Société Française de Pharmacologie et de Thérapeutique

Groupe de Travail Méthodologie

Livre blanc SFPT

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

Document compagnon

Dossier 19 – Les essais bayésiens

V2, 8 avril 2022

Comité de rédaction et relecture (par ordre alphabétique)

Jean Luc Cracowski

Michel Cucherat

Dominique Deplanque

Behrouz Kassai

Silvy Laporte

Clara Locher

Florian Naudet

Edouard Ollier

Matthieu Roustit

Table des matières

1 Introduction

Deux grands cadres (*framework*) conceptuels s'affrontent en statistique : l'approche classique dite fréquentiste et l'approche bayésienne. Les essais thérapeutiques n'utilisaient pas l'approche bayésienne jusqu'à présent, car il est nécessaire d'injecter dans la méthode une estimation à priori de l'effet du traitement pour obtenir les résultats de l'essai. Ainsi il est nécessaire d'introduire une idée préconçue de ce que l'étude est supposée donner comme résultat. Ce côté arbitraire a été le frein à l'adoption de l'approche. Le risque (cf. 4.4) est alors de fortement conditionner le résultat de l'étude par la croyance qu'a l'investigateur sur la taille de l'effet du traitement évalué au détriment de ce qui est observé dans l'étude elle-même. Cependant l'approche bayésienne a l'avantage de fournir des résultats en termes de probabilités ou d'intervalles de crédibilité qui sont d'interprétation intuitive et il est possible d'utiliser des à priori non informatifs qui n'introduisent aucun arbitraire dans les résultats.

> *L'approche bayésienne donne ses résultats suivant des concepts simples de probabilité et d'intervalle de crédibilité, d'interprétation intuitive contrairement aux concepts de l'approche fréquentiste. Ainsi les résultats bayésiens ont le sens des interprétations erronées courantes des résultats fréquentistes : le bayésien donne la probabilité que le traitement soit sans effet (qui est l'interprétation erronée courante de la p value) et donne l'intervalle de crédibilité dans lequel le vrai effet du traitement à 95% de chance de se trouver (et qui l'interprétation courante erronée de l'intervalle de confiance).*

Les essais randomisés dont l'analyse primaire repose sur une approche bayésienne sont d'abord apparus dans le domaine du dispositif [1, 2] et très timidement dans celui du médicament [3]. La crise de la COVID a fait accélérer le mouvement [4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. En septembre 2021, cette approche reste toutefois largement minoritaire, avec quelques dizaines d'essais pivots publiés ces dernières années. Mais le fait que le mouvement soit amorcé, que les résultats produits se lisent et s'interprètent de manière complètement différente par rapport aux résultats classiques et qu'il existe un risque d'emploi inapproprié de cette approche avec l'utilisation d'un apriori conditionnant le résultat final justifie amplement l'existence de ce chapitre.

2 Introduction à la logique bayésienne

2.1 Limite conceptuelle de la p value et de l'approche fréquentiste

L'approche classique des statistiques (test d'hypothèse, intervalle de confiance, p value) est l'approche fréquentiste. Traditionnellement, les essais cliniques adoptent cette approche et la p value est utilisée pour apprécier la « plausibilité statistique » de l'effet du traitement étudié. Cependant la p value ne donne qu'une appréciation très indirecte de cette plausibilité.

Devant la possibilité que le résultat d'un l'essai soit dû au hasard (cf. dossier 1 – risque alpha global), la question qui vient naturellement à l'esprit est « quelle est alors la probabilité que le traitement soit efficace compte tenu du résultat obtenu ? ». Compte tenu des incertitudes dues aux fluctuations aléatoires qui affectent tout résultat observé, que peut-on conclure à partir de ces résultats quant à la plausibilité de l'effet du traitement ? Finalement, à l'issue de cet essai, quelle est la probabilité que le traitement apporte un bénéfice (ou quelle est la probabilité qu'il n'apporte pas de bénéfice) ? Mais la p value ne va pas répondre à cette question. Elle va apporter une tout autre information qui n'est qu'une appréciation très indirecte de la plausibilité de l'effet du traitement à l'issue des résultats obtenus dans l'essai.

La p value apprécie la probabilité d'obtenir les résultats effectivement observés sous l'hypothèse qu'il n'y a pas d'effet du traitement (l'hypothèse nulle, H0). En fait la p value donne une réponse qui est à l'opposé de la question que l'on se pose. En termes de probabilité conditionnelle, cette question est la probabilité que le traitement « marche » (ou « ne marche pas ») en fonction des résultats obtenus. Avec les notations suivantes : H1 le traitement apporte un bénéfice, H0 il n'apporte pas de bénéfice (H1 = 1- H0) et R les résultats effectivement obtenus dans l'essai, la question naturelle est quelle est la probabilité de H1 si R (ou H0 si R) : Pr (H0 | R). À la place de cela, la p value est connectée à la probabilité des résultats si H0 : Pr (R | H0).

Mais pourquoi donne-t-on une réponse qui est l'inverse de la question posée ? Cela provient du fait que l'on ne sait pas calculer directement à partir des résultats la probabilité conditionnelle Pr(H0|R).

Ainsi on aimerait connaitre Pr(H0|R) (ou Pr(H1|R)) mais seul Pr(R|H0) est accessible. Cela s'avère possible grâce au théorème de Bayes :

$$Pr(A/B) = \frac{Pr(A)\,Pr(B/A)}{Pr(A)\,Pr(B/A) + Pr(\bar{A})\,Pr(B/\bar{A})}$$

Cependant ce calcul implique de connaitre Pr(A), c'est-à-dire Pr(H0) ou Pr(H1). C'est toute la difficulté de cette approche : pour calculer à l'issue de l'essai la probabilité que le traitement apporte un bénéfice, il faut introduire la probabilité à priori que le traitement apporte un bénéfice[1].

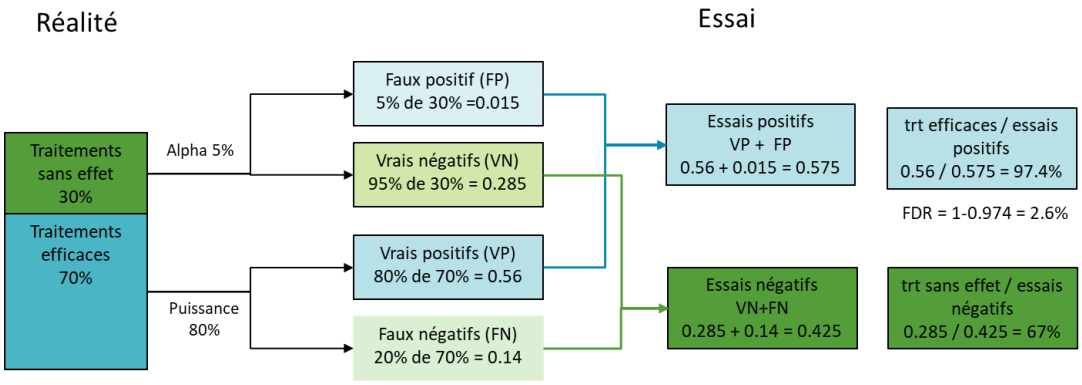
[1] C'est la même situation qu'avec les tests diagnostiques où pour calculer la VPP (probabilité que le patient soit malade en fonction d'un test positif) à partir de la sensibilité (probabilité d'avoir un test positif si le patient est malade) il faut connaitre la prévalence de la maladie à diagnostiquer (la probabilité à priori que le patient soit malade).

2.2 p value et probabilité que le traitement soit efficace

Hormis l'aspect inélégant de la réponse apportée par la p value, elle a aussi une valeur de preuve toute relative en fonction de paramètres autres que le résultat de l'essai : la puissance de l'étude et le degré spéculatif de l'hypothèse thérapeutique testée. Ainsi un « p<0.05 » ne conduit pas toujours au même degré de certitude. Il existe des circonstances où le risque qu'un résultat statistiquement significatif (p<0.05) soit un résultat faussement positif est très élevé [14, 15, 16].

Le calcul du risque qu'un résultat soit faussement positif peut être présenté de manière empirique, sans faire appel au calcul des probabilités, en faisant appel au dénombrement. Pour un ensemble fini de traitements développés, il y a p% de ces traitements qui apporte un bénéfice et 1-p% de traitement sans intérêt. Le risque alpha est la fréquence avec laquelle les essais des traitements sans intérêt donneront un résultat significatif (soit 5%). La puissance est a fréquence avec laquelle les essais des traitements apportant un bénéfice donneront un résultat positif (statistiquement significatif). La fréquence des résultats faussement positifs dans une situation donnée sera le rapport entre le nombre d'essais positifs (statistiquement significatifs) obtenus avec un traitement sans intérêt divisé par le nombre total d'essais positifs obtenus sur l'ensemble des traitements testés (ceux avec bénéfice et ceux sans intérêt).

Les calculs nécessaires sont illustrés avec une approche bilatérale sur le graphique suivant.



Dans cet exemple la probabilité qu'un traitement apporte un bénéfice est de 70%. Sur l'ensemble des traitements testés (un seul essai par traitement, 70% des essais seront donc conduits sur des traitements apportant un bénéfice et 30% sur un traitement sans intérêt. Avec un risque alpha[2] de 5%, 5% de ces essais donneront à tort un résultat positif. Ces résultats seront des faux positifs (FP) et 95% de ces essais seront négatifs et sont des vrais négatifs. Soixante-dix pour cent des essais seront conduits avec un traitement apportant un bénéfice. Si ces essais ont une puissance de 80%, sur l'ensemble des essais réalisés, 80% des 70% conduiront à un résultat vrai positif. La fréquence des essais faux négatifs sera 20% de 70%.

Au total, 57.5% des essais auront produit un résultat apparemment positif (statistiquement significatif). Mais parmi ces essais positifs, seule une partie correspond à un traitement apportant réellement un bénéfice. Ainsi quand l'essai est positif (statistiquement significatif), le traitement

[2] Unilatéral en faveur de l'hypothèse de supériorité, mais ce point est sans importance pour la suite de l'explicitation du raisonnement

apporte un bénéfice que dans 97.4% des cas. Dans cette situation numérique, la probabilité que le traitement apporte un bénéfice quand l'essai est positif est de 97.4%. Le taux de fausse découverte est donc 100%-97.4%= 2.6%.

La probabilité que le traitement soit efficace à l'issue d'un essai concluant (appelée probabilité à postériori) est donc $\frac{VP}{VP+FP}$, avec VP qui dépend de la probabilité à priori que le traitement a un intérêt et de la puissance, tandis que FP dépend du risque alpha et de la probabilité à priori.

Dans l'exemple numérique précédent, la probabilité que le traitement apporte un bénéfice après un essai significatif était très élevée ce qui semble donner de la valeur à p<0.05. Mais ce n'est pas toujours le cas dans d'autres configurations de puissance ou de probabilité à priori. Le tableau suivant explore quelques situations différentes.

Situation	**Probabilité à priori**	**Puissance**	**Alpha (unilatéral)**	**Probabilité de bénéfice réel à la suite d'un essai significatif**
1	70%	80%	2.5%	98.7%
2	70%	50%	2.5%	97.9%
3	70%	50%	30%	79.5%
4	20%	80%	2.5%	88.9%
5	20%	80%	40%	33.3%
6	20%	50%	2.5%	83.3%
7	20%	50%	20%	38.5%

Dans la situation n° 2, la puissance est réduite (50%). Dans ce cas la probabilité que le traitement ait un bénéfice avec un résultat significatif diminue légèrement. Les études peu puissantes sont donc moins probantes même si elles sont significatives. En n° 3, le risque alpha est augmenté par exemple en raison d'une multiplicité non contrôlée. La probabilité à postériori chute alors à moins de 80%, montrant l'importance du contrôle strict du risque alpha.

Pour les autres situations (4 à 7), la probabilité à priori est bien plus faible (seulement 20%), car il s'agit, par exemple, d'un nouveau mécanisme d'action spéculatif, testé pour la première fois. Seul un essai avec une forte puissance et un strict contrôle du risque alpha (n° 4) sera relativement probant, avec une probabilité à postériori assez faible de l'ordre de 90% (par analogie à l'approche fréquentiste standard, le seuil à atteindre pourrait être fixé à 97.5%). Dans toutes les autres situations, la probabilité à postériori est faible, voire très faible. Cela illustre bien la faible valeur probante que peut avoir un essai significatif dans certaines situations.

En conclusion, défaut de puissance et risque alpha non contrôlé enlèvent tout degré de certitude dans un résultat statistiquement significatif.

3 Les essais pivots 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 à postériori d'efficacité, intervalle de crédibilité) et qui correspondent directement à la question du chercheur (quelle est la probabilité que le traitement « marche » compte-tenu des résultats observés ?). Les résultats familiers bien connus comme la p value ou l'intervalle de confiance, qui sont spécifiques de l'approche fréquentiste, ne sont pas estimés avec cette méthode.

En revanche, 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 à 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. document sur les nouvelles méthodologies).

> *N.B. Les essais bayésiens sont à distinguer des études qui utilisent une **méthode bayésienne**. Les méthodes bayésiennes, comme celle utilisée dans de très nombreuses méta-analyses en réseau, sont des méthodes d'estimation particulières reposant sur le principe bayésien. Avec des modèles complexes, il est souvent impossible d'estimer les paramètres d'intérêt avec les approches statistiques classiques (fréquentiste). Cependant, ces problèmes deviennent solutionnable par des méthodes, dites bayésiennes, basées sur les principes de dérivation d'une distribution à posteriori à partir d'une distribution apriori et de données. Le principe bayésien n'est pas utilisé pour lui-même (comme dans les essais bayésiens), mais simplement comme outil pour solutionner un problème calculatoire. De plus ces méthodes s'appuient sur des méthodes de calcul intensif (Monte Carlo Markov Chain) extrêmement flexibles. Les apriori utilisés sont non informatifs, ce qui fait que les résultats produits ne dépendent que des données comme dans le cadre des méthodes d'estimation habituelles (fréquentistes).*

4 Principes des essais bayésiens

4.1 Distribution à posteriori de l'effet

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).

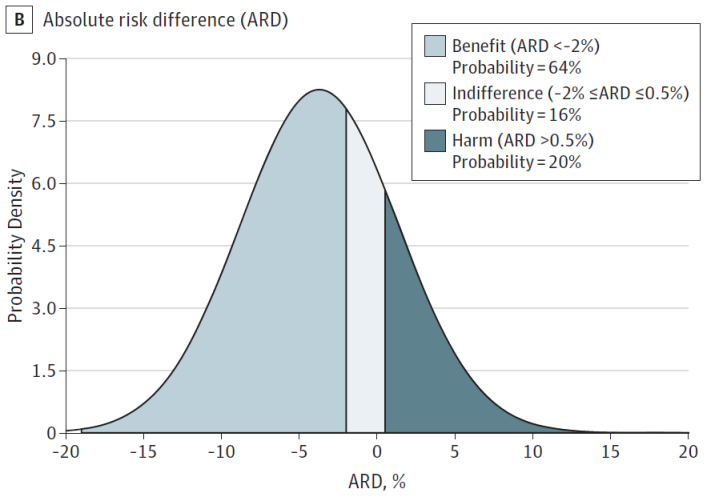


Figure 1 – Exemple de présentation de la distribution de l'effet traitement produite par une approche d'inférence bayésienne
Ici l'effet traitement est mesuré par la différence des risques (ARD). L'absence d'effet correspond à la valeur zéro. Les valeurs négatives correspondent à un bénéfice et les valeurs positives à un effet délétère.

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

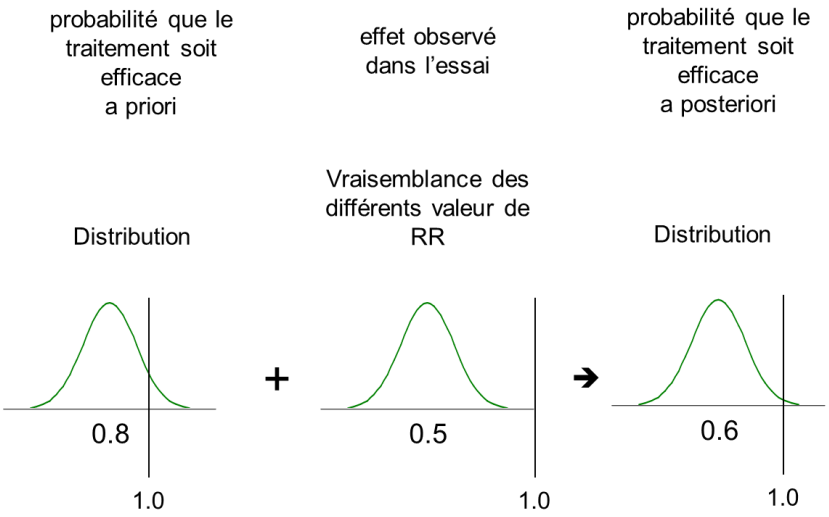


Figure 2 – Illustration du processus de production du résultat à postériori dans l'inférence bayésienne
Le résultat est la combinaison de l'information apportée par l'essai (résultat de l'essai) avec une idée à priori de la distribution de l'effet du traitement (souvent arbitraire)

4.2 Probabilité à posteriori d'efficacité, intervalle de crédibilité

À partir de cette distribution « à postériori » de l'effet traitement, plusieurs mesures sont dérivées, principalement une estimation ponctuelle (médiane ou moyenne), l'intervalle de crédibilité à 95% et la probabilité à postériori 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 3) 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). Elle est dénommée **probabilité à posteriori d'efficacité** (ou de supériorité) (« *posterior probabilities of efficacy*», « *posterior probabilities of superiority to control* »).

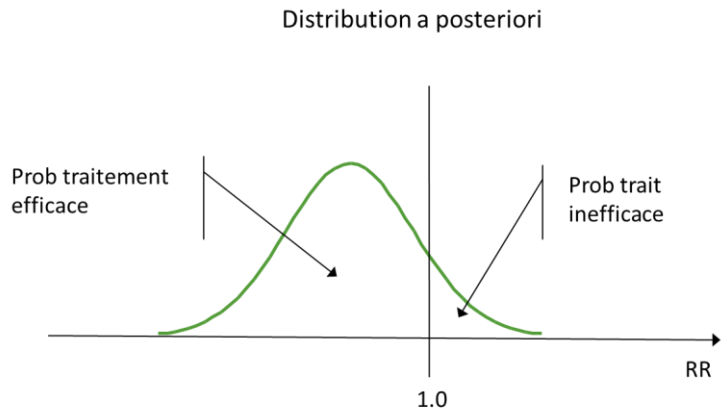


Figure 3 – Illustration du calcul de la probabilité à postériori que le traitement soit efficace
Il s'agit de l'aire sous la courbe en dessous de la valeur de l'absence d'effet. Ici inférieure à 1 comme il s'agit d'une distribution de risque ratio (RR). La distribution représente l'incertitude sur l'estimation 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 (supériorité) à partir de la probabilité à postériori, il est nécessaire de fixer un **seuil de décision**[3], sinon la conclusion serait arbitraire, décidée au cas par cas et très influencée par l'intérêt de conclure au bénéfice du traitement. Sans ce seuil chacun pourra voir dans la valeur obtenue de probabilité à posteriori d'efficacité ce qu'il souhaite voir. 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

[3] Certain essai utilise aussi un seuil de décision pour conclure à la futilité du traitement (et éventuellement à son infériorité).

souvent faite[4], 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.

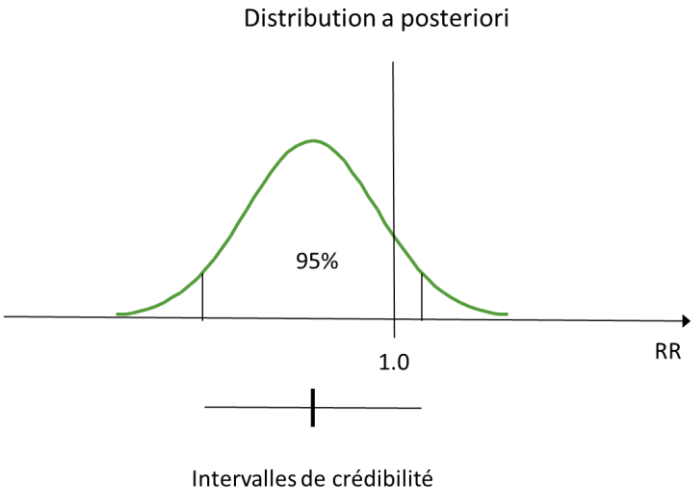


Figure 4 – Illustration de la détermination de l'intervalle de crédibilité

A la place de la probabilité d'être efficace, Il est aussi possible de calculer une probabilité que l'effet du traitement soit supérieur à une valeur de bénéfice cliniquement pertinent minimal (MCID) (cf. Figure 1).

4.3 Interprétation de la probabilité à posteriori d'efficacité

L'interprétation correcte de la probabilité à posteriori d'efficacité nécessite quelques points de repère.

La probabilité de 50% correspond à un traitement dont la distribution de l'effet est centrée sur l'absence d'effet, quelle que soit sa précision (Figure 1). C'est donc la valeur de « base » qui correspond à l'absence d'argument en faveur de l'effet du traitement. Le seuil de décision doit ainsi être bien au-dessus de cette valeur.

[4] 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'autres termes, si l'on réplique (hypothétiquement) un grand nombre de fois le même essai, 95% des IC ainsi générés contiendront la vraie valeur.

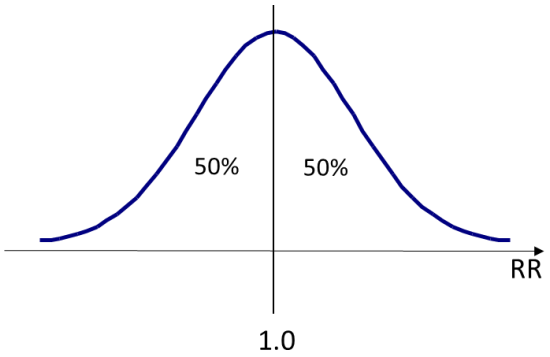


Figure 5 – Probabilité d'être efficace dans le cas d'absence d'effet.
La distribution du risque relatif est alors centrée, par définition, sur la valeur 1. La moitié de l'aire sous la courbe est du côté des risques relatifs inférieurs à 1, en faveur de la supériorité. La probabilité à posteriori d'être efficace (supérieur) d'un traitement sans effet est donc de 50%.

La probabilité à posteriori d'efficacité n'est pas la probabilité qu'un patient bénéficie du traitement. Une probabilité de 99% ne signifie pas que 99% des patients traités bénéficieront du traitement. Il s'agit uniquement de la probabilité que le traitement est un effet non nul. Ensuite l'importance du bénéfice apporté aux patients s'évalue avec la valeur de l'indice d'efficacité utilisé (risque ratio, hazard ratio, etc…) et son intervalle de crédibilité, de la même façon que dans un essai fréquentiste.

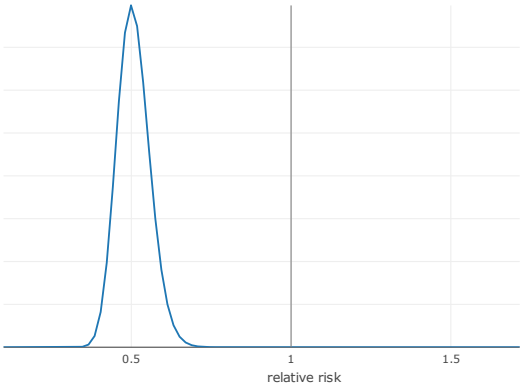
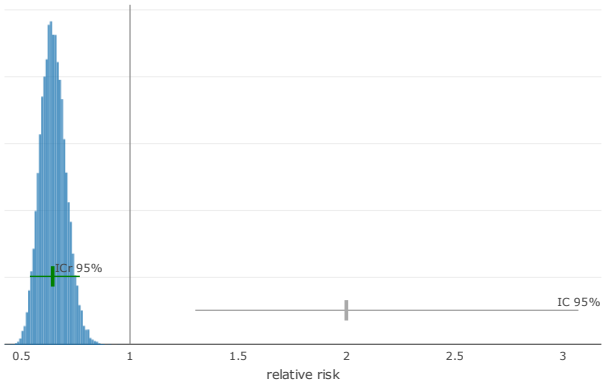
4.4 Dépendance des résultats à l'apriori

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 (à postériori) 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 à 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 à posteriori est conditionné par les données observées.*

Ce risque est illustré par l'exemple présenté Figure 6.

Figure 6 – Illustration de la dépendance des résultats bayésiens à l'apriori

Distribution apriori	**Distribution à posteriori** (et résultat fréquentiste IC 95%)
	
Distribution à 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 à postériori (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 à 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 à postériori 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 [12]

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'à priori, le traitement n'est pas efficace avec une assez forte certitude (distribution étroite). Si malgré cet apriori, 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 [17]).

Il est aussi possible d'utiliser comme « à 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 « à 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.

4.5 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, 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 à postériori 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**. [3]

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 [6].

	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**

5 Études de cas - Exemples de présentation de résultats bayésiens

5.1 Exemple 1

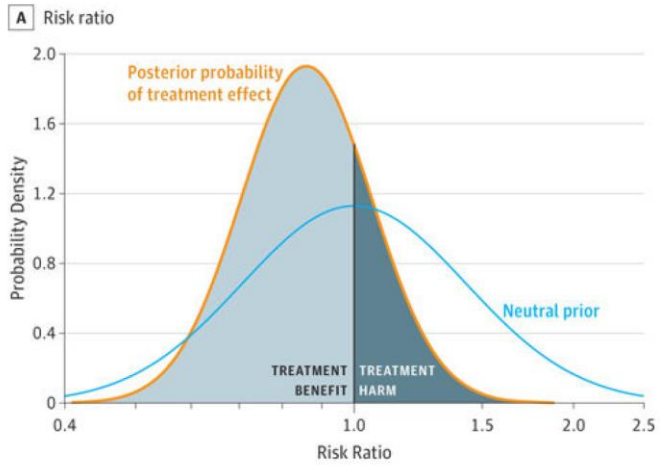
L'essai C3PO a évalué l'intérêt du plasma de patients convalescents dans la COVID -19 [12]. Une approche bayésienne a été utilisée et les résultats sont présentés de la façon suivante.

> *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).*

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

5.2 Exemple 2

Dans cet exemple [2], plusieurs types d'apriori ont été utilisés donnant chacun un résultat différent



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

6 Synthèse des 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 4.4)
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 à priori (cf. ci-dessous section **Erreur ! S ource du renvoi introuvable.**)
Choix arbitraire du seuil de probabilité à postériori pour définir la « positivité » de l'essai	Sans utilisation d'un seuil standard, l'interprétation de la probabilité à postériori 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 à posteriori)
Risque alpha et multiplicité non pris en compte	Définition du seuil de « positivité » de la probabilité a postériori 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 l'ajustement du seuil de « positivité » de la probabilité a postériori afin de contrôler le risque alpha global

AI : analyse intermédiaire

Plusieurs études de méta-épidémiologie ont observé que les éléments clés de la lecture critique des études bayésiennes (ou des études ayant mis en œuvre de méthodes bayésiennes) sont fréquemment mal documentés dans les publications [18, 19]. Des recommandations de publication ont été établies devant ce constant et sont maintenant disponibles comme [20].

7 Méta-recherche

Aucune étude de méta-recherche n'a été trouvée. Le nombre d'essais bayésiens publiés jusqu'à présent (novembre 2021) semble restreint (une recherche non exhaustive exhaustive rapporte quelques dizaines d'essais contrôlés randomisés) [1, 2, 4, 5, 6, 8, 9, 10, 11, 12, 13]. Cette approche a été utilisée à plusieurs reprises pour des traitements de la COVID. Antérieurement la majorité des essais concernait des dispositifs médicaux.

La plupart de ces essais a utilisé des « apriori » non informatif et des seuils de probabilité à postériori 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 [12]	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 [11]	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) [4]	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 [10]	Prior distributions for individual treatment effects were neutral Pas plus de précision dans l'article
REMAP-CAP protocol [7]	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 [6]	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 [5]	The primary model incorporated weakly informative Dirichlet prior distributions for the number of days without organ support

8 Utilisation d'un apriori informatif

L'inférence bayésienne est très souvent présentée avec l'intérêt de pouvoir prendre en compte les connaissances ou les données préexistantes à la réalisation de l'étude.

Dans le champ du développement des médicaments, une phase 3 de confirmation est entreprise à la suite de l'obtention de « bons » résultats lors d'une phase 2 ; les essais randomisés académiques sont réalisés après que des études exploratoires, souvent observationnelles, aient suggéré le potentiel bénéfice d'un traitement. Dans ces situations, il pourrait être envisagé d'utiliser ces résultats préliminaires comme apriori d'un nouvel essai. Ainsi, l'apriori utilisé ne serait pas arbitraire, mais bien dérivé d'observations réelles, considérées comme apportant un « début » d'information sur l'effet du traitement étudié. Le résultat de l'essai bayésien intègrerait ainsi les nouvelles données obtenues dans cette étude avec un effet du traitement apriori estimé lors de la phase 2, entrainant un besoin moindre en termes d'effectif pour la nouvelle étude.

Bien que ne reposant pas sur un choix arbitraire de l'apriori, cette approche soulève néanmoins plusieurs points de discussion.

Actuellement, il est attendu que la preuve de l'intérêt d'un traitement soit apportée par une étude de confirmation, spécialement entreprise pour confirmer les résultats préliminaires. Cette confirmation est indépendante, apportant par elle-même la preuve de l'intérêt du traitement. Cette démarche s'inscrit dans le respect de la démarche hypothético-déductive où la vérification d'une hypothèse porte sur de toutes nouvelles données, complètement différentes que celles ayant fait générer l'hypothèse. Ce principe de confirmation indépendante et de reproductibilité des résultats exploratoires est universellement utilisé en recherche clinique (par exemple dans l'attente d'une réelle validation externe pour un outil pronostique ou diagnostique) et garantit la solidité des résultats obtenus (cf. dossier n° 1).

L'intégration des résultats faisant générer l'hypothèse, et motivant peut-être à elles seules l'intérêt que l'on porte au traitement (dans le cas d'une découverte fortuite par exemple), entraine une rupture de ce principe. La preuve ne sera pas apportée par de toutes nouvelles données indépendantes, mais par un résultat qui dépend en partie des premiers résultats. Le résultat final ne pourra pas être considéré comme une confirmation indépendante et n'apportera pas vraiment la preuve de la reproductibilité des premiers résultats. Il n'y aura pas respect d'un des principes fondamentaux de la démarche scientifique, le principe de reproductibilité des résultats.[5]

De plus, dans cette démarche, la nouvelle étude ne sera pas calibrée (en nombre de sujets par exemple) pour être autosuffisante, mais pour simplement apporter le complément d'information nécessaire pour que le résultat final ait la puissance statistique suffisante. Cette nouvelle étude sera ainsi plus petite que celle qui aurait dû être entreprise pour apporter une confirmation indépendante. Dans le cas où l'apriori est déjà par lui-même très en faveur de l'intérêt du traitement, un nombre très faible de sujets dans la nouvelle étude pourraient conduire à un résultat final concluant. Dans ce cas, la nouvelle étude n'apporterait qu'une faible quantité de nouvelle information et serait encore moins une étude de confirmation, car la part des nouvelles données dans le résultat final de l'essai bayésien sera marginale. D'ailleurs, dans ce cas de figure, il y aurait presque un enjeu pour le développeur à faire une nouvelle étude la plus petite possible pour ne pas risquer d'estomper la tendance initiale par

[5] https://en.wikipedia.org/wiki/Reproducibility

de nouvelles données allant dans l'autre sens (si par exemple les données initiales correspondent à une découverte fortuite sans réalité).

Au niveau de l'enregistrement des nouveaux traitements, ce principe de reproductibilité conduit à exiger deux essais de phase 3 avant de statuer sur l'intérêt clinique du traitement (cf. [21] pour un exemple). Cette exigence n'est cependant pas toujours de mise, en particulier avec les grands essais de morbi mortalité en cardiologie ou en oncologie.

Sans passer par l'inférence bayésienne, il est d'ores déjà possible d'intégrer des résultats antérieurs avec ceux d'une nouvelle étude en procédant à une méta-analyse. Si aucune étude n'est concluante par elle-même et que seul leur regroupement permet d'obtenir un résultat en faveur du bénéfice du traitement, ce type de résultat est considéré actuellement comme insuffisant pour adopter le traitement[6] [23, 24], sauf peut-être dans le cadre d'une méta-analyse prospective. Il apparait aussi que méta-analyse et essai bayésien basé sur une étude précédente sont deux approches similaires [25] conduisant à des résultats comparables. Un petit avantage de la méta-analyse est de documenter l'hétérogénéité entre les données « apriori » et les données produites par la nouvelle étude et de détecter ainsi d'éventuelles discordances (mais l'approche de l'hétérogénéité en méta-analyse a une faible puissance).

L'autre problème que pose la réalisation d'essai bayésien basée sur les résultats d'une étude précédente est celui du risque de **canonisation des résultats faux positifs** [26], produits par une découverte fortuite issue d'études purement exploratoires par exemple. En effet, dans ce cas, le résultat qui déclenchera la démarche peut être simplement artéfactuel, sans réalité (cf. dossier n°1 et section 3.2 du livre blanc). La réalisation de l'essai bayésien sera donc entièrement conditionnée par le résultat obtenu (sans ce résultat, personne n'aurait peut-être fait cette hypothèse[7]). Comme l'apriori est déjà très en faveur du résultat, le résultat de l'essai ne sera que la répercussion assez directe de celui-ci et sera donc automatiquement en faveur du premier résultat. Si l'essai bayésien réalisé est petit, il ne sera pas en mesure de récuser les premières données et le résultat final, à posteriori, sera conforme au premier résultat d'où le processus de canonisation de faux positifs.

Même en dehors des situations à fort risque de découverte fortuite, il arrive fréquemment qu'une étude de confirmation indépendante ne reproduise pas de premiers résultats positifs (cf. la crise de la reproductibilité des résultats [28, 29] [30] de plus en plus discutée dans de nombreux champs de la recherche scientifique[8]). Par exemple, une phase 3 de confirmation sur 2 est un échec alors que les phases 2 précédentes avaient été suffisamment positives pour conduire à la réalisation de ces études de confirmation [31].

Il a aussi été montré empiriquement que les résultats préliminaires de grande taille n'étaient pas prédictifs de la réalité de l'effet et des résultats des études de confirmation entreprises pour les confirmer [32]. Pourtant l'intuition pourrait faire penser que des résultats de grande taille (*very large effect*) ne peuvent pas provenir de biais ou de du hasard, ce qui garantit leur fiabilité. Utilisés comme apriori d'un essai bayésien, ces effets de grandes tailles conditionneront fortement les résultats finals malgré leur faible fiabilité initiale, exposant au même risque de canonisation des faux positifs.

---

[6] ICH E9 mentionne l'intérêt de la méta-analyse pour résumé les résultats des études, mais pas pour apporter la preuve de l'intérêt 22.

[7] Comme dans le cas du canakinumab et la prévention des cancer du poumons [27.]

[8] https://en.wikipedia.org/wiki/Replication_crisis

Le recours à une démarche bayésienne basée sur des apriori informatifs issus d'études initiales entrainerait l'abandon de facto d'un des piliers fondamentaux de la recherche des preuves scientifiques, sans les compenser.

L'approche bayésienne est aussi de plus en plus utilisée pour faire des réanalyses post hoc de résultat exploratoire controversé [33].

> La pratique de réanalyse bayésienne post hoc d'essais non concluants à l'aide d'un apriori informatif commence à se rependre. L'essai TAILOR-PCI n'a pas permis de conclure à l'intérêt du génotypage pour guider le choix de l'antiagrégant plaquettaire après PCI avec un risque relatif sur les MACE de 0.78 (95% CrI, 0.55–1.07). La réanalyse à l'aide d'un apriori informatif débouche sur une probabilité à posteriori de 99% (RR, 0.69 [95% CrI, 0.57–0.84]) [21] utilisée pour conclure à l'intérêt de l'intervention.

L'idée est globalement de faire des analyses de sensibilité (*tipping point analysis*) pour montrer que ce résultat se maintient même en utilisant des aprioris arbitraires très pessimistes (informatifs et contre l'hypothèse soulevée par le résultat). L'objectif de cette approche est de dire que, même si « l'on ne croit pas au résultat », celui-ci doit cependant remporter la conviction, car il résiste aux hypothèses les plus pessimistes. Ce raisonnement est cependant spécieux, car la réanalyse bayésienne ne repose sur aucun élément de confirmation factuelle, l'apriori étant purement arbitraire. Les résultats initiaux de très grande taille non retrouvés par les études de confirmation (cf. supra) seraient robustes dans de telles réanalyses, ce qui montre que seules de nouvelles données indépendantes permettent d'éprouver avec fiabilité des résultats inattendus de grandes tailles.

> Le cas de la réduction de mortalité totale observée avec l'empagliflozine dans le diabète de type 2 dans l'essai EMPAREG OUTCOME. Ce résultat est inattendu, la mortalité étant un critère purement exploratoire et aucun élément du rationnel présenté n'oriente vers une telle possibilité. Une réanalyse bayésienne montre cependant que ce résultat se maintient même sous une hypothèse apriori très pessimiste [33]. Cependant le résultat d'EMPAREG n'a jamais était retrouvé dans 3 autres essais subséquents, avec les 3 autres molécules de la classe, remettant ainsi en cause la conclusion de la réanalyse bayésienne (sauf à faire l'hypothèse que seule l'empagliflozine apporte une réduction de mortalité qui est donc indépendante du mécanisme de classe, hypothèse qui serait alors élaborée entièrement de manière post hoc, uniquement d'après les résultats observés et qui n'a jamais était évoqué apriori avant d'avoir connaissance de ces 4 résultats !).

9 Recevabilité

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

- L'utilisation d'un « à priori » strictement non informatif (voire sceptique)
- L'utilisation d'un seuil pour la probabilité à postériori 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 à 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).

La conclusion à la démonstration du bénéfice clinique repose sur la comparaison de la probabilité à postériori 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%).

Références

1 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;

2 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;

3 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;

4 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;

5 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;

6 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;

7 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;

8 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;

9 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;

10 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;

11 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;

12 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;

13 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;

14 Ioannidis JPA. Why most published research findings are false. *PLOS Medicine* 2005;2:e124 doi:10.1371/journal.pmed.0020124; PMID:16060722;

15 Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ* 2001;322:226–31 doi:10.1136/bmj.322.7280.226; PMID:11159626;

16 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;

17 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;

18 Pibouleau L, Chevret S. Bayesian statistical method was underused despite its advantages in the assessment of implantable medical devices. *Journal of Clinical Epidemiology* 2011;64:270–79 doi:10.1016/j.jclinepi.2010.03.018; PMID:20800443;

19 Sobieraj DM, Cappelleri JC, Baker WL, et al. Methods used to conduct and report Bayesian mixed treatment comparisons published in the medical literature: a systematic review. *BMJ open* 2013;3 doi:10.1136/bmjopen-2013-003111;

20 Kruschke JK. Bayesian Analysis Reporting Guidelines. *Nat Hum Behav* 2021;5:1282–91 doi:10.1038/s41562-021-01177-7; PMID:34400814;

21 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;

22 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).

23 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;

24 Bailar JC. The promise and problems of meta-analysis. *New Engl J Med* 1997;337:559–61 doi:10.1056/NEJM199708213370810; PMID:9262502;

25 Ibrahim JG, Chen M-H, Sinha D. On Optimality Properties of the Power Prior. *Journal of the American Statistical Association* 2003;98:204–13 doi:10.1198/016214503388619229;

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

27 Ridker PM, MacFadyen JG, Thuren T, et al. Effect of interleukin-1β inhibition with canakinumab on incident lung cancer in patients with atherosclerosis: exploratory results from a randomised, double-blind, placebo-controlled trial. *Lancet* 2017;390:1833–42 doi:10.1016/S0140-6736(17)32247-X; PMID:28855077;

28 Absolutely Maybe. Reproducibility Crisis Timeline: Milestones in Tackling Research Reliability - Absolutely Maybe 2016 Accessed March 24, 2022.

29 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;

30 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;

31 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;

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

33 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;