching-adjusted indirect comparisons* (MAIC) [176, 177], « ajustement » traditionnel), simulation du taux de référence (*Simulated Treatment Comparison*), etc.). En d'autres termes, une étude non comparative purement descriptive[24] ne sert à rien pour l'évaluation du bénéfice d'un traitement. Pour être exploitable dans ce but, l'étude doit être une étude de comparaison externe et doit donc le prévoir explicitement dans son objectif et sa méthode. Ainsi ICH E10 (version de juillet 2000[25]) ne mentionne que les études à comparateur externe (section 2.5) et ne parle pas d'études mono-bras, non comparatives.

Une telle étude doit donc mise en place prospectivement et prévue au plan de développement comme étant l'étude « finale » devant apporter les preuves cliniques de l'intérêt du traitement. Il s'agit donc d'une phase 3 stricto sensu. En effet rien n'empêche de prévoir dans un plan de développement que

[24] Souvent le sponsor industriel de l'étude non comparative insiste sur cet objectif purement descriptif afin de prévenir toute utilisation de cette étude en défaveur du traitement si, par exemple, les résultats pouvaient être utilisés pour conclure à une faible ou à une non-efficacité ou à un surcroit d'effet indésirable. Il y a donc une reconnaissance de facto par le monde industriel de l'insuffisance de ces études avec cet objectif de conduire à une évaluation des effets des traitements. En fait ces études contribuent à un jeu de dupe : si elles donnent des valeurs peu flatteuses pour le traitement évalué, l'objectif purement descriptif est mis en avant pour prévenir toute exploitation des résultats en défaveur du traitement. En revanche, si les résultats peuvent être exploités en faveur du traitement, elles sont alors proposées comme étude d'évaluation du bénéfice, donc comme base à une comparaison externe. Tout cela de façon purement post hoc, uniquement en fonction des résultats obtenus. D'où l'importance du respect des objectifs initiaux des études

[25] http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/choice-of-control-group-and-related-issues-in-clinical-trials.html

la démonstration ultime du bénéfice du traitement sera apportée par un essai non comparatif avec un comparateur externe (cf. ICH E10).

17.3.2 La comparaison externe doit être clairement explicitée

La référence de comparaison doit être une étude ou une valeur déduite d'une étude et non pas une valeur arbitraire issue d'un avis d'expert sans justification factuelle.

La base de la comparaison doit être clairement explicitée :

- Pour la comparaison à une norme, la valeur de norme retenue doit être justifiée à partir de données factuelles (études des traitements précédents, études d'histoire naturelle, etc.), nonobstant les difficultés à fixer une norme représentative des patients recrutés dans l'étude
- Pour les autres types de comparaisons externes (MAIC, etc.) : la ou les études de référence qui serviront de base à la comparaison

17.3.3 Il doit être possible d'écarter un choix arbitraire de la référence de comparaison, destiné à favoriser le traitement évalué

Il doit être possible d'écarter un choix post hoc de la base de la comparaison externe, drivé par la valeur obtenue par l'étude mono-bras, par un choix clairement effectué avant la date de prise de connaissance des résultats (analyse intermédiaires ou analyse finale).

Il doit être démontré que la référence finalement retenue n'est pas celle la plus favorable à la mise en évidence du bénéfice du traitement évalué. Fréquemment il existe plusieurs études pouvant servir de base à la comparaison externe. Une forte variabilité des résultats entre études est aussi fréquemment rencontrée. Il est donc possible de choisir, arbitrairement, parmi toutes les études disponibles, celle ayant donné les pires valeurs, ce qui in fine favorisera le plus le nouveau traitement. Ce problème existe aussi si ce choix est fait *a priori*, avant d'avoir les résultats de la mono-bras, en prenant le groupe contrôle avec les plus mauvais résultats.

La seule possibilité permettant d'exclure formellement une telle démarche est de disposer de la revue systématique de toutes les études pouvant potentiellement servir de base de comparaison. Ensuite, l'étude retenue doit être celle la plus défavorable au nouveau traitement, dans le but d'être conservateur. Une attitude conservatrice est indispensable afin d'apporter des garanties solides avec un design fragile : il doit donc être possible d'exclure tous les points faisant la fragilité de ces comparaisons externes, par exemple le choix d'un comparateur favorable.

Il est aussi possible que des analyses de sensibilité soient réalisées avec toutes les autres études pouvant servir potentiellement de référence et que toutes ces analyses de sensibilité montrent la supériorité du nouveau traitement.

La revue systématique des études de référence doit être loyale (cf. ci-dessous).

L'approche de constitution de **groupe contrôle synthétique** par sélection dans les bases historiques de patients similaires à ceux traités avec le nouveau traitement dans l'étude mono-bras expose à un risque important et non contrôlable de « data dregging ». En effet, cette approche consiste à extraire, dans une base contenant de nombreux cas, quelques patients bien sélectionnés similaires aux patients traités pour constituer un pseudo groupe contrôle. Il est facile de construire un processus de sélection

testant toutes les possibilités de constitution de groupe contrôle de taille n qu'offre la base de données. Plusieurs bases peuvent être testées à la recherche d'un groupe contrôle optimal. L'acceptabilité de ce type d'approche n'est possible que si des garanties formelles d'absence de processus de data dregging de ce type sont apportées par un protocole strict, ne laissant aucune marge de manœuvre pour faire une sélection des contrôles drivée par les résultats.

17.3.4 Les ajustements effectués doivent permettre d'écarter un biais de confusion.

Pour prétendre avoir un niveau de preuve similaire à celui des essais randomisés, l'étude doit pouvoir apporter la preuve de l'absence d'un biais de confusion résiduel.

Pour cela, une revue systématique de bonne qualité doit avoir été entreprise pour déterminer cette liste des facteurs pronostiques du/des critères de jugements considérés (les facteurs pronostiques peuvent être différents en fonction du critère de jugement) des facteurs modifiant l'effet du traitement.

Ensuite, les ajustements statitiques réalisés doivent prendre en compte non seulement l'ensemble des facteurs pronostiques mais également les modificateurs d'effet.

Les modificateurs de l'effet doivent avoir été aussi inclus.

Pour prétendre avoir un niveau de preuve similaire à celui des essais randomisés, ces approches doivent apporter la preuve de l'absence de biais de confusion résiduelle.

Pour juger si cette condition de validée est vérifiée, il convient qu'une recherche préalable de tous les facteurs pronostiques et modifiants les effets a été conduite par des revues systématiques. Ensuite, il convient que les ajustements aient pris en compte tous ces facteurs. Le plus souvent cela est impossible en raison de l'indisponibilité de certaines de ces covariables (absence dans la publication de l'étude de référence pour une MAIC par exemple ou par absence de recueil dans l'étude).

Si un ou plusieurs de ces facteurs n'a pas pu être pris en compte, il convient que des analyses soient réalisées afin de quantifier par simulation le biais de confusion résiduelle (« *quantitative bias analysis* ») et de montrer que la différence obtenue après ajustement ne peut pas provenir d'un tel biais [178, 179, 180]. L'utilisation de contrôles négatifs est aussi possible. Dans tous les cas, une discussion approfondie du biais de confusion résiduelle soit être avoir été effectuée dans le rapport de l'étude en se basant sur les approches actuelles de quantification et de simulation.

Une partie du biais de confusion peut provenir de variables non liées aux patients mais liées au contexte des études : évolution séculaire et différence de risque intrinsèque des populations, des déterminants génétiques, des contextes de soins, des traitements concomitants, des stratégies de recours aux soins palliatifs, etc. Il s'agit de covariables pour lesquelles les ajustements ne peuvent pas être effectués. La référence de comparaison doit donc être le plus proche possible du contexte de soins de l'étude mono-bras.

La disponibilité d'une revue systématique est indispensable pour permettre de vérifier que tous les facteurs pronostiques ont bien été pris en compte. Une liste déclarative obtenue par avis d'experts est très insuffisante, car rien ne permet d'exclure que celle-ci n'ait pas été conditionnée, entre autres, par la disponibilité des variables dans les différentes études.

17.3.5 La référence de comparaison doit être cliniquement pertinente et loyale

Les patients utilisés comme référence de comparaison doivent avoir reçu les meilleurs traitements actuellement disponibles (« standard of care ») afin de représenter une base de comparaison loyale. La problématique est la même que celle concernant le choix du comparateur pour toutes études versus traitement actif.

Les patients doivent avoir reçu les traitements contemporains et ne peuvent pas être considérés comme sous-traités au regard des recommandations, pratiques et données actuelles.

Les groupes traités des essais randomisés de validation des traitements actuels apportent cette garantie.Ainsi ils doivent avoir été considérés de manière systématique.

Il est à noter que toutes les approches de formalisation des comparaisons externes peuvent aussi être utilisées pour comparer le nouveau traitement à plusieurs traitements de références (efficacité et sécurité relative), à la manière des méta-analyse en réseau [181].

17.3.6 Les revues systématiques doivent être de bonne qualité

Les revues systématiques entreprises pour identifier toutes les études de référence potentielles et lister les facteurs pronostiques du/des critères de jugement doivent être aux standards actuels (par exemple *Cochrane handbook*).

Une attention toute particulière sera apportée aux critères d'exclusion et à la liste des études exclues et des raisons de ces exclusions qui représentent le principal moyen de sélectionner les études favorisant le traitement étudié.

Pour prétendre à l'exhaustivité, la revue systématique doit être basée sur plusieurs bases bibliographiques (Pubmed plus au moins une autre). Il est à noter que le registre CENTRAL de la collaboration Cochrane, étant focalisé sur les essais randomisés, n'a qu'une très faible pertinence pour ce type de revue systématique.

La qualité méthodologique des études doit être prise en compte avec des outils d'évaluation du risque de biais approprié (comme PROBAST pour les études pronostiques(13))

17.3.7 L'exposition potentielle aux biais de l'étude mono-bras et des études de références doit être acceptable

Les méthodologies des 2 études doivent être comparables en termes de sélection, suivi des patients et mesure du critère de jugement pour exclure la possibilité d'une comparaison biaisée par des différences de méthode entre les études.

La taille d'effet doit être très importante pour qu'elle ne puisse pas provenir des biais (liés aux différences des méthodes entre les 2 études)

L'étude de comparaison externe doit discuter soigneusement tous les biais potentiels et apporter une argumentation convaincante comme quoi, quantitativement, les biais ne peuvent pas expliquer la taille de la différence.

La question des biais dans ce contexte est particulière, car il ne s'agit pas des biais traditionnels des études descriptives (aptitudes à estimer le paramètre de la population), mais bien des biais pouvant affecter la comparaison externe. Il convient donc de juger de la méthodologie des 2 études, l'étude

mono-bras et l'étude de référence, dans la perspective d'un biais dans la comparaison externe qui sera effectuée à partir de ces 2 études[26].

Un **biais de mesure** surviendra quand la façon de mesurer le critère de jugement sera différente entre les 2 études, conduisant à une sous-estimation relative de la fréquence du critère de jugement dans l'étude mono-bras par rapport à l'étude de référence. Par exemple avec une définition des évènements plus restrictive dans l'étude mono-bras que dans l'étude de référence ou une méthode de recherche des évènements relativement plus sensible dans l'étude de référence ou relativement plus spécifique dans l'étude mono-bras. Une analyse comparative soigneuse des définitions et des méthodes est donc nécessaire. Cette description doit être discutée dans l'étude et toutes les analyses de sensibilité nécessaires réalisées. Une attention toute particulière doit être portée sur la définition des censures qui peuvent impacter les estimations de PFS par exemple [182, 183]. Finalement il n'y a guère que les critères purement objectifs (donc la mortalité totale) qui sont d'emblée à l'abri de ce type de biais.

L'horizon d'analyse (évènements enregistrés uniquement durant la période de traitement ou durant tout le suivi) ainsi que la durée moyenne de suivi doivent aussi être comparables entre les études, car ces 2 paramètres influencent les fréquences absolues d'évènements. L'utilisation d'un taux (densité d'incidence, en patients*mois par exemple) solutionne les différences de durées de suivi moyennes, mais pas celui de l'horizon d'analyse.

Le **biais d'information** est en général facilement excluable, mais peut survenir au cours du temps si la dynamique d'arrêt du traitement est différente entre les 2 études, et ce indépendamment des conséquences de l'efficacité ou des effets indésirables sur la maintenance des traitements).

Un **biais d'attrition** est fréquemment possible comme avec une étude mono-bras adoptant une logique d'analyse per protocole comparée à une étude de référence dont l'objectif était purement descriptif avec une logique d'analyse en ITT. Par exemple, avec un dispositif ou un produit de thérapie cellulaire, l'analyse de l'étude mono -bras peut être restreinte aux patients chez lesquels le traitement a pu être mis en œuvre avec succès alors que l'étude de référence, dont le but était de décrire les pratiques, a intégré tous les patients.

Ce type de biais peut aussi prendre la forme de **biais de sélection** par exemple avec les décès précoces survenant entre l'inclusion dans 'étude et la mise en œuvre du traitement qui seront exclus que dans l'étude mono-bras et non pas dans l'étude de référence descriptive

Il est à noter qu'il est extrêmement difficile pour une étude de comparaison externe de se mettre à l'abri des biais. Cela est bien plus difficile que dans une étude comparative avec comparateur interne où, par exemple, la définition et la méthode de mesure du critère de jugement sont les mêmes entre les 2 groupes par essence. Une fois de plus, il faut rappeler qu'historiquement la méthode expérimentale au sens large, et plus particulièrement l'essai contrôlé randomisé en double aveugle, ont été construits afin d'apporter des solutions simples et efficaces à l'ensemble de ces problèmes rencontrés lors des comparaisons externes [184].

On peut donc conclure que le contrôle des biais dans une étude à comparateur externe nécessiterait de construire l'étude mono-bras en fonction de l'étude de référence : utilisation de la même définition

[26] Il convient aussi de remarquer que la problématique des biais est aussi importante que celle des différences de patients entre études mono-bras et de référence et que les méthodes d'analyse utilisée pour corriger du biais de confusion ne corrigent pas les autres biais.

du critère de jugement, de la même méthode de mesure, de la même durée de suivi, des mêmes règles d'arrêt de traitement, etc.

17.3.8 Le résultat suggéré par la comparaison externe doit être cliniquement pertinent

Le résultat doit être obtenu sur un critère clinique pertinent. La taille de l'effet doit elle aussi être pertinente. La balance bénéfice risque doit être appréciable et favorable qualitativement et quantitativement. La généralisabilité du résultat doit être assurée.

Comme pour toute évaluation critique de résultat d'étude clinique, la pertinence clinique est tout aussi importante que la fiabilité méthodologique du résultat. L'évaluation de la pertinence clinique d'un résultat issu d'une comparaison externe formalisée doit être évaluée avec les mêmes critères que ceux appliqués à un résultat d'essai randomisé.

17.4 Synthèses des problématiques et de leurs solutions

Problématique méthodologique spécifique (Exposant à un risque de production de résultat favorable à tort au traitement étudié)	**Démonstration que doivent apporter les solutions envisagée (pour garantir la disparition du risque de conclure à tort)**
Le groupe contrôle externe n'est pas un véritable contrefait documentant ce qu'aurait dû être le critère de jugement chez les mêmes patients en l'absence du traitement évaluer (sous traitement de référence). C'est une problématique de biais de confusion impliquant facteurs pronostiques et modificateurs d'effet de chaque critère de jugement considéré	Les ajustements réalisés doivent donner la garantie d'une absence de confusion résiduelle. Pour cela il doit être montré que tous les facteurs pronostiques et les modificateurs de l'effet ont été pris en considération.
Malgré la réalisation d'ajustements, un biais de confusion résiduelle persiste et est à l'origine du résultat obtenu	• L'ajustement a porté sur l'ensemble des facteurs de confusion de chacun des critères de jugement identifié par une revue systématique des facteurs pronostiques conduite suivant les standards actuels de ce type de méta-analyse[27] [185]. • L'ajustement a porté sur l'ensemble des modificateurs d'effet des traitements comparés [181] • Et/ou une analyse quantitative de biais démontre que la taille de l'effet obtenu ne peut pas s'expliquer par le biais de confusion résiduelle • ET/ou l'utilisation de contrôle négatif (avec éventuellement recalibration) démontre l'absence de biais de confusion résiduel
Le groupe contrôle externe a été choisi par le promoteur pour favoriser le traitement évalué (choix post hoc déterminé à partir des résultats)	Choix *a priori* (avant recrutement des patients traités avec le nouveau traitement), fixé au protocole (véritable étude comparative à contrôle externe) donnant la garantie qu'il ne s'agit pas de l'étude ayant obtenu le moindre effet

[27] Voir https://methods.cochrane.org/prognosis/

	Si choix post hoc (en ayant connaissance des résultats chez les patients traités), démontrer que le choix n'est pas déterminé par les résultats en faisant une revue systématique de toutes les sources de données pouvant donner un groupe contrôle, montrer que celui utilisé était le seul possible ou montrer qu'il ne favorise pas le résultat obtenu (analyse de sensibilité avec toutes les sources possibles)
Biais de mesure Extrêmement probable du fait qu'il s'agit d'une comparaison entre 2 études différentes	Aligner le protocole de l'étude à contrôle externe sur celui de l'étude servant de comparateur (même définition du critère, même méthode de mesure, etc.)
Biais de sélection	Synchroniser les débuts de suivi par rapport à l'évolution naturelle de la maladie (pour éviter un biais de temps d'immortalité) Éviter les comparaisons avant après
Biais d'attrition L'étude de comparaison est une étude descriptive où la problématique des données manquantes est moins aigue que pour les études destinées à comparer deux traitements	Utiliser un contrôle externe de qualité, ayant assuré un suivi exhaustif des patients (faible attrition)
Pertinence clinique Les études mono-bras portent souvent sur des critères intermédiaires. Les contrôles externes peuvent être trop anciens et ne pas avoir été traités avec les traitements actuellement utilisés	Le ou les critères de jugement doivent être les critères cliniques standards de la situation clinique. Le ou les traitements contrôles doivent être pertinent.
Balance bénéfice risque	La comparaison externe doit documenter la balance bénéfice risque

Aucun type de contrôles externes (cohorte historique, cohorte ad hoc, bras d'un essai randomisé, groupe contrôle synthétique, etc.) n'est *a priori* plus fiable. Seules comptent les réponses apportées aux problématiques méthodologiques qui entachent ces études.

17.5 Méta-recherche

Une étude par Carrigan et al. dans le cancer du poumon montre sur quelques cas sélectionnés une bonne aptitude des groupes contrôles synthétiques pour reproduire les résultats des essais randomisés (Figure 10) [186].

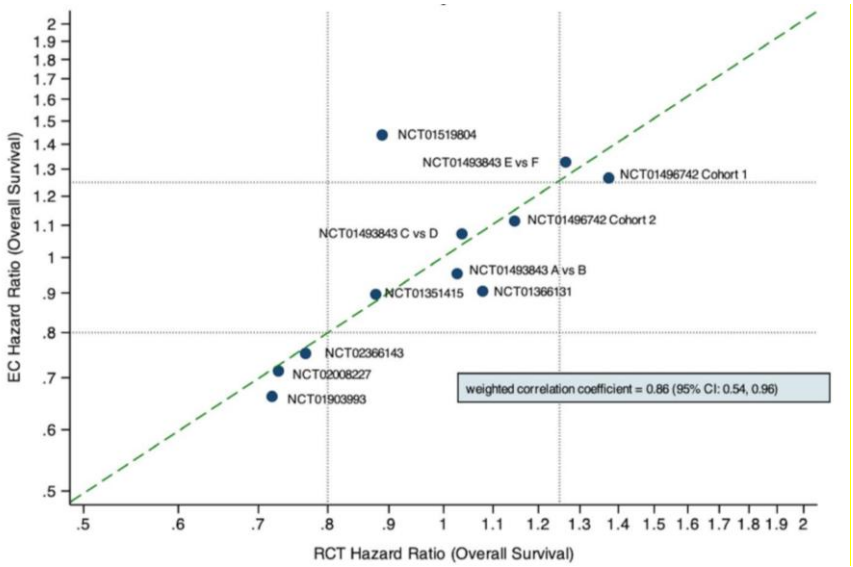


Figure 10 – Corrélation entre les estimations produites par les groupes contrôles synthétiques synthétiques (« EC Hazard Ratio ») et les essais randomisés (« RCT Hazard Ratio ») (d'après [186]).

Comme le fait remarquer Larrouquere et al. [187], cette concordance n'apporte pas une démonstration complète de l'aptitude des groupes contrôles synthétiques à suppléer les essais randomisés, car parfois, même si les effets relatifs mesurés sont similaires, les ajustements réalisés ne reproduisent pas strictement les résultats de l'essai randomisé comme le montre la Figure 11.

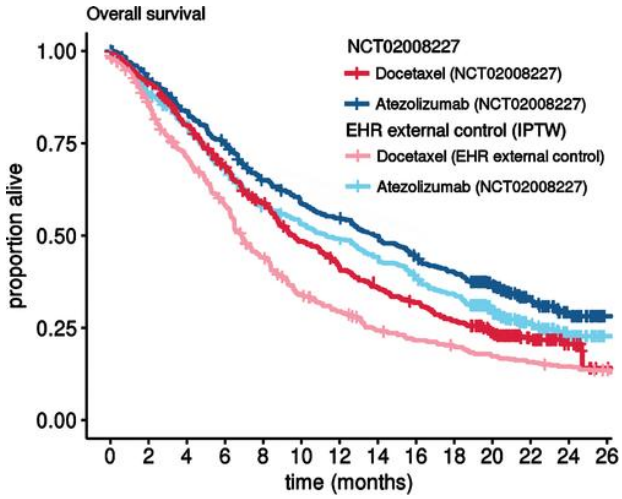


Figure 11 – Exemple des résultats obtenus par groupe contrôle synthétique comparés à ceux de l'essai randomisé (d'après [187]).

17.6 Avis de la SFPT

L'acceptabilité d'une comparaison externe comme démonstration du bénéfice d'un nouveau traitement va dépendre de plusieurs conditions. La vérification de ces conditions va être indispensable pour conclure que les résultats ainsi produits sont suffisamment fiables, proches de ceux qui auraient été obtenus par un essai randomisé et permettent ainsi d'intégrer le nouveau traitement dans la stratégie thérapeutique.

- La comparaison externe doit être formalisée, avec l'utilisation d'une méthode de comparaisons indirecte non ancrée permettant un ajustement et débouchant sur la quantification de la taille du bénéfice
- La comparaison externe doit avoir été planifiée a priori :
    - Le jugement de l'intérêt du traitement évalué est basé sur une méthode de comparaison externe définie dans la méthode de l'étude (Il s'agit d'une véritable étude à comparateur externe et non pas d'une étude mono-bras, purement descriptive, sans objectif d'évaluation du bénéfice du traitement évalué
- Le contrôle externe doit être justifié factuellement et non pas choisi arbitrairement :
    - La base de comparaison est une étude ou une valeur issue d'une étude (et non pas une valeur arbitraire issue d'un avis d'expert sans justification factuelle)
- Il estpossible d'écarter un choix arbitraire de la référence de comparaison destiné à favoriser le traitement évalué
    - Une revue systématique de toutes les études pouvant servir potentiellement de référence est disponible
    - La référence finalement retenue n'est pas la possibilité la plus favorable au traitement évalué
    - Le choix de la base de la comparaison externe (norme ou études de référence) a été effectué indépendamment de la connaissance des résultats de l'étude mono-bras
    - Les analyses de sensibilité réalisées avec les autres études pouvant servir potentiellement de référence montrent toutes la supériorité du traitement évalué
- Les ajustements effectués permettent d'écarter un biais de confusion. L'étude donne l'assurance de l'absence d'un biais de confusion résiduel :
    - L'ajustement réalisé (MAIC ou autres) a pris en compte l'ensemble des facteurs pronostiques et des modificateurs d'effet connus pour chaque critère de jugement.
    - L'exhaustivité de la prise en compte des facteurs pronostiques est justifiée sur la base d'une revue systématique des études de facteurs pronostiques
    - L'exhaustivité de la prise en compte des modificateurs de l'effet est justifiée de manière satisfaisante (sur la base des essais comparatifs des traitements de références et de la plausibilité biologique pour le traitement évalué)
- La référence utilisée pour la comparaison externe est cliniquement loyale :
    - Cette référence correspond à des patients dont la prise en charge et les traitements reçus sont au standard actuel
- Les revues systématiques (à la recherche des groupes contrôles possible et facteurs de confusion potentiels) sont de bonne qualité
    - Les revues systématiques des potentielles études de référence et des facteurs pronostiques des critères de jugement sont satisfaisantes (recherche exhaustive, critère de sélection non arbitraire, pas d'exclusion arbitraire d'étude, évaluation du risque de biais)
- L'exposition potentielle au biais de la comparaison externe est réduite :
    - Les méthodologies des 2 études (étude monobras et étude servant de comparateur exeterne) sont suffisamment comparables en termes de sélection, suivi des patients et de mesure du critère de jugement pour exclure la possibilité d'une comparaison biaisée par des différences de méthode entre les études.
    - L'étude discute soigneusement tous les biais potentiels (mesure, information ou attrition) et apporte une argumentation convaincante comme quoi, quantitativement, les biais ne peuvent pas expliquer la taille de la différence observée
- Le résultat suggéré par la comparaison externe est cliniquement pertinent :

- Le résultat est obtenu sur un critère clinique pertinent
- la taille de l'effet est pertinente
- La balance bénéfice risque est favorable qualitativement et quantitativement
- la généralisabilité du résultat est assurée

18 L'emprunt d'information

L'emprunt d'information (*historical data borrowing*) consiste à enrichir les données apportées par un petit essai randomisé avec davantage de données contrôles empruntés à d'autres sources (historique le plus souvent). Les études de comparaisons externes correspondent à des situations où 100% des données contrôles sont empruntées.

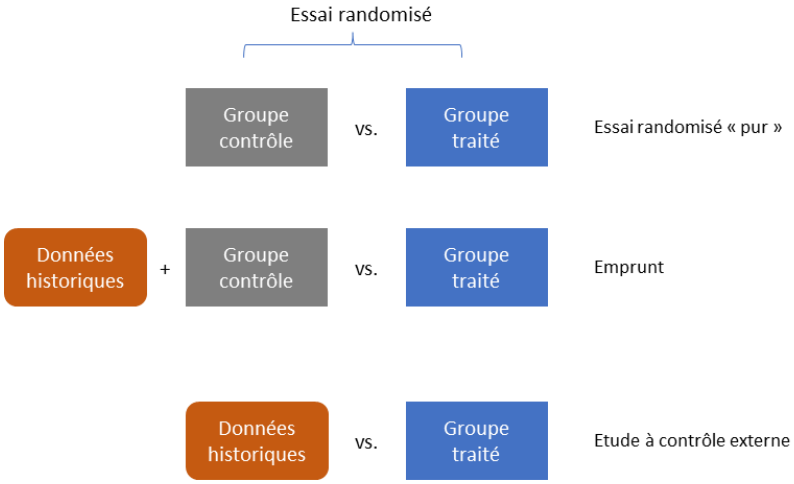


Figure 12 – Illustration du principe de l'enrichissement du groupe contrôle d'un essai randomisé par un emprunt de données historiques (ligne du milieu)
Cette approche se situe entre l'essai randomisé classique (où il n'y a aucun emprunt de données pour le groupe contrôle) et les études de comparaisons externes (où 100% des données contrôles proviennent sont des données historiques).

De nombreuses méthodes statistiques ont été proposées, bayésiennes ou fréquentistes. Leur principe générique est simple. Seront comparés les résultats obtenus dans le groupe traité de l'essai randomisé avec le résultat du « pooling » des données du groupe contrôle de l'essai et les données historiques empruntées (comme si ces données avaient été produites par l'étude elle-même). Ce « pooling » peut être vu comme une méta-analyse des données des groupes contrôles. En bayésien, les données historiques peuvent être utilisées comme apriori informatif pour le groupe contrôle (et un apriori non informatif est utilisé pour le groupe traité). Un paramètre arbitraire règle le poids relatif des données empruntées par rapport aux données du groupe contrôle de l'essai randomisé dans ce « pooling ».

Même si cette approche n'augmente que l'information (les données) contrôle, elle conduit à une augmentation de la puissance statistique de la comparaison si les données historiques ne sont pas discordantes avec celle de l'étude. Moins de patients ont donc besoin d'être inclus dans l'essai randomisé conduisant à un cout plus faible de l'étude et à une durée d'inclusion plus courte.

La validité des résultats produits dépend d'une hypothèse fondamentale qui, en terme bayésien, est le caractère échangeable des études apportant les données contrôles. Les données des contrôles historiques estiment correctement les résultats que devrait obtenir le groupe contrôle de l'essai randomisé. Les patients des groupes contrôles historiques doivent donc comparables en termes de distributions de tous les facteurs pronostiques et des modificateurs d'effets (cf. hypothèse des comparaisons indirectes non ancrées [188]). Aucune méthode n'a été proposée jusqu'à présent pour ajuster sur les caractéristiques des patients.

Ces techniques sont principalement proposées pour des phases précoces (phase 2) où l'aspect spéculatif de l'hypothèse fondamentale de validité expose seulement au risque pour l'industriel de financer et réaliser une phase 3 à tort et non pas à une utilisation indue en pratique du nouveau traitement. Cependant ces approches sont aussi de plus en plus envisagées pour a réalisation d'études pivots (comme dans les maladies rares) et apparaissent ainsi dans des guidelines réglementaires :

- 2019 FDA guidance for Interacting with the FDA on Complex Innovative Clinical Trial Designs for Drugs and Biological Products [189].
- 2019 FDA guidance for Rare Diseases: Common Issues in Drug Development [190] : « The potential use of natural history data as a historical comparator for patients treated in clinical trial is often of interest… in general studies using historical controls are credible only when the effect is large in comparison to variability in disease course »
- EMA: Guideline on Clinical Trials in Small Populations [191].

Historiquement cette approche a été proposée dès 1976 par Stuart Pocock [192]. D'autres propositions et variantes ont été faites plus récemment comme les « *power priors* » [193], ou les « Meta Analytic Predictive (MAP) Prior » [188, 194]. Ces approches sont très techniques et leur description détaillée dépasse largement le cadre de ce document.

18.1 Problématiques méthodologiques spécifiques

L'approche d'emprunt de données historiques pour enrichir le groupe contrôle d'un petit essai randomisé débouche sur les mêmes problématiques que les essais à contrôle externe en termes de choix arbitraire d'une information qui peut conditionner en grande partie le résultat final de l'essai (cf. section 17.1). Elles posent aussi les problèmes de l'inférence bayésienne avec des apriori informatifs (cf. section 13.1.3).

De plus leur technicité rend leur lecture critique compliquée, nécessitant une expertise statistique. Même si ces approches proposent des solutions techniques pour diagnostiquer et corriger les situations à risque (hétérogénéité des données historiques par rapport aux données de l'essai, modèle hiérarchique), il est difficile de connaitre leur aptitude à détecter ces situations et à corriger le résultat. La technicité pouvant aussi conduire à croire que les problèmes sont automatiquement solutionnés.

Ces limites sont donc

Problématique méthodologique ***et particularité spécifique à la nouvelle méthodologie.***	**Solution spécifique à apporter avec cette nouvelle méthodologie** ***pour garantir l'obtention du même degré de certitude qu'avec la méthodologie classique***
Choix arbitraire de données historiques favorisant le traitement étudié	Revue systématique de toutes les sources de données utilisables. Certaines approches passent par une méta-analyse des données historiques et s'inscrivent donc dans cette logique
Choix post hoc après connaissance des résultats des résultats de l'essai randomisé (Peu de risque sur ce point, car le calcul des effectifs à recruter dans l'essai randomisé nécessite de connaitre l'information emprunté)	Choix *a priori* des données contrôles, intégré dans le protocole de l'étude

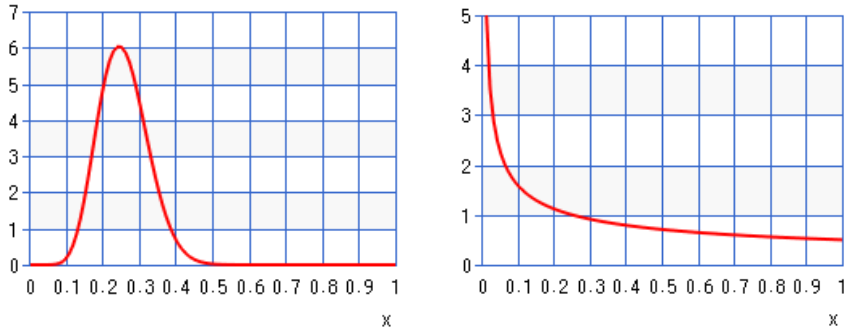
Données historiques non représentatives du groupe contrôle et introduisant un biais de confusion dans la comparaison	Démonstration que les données historiques apportent une estimation correcte de la valeur du critère de jugement dans le groupe contrôle de l'essai clinique
Impossibilité de faire des ajustements sur les caractéristiques des patients dans les méthodes disponibles jusqu'à présent	
Manque de pertinence des traitements reçus par les patients des données historiques (non-respect de l'hypothèse d'échangeabilité entre groupes contrôles)	Données historiques correspondant aux traitements standards actuels
Faible pertinence clinique du critère de jugement	Utilisation d'un critère cliniquement pertinents
Autres limites des comparaisons externes en termes de biais de mesure, réalisation, pertinence clinique, etc.	Idem comparaisons externes

18.2 Études de cas

Une approche d'emprunt de données à des groupes contrôles historiques a été utilisée dans une étude de preuve de concept du secukinumab dans la spondylarthrite ankylosante chez l'adulte [195]. Dans l'essai randomisé versus placebo, 24 patients ont été inclus dans le groupe secukinumab et seulement 6 dans le groupe placebo (randomisation 4:1). Le critère de jugement principal était le pourcentage de patients ASA20 à la semaine 6.

Le taux de réponse ASAS20 sous placebo a été estimé en empruntant les données de 533 patients issus de 8 précédents essais versus placebo dans la spondylarthrite ankylosante. Pour tenir compte de l'hétérogénéité entre les groupes contrôles, les données empruntées aux groupes contrôles historiques ont été sous pondérées par la méthode et correspondes à l'équivalent de 43 patients.

> *L'utilisation de cette information conduit à l'utilisation d'un apriori informatif dans l'estimation bayésienne du taux de réponse sous placebo correspondant à une distribution beta de paramètre 11 et 32. Pour le groupe secukinumab, l'apriori on informatif correspondait à une distribution beta de paramètre 0.5 et 1.*



Les résultats suivants ont été obtenus conduisant à une conclusion positive de l'essai

	Responders, n (%)	Response rate*	Difference *vs* placebo†	95% credibility interval†	Probability
Secukinumab‡§	14 (60·9%)	59·2%	34·7%	11·5–56·4%	99·8%
Placebo	1 (16·7%)	24·5%	..	..	..

ASAS=Assessment of SpondyloArthritis international Society criteria. *Means from the posterior beta ($0{\cdot}5+x$, $1+n-x$) distribution for secukinumab and beta ($11+x$, $32+n-x$) distribution for placebo, where x represents the number of responders and $n-x$ represents the number of non-responders in the corresponding treatment group. †Difference in response rates simulated from the posterior probability distributions of secukinumab and placebo. ‡Secukinumab: 2×10 mg/kg. §The efficacy dataset included only 23 of 24 patients in the secukinumab group, since one patient was excluded due to a dosing error.

***Table 2:* Primary endpoint Bayesian analysis of ASAS20 responders at week 6**

Il apparait que l'emprunt d'information fait passer l'estimation du taux de réponse de 16.7% à 24.5%, changement en défaveur du traitement testé.

18.3 Avis de la SFPT

Les critères d'acceptabilité de ces approches d'emprunt de données historiques pour enrichir le groupe contrôle sont identiques à celle des comparaisons externes :

- Le choix des groupes contrôles historiques emprunté a été effectué de manière non arbitraire et non drivé par les résultats
- Le choix des groupes contrôles historiques a été effectué *a priori*, lors de l'élaboration du protocole
- Le traitement reçu par les groupes contrôles historiques est identique à celui utilisé dans l'essai randomisé
- Les groupes contrôles historiques sont comparables à l'essai randomisé et n'introduisent pas de biais de confusion
- Le critère de jugement et sa méthode d'évaluation sont identiques à celui de l'essai randomisé

S'adjoignent aussi des critères liés à l'approche bayésienne utilisant des aprioris informatifs :

Définition apriori d'un seuil de probabilité *a posteriori* d'efficacité pour conclure à l'intérêt du traitement en fonction de la multiplicité de l'étude et garantissant un contrôle du risque alpha global

Dans son papier de 1979 [192], Stuart Pocock mentionnait aussi, en plus de ces critères, la nécessité que les contrôles historiques proviennent largement des mêmes organisations et des mêmes investigateurs.

19 Les *surrogates* (critères de substitution)

Un biomarqueur est un paramètre objectivement mesuré comme un indicateur de processus biologique normal ou pathologique, ou de réponse pharmacologique [196]. À noter que le terme « biomarqueur » n'est pas restreint aux marqueurs biologiques (marqueur biologique, d'imagerie, critère clinique fonctionnel, etc.). Les biomarqueurs sont des critères de jugement intermédiaires souvent utilisés à la place de critères cliniques « durs », car ils permettent des essais plus petits en taille et avec des durées de suivi plus courtes. Cette approche repose alors sur l'hypothèse que la mise en évidence d'un effet du traitement sur ce critère vaut démonstration du bénéfice sur le critère clinique. Mais cela n'est vraiment le cas que si le critère intermédiaire a valeur de critère de substitution (*surrogate*). Par conséquent, **si tous les critères de substitution sont des biomarqueurs, l'inverse n'est pas vrai**.

19.1 Problématiques méthodologiques

L'existence d'une association (*subject level correlation*) entre un critère intermédiaire (biomarqueurs) et un critère clinique (décès, évènement) n'est pas suffisant pour faire de ce biomarqueur un surrogate, car cette association (lien épidémiologique) ne garantit pas d'une modification provoquée du biomarqueur par le traitement se traduit par un bénéfice proportionné au niveau du critère clinique. Il existe de nombreux exemples de cette limitation (cf. section 19.3.1 par exemple).

> La réponse pathologique complète (pCR) est un facteur de risque de récidive dans le cancer du sein précoce. Cependant il n'a pas été possible de démontrer qu'il s'agissait d'un surrogate dans le cadre du traitement néoadjuvant [63]. L'observation qu'un traitement impacte favorablement la pCR ne permet pas d'induire avec certitude qu'il impacte favorablement la récidive.

19.2 Solution

La démonstration qu'un biomarqueur est un surrogate utilisable à la place d'un critère clinique repose sur la mise en évidence d'un haut degré de corrélation entre l'effet du traitement sur le biomarqueur et l'effet du traitement sur le critère clinique (*trial level correlation* ou R_{trial}) [197, 198]. Cette démonstration passe par la réalisation d'une régression montrant la relation entre l'effet sur le critère intermédiaire et l'effet sur le critère clinique pertinent. Il a été proposé une borne inférieure de l'IC à 95% du coefficient de corrélation $R > 0.85$ [199] pour la démonstration de la surrogacy.

Cette démonstration nécessite donc que soit disponibles des essais randomisés documentant l'effet du traitement sur les critères cliniques et sur le surrogate. D'une certaine façon la démonstration qu'un biomarqueur est un surrogate valide s'obtient paradoxalement alors que les essais ont démontré le bénéfice clinique du traitement sur les critères cliniques eux-mêmes.

La corrélation entre le biomarqueur et le critère clinique doit aussi être vérifiée. Cette corrélation « au niveau des individus » (R_{indiv}) est habituellement évaluée à partir d'études observationnelles ou randomisées, mais ne suffit pas à établir la surrogacy.

Pour avoir un intérêt en pratique, la démonstration qu'un biomarqueur est un surrogate doit pouvoir s'extrapoler aux futurs médicaments d'un autre mécanisme d'action ou d'une autre classe pharmacologique. À titre d'exemple, l'effet des glifozines pourraient être liés à deux effets différents, l'un passant par un effet hypoglycémique, l'autre par un effet potentiel effet antihypertenseur ou

diurétique. Cette universalité n'est pas démontrable en elle-même et plusieurs limites sont envisageables :

- Le biomarqueur n'est pas sur la voie du nouveau mécanisme d'action
- Le biomarqueur n'est pas un intermédiaire de l'action du traitement mais un effet parallèle
- Les effets secondaires du nouveau mécanisme d'action n'impactent pas le critère clinique de la même façon que ceux des traitements utilisés pour l'étude de « validation » du surrogate (par exemple décès toxiques et overall survival en oncologie)

Au-delà de la démonstration de la corrélation des effets, la question de la généralisabilité de la démonstration de la surrogacy est un élément clés de l'utilisabilité en pratique d'un candidat surrogate.

Lors de leur utilisation, la démonstration d'un effet non nul du traitement sur un surrogate ne permet pas d'en déduire de facto un effet non nul du traitement sur le critère clinique, car la relation entre les 2 effets n'est jamais une relation d'identité. De manière régulière, les tailles des effets sur le critère clinique sont plus petites que celle sur le surrogate. L'analyse de corrélation doit donc déterminer le seuil de l'effet minimal sur le surrogate (« *surrogate threshold effect* ») [200] qui conduit à un effet non nul avec certitude sur le critère clinique si la valeur de surrogacy est démontré. Ce seuil est déterminé à partir de la limite péjorative de l'intervalle de prédiction de la relation entre les effets sur le surrogate et les effets sur le critère clinique. Pour un nouveau traitement, l'éventuelle conclusion à son bénéfice sur le critère clinique ne pourra être effectuée que si l'effet observé dans son essai sur le surrogate est plus important que ce seuil. La taille de l'effet extrapolé sur le critère clinique est estimée à l'aide de l'intervalle de prédiction de la relation de surrogacy obtenue à partir du modèle de régression linéaire. En effet, la disponibilité d'une estimation de la taille de l'effet sur le critère clinique est indispensable pour juger de la pertinence clinique et pour positionner le nouveau traitement dans la stratégie thérapeutique.

> *L'estimation du STE peut aussi être utilisée comme une aide au plan de développement pour optimiser les tailles des études de confirmation sur critère clinique, à partir des résultats initiaux sur le critère de substitution.*

Même avec un surrogate valide dont la généralisabilité serait démontrée (autant que cela est possible), des limites méthodologiques d'un autre type apparaissent. L'utilisation d'un surrogate permet de réaliser des essais plus courts et/ou avec moins de sujets. De ce fait, ces petites études ne permettent de documenter avec fiabilité le risque de ces nouveaux médicaments et d'établir avec certitude la balance bénéfice risque. L'idée de rechercher des surrogates des risques (des effets indésirables) n'est pas envisageable compte tenu de la spécificité des effets indésirables et de leur caractère inattendu.

Bien que l'approche proposée par Buyse et al. [56, 59] soit devenue la méthode standard, d'autres propositions [201, 202, 203] ont été faites depuis, mais ne sont pas employées en pratique.

19.3 Études de cas

19.3.1 Le LDL cholestérol, un contre-exemple

L'exemple du LDL cholestérol et des traitements hypocholestérolémiants illustre bien la difficulté de trouver des surrogate universels et les limites des approches habituelles implicites.

La réduction de LDL cholestérol a été proposée comme surrogate pour les médicaments « hypocholesterolémiants » en prévention des évènements cardiovasculaires. La méta-analyse des essais de statines [204] donne un argument en faveur d'une relation entre la baisse de LDL et la réduction de fréquence des évènements cardiovasculaires (Figure 13).

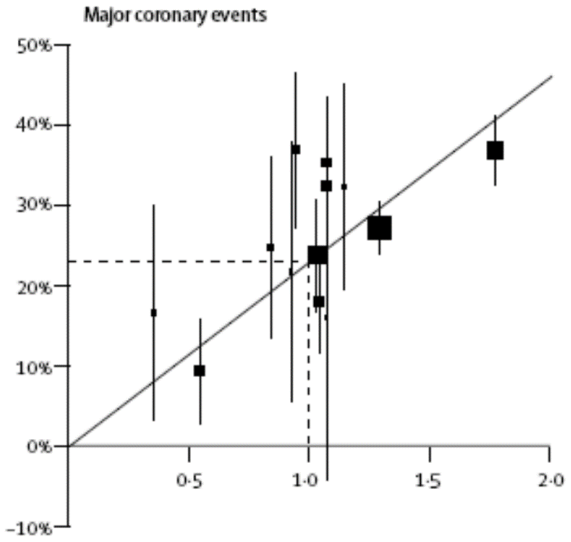


Figure 13 – Relation entre la baisse du LDL (en abscisse) et la réduction relative de la fréquence ce des évènements coronariens (en ordonnée) obtenus par méta-régression dans la méta-analyse de la Cholesterol Treatment Trialists (CTT) Collaborators [204].

Cette relation semble faire du LDL cholestérol un surrogate pour les traitements hypocholestérolémiants en prévention cardiovasculaire (secondaire) ; une baisse de 1mmol/L entrainant une réduction de 22% des évènements.

Cette approche n'est cependant pas une véritable étude de la surrogacy du LDL telle que présentée dans la section précédente. Il s'agit d'une simple d'analyse explicative de l'hétérogénéité des résultats des essais statines par méta-régression. Il manque l'évaluation du R_{trial} et le calcul du STC. De plus la régression a été forcée pour passer par l'origine. Il s'agit donc d'une analyse qui présuppose une relation entre ces 2 effets et cherche seulement à estimer la pente. Une vraie validation de surrogate doit, avant tout, chercher s'il existe une relation entre les effets, sans la préfixer.

L'acide nicotinique est un hypocholestérolémiant qui entraine d'une baisse substantielle du LDL cholestérol. Son bénéfice sur les évènements cardiovasculaires, qui semblait d'emblée acquis compte tenu de cette action pharmacologique et des bénéfices apportés par les statines par exemple, a cependant été recherché dans l'essai de morbi-mortalité de grande taille, HPS2-THRIVE [205].

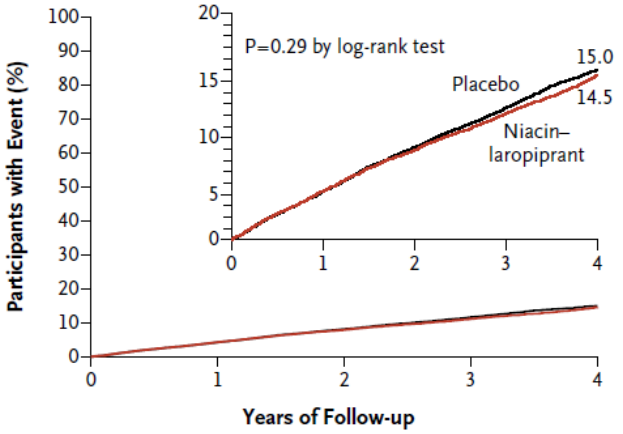
Dans cet essai l'effet hypocholestérolémiant est observé tout au long des 4 années de l'essai :

Table S2: Effects of niacin/laropiprant on lipids

	Lipid parameter (mg/dL)					
Year of follow-up	**Total cholesterol**	**LDL cholesterol**	**HDL cholesterol**	**Triglycerides**	**Apo A1**	**Apo B**
<1	-8 (1.3)	-12 (1.1)	6 (0.4)	-35 (3.4)	5 (0.8)	-9 (0.8)
≥1 <2	-6 (0.4)	-10 (0.4)	7 (0.2)	-33 (1.3)	7 (0.3)	-8 (0.3)
≥2 <3	-5 (0.9)	-9 (0.7)	6 (0.3)	-31 (2.3)	7 (0.6)	-7 (0.5)
≥3 <4	-3 (0.9)	-7 (0.7)	6 (0.3)	-33 (2.6)	7 (0.6)	-6 (0.5)
≥4	-3 (0.6)	-7 (0.5)	6 (0.2)	-32 (1.8)	7 (0.4)	-5 (0.4)
Study average	**-5 (0.4)**	**-10 (0.3)**	**6 (0.1)**	**-33 (1.1)**	**7 (0.3)**	**-7 (0.2)**

Absolute differences (with standard error) in lipid fraction during the trial. Study averages are weighted for participant-years at risk within region. Follow-up periods are as in supplementary table 1. LDL: low density lipoprotein. HDL: high density lipoprotein. Apo A1: apolipoprotein A1. Apo B: apolipoprotein B.

Malgré cette action pharmacologique, aucun bénéfice sur les évènements cliniques n'a été observé :



No. at Risk					
Niacin–laropiprant	12,838	12,232	11,517	7672	4978
Placebo	12,835	12,247	11,523	7643	5036

Ce résultat réfute donc la valeur « universelle » de substitution (surrogacy) du LDL à travers les mécanismes d'action.

Pour la dernière nouvelle génération « d'hypocholestérolémiant », les anti-PSK9, la réalisation d'essais de morbi-mortalité (portant sur plus de 27 000 patients) était donc indispensable compte tenu de cette expérience et a été exigée [206]. Seul un bénéfice sur les évènements cardiovasculaires mortels et non mortels a été montré alors que certains travaux montrent une valeur de surrogacy du LDL sur la mortalité coronarienne.

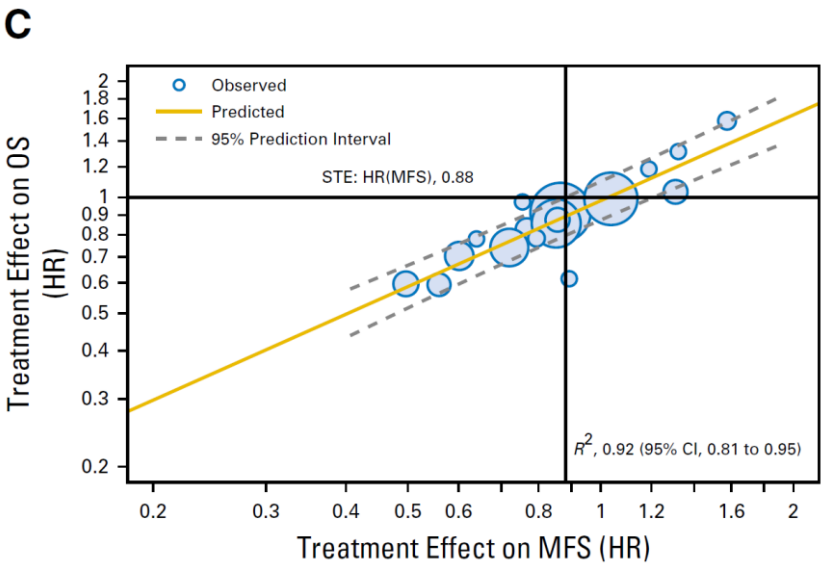
19.3.2 PFS et OS, un autre contre-exemple

En oncologie, la PFS (*progression free survival*) est souvent considérée comme un surrogate implicite de l'OS (*overall survival*), en prétextant une relation structurelle temporelle entre les deux critères : « retarder la progression repousse mécaniquement le décès ». Dans le domaine des chimiothérapies par exemple, tous les produits qui impactent la survie ont eu un effet favorable sur la PFS. Cependant

avec les immunothérapies sont apparues des situations [207] où l'effet traitement sur la PFS est apparent délétère (HR>1), mais où la survie est augmentée. Ce résultat paradoxal a conduit au concept de pseudoprogression liée à l'infiltration lymphocytaire de la tumeur consécutive à la levée par le traitement de la freination du système immunitaire. Cet exemple illustre bien la possible dépendance au mécanisme d'action de la valeur de substitution (surrogacy).

19.3.3 Metastasis-Free Survival dans le cancer de la prostate

Dans le cancer de la prostate, la survie sans métastase (MFS, Metastasis-Free Survival) à 5 ans apparait être un surrogate de la survie (OS) à 8 ans en raison d'une corrélation des effets élevée (R^2 à 0.92, IC95% entre 0.81 et 0.95) [208]. Le seuil STE est à 0.88 signifiant d'un nouveau traitement devra montrer un effet sur la MFS se traduisant par une borne supérieure de l'intervalle de confiance du HR inférieure à 0.88.



Le tableau 1 de la publication (cf. ci-dessous) apporte les différents résultats avec la corrélation au niveau individuel et la corrélation entre les effets au niveau essai.

Table 1. Two-Condition Surrogacy Analysis

					Condition 1 (TE and ICE are correlated)		Condition 2 (treatment effects on both end points are correlated)	
							Regression of Log(HR)-TE *v* Log(HR)-ICE (by trial)	
TE	ICE	No. of Trials	No. of Units*	No. of Patients	Correlation at the Patient Level, Kendall's τ (95% CI)	Regression of 8-Year TE Rate *v* 5-Year ICE Rate† by Trial and Arm, R^2 (95% CI)	R^2 (95% CI)	Regression Equation
OS	DFS	24	31	21,140	0.85 (0.85 to 0.86)	0.86 (0.78 to 0.90)	0.73 (0.53 to 0.82)	$\text{Log(HR)}_{OS} = 0.035 + 0.605 \times \text{Log(HR)}_{DFS}$
DSS	TDR	21‡	28	20,496‡	0.68 (0.67 to 0.69)	0.80 (0.70 to 0.85)	0.63 (0.36 to 0.75)	$\text{Log(HR)}_{DSS} = 0.027 + 0.809 \times \text{Log(HR)}_{TDR}$
OS	MFS	19	21	12,712	0.91 (0.91 to 0.91)	0.83 (0.71 to 0.88)	0.92 (0.81 to 0.95)	$\text{Log(HR)}_{OS} = -0.021 + 0.740 \times \text{Log(HR)}_{MFS}$
DSS	TTM	16‡	18	12,068‡	0.91 (0.91 to 0.92)	0.86 (0.75 to 0.90)	0.89 (0.72 to 0.93)	$\text{Log(HR)}_{DSS} = -0.072 + 0.880 \times \text{Log(HR)}_{TTM}$

Abbreviations: DFS, disease-free survival; DSS, disease-specific survival; HR, hazard ratio; ICE, intermediate clinical end point; MFS, metastasis-free survival; OS, overall survival; TDR, time to disease recurrence; TE, true end point; TTM, time to metastasis.
*Five trials were split according to the type of primary therapy or experimental arm (if two or more experimental arms).
†Eight-year TE rates and 5-year ICE rates were Kaplan-Meier estimates by trial and treatment arm, excluding three studies with median follow-up < 6 years.
‡Excluding three studies with the number of prostate cancer deaths < 3.

Ce travail illustre aussi les limites méthodologiques de la validation des surrogate qui passe forcément par une démarche rétrospective. Plusieurs candidats surrogate ont été testés, conduisant à des conclusions variables. La MFS est mise en avant dans une démarche purement inductive, car elle a obtenu le meilleur niveau de corrélation au niveau essai. La réalisation du travail à partir des données individuelles peut amener à discuter la nature exploratoire des réanalyses des études pour produire les points de la régression et l'exhaustivité des données. Il n'y a pas non plus d'argumentation sur l'universalité de cette relation à travers les classes thérapeutiques. Il convient aussi de noter qu'il s'agit de la survie sans métastase à 5 ans et rien ne garantit que cette relation soit aussi valable pour des suivis plus courts.

Cet exemple montre ainsi la difficulté de produire une démonstration formelle « au de la de tout doute raisonnable » qu'un critère intermédiaire est un surrogate valide pouvant dispenser de la réalisation des essais sur le critère clinique même lorsqu'une approche rigoureuse et conforme aux standards actuels est employée.

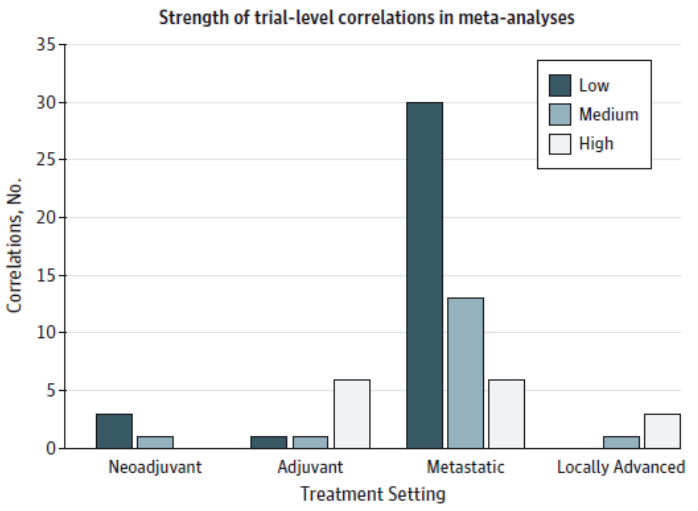
19.4 Méta-recherche

Depuis une dizaine d'années, de nouveaux traitements, particulièrement en oncologie, sont seulement évalués sur des « surrogates » qui n'ont pas fait l'objet de validation, à la suite de la possibilité offerte par la FDA[28]. De ce fait un doute persiste souvent sur le réel intérêt de ces traitements. Plusieurs auteurs se sont émues de cette situation [18, 23, 38, 209, 210, 211, 212, 213] qui a été bien étudié par diverses études de méta-épidémiologie.

En oncologie le critère clinique est la survie (ou la qualité de vie en montrant que le nouveau traitement s'accompagne d'une meilleure qualité de vie avec une survie identique au traitement standard). La PFS plus fréquente et plus facilement impactée que l'OS est souvent avancé comme étant un surrogate de l'OS. Les travaux étudiants la corrélation des effets sur la PFS avec ceux sur l'OS dans différent type de cancer et/ou différents types de traitements ont été colligés par Vinay Prasad et ses collègues en 2015 [214]. En majorité, ces travaux montrent une corrélation faible ou moyenne ne permettant pas de considérer la PFS comme surrogate de l'OS dans les situations correspondantes.

28 Sec. 314.510 approval based on a surrogate endpoint or on an effect on a clinical endpoint other than survival or irreversible morbidity (FDA TITLE 21, PART 314, SUBPART H) "FDA may grant marketing approval for a new drug product on the basis of adequate and well-controlled clinical trials establishing that the drug product has an effect on a surrogate endpoint that is reasonably likely … to predict clinical benefit…"

Figure 3. Correlations by Treatment Setting



We scored strength of trial-level correlation according to a modification to surrogate criteria proposed by the Institute of Quality and Efficiency in Health Care[34]: low correlation ($r \leq 0.7$), medium strength correlation ($r > 0.7$ to $r < 0.85$), and high correlation ($r \geq 0.85$).

Les situations où la PFS peut être considéré comme un surrogate fiable de l'OS sont très peu nombreux et peuvent éventuellement être remis en cause quand des essais supplémentaires deviennent disponibles. Ce fût le cas avec les traitements du cancer du côlon métastatique pour lesquels la PFS semblait être dans un 1er temps un surrogate acceptable [215] mais un travail ultérieur a récusé ce résultat[216].

La conséquence de cela est que, 5 ans après leur enregistrement, des preuves du bénéfice sur l'OS ne sont disponibles que pour une faible proportion des nouveaux produits enregistrés sur des essais portant sur ces critères présentés comme étant des « surrogate » [22] ; et que pour un nombre non négligeable de produits, les essais de mortalité réalisés ne montrent pas de bénéfice sur l'OS.

La FDA (Food and Drug Administration) a listé les critères intermédiaires retenus comme des « surrogates » acceptables pour eux[29], sans que la justification ne soit présente. Les agences de régulation européennes, dont l'EMA (European Medecines Agency), n'ont pas produit un tel document à notre connaissance. Une revue systématique des enregistrements réalisés par la FDA sur la base d'une « surrogate » a montré que la discussion du rationnel de la surrogacy n'est réalisée que dans 26% des évaluations seulement [217].

[29] https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure

19.5 Avis de la SFPT

L'utilisation d'un critère de substitution sera en mesure de documenter avec un haut degré de certitude le bénéfice du nouveau traitement sur le critère clinique (dispensant de réaliser un essai sur ce critère clinique) uniquement lorsque :

Une démonstration de la valeur de substitution du critère de jugement intermédiaire utilisé a été apportée par la méthodologie standard [197, 198, 200, 218, 219, 220, 221] ou par une analyse de médiation suffisamment probante (précision de la capture)

Démonstration de la relation patient-level R_{indiv} dans des études non biaisées

Démonstration de la relation study level R_{trial}

Analyses de sensibilité appropriées confirmant toutes les résultats de l'analyse principale (pondérée et non pondérée, avec et sans prise en compte de l'erreur de mesure sur la variable explicative, qualité des études, *small study effect*, etc.)

Pour l'approche par corrélation, la démonstration du haut degré de corrélation avec les critères IQWIG [221, 222] (borne inférieure de l'IC du R supérieure à 0.85)

Absence de data dredging / p hacking en particulier au niveau du choix du surrogate (très difficile à assurer du fait de la nature rétrospective de la démarche)

Une démonstration spécifique que la valeur de substitution (surrogacy) s'applique à la (nouvelle) classe pharmacologique du nouveau traitement

L'étude de validation du surrogate repose sur un travail de revue systématique acceptable, pour lequel il peut être exclu avec certitude une sélection des essais contributifs biaisée par la connaissance des résultats :

Recherche exhaustive de tous les essais publiés et non publiés (au moins 2 bases, recherche active des essais non publiés)

Possibilité de conclure à l'absence de biais de publication

Critère de sélection des essais non arbitraire, logique et approprié

Et pour **le nouveau traitement**, la démonstration d'un effet sur le surrogate supérieur au « Surrogate threshold effect »

L'essai du **nouveau traitement** sur le surrogate est suffisamment long et inclus suffisamment de patients pour assure une certaine représentativité et apprécier assez correctement la safety

20 Les essais basket

Le principe des essais basket est de démontrer le **bénéfice du traitement pour plusieurs pathologies avec un seul et unique essai** mélangeant toutes ces pathologies. En oncologie, il s'agit de démontrer le bénéfice d'une thérapie ciblée sur des tumeurs associées à la même altération moléculaire, quel que soit le tissu ou l'organe. La finalité de cette approche est d'éviter de devoir valider l'efficacité pathologie par pathologie (organe par organe) en partant de l'hypothèse que le traitement apporte un bénéfice similaire dans toutes ces pathologies (hypothèse d'homogénéité). Si cette hypothèse est exacte, il n'y a donc plus besoin de distinguer les pathologies et le résultat global permet de conclure pour chaque pathologie individuellement (sans avoir une démonstration spécifique pour chacune de ces pathologies).

20.1 Problématiques méthodologiques

La clé de voute de cette approche est donc **l'hypothèse d'homogénéité**. Si elle n'est pas vérifiée (le traitement n'apporte pas le même bénéfice dans toutes les pathologies et n'en apporte pas dans certaines d'entre elles), l'approche basket conduira à considérer à tort le traitement comme bénéfique dans certaines pathologies. Il y aura validation abusive d'une population de dissémination plus large, ce qu'elle devrait être avec des types de patients traités à tort.

L'hypothèse d'homogénéité ne peut pas être démontrée par l'essai basket lui-même, car cela nécessiterait de concevoir un essai assurant la démonstration pour chaque pathologie, ce que l'on cherche à éviter par principe dans l'essai basket. Cependant il est possible de réfuter cette hypothèse si des effets traitement hétérogènes, avec un test d'interaction concluant, sont observés entre les strates. Mais l'absence de réfutation de l'hypothèse d'homogénéité ne permet pas de conclure à sa démonstration, car un résultat non significatif ne signifie pas l'absence, surtout dans cette situation où la puissance du test d'interaction est très faible.

L'hypothèse d'homogénéité doit donc avoir été démontrée en amont de l'essai basket dont la logique est finalement : partant du principe que le traitement, s'il est efficace, va apporter le même bénéfice dans toutes les pathologies, démontrons ce bénéfice commun à partir d'un essai mélangeant toutes ces pathologies. Un essai basket ne permet pas de monter que l'effet du traitement est le même pour toutes les pathologies, mais bien d'estimer l'effet du traitement commun à toutes ces pathologies partant du principe que l'hypothèse d'homogénéité est vérifiée.

La problématique est donc de démontrer en amont que l'hypothèse d'homogénéité est valide. Cette démonstration est impossible (cf. ci-dessus) et seule sa plausibilité est argumentable sur la base des mécanismes d'actions, de similitudes à d'autres situations ayant acceptées/démontrées l'hypothèse d'homogénéité, etc.

20.2 Étude de cas

L'essai CAPRIE a comparé le clopidogrel à l'aspirine en prévention cardiovasculaire secondaire [223]. Trois types de patients qui antérieurement à cette étude étaient étudiés séparément ont été inclus : des patients avec un AVC ischémique récent, un infarctus du myocarde récent ou une maladie artérielle périphérique symptomatique. Ce regroupement de différentes pathologies d'organes était justifié, car physio-pathologiquement il s'agit dans les trois cas d'une maladie athéromateuse des

artères et une grande méta-analyse avait montré que l'aspirine et d'autres antiagrégants plaquettaires apportaient le même bénéfice dans ces 3 pathologies [224, 224, 225]. Le but était de montrer la supériorité du clopidogrel sur l'aspirine, globalement pour justifier l'usage du clopidogrel dans ces 3 pathologies, sans devoir apporter une démonstration pathologie par pathologie.

L'essai CAPRIE est donc un essai basket avant l'heure dont l'hypothèse fondamentale d'homogénéité était justifiée par la physiopathologie et les résultats des méta-analyses.

Malgré la solidité de la justification, les résultats ne parvinrent pas à convaincre en raison d'une hétérogénéité graphique (Figure 14) et statistiquement significative (p=0.042).

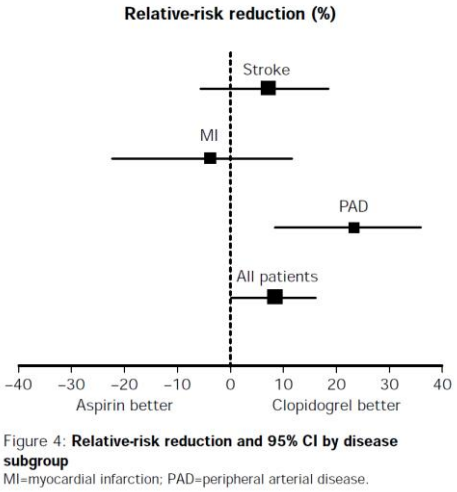


Figure 4: **Relative-risk reduction and 95% CI by disease subgroup**
MI=myocardial infarction; PAD=peripheral arterial disease.

Figure 14 – Résultats de l'essai CAPRIE par types de patients

Cet exemple montre les difficultés d'apporter la démonstration de l'hypothèse d'homogénéité et les limites de ces approches.

20.3 Avis de la SFPT

L'acceptabilité d'un essai basket pour intégrer le nouveau traitement dans la stratégie thérapeutique de façon agnostique nécessite sur les aspects spécifiques :

La démonstration au préalable que l'hypothèse fondamentale d'homogénéité était vérifiée

Et la non remise en cause de cette démonstration par les résultats obtenus

21 Les analyses poolées, les méta-analyses

La place de la méta-analyse dans la démonstration de l'intérêt clinique d'un traitement est discutée depuis très longtemps [186, 226, 227, 228, 229, 230].

Dans ce contexte, l'intérêt de la méta-analyse serait, en regroupant plusieurs essais, de montrer le bénéfice du traitement sur des critères plus pertinents que ceux utilisés dans les essais sources et d'éviter ainsi le recours à des essais plus importants (méga-essais).

Par exemple, en regroupant plusieurs essais d'étidronate réalisés pour mesure l'effet du produit sur la densité osseuse, pouvoir documenter l'effet du traitement sur les fractures vertébrales [231]. Aucun des essais n'avait comme critère principal ces fractures et aucun n'avait la puissance statistique nécessaire. Cependant dans certains essais (pas tous) des données sur les fractures vertébrales sont disponibles. En les agrégeant par une méta-analyse il est possible d'augmenter la puissance et peut-être de mettre en évidence de manière statistiquement significative un bénéfice du produit sur la prévention des fractures vertébrales. La preuve de l'intérêt du traitement serait apportée par la méta-analyse et non pas par un essai.

21.1 Problématiques méthodologiques

La méta-analyse classique est une approche rétrospective dont elle hérite des limites méthodologiques : comme les résultats des études sont disponibles lorsque la méta-analyse est réalisée, il est possible de choisir les essais à inclure en fonction du résultat de la méta-analyse qu'ils produisent.

Même si le travail de méta-analyse n'effectue pas une sélection des études, celle-ci a peut-être eu lieu en amont, lors de la publication ou non des études en fonction de leur résultat (biais de publication).

De même, il peut être décidé de répondre à une question (bénéfice du traitement sur la mortalité par exemple) par la voie de la méta-analyse en fonction du résultat produit. L'approche rétrospective ne s'inscrit pas dans la démarche hypothético déductive nécessaire à l'obtention d'un haut degré de certitude des résultats.

L'aptitude à trouver un résultat potentiellement intéressant à tort est aussi amplifiée par la multiplicité non contrôlée des analyses que l'on peut faire dans une méta-analyse : nombreux critères de jugement, sous-groupes en fonction de la qualité des études, de leur contemporanéité, des traitements, des patients, etc. L'utilisation d'un de ces multiples résultats pour soutenir une revendication sera alors entièrement déterminée par le résultat lui-même.

La méta-analyse est aussi sensible au biais des études elles-mêmes. La présence d'études exposées aux biais (essais randomisés ou études observationnelles) conduit à un résultat de méta-analyse lui-aussi à risque de biais. Pour cette raison, une restriction aux études à faible risque de biais lorsqu'elle existe peut-être préférable par rapport à une méta-analyse incluant toutes les études y compris celles à risque de biais. Même si ce principe ne conduit pas à l'exhaustivité, il évite de dégrader le résultat de la méta-analyse du fait de la présence des études à risque de biais alors qu'existent des études apportant un haut degré de certitude.

La dernière problématique de la méta-analyse est le sens médical du regroupement des études qui peuvent porter sur des populations de patients différentes ou avoir utilisé des modalités de traitements différents. Une méta-analyse peut faire conclure à tort à l'intérêt du traitement pour une

sous-population de patients, car la méta-analyse regroupant les essais réalisés chez ces patients avec d'autres essais montre globalement un bénéfice, mais qui est conditionné par le bon résultat du traitement obtenu chez les autres sous-populations. De même, un effet délétère spécifique d'une population ou d'un traitement peut disparaitre en méta-analyse où les résultats le montrant sont dilués par ceux d'autres études réalisées avec des patients ou des traitements ne présentant pas cet effet indésirable.

> *Par exemple, une méta-analyse peut amener à faire conclure à un effet de classe, donc avec son résultat applicable à tous les représentants de la classe, alors qu'en réalité l'effet n'a été obtenu ou vraiment documenté que pour une molécule de ladite classe (sans que cela ne soit détecter par le test d'hétérogénéité compte tenu de sa faible puissance ou de sa non-prise en considération).*

Au total ces limites rendent rédhibitoire l'utilisation d'une telle méta-analyse pour apporter la preuve d'un bénéfice clinique d'un traitement en l'absence d'études concluantes par elle-même ; la méta-analyse servant principalement soit à vérifier la cohérence externe du résultat d'un essai pivot, soit à générer de nouvelles hypothèses à tester dans un nouvel essai.

La solution est la planification *a priori* de la méta-analyse (analyse conjointe) avant la réalisation des essais. Cette approche est parfois connue sous le nom de méta-analyse prospective. Ainsi disparait la problématique liée au choix rétrospectif des études participant à la méta-analyse. La démarche hypothético-déductive est parfaitement respectée, car l'objectif de la méta-analyse est aussi établi *a priori*.

> Le candersartan dans l'insuffisance cardiaque a fait l'objet de 3 essais CHARM [150, 232, 233] dont le critère de jugement était le critère composite hospitalisations ou décès cardiovasculaires. Pour documenter l'efficacité du produit sur les décès, le plan de développement prévoyait de répondre à cette question par la méta-analyse des 3 essais, ce regroupement permettant (CHARM program) d'obtenir le nombre de sujets nécessaires pour garantir la puissance statistique sur ce critère [234].

Une limite persiste, celle de l'hétérogénéité statistiques des résultats des études qui peut rendre difficile l'interprétation de la méta-analyse et compromettre la démonstration du bénéfice. Il faut bien sûr aussi que le regroupement de ces études ait un sens clinique avec des populations des essais qui correspondent à une même entité clinique. Un essai entrepris pour répondre à la même question aurait regroupé aussi ces différents patients.

21.2 Étude de cas

Dans la sclérose en plaques, les exigences règlementaires demandent la réalisation de 2 essais pivots sur le taux annualisé de poussées. L'enjeu actuel des traitements est de montre qu'ils retardent l'apparition des déficits. Mais ce critère demande plus de patients pour garantir la puissance.

L'ofatumumab a été évalué dans deux essais randomisés identiques ASCLEPIOS I et ASCLEPIOS II [235]. Le critère de jugement principal de ces 2 essais était le taux annualisé de poussées. Il a été prévu d'emblée un regroupement de ces études pour répondre à la question de l'impact de ce traitement sur la progression du handicap. Cet objectif a de plus été intégré dans le plan de contrôle du risque alpha global et ne pouvait être testé que si les 2 études étaient concluantes sur leur critère de jugement principal.

Secondary clinical end points were disability worsening confirmed at 3 months, disability worsening confirmed at 6 months, and disability improvement (i.e., lessening of disability) confirmed at 6 months; a prespecified meta-analysis of these end points used the combined data from both trials.

The type I error was controlled by a statistical testing procedure, with seven prespecified secondary end points tested; disability worsening confirmed at 3 months or 6 months and disability improvement confirmed at 6 months were tested in preplanned meta-analyses of the combined trials only if the primary null hypothesis for the annualized relapse rate was rejected in both trials independently.

21.3 Avis de la SFPT

L'acceptabilité d'un résultat de méta-analyse ou d'analyse poolée pour positionner un nouveau traitement dans la stratégie thérapeutique nécessite :

Une approche de méta-analyse prospective, avec la garantie qu'elle était bien prévue avant la réalisation des essais regroupée (mentionnée dans le protocole des essais)

L'inclusion uniquement d'études à faible risque de biais

Un sens médical au regroupement des études effectuées en termes de patients, traitements, comparateurs

L'absence d'hétérogénéité statistique des résultats des études

Une gestion de la multiplicité (par exemple avec un plan de contrôle du risque alpha global englobant les études et la méta-analyse)

22 Les comparaisons indirectes en remplacement d'études « head to head » manquantes

Le grand nombre de traitements développé simultanément conduit fréquemment, lorsque les essais sont terminés, au constat que le traitement comparateur utilisé n'est plus le meilleur traitement disponible à la date de prise de décision. L'essai n'apporte alors pas la démonstration que le nouveau traitement sur-performe par rapport au traitement de la stratégie thérapeutique. Cette situation survient quand, dans l'intervalle de réalisation de l'essai du nouveau traitement considéré, un autre traitement a montré plus rapidement sa supériorité sur le même critère par rapport au même comparateur. Cela peut aussi survenir lors de développement simultané de plusieurs molécules du même mécanisme d'action. Les essais s'initient versus le même comparateur (le traitement optimal au moment de la mise en place des essais) de manière assez contemporaine, mais quand même un peu étalée dans le temps produisant finalement des résultats eux-mêmes décalés dans le temps. Cette situation conduit, là aussi, à des problématiques de décision pour les résultats disponibles après le premier. S'il n'est pas possible de différentier les études du fait de leurs résultats (concluant ou non concluant), leur degré de certitude ou leur pertinence clinique, se posent alors la question du bénéfice relatif et de la sécurité relative de ces traitements entre eux. Réaliser un essai de comparaisons des traitements entre eux et attendre ses résultats pour décider n'est pas possible.

Les comparaisons indirectes (ancrées) peuvent alors être envisagées dans cette situation pour extrapoler ce qu'auraient pu être les résultats d'un essai de comparaison directe (« head to head ») entre les nouveaux traitements.

22.1 Problématiques méthodologiques

La fiabilité des comparaisons indirectes repose sur une hypothèse fondamentale, l'hypothèse de transitivité [236, 237]. Cette hypothèse est une hypothèse d'échangeabilité des traitements entre les essais. Cette hypothèse ne peut pas être testée et seule sa plausibilité peut être évaluée. Celle-ci repose sur l'identification d'éventuels modificateurs de l'effet des traitements comparés et par la vérification, si de tels modificateurs existent, que leur distribution est identique entre les essais contribuant aux comparaisons indirectes (identique en moyenne si plusieurs essais effectuant la même comparaison existent).

Les comparaisons indirectes sont souvent réalisées dans le cadre de grande méta-analyse en réseau regroupant tous les traitements de la condition clinique considérée. Ces méta-analyses en réseau sont entreprises avec d'autres objectifs que de pallier au manque de comparaison directe d'intérêt (modélisation médico-économique, élaboration de recommandations, etc.). Leur standard de réalisation n'assure pas forcément le degré de certitude des résultats nécessaire pour remplacer un essai de comparaison « head to head ».

Problématique méthodologique spécifique (exposant à un risque de production de résultat favorable à tort au traitement étudié)	**Démonstration que doivent apporter les solutions à ces problématiques (pour garantir la disparition du risque de conclure à tort)**
Hypothèse de transitivité	Démonstration de la plausibilité de l'hypothèse de transitivité 1. Soit en montrant qu'il n'y a pas de facteurs modifiant l'effet des traitements à partir des

	analyses en sous-groupes, mais cette recherche est limitée par la disponibilité des résultats en sous-groupes pertinents et la recherche d'une conclusion à l'absence de modification d'effet (faible puissance des tests d'interaction) 2. Soit, si des modificateurs de l'effet ont été trouvés, en montrant que les populations des essais sont comparables en moyenne sur ces variables
Manque de puissance et dilution de la précision	Aucune, la comparaison indirecte est complètement tributaire d'essais déjà réalisés et qui n'ont pas été calibrés pour une comparaison avec un traitement control. Impossibilité de conclure à l'équivalence des produits

22.2 Méta-recherche

La concordance des résultats des comparaisons indirectes avec ceux des comparaisons directes a été étudiée plusieurs fois [238, 239, 240, 241]. Le dernier travail trouve un taux de discordance plus importante que les précédents avec une fréquence dans leur échantillon de méta-analyse de différences statistiquement significatives entre les 2 approches de 14%. Il n'existe donc pas de validation empirique implicite des comparaisons indirectes et la validité doit être discutée cas par cas.

22.3 Avis de la SFPT

L'acceptabilité d'un résultat d'une comparaison indirecte pour positionner un nouveau traitement dans la stratégie thérapeutique nécessite que les critères suivants soient vérifiés.

- Cette comparaison indirecte a été entreprise spécifiquement pour pallier l'absence d'un essai de comparaison directe de bonne qualité
- Il ne s'agit pas d'une méta-analyse en réseau réalisée pour d'autres objectifs (recommendations, modélisation médico économique)
- Une démonstration de la plausibilité de l'hypothèse de transitivité comprenant une recherche soigneuse des modificateurs d'effets et la démonstration de l'absence de différence, en moyenne, entre les essais de la distribution des modificateurs. En cas de rejet globalement de cette hypothèse, les résultats ajustés ou obtenus en splitting ne sont pas recevables et la question de la comparaison des 2 molécules doit être considérée comme insoluble par l'approche des comparaisons indirectes.
- Des essais contribuant à la comparaison indirecte ont été réalisés de façon relativement contemporaine, à partir d'une guideline standardisant leurs méthodes (critères, population, durées, etc.).
- Les centres investigateurs des essais sont en grande partie les mêmes
- Tous les essais du plan de développement pertinents pour la comparaison indirecte ont été utilisés (méta-analyse)

En cas de « méta-analyse » réalisée en dehors d'un plan de développement, ces critères s'ajoutent à ceux des méta-analyses « classiques », inhérent au processus de revue systématique

et d'analyses poolées. En cas d'approches par réseau global et/ou d'intégration de comparaisons indirectes et directes, des critères supplémentaires de validités seront nécessaires (notamment l'absence d'incohérence entre les estimations directes versus indirectes, entre les différents chemins du réseau). Enfin, les estimations de « ranking » des traitements d'une méta-analyse en réseau doivent être interprétées avec la plus grande prudence [242] et ne peuvent en l'état actuel suppléer les estimations des effets des traitements.

23 Les maladies rares

Lorsque le nombre de sujets est très faible, il est impossible de mettre en évidence le bénéfice d'un traitement sauf si cet effet est très important. Il s'agit d'une impossibilité mathématique de séparer un effet faible du bruit de fond induit par la variabilité du vivant. Ces situations ne permettent donc pas, et ne permettront jamais, d'obtenir de preuve du bénéfice des traitements du fait de l'extrême rareté des patients (maladies rares ou variations moléculaires de pathologies fréquentes) et du faible bénéfice des traitements proposés, quelle que soit la méthodologie utilisée[30].

Dans ces situations, il est donc impossible de garantir aux patients le bénéfice clinique réel des traitements utilisés [243]. La solution n'est pas de tenter de trouver une méthode magique permettant de conclure avec un haut degré de certitude à partir de rien ou d'espérer que de nouvelles méthodes apporteraient le niveau de certitude recherchée (inatteignable compte tenu de l'importance de la variabilité biologique par rapport à la quantité d'unités statistiques observables), mais d'acter cette impuissance, d'en informer les patients et d'imaginer de nouvelles formes de régulation pour empêcher une approche prédatrice de ces nouveaux marchés [244].

Ces situations sont cependant exceptionnelles et ne concernent pas toutes les maladies rares. La réalisation d'essais randomisés reste faisable dans de nombreuses situations et doit être l'approche de base [245]. La FDA dans un document de guidance, non encore approuvé en novembre 2021, réaffirme importance des essais randomisés avec une ouverture possible vers les études mono-bras [190]. La problématique des études mono-bras et des essais à contrôle externe est discutée section 16 et section 17. Dans le cadre des maladies rares aussi, ces études se révèlent difficiles à réaliser, par manque de contrôles historiques appropriés (sélectionnés sur le biomarqueur spécifique du nouveau traitement par exemple) et bien souvent le développement d'un traitement de ce type commence par documenter l'histoire naturelle de la maladie (cf. guide FDA déjà cité).

Des éléments de design permettent d'optimiser la puissance des essais randomisés sans compromettre leur fiabilité en jouant principalement sur une réduction de la variabilité : design adaptatif, analyse séquentielle, design combiné (*seamless*) phase 2/3, étude en cross-over, et pour les critères continus suivi longitudinal permettant de répéter les mesures du critère de jugement et de modéliser, éviter la dichotomisation, etc. [246]. Malgré cela la puissance des études reste en général faible et conduit à une valeur prédictive positive faible des études statistiquement significatives [247, 248].

Pour pallier ce manque de puissance structurelle, apparait parfois une certaine dérive méthodologique pour obtenir malgré tous des résultats d'essais randomisés présentés comme « concluant » : augmentation du seuil de la signification statistique [246, 249], non-infériorité avec une limite excessive (et arbitraire), utilisation de critères intermédiaires présentés comme étant des critères cliniques, etc. [248, 250, 251]

Une autre voie envisagée pour remédier à la difficulté de recruter les effectifs suffisants est celle de l'emprunt d'information [252, 253, 254], avec principalement l'utilisation d'apriori informatif en inférence bayésienne (cf. section 18). La principale limite de cette approche est l'utilisation d'un apriori arbitraire, plus basé sur les croyances de l'investigateur que sur des données adaptées.

30 La quantité d'information (entropie) possible de produire par une étude est insuffisante pour distinguer l'effet du traitement du bruit.

À noter que, même pour les maladies rares, quand les traitements apportent un bénéfice substantiel il devient tout à fait possible de le démontrer avec les méthodologies standards (comme par exemple avec le *nusinersen* dans l'atrophie musculaire spinale [255] qui a démontré dans un essai en double aveugle par une sham-intervention, avec un contrôle du risque alpha global, un gain en mortalité lors d'une analyse intermédiaire conduisant à l'arrêt prématuré de l'essai).

Les critères d'acceptabilités des résultats dépendent du type de méthodologie utilisée, l'aspect maladie rare n'introduisant aucune spécificité dans l'interprétation des résultats, et ne rend pas non plus les hypothèses simplificatrices plus valides *a priori* que dans les autres domaines. Nous renvoyons donc le lecteur aux parties de ce document qui correspondent aux méthodes utilisées.

Références

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

2 Hwang TJ, Carpenter D, Lauffenburger JC, et al. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Intern Med* 2016;176:1826–33 doi:10.1001/jamainternmed.2016.6008; PMID:27723879;

3 Prasad V, Cifu A, Ioannidis JPA. Reversals of established medical practices: evidence to abandon ship. *JAMA* 2012;307:37–38 doi:10.1001/jama.2011.1960; PMID:22215160;

4 Prasad V, Gall V, Cifu A. The frequency of medical reversal. *Arch Intern Med* 2011;171:1675–76 doi:10.1001/archinternmed.2011.295; PMID:21747003;

5 Sutton D, Qureshi R, Martin J. Evidence reversal-when new evidence contradicts current claims: a systematic overview review of definitions and terms. *J Clin Epidemiol* 2018;94:76–84 doi:10.1016/j.jclinepi.2017.10.004;

6 Tajika A, Ogawa Y, Takeshima N, et al. Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *Br J Psychiatry* 2015;207:357–62 doi:10.1192/bjp.bp.113.143701; PMID:26159600;

7 Tatsioni A, Bonitsis NG, Ioannidis JPA. Persistence of contradicted claims in the literature. *JAMA* 2007;298:2517–26 doi:10.1001/jama.298.21.2517; PMID:18056905;

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

9 The Centre for Evidence-Based Medicine. The ethics of COVID-19 treatment studies: too many are open, too few are double-masked - The Centre for Evidence-Based Medicine 2020. Available at: https://www.cebm.net/covid-19/the-ethics-of-covid-19-treatment-studies-too-many-are-open-too-few-are-double-masked/ Accessed August 23, 2021.

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

11 Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. *Trials* 2017;18:280 doi:10.1186/s13063-017-1978-4; PMID:28681707;

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

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

14 Nikolaidis GF, Woods B, Palmer S, et al. Classifying information-sharing methods. *BMC Med Res Methodol* 2021;21:107 doi:10.1186/s12874-021-01292-z; PMID:34022810;

15 Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med* 1996;15:2733–49 doi:10.1002/(SICI)1097-0258(19961230)15:24<2733:AID-SIM562>3.0.CO;2-0; PMID:8981683;

16 Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. *N Engl J Med* 2017;377:62–70 doi:10.1056/NEJMra1510062; PMID:28679092;

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

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

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

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

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

22 Kim C, Prasad V. Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. *JAMA Internal Medicine* 2015;175:1992–94 doi:10.1001/jamainternmed.2015.5868; PMID:26502403;

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

24 BEECHER HK. Surgery as placebo. A quantitative study of bias. *JAMA* 1961;176:1102–07 doi:10.1001/jama.1961.63040260007008; PMID:13688614;

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

26 Heckerling PS. The Ethics of Single Blind Trials. *IRB: Ethics and Human Research* 2005;27:12 doi:10.2307/3563956;

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

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

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

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

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

32 Cohen D. Cancer drugs: high price, uncertain value. *BMJ* 2017;359:j4543 doi:10.1136/bmj.j4543; PMID:28978508;

33 Sachs RE, Gavulic KA, Donohue JM, et al. Recent Trends in Medicaid Spending and Use of Drugs With US Food and Drug Administration Accelerated Approval. *JAMA Health Forum* 2021;2:e213177 doi:10.1001/jamahealthforum.2021.3177;

34 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;

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

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

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

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

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

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

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

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

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

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

45 Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19:766–79

doi:10.1097/EDE.0b013e3181875e61; PMID:18854702;
46 Lash TL, VanderWeele TJ, Haneuse S, et al. Modern epidemiology. Philadelphia etc.: Wolters Kluwer 2021 ISBN:1451193289;
47 Schuemie MJ, Ryan PB, Pratt N, et al. Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND). *J Am Med Inform Assoc* 2020;27:1331–37 doi:10.1093/jamia/ocaa103; PMID:32909033;
48 Bruns SB, Ioannidis JPA. p-Curve and p-Hacking in Observational Research. *PLoS ONE* 2016;11:e0149144 doi:10.1371/journal.pone.0149144; PMID:26886098;
49 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;
50 Head ML, Holman L, Lanfear R, et al. The extent and consequences of p-hacking in science. *PLoS Biology* 2015;13:e1002106 doi:10.1371/journal.pbio.1002106; PMID:25768323;
51 Silberzahn R, Uhlmann EL, Martin DP, et al. Many analysts, one dataset: Making transparent how variations in analytical choices affect results 2017.
52 Chuard PJC, Vrtílek M, Head ML, et al. Evidence that nonsignificant results are sometimes preferred: Reverse P-hacking or selective reporting? *PLoS Biol* 2019;17:e3000127 doi:10.1371/journal.pbio.3000127; PMID:30682013;
53 Michels KB, Rosner BA. Data trawling: to fish or not to fish. *The Lancet* 1996;348:1152–53 doi:10.1016/S0140-6736(96)05418-9;
54 Data dredging - Wikipedia 2021. Available at: https://en.wikipedia.org/wiki/Data_dredging Accessed August 30, 2021.
55 Berger ML, Sox H, Willke RJ, et al. Good Practices for Real-World Data Studies of Treatment and/or Comparative Effectiveness: Recommendations from the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. *Value Health* 2017;20:1003–08 doi:10.1016/j.jval.2017.08.3019; PMID:28964430;
56 Orsini LS, Monz B, Mullins CD, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing-Why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Pharmacoepidemiol Drug Saf* 2020;29:1504–13 doi:10.1002/pds.5079; PMID:32924243;
57 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;
58 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;
59 Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health* 2005;95 Suppl 1:S144-50 doi:10.2105/AJPH.2004.059204; PMID:16030331;
60 Pearl J. An introduction to causal inference. *The International Journal of Biostatistics* 2010;6:Article 7 doi:10.2202/1557-4679.1203; PMID:20305706;
61 Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;60:578–86 doi:10.1136/jech.2004.029496; PMID:16790829;
62 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;
63 Belas N. P-hacking in Clinical Trials: A Meta-Analytical Approach ;
64 Hripcsak G, Suchard MA, Shea S, et al. Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension. *JAMA Internal Medicine* 2020 doi:10.1001/jamainternmed.2019.7454; PMID:32065600;
65 Nyström T, Bodegard J, Nathanson D, et al. Second line initiation of insulin compared with DPP-4 inhibitors after metformin monotherapy is associated with increased risk of all-cause mortality, cardiovascular events, and severe hypoglycemia. *Diabetes Research and Clinical Practice* 2017;123:199–208 doi:10.1016/j.diabres.2016.12.004; PMID:28056431;
66 Gerstein HC, Bosch J, Dagenais GR, et al. Basal insulin and cardiovascular and other outcomes in dysglycemia. *N Engl J Med* 2012;367:319–28 doi:10.1056/NEJMoa1203858; PMID:22686416;
67 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England journal of medicine* 2000;342:1887–92 doi:10.1056/nejm200006223422507;
68 Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286:821–30 ;

69 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *New Engl J Med* 2000;342:1878–86 doi:10.1056/NEJM200006223422506; PMID:10861324;
70 Banerjee R, Prasad V. Are Observational, Real-World Studies Suitable to Make Cancer Treatment Recommendations? *JAMA Netw Open* 2020;3:e2012119 doi:10.1001/jamanetworkopen.2020.12119; PMID:32729916;
71 Concato J. Observational versus experimental studies: what's the evidence for a hierarchy? *NeuroRx the journal of the American Society for Experimental NeuroTherapeutics* 2004;1:341–47 doi:10.1602/neurorx.1.3.341;
72 Dahabreh IJ, Kent DM. Can the Learning Health Care System Be Educated With Observational Data? *JAMA* 2014;312:129–30 doi:10.1001/jama.2014.4364;
73 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;
74 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;
75 Naudet F, Maria AS, Falissard B. Antidepressant response in major depressive disorder: a meta-regression comparison of randomized controlled trials and observational studies. *PLoS ONE* 2011;6:e20811 doi:10.1371/journal.pone.0020811; PMID:21687681;
76 Oliver S, Bagnall AM, Thomas J, et al. Randomised controlled trials for policy interventions: a review of reviews and meta-regression. *Health Technol Assess* 2010;14:1-165, iii doi:10.3310/hta14160; PMID:20338119;
77 Papanikolaou PN, Christidi GD, Ioannidis JPA. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ* 2006;174:635–41 doi:10.1503/cmaj.050873; PMID:16505459;
78 Shikata S, Nakayama T, Noguchi Y, et al. Comparison of effects in randomized controlled trials with observational studies in digestive surgery. *Ann Surg* 2006;244:668–76 doi:10.1097/01.sla.0000225356.04304.bc; PMID:17060757;
79 Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. *PLOS Medicine* 2011;8:e1001026 doi:10.1371/journal.pmed.1001026; PMID:21559325;
80 Kuss O, Legler T, Börgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. *Journal of Clinical Epidemiology* 2011;64:1076–84 doi:10.1016/j.jclinepi.2011.01.005; PMID:21482068;
81 Bhandari M, Tornetta P, Ellis T, et al. Hierarchy of evidence: differences in results between non-randomized studies and randomized trials in patients with femoral neck fractures. *Arch Orthop Trauma Surg* 2004;124:10–16 doi:10.1007/s00402-003-0559-z; PMID:14576955;
82 Edwards JP, Kelly EJ, Lin Y, et al. Meta-analytic comparison of randomized and nonrandomized studies of breast cancer surgery. *Can J Surg* 2012;55:155–62 doi:10.1503/cjs.023410; PMID:22449722;
83 Furlan AD, Tomlinson G, Jadad AAR, et al. Examining heterogeneity in meta-analysis: comparing results of randomized trials and nonrandomized studies of interventions for low back pain. *Spine (Phila Pa 1976)* 2008;33:339–48 doi:10.1097/BRS.0b013e31816233b5; PMID:18303468;
84 Müeller D, Sauerland S, Neugebauer EAM, et al. Reported effects in randomized controlled trials were compared with those of nonrandomized trials in cholecystectomy. *Journal of Clinical Epidemiology* 2010;63:1082–90 doi:10.1016/j.jclinepi.2009.12.009; PMID:20346627;
85 Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *BMJ* 2009;338:b81 doi:10.1136/bmj.b81; PMID:19174434;
86 Dahabreh IJ, Sheldrick RC, Paulus JK, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur. Heart J.* 2012;33:1893–901 doi:10.1093/eurheartj/ehs114;
87 Lonjon G, Boutron I, Trinquart L, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Ann Surg* 2014;259:18–25 doi:10.1097/SLA.0000000000000256;
88 Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev* 2014:MR000034

doi:10.1002/14651858.MR000034.pub2; PMID:24782322;
89 Califf RM, Hernandez AF, Landray M. Weighing the Benefits and Risks of Proliferating Observational Treatment Assessments: Observational Cacophony, Randomized Harmony. *JAMA* 2020;324:625–26 doi:10.1001/jama.2020.13319; PMID:32735313;
90 Rush CJ, Campbell RT, Jhund PS, et al. Association is not causation: treatment effects cannot be estimated from observational data in heart failure. *Eur Heart J* 2018;39:3417–38 doi:10.1093/eurheartj/ehy407; PMID:30085087;
91 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;
92 Klassen SA, Senefeld J, Johnson PW, et al. The Effect of Convalescent Plasma Therapy on COVID-19 Patient Mortality: Systematic Review and Meta-analysis. *medRxiv* 2021 doi:10.1101/2020.07.29.20162917; PMID:33140056;
93 Janiaud P, Axfors C, Schmitt AM, et al. Association of Convalescent Plasma Treatment With Clinical Outcomes in Patients With COVID-19: A Systematic Review and Meta-analysis. *JAMA* 2021;325:1185–95 doi:10.1001/jama.2021.2747; PMID:33635310;
94 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 2013. Available at: http://journals.sagepub.com.docelec.univ-lyon1.fr/doi/pdf/10.1177/0962280211403603.
95 Dickerman BA, García-Albéniz X, Logan RW, et al. Avoidable flaws in observational analyses: an application to statins and cancer. *Nat Med* 2019;25:1601–06 doi:10.1038/s41591-019-0597-x; PMID:31591592;
96 Webster-Clark M, Lund JL, Stürmer T, et al. Reweighting Oranges to Apples: Transported RE-LY Trial Versus Nonexperimental Effect Estimates of Anticoagulation in Atrial Fibrillation. *Epidemiology* 2020;31:605–13 doi:10.1097/EDE.0000000000001230; PMID:32740469;
97 Cain LE, Logan R, Robins JM, et al. When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. *Ann. Intern. Med.* 2011;154:509–15 doi:10.7326/0003-4819-154-8-201104190-00001; PMID:21502648;
98 Franklin JM, Patorno E, Desai RJ, et al. Emulating Randomized Clinical Trials With Nonrandomized Real-World Evidence Studies: First Results From the RCT DUPLICATE Initiative. *Circulation* 2021;143:1002–13 doi:10.1161/CIRCULATIONAHA.120.051718; PMID:33327727;
99 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;
100 European Medicines Agency. Guideline on registry-based studies ;
101 Mathes T, Pieper D. Study design classification of registry-based studies in systematic reviews. *Journal of Clinical Epidemiology* 2018;93:84–87 doi:10.1016/j.jclinepi.2017.09.016; PMID:28951107;
102 Karanatsios B, Prang K-H, Verbunt E, et al. Defining key design elements of registry-based randomised controlled trials: a scoping review. *Trials* 2020;21:552 doi:10.1186/s13063-020-04459-z; PMID:32571382;
103 Li G, Sajobi TT, Menon BK, et al. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? *Journal of Clinical Epidemiology* 2016;80:16–24 doi:10.1016/j.jclinepi.2016.08.003; PMID:27555082;
104 Lauer MS, D'Agostino RB. The randomized registry trial--the next disruptive technology in clinical research? *N Engl J Med* 2013;369:1579–81 doi:10.1056/NEJMp1310102; PMID:23991657;
105 Nicholls SG, Carroll K, Hey SP, et al. A review of pragmatic trials found a high degree of diversity in design and scope, deficiencies in reporting and trial registry data, and poor indexing. *Journal of Clinical Epidemiology* 2021;137:45–57 doi:10.1016/j.jclinepi.2021.03.021; PMID:33789151;
106 Loudon K, Treweek S, Sullivan F, et al. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* 2015;350:h2147 doi:10.1136/bmj.h2147; PMID:25956159;
107 Zuidgeest MGP, Goetz I, Groenwold RHH, et al. Series: Pragmatic trials and real world evidence: Paper 1. Introduction. *Journal of Clinical Epidemiology* 2017;88:7–13 doi:10.1016/j.jclinepi.2016.12.023; PMID:28549929;
108 Ford I, Norrie J. Pragmatic Trials. *N Engl J Med* 2016;375:454–63 doi:10.1056/NEJMra1510059; PMID:27518663;
109 Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. *N Engl J Med* 2017;377:62–70 doi:10.1056/NEJMra1510062; PMID:28679092;
110 Berry SM, Connor JT, Lewis RJ. The platform trial: an efficient strategy for evaluating multiple

treatments. *JAMA* 2015;313:1619–20 doi:10.1001/jama.2015.2316; PMID:25799162;
111 Park JJH, Harari O, Dron L, et al. An overview of platform trials with a checklist for clinical readers. *Journal of Clinical Epidemiology* 2020;125:1–8 doi:10.1016/j.jclinepi.2020.04.025; PMID:32416336;
112 Park JJH, Harari O, Dron L, et al. An overview of platform trials with a checklist for clinical readers. *J Clin Epidemiol* 2020;125:1–8 doi:10.1016/j.jclinepi.2020.04.025;
113 Barker A, Sigman C, Kelloff G, et al. I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy. *Clin Pharmacol Ther* 2009;86:97–100 doi:10.1038/clpt.2009.68;
114 Park JJH, Siden E, Zoratti MJ, et al. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials* 2019;20:572 doi:10.1186/s13063-019-3664-1; PMID:31533793;
115 James ND, Sydes MR, Clarke NW, et al. STAMPEDE: Systemic Therapy for Advancing or Metastatic Prostate Cancer — A Multi-Arm Multi-Stage Randomised Controlled Trial. *Clinical Oncology* 2008;20:577–81 doi:10.1016/j.clon.2008.07.002;
116 LaVange L, Adam SJ, Currier JS, et al. Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV): Designing Master Protocols for Evaluation of Candidate COVID-19 Therapeutics. *Ann Intern Med* 2021;174:1293–300 doi:10.7326/M21-1269;
117 Angus DC, Berry S, Lewis RJ, et al. The REMAP-CAP (Randomized Embedded Multifactorial Adaptive Platform for Community-acquired Pneumonia) Study. Rationale and Design. *Ann Am Thorac Soc* 2020;17:879–91 doi:10.1513/AnnalsATS.202003-192SD; PMID:32267771;
118 Pessoa-Amorim G, Campbell M, Fletcher L, et al. Making trials part of good clinical care: lessons from the RECOVERY trial. *Future Healthc J* 2021;8:e243-e250 doi:10.7861/fhj.2021-0083; PMID:34286192;
119 Dodd LE, Freidlin B, Korn EL. Platform Trials - Beware the Noncomparable Control Group. *N Engl J Med* 2021;384:1572–73 doi:10.1056/NEJMc2102446; PMID:33882210;
120 Park JJH, Harari O, Dron L, et al. An overview of platform trials with a checklist for clinical readers. *J Clin Epidemiol* 2020;125:1–8 doi:10.1016/j.jclinepi.2020.04.025;
121 Collignon O, Gartner C, Haidich A-B, et al. Current Statistical Considerations and Regulatory Perspectives on the Planning of Confirmatory Basket, Umbrella, and Platform Trials. *Clin Pharmacol Ther* 2020;107:1059–67 doi:10.1002/cpt.1804;
122 Choodari-Oskooei B, Bratton DJ, Gannon MR, et al. Adding new experimental arms to randomised clinical trials: Impact on error rates. *Clin Trials* 2020;17:273–84 doi:10.1177/1740774520904346;
123 James ND, Bono JS de, Spears MR, et al. Abiraterone for Prostate Cancer Not Previously Treated with Hormone Therapy. *N Engl J Med* 2017;377:338–51 doi:10.1056/NEJMoa1702900; PMID:28578639;
124 Sydes MR, Parmar MKB, James ND, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials* 2009;10:39 doi:10.1186/1745-6215-10-39; PMID:19519885;
125 James ND, Sydes MR, Clarke NW, et al. Addition of docetaxel, zoledronic acid, or both to first-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *The Lancet* 2016;387:1163–77 doi:10.1016/S0140-6736(15)01037-5;
126 Parmar MKB, Sydes MR, Cafferty FH, et al. Testing many treatments within a single protocol over 10 years at MRC CTU at UCL: Multi-arm, multi stage platform, umbrella and basket protocols. *Clin Trials* 2017;14:451–61 doi:10.1177/1740774517725697;
127 Normand S-LT. The RECOVERY Platform. *New Engl J Med* 2021;384:757–58 doi:10.1056/NEJMe2025674;
128 Horby PW, Mafham M, Bell JL, et al. Lopinavir–ritonavir in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *The Lancet* 2020;396:1345–52 doi:10.1016/S0140-6736(20)32013-4;
129 Horby P, Lim WS, Emberson JR, et al. Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report. *N Engl J Med* 2020 doi:10.1056/NEJMoa2021436; PMID:32678530;
130 Siden EG, Park JJ, Zoratti MJ, et al. Reporting of master protocols towards a standardized approach: A systematic review. *Contemp Clin Trials Commun* 2019;15:100406 doi:10.1016/j.conctc.2019.100406; PMID:31334382;
131 Kapur J, Elm J, Chamberlain JM, et al. Randomized Trial of Three Anticonvulsant Medications for Status Epilepticus. *N Engl J Med* 2019;381:2103–13 doi:10.1056/NEJMoa1905795; PMID:31774955;
132 Azithromycin for community treatment of suspected COVID-19 in people at increased risk of an adverse clinical course in the UK (PRINCIPLE): a randomised, controlled, open-label, adaptive platform trial. *Lancet* 2021;397:1063–74 doi:10.1016/S0140-6736(21)00461-X; PMID:33676597;

133 Korley FK, Durkalski-Mauldin V, Yeatts SD, et al. Early Convalescent Plasma for High-Risk Outpatients with Covid-19. *N Engl J Med* 2021 doi:10.1056/NEJMoa2103784; PMID:34407339;
134 Kaul S. Is the Mortality Benefit With Empagliflozin in Type 2 Diabetes Mellitus Too Good To Be True? *Circulation* 2016;134:94–96 doi:10.1161/CIRCULATIONAHA.116.022537; PMID:27400894;
135 Laptook AR, Shankaran S, Tyson JE, et al. Effect of Therapeutic Hypothermia Initiated After 6 Hours of Age on Death or Disability Among Newborns With Hypoxic-Ischemic Encephalopathy: A Randomized Clinical Trial. *JAMA* 2017;318:1550–60 doi:10.1001/jama.2017.14972; PMID:29067428;
136 Reardon MJ, van Mieghem NM, Popma JJ, et al. Surgical or Transcatheter Aortic-Valve Replacement in Intermediate-Risk Patients. *N Engl J Med* 2017;376:1321–31 doi:10.1056/NEJMoa1700456; PMID:28304219;
137 Lawler PR, Goligher EC, Berger JS, et al. Therapeutic Anticoagulation with Heparin in Noncritically Ill Patients with Covid-19. *N Engl J Med* 2021 doi:10.1056/NEJMoa2105911; PMID:34351721;
138 Effect of anakinra versus usual care in adults in hospital with COVID-19 and mild-to-moderate pneumonia (CORIMUNO-ANA-1): a randomised controlled trial. *Lancet Respir Med* 2021;9:295–304 doi:10.1016/S2213-2600(20)30556-7; PMID:33493450;
139 Angus DC, Derde L, Al-Beidh F, et al. Effect of Hydrocortisone on Mortality and Organ Support in Patients With Severe COVID-19: The REMAP-CAP COVID-19 Corticosteroid Domain Randomized Clinical Trial. *JAMA* 2020;324:1317–29 doi:10.1001/jama.2020.17022; PMID:32876697;
140 Arabi YM, Gordon AC, Derde LPG, et al. Lopinavir-ritonavir and hydroxychloroquine for critically ill patients with COVID-19: REMAP-CAP randomized controlled trial. *Intensive Care Med* 2021;47:867–86 doi:10.1007/s00134-021-06448-5; PMID:34251506;
141 Gordon AC, Mouncey PR, Al-Beidh F, et al. Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19. *N Engl J Med* 2021;384:1491–502 doi:10.1056/NEJMoa2100433; PMID:33631065;
142 Hermine O, Mariette X, Tharaux P-L, et al. Effect of Tocilizumab vs Usual Care in Adults Hospitalized With COVID-19 and Moderate or Severe Pneumonia: A Randomized Clinical Trial. *JAMA Intern Med* 2021;181:32–40 doi:10.1001/jamainternmed.2020.6820; PMID:33080017;
143 Houston BL, Lawler PR, Goligher EC, et al. Anti-Thrombotic Therapy to Ameliorate Complications of COVID-19 (ATTACC): Study design and methodology for an international, adaptive Bayesian randomized controlled trial. *Clin Trials* 2020:1740774520943846 doi:10.1177/1740774520943846; PMID:32815416;
144 U.S. Department of Health, Human Services - Food, Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics - Guidance for Industry 2019 ;
145 Bhatt DL, Mehta C, Drazen JM, et al. Adaptive Designs for Clinical Trials. *New Engl J Med* 2016;375:65–74 doi:10.1056/NEJMra1510061;
146 Porcher R, Lecocq B, Vray M, et al. Les méthodes adaptatives quand et comment les utiliser dans les essais cliniques ? *Thérapie* 2011;66:309–17 doi:10.2515/therapie/2011042;
147 Lachin JM. A review of methods for futility stopping based on conditional power. *Stat Med* 2005;24:2747–64 doi:10.1002/sim.2151;
148 Thall P, Fox P, Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology* 2015;26:1621–28 doi:10.1093/annonc/mdv238;
149 McMurray JJV, Packer M, Desai AS, et al. Angiotensin–Neprilysin Inhibition versus Enalapril in Heart Failure. *New Engl J Med* 2014 doi:10.1056/NEJMoa1409077;
150 McMurray JJV, Ostergren J, Swedberg K, et al. Effects of candesartan in patients with chronic heart failure and reduced left-ventricular systolic function taking angiotensin-converting-enzyme inhibitors: the CHARM-Added trial. *The Lancet* 2003;362:767–71 doi:10.1016/S0140-6736(03)14283-3; PMID:13678869;
151 Boulware DR, Pullen MF, Bangdiwala AS, et al. A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19. *New Engl J Med* 2020 doi:10.1056/NEJMoa2016638;
152 van Eijk RPA, Nikolakopoulos S, Ferguson TA, et al. Increasing the efficiency of clinical trials in neurodegenerative disorders using group sequential trial designs. *Journal of Clinical Epidemiology* 2018;98:80–88 doi:10.1016/j.jclinepi.2018.02.013;
153 Wilson N, Biggs K, Bowden S, et al. Costs and staffing resource requirements for adaptive clinical trials: quantitative and qualitative results from the Costing Adaptive Trials project. *BMC Med* 2021;19:1–17 doi:10.1186/s12916-021-02124-z;
154 Prowell TM, Theoret MR, Pazdur R. Seamless Oncology-Drug Development. *N Engl J Med* 2016;374:2001–03 doi:10.1056/NEJMp1603747;
155 Mulligan MJ, Lyke KE, Kitchin N, et al. Phase I/II study of COVID-19 RNA vaccine BNT162b1 in adults. *Nature* 2020;586:589–93

doi:10.1038/s41586-020-2639-4; PMID:32785213;
156 Walsh EE, Frenck RW, Falsey AR, et al. Safety and Immunogenicity of Two RNA-Based Covid-19 Vaccine Candidates. *N Engl J Med* 2020;383:2439–50 doi:10.1056/NEJMoa2027906; PMID:33053279;
157 Polack FP, Thomas SJ, Kitchin N, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med* 2020;383:2603–15 doi:10.1056/NEJMoa2034577; PMID:33301246;
158 Thomas SJ, Moreira ED, Kitchin N, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine through 6 Months. *N Engl J Med* 2021 doi:10.1056/NEJMoa2110345; PMID:34525277;
159 Hobbs BP, Barata PC, Kanjanapan Y, et al. Seamless Designs: Current Practice and Considerations for Early-Phase Drug Development in Oncology. *JNCI Journal of the National Cancer Institute* 2019;111:118–28 doi:10.1093/jnci/djy196;
160 Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *The Lancet* 2014;383:156–65 doi:10.1016/S0140-6736(13)62229-1;
161 Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *The Lancet* 2009;374:86–89 doi:10.1016/S0140-6736(09)60329-9;
162 Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet* 2014;383:267–76 doi:10.1016/S0140-6736(13)62228-X;
163 Glasziou P, Chalmers I. Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ* 2018:k4645 doi:10.1136/bmj.k4645;
164 Lu J, Xu B, Shen L, et al. Characteristics and Research Waste Among Randomized Clinical Trials in Gastric Cancer. *JAMA Netw Open* 2021;4:e2124760 doi:10.1001/jamanetworkopen.2021.24760; PMID:34533573;
165 The BMJ. Paul Glasziou and Iain Chalmers: Can it really be true that 50% of research is unpublished? - The BMJ 2017. Available at: https://blogs.bmj.com/bmj/2017/06/05/paul-glasziou-and-iain-chalmers-can-it-really-be-true-that-50-of-research-is-unpublished/ Accessed August 22, 2021.
166 Yordanov Y, Dechartres A, Porcher R, et al. Avoidable waste of research related to inadequate methods in clinical trials. *BMJ* 2015;350:h809 doi:10.1136/bmj.h809; PMID:25804210;
167 Downing NS, Aminawung JA, Shah ND, et al. Clinical trial evidence supporting FDA approval of novel therapeutic agents, 2005-2012. *JAMA* 2014;311:368–77 doi:10.1001/jama.2013.282034; PMID:24449315;
168 Bours MJL. A nontechnical explanation of the counterfactual definition of confounding. *J Clin Epidemiol* 2020;121:91–100 doi:10.1016/j.jclinepi.2020.01.021; PMID:32068101;
169 International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use. CHOICE OF CONTROL GROUP AND RELATED ISSUES IN CLINICAL TRIALS E10.
170 INTERNATIONAL COUNCIL FOR HARMONISATION OF TECHNICAL REQUIREMENTS FOR PHARMACEUTICALS FOR HUMAN USE. ADDENDUM ON ESTIMANDS AND SENSITIVITY ANALYSIS IN CLINICAL TRIALS TO THE GUIDELINE ON STATISTICAL PRINCIPLES FOR CLINICAL TRIALS E9(R1).
171 Grein J, Ohmagari N, Shin D, et al. Compassionate Use of Remdesivir for Patients with Severe Covid-19. *N Engl J Med* 2020;382:2327–36 doi:10.1056/NEJMoa2007016; PMID:32275812;
172 Rittberg R, Czaykowski P, Niraula S. Feasibility of Randomized Controlled Trials for Cancer Drugs Approved by the Food and Drug Administration Based on Single-Arm Studies. *JNCI Cancer Spectrum* 2021;5:pkab061 doi:10.1093/jncics/pkab061; PMID:34409254;
173 International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use. Choice of control group and related issues in clinical trials E10 2000.
174 Rosenberg JE, Hoffman-Censits J, Powles T, et al. Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *The Lancet* 2016;387:1909–20 doi:10.1016/S0140-6736(16)00561-4;
175 Powles T, Durán I, van der Heijden MS, et al. Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (IMvigor211): a multicentre, open-label, phase 3 randomised controlled trial. *The Lancet* 2018;391:748–57 doi:10.1016/S0140-6736(17)33297-X; PMID:29268948;
176 Signorovitch JE, Sikirica V, Erder MH, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value Health* 2012;15:940–47 doi:10.1016/j.jval.2012.05.004; PMID:22999145;
177 Signorovitch JE, Wu EQ, Yu AP, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *PharmacoEconomics*

2010;28:935–45 doi:10.2165/11538370-000000000-00000; PMID:20831302;
178 Fox MP, Lash TL. On the Need for Quantitative Bias Analysis in the Peer-Review Process. *Am J Epidemiol* 2017;185:865–68 doi:10.1093/aje/kwx057; PMID:28430833;
179 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;
180 Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43:1969–85 doi:10.1093/ije/dyu149; PMID:25080530;
181 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.
182 Sridhara R, Mandrekar SJ, Dodd LE. Missing data and measurement variability in assessing progression-free survival endpoint in randomized clinical trials. *Clin Cancer Res* 2013;19:2613–20 doi:10.1158/1078-0432.CCR-12-2938; PMID:23669421;
183 Denne JS, Stone AM, Bailey-Iacona R, et al. Missing data and censoring in the analysis of progression-free survival in oncology clinical trials. *Journal of biopharmaceutical statistics* 2013;23:951–70 doi:10.1080/10543406.2013.813515; PMID:23957509;
184 Marks HM. A rational therapeutics: Science and the reform of therapeutics in the United States, 1900-1990. Cambridge England, New York: Cambridge University Press 1997 ISBN:0521581427;
185 Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;356:i6460 doi:10.1136/bmj.i6460; PMID:28057641;
186 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;
187 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;
188 Neuenschwander B, Capkun-Niggli G, Branson M, et al. Summarizing historical information on controls in clinical trials. *Clin Trials* 2010;7:5–18 doi:10.1177/1740774509356002; PMID:20156954;
189 FDA/CBER. Interacting with the Food and Drug Administration on Complex Innovative Clinical Trial Designs for Drugs and Biological Products, Guidance for Industry ;
190 FDA/CDER/mccrayk. Rare Diseases: Common Issues in Drug Development: Guidance for Industry ;
191 EWP. Committee for Medicinal Product for Human Use (CHMP). In: D'Agostino RB, Sullivan L, Massaro J, eds. Wiley Encyclopedia of Clinical Trials. Hoboken, NJ, USA: John Wiley & Sons, Inc 2007.
192 Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis* 1976;29:175–88 doi:10.1016/0021-9681(76)90044-8; PMID:770493;
193 Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Stat Med* 2009;28:3562–66 doi:10.1002/sim.3722; PMID:19735071;
194 Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014;70:1023–32 doi:10.1111/biom.12242; PMID:25355546;
195 Baeten D, Baraliakos X, Braun J, et al. Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial. *The Lancet* 2013;382:1705–13 doi:10.1016/S0140-6736(13)61134-4;
196 Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* 2001;69:89–95 doi:10.1067/mcp.2001.113989; PMID:11240971;
197 Buyse M, Molenberghs G, Burzykowski T, et al. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000;1:49–67 doi:10.1093/biostatistics/1.1.49; PMID:12933525;
198 Buyse M, Molenberghs G, Paoletti X, et al. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials: Statistical evaluation of surrogate endpoints. *Biometrical Journal* 2016;58:104–32 doi:10.1002/bimj.201400049;
199 IQWiG. Validity of surrogate endpoints in oncology. Cologne 2011.
200 Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharm Stat* 2006;5:173–86 doi:10.1002/pst.207; PMID:17080751;
201 Baker SG. Five criteria for using a surrogate endpoint to predict treatment effect based on data from multiple previous trials. *Stat Med* 2018;37:507–18 doi:10.1002/sim.7561; PMID:29164641;
202 Papanikos T, Thompson JR, Abrams KR, et al. A novel approach to bivariate meta-analysis of

binary outcomes and its application in the context of surrogate endpoints 2020.
203 Bujkiewicz S, Jackson D, Thompson JR, et al. Bivariate network meta-analysis for surrogate endpoint evaluation. *Stat Med* 2019;38:3322–41 doi:10.1002/sim.8187; PMID:31131475;
204 Baigent C, Keech A, Kearney PM, et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. *The Lancet* 2005;366:1267–78 doi:10.1016/S0140-6736(05)67394-1; PMID:16214597;
205 Landray MJ, Haynes R, Hopewell JC, et al. Effects of extended-release niacin with laropiprant in high-risk patients. *N Engl J Med* 2014;371:203–12 doi:10.1056/NEJMoa1300955; PMID:25014686;
206 Sabatine MS, Giugliano RP, Keech AC, et al. Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *N Engl J Med* 2017;376:1713–22 doi:10.1056/NEJMoa1615664; PMID:28304224;
207 Mok TSK, Wu Y-L, Kudaba I, et al. Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. *The Lancet* 2019;393:1819–30 doi:10.1016/S0140-6736(18)32409-7;
208 Xie W, Regan MM, Buyse M, et al. Metastasis-Free Survival Is a Strong Surrogate of Overall Survival in Localized Prostate Cancer. *Journal of Clinical Oncology* 2017:JCO.2017.73.9987 doi:10.1200/JCO.2017.73.9987;
209 Naci H, Davis C. Inappropriate use of progression-free survival in cancer drug approvals. *BMJ-BRITISH MEDICAL JOURNAL* 2020;368:m770 doi:10.1136/bmj.m770; PMID:32156802;
210 Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC Med* 2017;15:134 doi:10.1186/s12916-017-0902-9; PMID:28728605;
211 Chen EY, Joshi SK, Tran A, et al. Estimation of Study Time Reduction Using Surrogate End Points Rather Than Overall Survival in Oncology Clinical Trials. *JAMA Internal Medicine* 2019;179:642–47 doi:10.1001/jamainternmed.2018.8351; PMID:30933235;
212 Downing NS, Aminawung JA, Shah ND, et al. Clinical trial evidence supporting FDA approval of novel therapeutic agents, 2005-2012. *JAMA* 2014;311:368–77 doi:10.1001/jama.2013.282034; PMID:24449315;
213 Pease AM, Krumholz HM, Downing NS, et al. Postapproval studies of drugs initially approved by the FDA on the basis of limited evidence: systematic review. *BMJ-BRITISH MEDICAL JOURNAL* 2017;357:j1680 doi:10.1136/bmj.j1680; PMID:28468750;
214 Prasad V, Kim C, Burotto M, et al. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. *JAMA Internal Medicine* 2015;175:1389–98 doi:10.1001/jamainternmed.2015.2829; PMID:26098871;
215 Buyse M, Burzykowski T, Carroll K, et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *J Clin Oncol* 2007;25:5218–24 doi:10.1200/JCO.2007.11.8836; PMID:18024867;
216 Ciani O, Buyse M, Garside R, et al. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. *J Clin Epidemiol* 2015;68:833–42 doi:10.1016/j.jclinepi.2015.02.016; PMID:25863582;
217 Yu T, Hsu Y-J, Fain KM, et al. Use of surrogate outcomes in US FDA drug approvals, 2003-2012: a survey. *BMJ open* 2015;5:e007960 doi:10.1136/bmjopen-2015-007960; PMID:26614616;
218 Ciani O, Buyse M, Garside R, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. *BMJ* 2013;346:f457 doi:10.1136/bmj.f457; PMID:23360719;
219 Ciani O, Buyse M, Drummond M, et al. Time to Review the Role of Surrogate End Points in Health Policy: State of the Art and the Way Forward. *Value in Health* 2017;20:487–95 doi:10.1016/j.jval.2016.10.011;
220 Ciani O, Buyse M, Drummond M, et al. Use of surrogate end points in healthcare policy: a proposal for adoption of a validation framework. *Nature Reviews Drug Discovery* 2016;15:516 doi:10.1038/nrd.2016.81;
221 Ciani O, Davis S, Tappenden P, et al. VALIDATION OF SURROGATE ENDPOINTS IN ADVANCED SOLID TUMORS: SYSTEMATIC REVIEW OF STATISTICAL METHODS, RESULTS, AND IMPLICATIONS FOR POLICY MAKERS. *International Journal of Technology Assessment in Health Care* 2014;30:312–24 doi:10.1017/S0266462314000300;
222 Validity of surrogate endpoints in oncology: IQWiG Reports – Commission No. A10-05.
223 A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *The Lancet* 1996;348:1329–39 doi:10.1016/s0140-6736(96)09457-3; PMID:8918275;
224 Collaborative overview of randomised trials of antiplatelet therapy--I: Prevention of death,

myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. Antiplatelet Trialists' Collaboration. *BMJ* 1994;308:81–106 ; PMID:8298418;

225 Collaborative overview of randomised trials of antiplatelet therapy--II: Maintenance of vascular graft or arterial patency by antiplatelet therapy. Antiplatelet Trialists' Collaboration. *BMJ* 1994;308:159–68 ; PMID:8312766;

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

227 Villar J, Carroli G, Belizán JM. Predictive ability of meta-analyses of randomised controlled trials. *The Lancet* 1995;345:772–76 doi:10.1016/s0140-6736(95)90646-0; PMID:7891492;

228 LeLorier J, Grégoire G, Benhaddad A, et al. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New Engl J Med* 1997;337:536–42 doi:10.1056/NEJM199708213370806; PMID:9262498;

229 Cappelleri JC, Ioannidis JP, Schmid CH, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? *JAMA* 1996;276:1332–38 ; PMID:8861993;

230 Meta-analysis under scrutiny. *The Lancet* 1997;350:675 ; PMID:9291895;

231 Cranney A, Welch V, Adachi JD, et al. Etidronate for treating and preventing postmenopausal osteoporosis. *Cochrane Database Syst Rev* 2001:CD003376 doi:10.1002/14651858.CD003376; PMID:11687195;

232 Granger CB, McMurray JJV, Yusuf S, et al. Effects of candesartan in patients with chronic heart failure and reduced left-ventricular systolic function intolerant to angiotensin-converting-enzyme inhibitors: the CHARM-Alternative trial. *The Lancet* 2003;362:772–76 doi:10.1016/S0140-6736(03)14284-5; PMID:13678870;

233 Yusuf S, Pfeffer MA, Swedberg K, et al. Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved Trial. *The Lancet* 2003;362:777–81 doi:10.1016/S0140-6736(03)14285-7; PMID:13678871;

234 Pfeffer MA, Swedberg K, Granger CB, et al. Effects of candesartan on mortality and morbidity in patients with chronic heart failure: the CHARM-Overall programme. *The Lancet* 2003;362:759–66 doi:10.1016/s0140-6736(03)14282-1; PMID:13678868;

235 Hauser SL, Bar-Or A, Cohen JA, et al. Ofatumumab versus Teriflunomide in Multiple Sclerosis. *N Engl J Med* 2020;383:546–57 doi:10.1056/NEJMoa1917246; PMID:32757523;

236 Salanti G, Del Giovane C, Chaimani A, et al. Evaluating the quality of evidence from a network meta-analysis. *PLoS ONE* 2014;9:e99682 doi:10.1371/journal.pone.0099682; PMID:24992266;

237 Salanti G, Higgins J, Ades AE, et al. Evaluation of networks of randomized trials. *Statistical methods in medical research* 2007 ;

238 Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ (Clinical research ed.)* 2011;343:d4909-d4909 doi:10.1136/bmj.d4909;

239 Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ* 2003;326:472 ;

240 Song F, Glenny AM, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. *Control Clin Trials* 2000;21:488–97 ;

241 Veroniki AA, Tsokani S, White IR, et al. Prevalence of evidence of inconsistency and its association with network structural characteristics in 201 published networks of interventions. *BMC Med Res Methodol* 2021;21:224 doi:10.1186/s12874-021-01401-y; PMID:34689743;

242 Salanti G, Nikolakopoulou A, Efthimou O, et al. Introducing the treatment hierarchy question in network meta-analysis 2020.

243 Pontes C, Fontanet JM, Vives R, et al. Evidence supporting regulatory-decision making on orphan medicinal products authorisation in Europe: methodological uncertainties. *Orphanet J Rare Dis* 2018;13:206 doi:10.1186/s13023-018-0926-z; PMID:30442155;

244 Meekings KN, Williams CSM, Arrowsmith JE. Orphan drug development: an economically viable strategy for biopharma R&D. *Drug Discov Today* 2012;17:660–64 doi:10.1016/j.drudis.2012.02.005; PMID:22366309;

245 Gaddipati H, Liu K, Pariser A, et al. Rare cancer trial design: lessons from FDA approvals. *Clin Cancer Res* 2012;18:5172–78 doi:10.1158/1078-0432.CCR-12-1135; PMID:22718862;

246 Day S, Jonker AH, Lau LPL, et al. Recommendations for the design of small population clinical trials. *Orphanet J Rare Dis* 2018;13:195 doi:10.1186/s13023-018-0931-2; PMID:30400970;

247 Cucherat M, Laporte S. Les résultats faux positifs ou quelle est la probabilité que le traitement soit efficace quand p < 0,05 ? *Thérapie* 2017;72:421–26 doi:10.1016/j.therap.2016.09.021; PMID:28577824;

248 Prasad V, Oseran A. Do we need randomised trials for rare cancers? *Eur J Cancer* 2015;51:1355–57 doi:10.1016/j.ejca.2015.04.015; PMID:25963018;
249 Mitsumoto J, Dorsey ER, Beck CA, et al. Pivotal studies of orphan drugs approved for neurological diseases. *Ann Neurol* 2009;66:184–90 doi:10.1002/ana.21676; PMID:19743448;
250 Hilgers R-D. Design and analysis of clinical trials for small rare disease populations. *J Rare Dis Res Treat* 2016;1:53–60 doi:10.29245/2572-9411/2016/3.1054;
251 Kesselheim AS, Myers JA, Avorn J. Characteristics of clinical trials to support approval of orphan vs nonorphan drugs for cancer. *JAMA* 2011;305:2320–26 doi:10.1001/jama.2011.769; PMID:21642684;
252 Jiao F, Tu W, Jimenez S, et al. Utilizing shared internal control arms and historical information in small-sized platform clinical trials. *Journal of biopharmaceutical statistics* 2019;29:845–59 doi:10.1080/10543406.2019.1657132; PMID:31462131;
253 Ursino M, Stallard N. Bayesian Approaches for Confirmatory Trials in Rare Diseases: Opportunities and Challenges. *Int J Environ Res Public Health* 2021;18 doi:10.3390/ijerph18031022; PMID:33498915;
254 Wandel S, Neuenschwander B, Röver C, et al. Using phase II data for the analysis of phase III studies: An application in rare diseases. *Clin Trials* 2017;14:277–85 doi:10.1177/1740774517699409; PMID:28387537;
255 Finkel RS, Mercuri E, Darras BT, et al. Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. *N Engl J Med* 2017;377:1723–32 doi:10.1056/NEJMoa1702752; PMID:29091570;

Index

A

accelerated approval, 21
adaptive enrichment, 60
ajustement, 27
analyse conditionnée, 27
analyses poolées, 97

B

bayésiens, 50
biais de confusion résiduel, 28
biais de publication, 31
boucle, 117

C

CAPRIE, 95
causal inference, 32
comparaison directe, 100
comparaison externe, 68
comparaisons indirectes, 100
comparateurs externes, 70
Confusion, 27
Continual Reassessment Methods, 60
contrôle historique, 70
contrôles négatifs, 28
critère de substitution, 87

D

data dredging, 30
découverte fortuite, 29, 55
déontologie, 11
distribution à postériori, 51
DSMB, 64

E

emprunt de données historiques, 83
émulation d'un essai cible, 29, 40
enregistrement accéléré, 20, 21
essais adaptatifs, 60
essais basket, 95
essais baskets, 18
essais bayésiens, 50
essais combinés, 60, 65
essais plateformes, 18, 45
essais pragmatiques, 43
essais umbrella, 18
estimand, 32
Éthique, 11
étude de confirmation, 29
Étude exploratoire, 29
études à contrôle externe, 70
études mono-bras, 68
études observationnelles, 25
externally controlled trial, 68

F

facteurs de confusion, 27
FDA, 20
fouille des données, 29

G

group sequential designs, 60
groupe contrôle synthétique, 74
groupes contrôles synthétiques, 70

H

historical data borrowing, 83
hypothèses simplificatrices, 17

I

Inférence causale, 32
infinite loop, 117
intégrité scientifique, 11
intention de traiter, analyse, 32
intervalle de crédibilité, 50

J

Janus, effet, 31

L

Lan et DeMets, 61

M

MAIC, 73
maladies rares, 103
master protocol, 45
matching, 27
medical reversals, 11
Meta Analytic Predictive, 84
méta-analyse, 97
méta-analyse prospective, 98
mono-bras, 68

O

O'Brien et Flemming, 61

P

p hacking, 30
p-hacking reverse, 30
phase 2, 11, 45
phase 3, 20
phases 3, 11
plan d'analyse statistique, 31
pondération, 27
power priors, 84
preuves au-delà de tout doute raisonnable, 20
prior, 51
probabilité à postériori d'efficacité, 50
PROBAST, 76

Q

quantitative bias analysis, 75

R

RCT DUPLICATE, 40
real world data, 24
real world evidence, 24, 43
registres, 42
registry, 42
registry based randomised controlled trials, 42
registry based study, 42
régression multivariée, 27
risque alpha global, 53
ROBINS-I, 29
RWD, 24
RWE, 24

S

SAP, 31
score de propension, 27
seamless, 60, 65
selective reporting, 31
Simulated Treatment Comparison, 73
single-arm study, 68
statistical analysis plan, 31
stratification, 27
surrogate, 87

U

unanchored indirect comparison, 71

V

Vérificationnisme, 12
vibration des résultats, 30