Utilité et pertinence des méthodes fondées sur la valeur de l'information pour les décisions prises dans la vie du dispositif médical

Alexandre Caron

UNIVERSITÉ DE LILLE
ECOLE DOCTORALE BIOLOGIE SANTE

Année : 2021

THESE

Présentée pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE DE LILLE

Spécialité : Santé Publique

Utilité et pertinence des méthodes fondées sur la valeur de l'information pour les décisions prises dans la vie du dispositif médical

Présentée et soutenue publiquement le 9 décembre 2021

Par Alexandre CARON

JURY

Président
Monsieur le Professeur Jérôme WITTWER – Université de Bordeaux
Examinateurs :
Madame le Professeur Lise ROCHAIX – Université Paris 1
Monsieur le Professeur Alain DUHAMEL – Université de Lille
Rapporteurs :
Madame le Docteur Isabelle BORGET – Université Paris-Saclay
Monsieur le Docteur Salah GHABRI – Haute Autorité de Santé
Invités :
Monsieur le Docteur Antoine BENARD – Université de Bordeaux
Directeur de thèse :
Monsieur le Docteur Benoît DERVAUX – Université de Lille

REMERCIEMENTS

A mon comité de suivi individuel,

Madame le Professeur Lise ROCHAIX

Monsieur le Professeur Alain DUHAMEL

Monsieur le Professeur Jérôme WITTWER

Monsieur le Docteur Antoine BENARD

Pour avoir accepté de juger ce travail, pour vos conseils avisés lors des comités de suivi individuel ainsi que pour l'intérêt que vous portez à cette thèse depuis ses débuts, je vous remercie et vous témoigne ma reconnaissance.

A mes rapporteurs de thèse,

Madame le Docteur Isabelle BORGET

Monsieur le Docteur Salah GHABRI

Pour l'intérêt que vous montrez pour ce travail en acceptant d'en être rapporteurs. Soyez assurés de ma reconnaissance et de mes sincères remerciements.

A mon directeur de thèse,

Monsieur le Docteur Benoît DERVAUX

Vous me faites l'honneur de diriger cette thèse. Je vous remercie pour votre enseignement, votre disponibilité et votre soutien durant ces années. A vos côtés, j'ai bien plus appris que ce qui peut être retranscrit dans ce manuscrit. Soyez assuré de mon profond respect.

A tous ceux qui par leur contribution ont rendu ce travail plus facile,

Fannette, Coralie, Maeva, Julien, Patrick, Alain, ainsi que l'ensemble du service de méthodologie et biostatistique, pour votre plaisante compagnie pendant toutes ces années, et vos conseils avisés sur tant de sujets. Travailler à vos côtés me manque.

Vincent, pour avoir accepté d'embarquer avec nous dans cette aventure. Ton expertise et la bonne humeur qui te caractérise ont contribué à bâtir d'improbables ponts avec les autres spécialités de notre équipe de recherche. Soit assuré de ma gratitude.

Mes amis, pour m'avoir permis de recharger les batteries au quotidien et m'avoir soutenu dans les moments difficiles.

Ma famille, et particulièrement **maman,** pour le temps que tu as consacré à la relecture de cette thèse. Je vous remercie pour vos encouragements.

Dori, pour ton soutien indéfectible durant toutes ses années, cette aventure a dû te sembler bien longue… Un page se tourne et une nouvelle s'ouvre avec notre petite **Carmen**. Il me tarde de poursuivre l'écriture de ce nouveau chapitre de notre vie à tes côtés.

Enfin, **Mamie Anne-Marie, Papi Hubert, Adoracion et Fernando**, qui étaient si fiers de moi lorsque j'ai commencé cette thèse. J'aurais tant aimé que vous puissiez en voir la fin…

TABLE DES MATIERES

ABRÉVIATIONS

AHRQ	US Agency for Healthcare Research and Quality
ANSM	Agence nationale de sécurité du médicament et des produits de santé
HAS	Haute Autorité de Santé
CCAM	Classification commune des actes médicaux
CIM10	10^{e} Classification Internationale des Maladies
CNEDiMTS	Commission nationale d'évaluation des dispositifs médicaux et des technologies de santé
CEESP	Commission évaluation économique et de santé publique
CEPS	Comité économique des produits de santé
EVPI	Expected Value of Perfect Information
GHM	Groupe homogène de malades
Hazard	Risque instantané
HR	Hazard ratio (rapport de risques instantanés)
LPPR	Liste des produits et prestations remboursables
NHS	United Kingdom National Health Service
NICE	National Institute for Health and Clinical Excellence
QALYs	Quality Adjusted Life Years

RÉSUMÉ

Introduction. Une information nouvelle possède une valeur si elle permet de réduire le risque de prendre une mauvaise décision. En utilisant une caractérisation bayésienne de l'incertitude, les méthodes fondées sur la valeur de l'information (VoI) estiment cette valeur au regard des conséquences associées à la mauvaise décision. Du fait de ses spécificités, le dispositif médical constitue un domaine de choix pour ces méthodes. L'objectif de cette thèse était, à partir d'exemples sélectionnés, d'illustrer dans quelle mesure les méthodes VoI pouvaient être utiles lors des décisions prises dans la vie du dispositif médical.

Méthodes. Deux cadres d'utilisation décrits dans la littérature ont été utilisés : la priorisation des efforts de recherche et l'optimisation des designs d'études. Ils ont été appliqués à deux temps distincts de l'évaluation du dispositif médical : la détermination des études post-inscription (EPI) demandées à l'occasion de la demande de remboursement et la détermination de la taille d'échantillon lors de l'évaluation précoce de l'utilisabilité réalisée par l'industriel en vue de l'obtention du marquage CE. Ces deux applications ont nécessité de préciser le contexte décisionnel (ensemble des choix, fonction-objectif des parties prenantes, etc.) de développer le modèle d'aide à la décision correspondant à la perspective décisionnelle adoptée et enfin, de caractériser l'incertitude. Les exemples concernaient respectivement les demandes d'EPI pour les endoprothèses aortiques et l'évaluation de l'utilisabilité d'un dispositif innovant d'auto-injection d'adrénaline. Ces exemples ont permis d'identifier les conditions de mise en œuvre des analyses VoI en termes de données requises, de délai, et de complexité.

Résultats. L'analyse menée sur l'exemple des endoprothèses aortiques a requis une re-paramétrisation de l'ensemble du modèle d'aide à la décision existant pour : incorporer les données françaises, intégrer l'incertitude relative à l'ensemble des paramètres en tenant compte des corrélations existantes, et prendre en compte l'hétérogénéité des effets selon les sous-groupes de patients. Sur l'ensemble de la population, l'EVPI élevée quel que soit le critère de jugement justifiait d'envisager l'acquisition de données supplémentaires. Les calculs de l'EVPPI confirmaient l'intérêt d'une d'EPI sur les paramètres d'efficacité, particulièrement sur le long terme. La prise en compte de l'hétérogénéité, originale tant d'un point de vue méthodologique

qu'applicatif, a montré qu'il aurait été pertinent de restreindre ces EPI aux patients jeunes et en bon état général. Ce qui n'était pas envisagé dans les EPI demandées aux industriels.

Dans le cadre des études d'utilisabilité, le développement de novo d'un modèle statistique de la découverte des erreurs d'usage a été rendu nécessaire au regard des insuffisances des modèles existants. En effet, notre approche reposant sur la modélisation de la matrice complète de découverte des erreurs dominait les modèles existants en termes de biais, de consistance et de probabilité de couverture de l'intervalle de confiance, notamment pour des petits échantillons. Cette approche originale a permis d'intégrer la fonction-objectif du décideur dans la détermination de la taille optimale des échantillons, dans une logique de gestion, plutôt que d'évitement du risque. Les tailles d'échantillon estimées étaient plus importantes que dans la littérature (environ 100 participants). L'implémentation de notre méthode est permise par la mise à disposition de l'outil de calcul en libre accès. Plusieurs enseignements émergent de ces applications : la nécessité de réinterroger le modèle d'aide à la décision, le processus de génération des données, et la caractérisation de l'incertitude, l'importance de prendre en compte l'existence de cadres décisionnels et de conclusions distinctes selon les parties prenantes, et de manière générale, des exigences méthodologiques propres conduisant à un accroissement du temps nécessaire à la mise en œuvre.

Conclusion. L'utilité de ces méthodes pour l'accès au marché devra donc être justifiée pour éviter d'imposer une contrainte inutile aux analystes. Enfin, le développement de recommandations par le régulateur est de nature à favoriser une meilleure diffusion dans le contexte français.

ABSTRACT

Introduction. New information has value if it reduces the risk of making a wrong decision. Using a Bayesian characterization of uncertainty, Value of Information (VOI) methods estimate this value in terms of the consequences associated with the wrong decision. Due to its specific features, the medical device is a field of choice for these methods. The objective of this thesis was, based on selected examples, to illustrate to what extent VoI methods could be useful in decisions made during the life of the medical device.

Methods. Two applications of VoI methods described in the literature were used in this thesis: prioritization of research efforts and optimization of study designs. They were applied to two distinct phases of the medical device evaluation: (i) the determination of the post-marketing studies (PMS) requested at the time of the reimbursement application and (ii) the determination of the sample size at the time of the early usability evaluation performed by the industry, in order to obtain the CE mark. These two applications required the specification of the decision-making context (alternatives, objective function of the stakeholders, etc.) to develop the decision model corresponding to the adopted decision-making perspective and finally, to characterize the uncertainty. The examples were respectively (i) PMS for aortic stents and (ii) usability assessment of an innovative adrenaline self-injection device. These examples were used to identify the conditions for implementing VoI analyses in terms of data requirements, timeframe, and complexity.

Results. The value-of-information analysis conducted on the aortic stenting example required a re-parameterization of the existing decision model to: incorporate French data, integrate uncertainty related to all parameters by taking into account existing correlations, and account for heterogeneity of effects across patient subgroups. In the overall population, the high EVPI for any perspective supported the collection of additional data. The EVPPI calculations confirmed the interest of an PMS on the efficacy parameters, particularly over the long term, as requested by the regulator. The consideration of heterogeneity, original both from a methodological and an applicative point of view, showed that it would have been relevant to restrict these PMSs to

young patients in good fitness. This was not suggested by the PMS requested to the manufaturers.

Within the framework of usability studies, the de novo development of a statistical model of the discovery of use errors was made necessary regarding the limits of the existing models. Indeed, our approach based on the modeling of the full discovery matrix dominated the existing models in terms of bias, consistency, and coverage probability of the confidence interval, especially for small samples. This original approach allowed the decision maker's objective function to be integrated into the determination of the optimal sample size, in a management, rather than a risk avoidance logic. The estimated sample sizes were larger than in the literature (about 100 participants). The implementation of our method is made possible since we made the computational tool available in open access.

Several lessons emerge from these two applications: the need to re-interrogate the decision model, the data generation process, and the characterization of uncertainty, the importance of considering the existence of different decision-making perspectives and conclusions depending on the stakeholders, and in general, the methodological requirements leading to an increase in the time needed for carrying out the analysis.

Conclusion. The usefulness of these methods for market access will therefore have to be justified to avoid imposing an unnecessary burden on analysts. Finally, the development of recommendations by the regulator is likely to promote more widespread use in the French context.

Partie 1. ETAT DE L'ART

Cette introduction a pour objectif de proposer un état de l'art des méthodes fondées sur la valeur de l'information. Dans le premier chapitre de cette introduction, nous détaillerons les fondements théoriques de ces méthodes. Nous présenterons ensuite les méthodes fondées sur la valeur de l'information d'un point de vue théorique puis calculatoire. Dans le troisième chapitre, nous envisagerons l'utilisation de ces méthodes dans le contexte du calcul médico-économique en santé et présenterons les principales applications. Enfin, nous présenterons leur articulation avec une méthode alternative de valorisation de l'incertitude, la valeur d'option réelle.

I LES ANALYSES DE LA VALEUR DE L'INFORMATION TROUVENT LEUR FONDATION DANS LA THEORIE DE LA DECISION

La plupart des décisions sont prises en situation d'incertitude sur leurs conséquences. Dans ce contexte, les méthodes fondées sur la valeur de l'information permettent de quantifier la valeur associée à la réduction de l'incertitude décisionnelle par l'acquisition d'une information supplémentaire (Claxton, Cohen, and Neumann 2005). Cette définition, couramment utilisée pour présenter ces méthodes, fait appel à deux paradigmes dont il convient de comprendre les principes pour appréhender au mieux la théorie de la valeur de l'information :

(i) Les méthodes fondées sur la valeur de l'information évaluent l'intérêt de la collecte de données supplémentaires, mais n'ont pas pour vocation de renseigner le choix de la meilleur stratégie ; la théorie sous-jacente est celle de la théorie de la décision.

(ii) La théorie des probabilités, et plus particulièrement le corpus bayésien, fournissent les outils permettant la caractérisation et l'analyse de l'incertitude.

Ces deux paradigmes font l'objet des deux premières sections de ce chapitre qui se conclura par une section clarifiant la typologie des concepts regroupés sous le terme d'incertitude. Précisons d'emblée que le terme incertitude tel qu'employé dans cette thèse (et plus généralement dans la littérature de la valeur de l'information) ne correspond pas exactement à la définition proposée

par Frank Knight (Knight 1921). En effet, dans une situation de connaissance imparfaite sur la réalisation d'évènements futurs, deux situations doivent être distinguées selon que l'on puisse ou non définir de manière objective la probabilité de réalisation des évènements. Lorsque c'est le cas, il s'agit d'une situation de « risque » au sens de Knight. A contrario, lorsque la probabilité de réalisation des évènements ne peut être définie objectivement, il s'agit d'une situation « d'incertitude ». Dans la suite de ce manuscrit, le terme « incertitude » doit être considéré comme recouvrant toutes les situations permettant l'attribution de probabilités aux évènements, qu'elles soient objectives (i.e. « risque » de Knight) ou subjectives (i.e. « incertitude » au sens de Knight).

I.A La théorie de la décision

La théorie de la décision est une théorie statistique constituée d'un corpus de méthodes quantitatives permettant de rationnaliser la prise de décision dans un contexte d'incertitude. Notons d'emblée qu'elle traite également de questions telles que l'interaction entre les décideurs, les décisions complexes, etc. qui dépassent largement le cadre de ce travail. La question du choix en situation d'incertitude, dont il est question ici, constitue néanmoins la pierre angulaire de la théorie de la décision. Après un bref historique, nous décrirons les grands principes la sous-tendant, notamment dans la perspective du calcul de la valeur de l'information. Le paragraphe suivant vise à offrir une vue synthétique de l'histoire ayant conduit à la théorie de la décision moderne, sur laquelle s'appuient les méthodes fondées sur la valeur de l'information.

Historiquement, la théorie de la décision est le fruit de plusieurs siècles de travaux sur la formalisation du hasard et l'étude des jeux. C'est au XVIIème siècle, à l'occasion d'une correspondance avec Pierre de Fermat sur le problème des partis, que Blaise Pascal introduit la notion d'espérance mathématique comme critère de décision dans le contexte des jeux de hasard (Pascal 1654). Cette notion est formalisée quelques années plus tard par Christiaan Huygens dans son traité mathématique sur la question des jeux de chances (Huygens 1657). C'est en 1738, soit près d'un siècle plus tard, que Daniel Bernoulli propose d'utiliser l'espérance de l'utilité du gain comme critère de décision (Bernoulli 1738). Ce critère fut ensuite théorisé par John von Neumann et Oskar Morgenstern pour décrire le comportement rationnel d'un décideur devant une situation de « risque », c'est-à-dire une situation dont les résultats sont incertains

mais peuvent être probabilisés (Von Neumann and Morgenstern 1947).[1] Le décideur respectant les axiomes de cette théorie dispose alors de l'utilité espérée, critère normatif sur lequel il peut baser sa décision (Kast 1993). C'est durant la seconde partie du XXème siècle que la théorie de la décision a connu un essor considérable, notamment sous l'effet de la volonté des entreprises de rationaliser les ressources et moyens de production à leur disposition (Kast 1993). C'est dans leur ouvrage de référence publié en 1961 qu'Howard Raiffa et Robert Schlaifer proposent la théorie moderne de l'utilité (et des probabilités subjectives), ainsi que de nombreux outils d'analyses permettant son application dans de multiples champs de l'économie (Raiffa and Schlaifer 1961). Bien que cet ouvrage proposait déjà les outils nécessaires au calcul, ce n'est que quelques années plus tard que Ronald Howard formalise la théorie de la valeur de l'information (Howard 1966). Enfin, les théories de la décision puis de la valeur de l'information ont largement bénéficié du corpus de la théorie des probabilités (Kolmogorov 1933), et plus particulièrement des déclinaisons bayésiennes (Savage 1954), qui fournissent les outils d'analyse nécessaires à leur mise en œuvre.

La formalisation du problème décisionnel est l'étape initiale de toute application de la théorie de la décision. Elle nécessite la définition exhaustive de trois éléments : (i) l'alternative décisionnelle à laquelle est confronté le décideur – le choix –, (ii) les évènements élémentaires représentant l'incertitude, et (iii) l'ensemble des conséquences possibles synthétisées sous la forme d'un critère de décision ou d'une fonction-objectif (Kast 1993). Un modèle d'aide à la décision plus ou moins complexe permet ensuite de décrire les relations entre ces différents éléments. Dans ce cadre, un arbre de décision permet une première illustration des concepts. Il constitue également un point de départ didactique à l'explication des méthodes fondées sur la valeur de l'information dans le contexte d'une distribution de probabilité discrète (Voir paragraphe II.A page 22). La Figure 1 présente un exemple de modèle d'aide à la décision représenté sous la forme d'un arbre : (i) le choix concerne les décisions A et B, (ii) les évènements élémentaires sont deux états

[1] La théorie de von Neumann et Morgenstern peut également être utilisée dans une situation d'incertitude (au sens de Knight), grâce à l'attribution de probabilités « subjectives » aux résultats incertains (Savage 1954). Néanmoins, ces probabilités ne correspondent pas stricto sensu à la répétition d'une expérience aléatoire, à l'inverse de la situation de « risque ».

du monde dont les probabilités de réalisation (dans le futur) sont connues, et (iii) les conséquences (feuilles terminales) sont valorisées par référence à l'utilité du décideur.[2] La théorie de la décision stipule que le décideur prendra sa décision ex ante en maximisant son utilité espérée (Von Neumann and Morgenstern 1947).

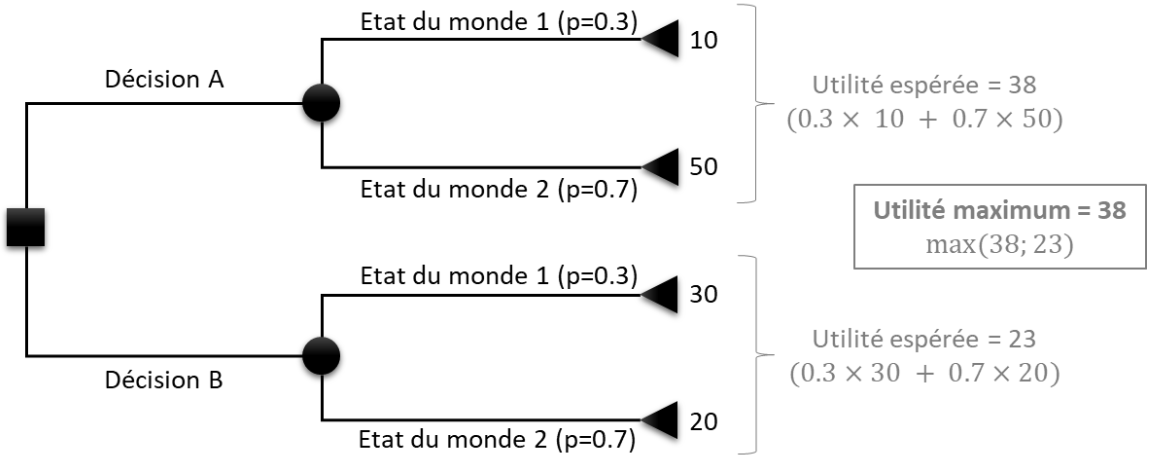


Figure 1 : Arbre de décision représentant un choix en situation d'incertitude. L'utilité associée à chacune des deux décisions possibles (A ou B) est liée à la réalisation de l'un des états du monde (1 ou 2). La probabilité de réalisation de chacun de ces deux états du monde est connue (respectivement 0.3 et 0.7). L'utilité espérée de ces deux décisions correspond à la moyenne des bénéfices pondérés par la probabilité de réalisation. D'après la théorie de la décision, le décideur choisira l'alternative maximisant l'utilité espérée, ici la décision A.

L'exemple proposé appelle plusieurs commentaires importants. Premièrement, les conséquences sont mesurées à l'aide du critère de l'utilité espérée. Pour l'utiliser, le comportement du décideur doit satisfaire aux axiomes de la théorie de von Neumann et Morgenstern. La fonction d'utilité intègre notamment l'aversion pour le risque (Arrow 1965; Pratt 1978). Notons d'emblée que la littérature de la valeur de l'information de manière générale, et en santé plus particulièrement considère, à de rares exceptions près (Eeckhoudt and Godfroid

[2] Les théories de la décision et de la valeur de l'information étant issues du champ de l'économie, les conséquences sont souvent présentées en termes monétaires. Bien que présentant une vertu didactique indéniable, l'utilisation d'un bénéfice monétaire limite souvent l'appropriation des méthodes dans le monde de la santé, où de nombreux critères de jugements sont des résultats de santé (années de vie sauvées, variation d'un paramètre biologique, etc.). Il nous apparait donc plus pertinent d'utiliser l'utilité pour refléter le fait que les méthodes fondées sur la valeur de l'information sont applicables à un large éventail de critères de jugement, facilitant ainsi leur pénétration dans le champ de la santé.

2000), le décideur comme risque neutre. L'hypothèse de neutralité vis-à-vis du risque peut être justifiée pour un acteur économique qui a la capacité de diversifier ses investissements ou de mutualiser les risques (par exemple dans le domaine de la santé, le régulateur). Un acteur économique qui n'a pas cette capacité montre plutôt de l'aversion au risque (par exemple, un industriel avec un portefeuille limité de produits). Deuxièmement, notre exemple est simple et doit être uniquement considéré à visée explicative. En pratique, les états du monde sont plus nombreux et leur probabilité de réalisation dépend de plusieurs variables non-indépendantes, le plus souvent continues, caractérisées par une distribution de probabilité jointe. Enfin, les modèles d'aide à la décision associant chacune des décisions à une fonction-objectif sont souvent complexes (ex : modèle multi-états, arbres décisionnels, etc.). Ces points seront détaillés dans la suite de ce manuscrit.

I.B La caractérisation de l'incertitude et le paradigme bayésien

L'objectif de cette section est de clarifier l'importance des outils bayésiens comme outil de description et d'analyse de l'incertitude, et de souligner leur importance pour la mise en œuvre de la théorie de la décision. En effet, l'appropriation des méthodes fondées sur la valeur de l'information nécessite de comprendre les grands principes du paradigme bayésien. C'est d'autant plus important qu'en recherche médicale, l'analyse bayésienne est souvent présentée, à tort, comme une méthode « subjective » d'analyse des données.

D'un point de vue purement théorique, le théorème de Bayes est un outil mathématique permettant de formaliser l'actualisation du degré d'incertitude concernant un évènement, à la suite de l'acquisition d'une information nouvelle. L'information connue a priori sur la probabilité d'un évènement d'intérêt E (notée $\Pr(E)$), est enrichie grâce à l'observation de nouvelles données (notées D), pour obtenir a posteriori une information mise à jour (notée $\Pr(E|D)$). La probabilité calculée a posteriori (de l'observation de nouvelles données) est la quantité d'intérêt. Elle est calculée grâce à la relation mathématique suivante (théorème de bayes) :

$$\Pr(E|D) = \frac{\Pr(D|E)}{\Pr(D)} \times \Pr(E)$$

Ici, la probabilité a priori est actualisée à l'aide de la vraisemblance de nouvelles données (notée $\Pr(D|E)$), indiquant le degré de compatibilité de la probabilité a priori avec les données observées, et la vraisemblance marginale (notée $\Pr(D)$).

Cette relation se généralise à tout (vecteur de) paramètre(s) θ :

$$p(\theta|\mathrm{Y}) = \frac{p(\mathrm{Y}|\theta) \times p(\theta)}{p(\mathrm{Y})} \propto p(\mathrm{Y}|\theta) \times p(\theta|\mathrm{Y})$$

Avec $p(\theta)$ la distribution a priori et $p(\mathrm{Y}|\theta)$ la distribution d'échantillonnage[3] (conditionnelle à θ). On retrouve la relation de Bayes, où la distribution a posteriori est proportionnelle à la vraisemblance multipliée par la distribution a priori.

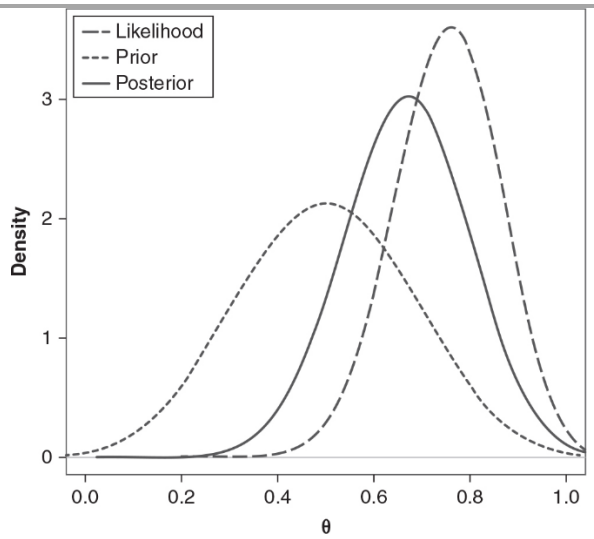


Figure 2 : Représentation graphique du paradigme bayésien (Frey 2018) : la densité de probabilité du paramètre θ a posteriori (posterior) est issue de la multiplication de la loi a priori (prior) et de la vraisemblance des données observées (likelihood) – Sage journals Pre-Approved Permission Requests.

L'une des principales critiques émises à l'encontre de l'approche bayésienne concerne le caractère subjectif de la probabilité a priori, $p(\theta)$. Ce faisant, l'analyse est décrite comme un moyen de montrer à quel point une « opinion » se trouve modifiée par les données observées. A ce titre, elle est parfois considérée comme une approche non-scientifique (Fisher 1996). Il est

[3] Également dénommé vraisemblance, notée L(θ|α).

vrai que la distribution a priori peut être paramétrée de manière très flexible, notamment à partir d'expériences in vitro, de phénomènes observés dans d'autres contextes similaires, de critères de jugement intermédiaires, voire des connaissances cliniques ou avis d'experts (Goodman 1999b). Cependant, la distribution a priori peut être paramétrée sur le résultat d'essais antérieurs ou de méta-analyses. De même, lorsque l'innovation est incrémentale comme pour les dispositifs médicaux, il est possible d'utiliser des données issues de versions antérieures. Dans ce cas, l'argument de la subjectivité devient caduc puisque la distribution a priori peut être considérée comme fondée sur les preuves. Dès lors, les outils de l'analyse bayésienne génèrent des résultats dont le niveau de preuve est identique à celui d'une analyse classique, tout en bénéficiant de « l'arsenal » d'analyse bayésien.

La difficulté réside alors dans la compréhension du paradigme. Pour ce faire, l'analogie avec l'approche fréquentiste dite « classique » est d'une grande utilité (O'Hagan et al. 2005). L'une des différences les plus évidentes entre les deux approches réside dans la notion de probabilité. Du point de vue fréquentiste, une probabilité est le résultat de la convergence de la fréquence empirique lorsque le nombre d'essais tend vers l'infini. Ainsi, cette probabilité (ou tout autre paramètre) est considérée comme une quantité inconnue mais certaine. La notion d'inférence se base ensuite sur le fait que l'échantillonnage de la population pourrait être répété. Cela permet le calcul de l'intervalle de confiance qui est construit de manière à avoir une probabilité de 95% de contenir la vraie valeur du paramètre étudié. En pratique, l'approche fréquentiste ne donne donc aucune information concernant la vraie valeur du paramètre sous-jacent. En effet, l'inférence fréquentiste est déductive : le chercheur part d'une hypothèse et prédit ce qu'il s'attend à observer si elle est vraie.[4] Bien qu'ayant l'avantage de l'objectivité, aucune conclusion ne peut être portée au-delà de l'hypothèse initiale.

A l'inverse de l'approche classique, le paradigme bayésien se fonde sur un raisonnement inductif. Tout paramètre y est considéré comme incertain et une distribution de probabilité lui est associé

[4] Partant de ce constat, Goodman propose un parallèle avec l'inférence en médecine (Goodman 1999a). L'approche déductive est comparée au raisonnement académique de l'étudiant qui énumère la fréquence des symptômes d'une maladie. A l'inverse, l'approche inductive est celle du médecin amené à diagnostiquer une maladie sur la base des symptômes, un « art » beaucoup plus complexe mais également beaucoup plus utile…

pour refléter un degré de connaissance (variable aléatoire). Cette distinction est représentée sur la Figure 3 à l'aide de graphes orientés acycliques (Whittaker 1990; Gilks, Richardson, and Spiegelhalter 1995). L'échantillon observé est considéré comme unique et constitue l'ensemble de l'information disponible sur la distribution du paramètre d'intérêt. L'incertitude, représentée par la distribution de probabilité du paramètre échantillonné, reflète d'une part l'incertitude stochastique (liée à la variabilité du résultat entre deux individus « identiques »), et d'autre part l'incertitude paramétrique (liée à la connaissance imparfaite du paramètre). L'intervalle de confiance peut alors être interprété comme ayant une probabilité de 95% de contenir le paramètre d'intérêt. Cette interprétation intuitive n'est possible que dans un contexte bayésien.

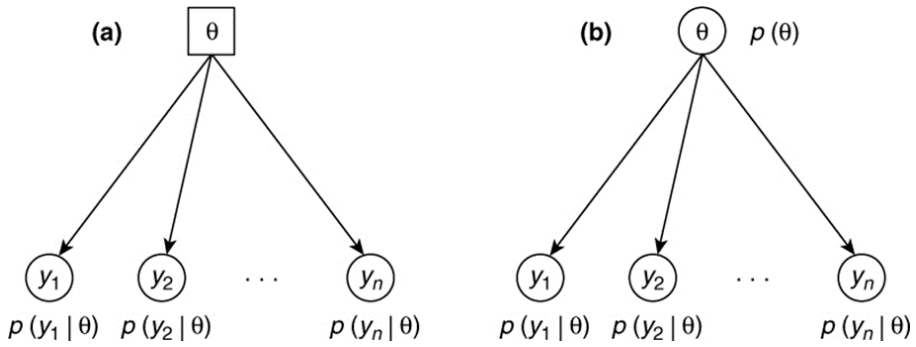


Figure 3 : Graphe orienté acyclique représentant le processus de génération des données du point de vue fréquentiste (a) et bayésien (b). D'un point de vue fréquentiste (a), le paramètre inconnu θ est une quantité fixe qui génère un échantillon d'observations dites indépendantes et identiquement distribuées. D'un point de vue bayésien (b), le paramètre inconnu θ est considéré comme une variable aléatoire (avec sa propre distribution de probabilité $p(\theta)$) et les échantillons observés sont dits échangeables). Dans les deux cas (a et b), les observations sont indépendantes conditionnellement à θ mais la nature des paramètres est distincte. (Osborne 2008) – Sage journals Pre-Approved Permission Requests

Dans les méthodes fondées sur la valeur de l'information, la nature des paramètres et la caractérisation de leur incertitude répond au paradigme bayésien. Ce dernier fournit les outils nécessaires à l'actualisation de la distribution de probabilité lorsque de nouvelles données sont disponibles (Figure 4)

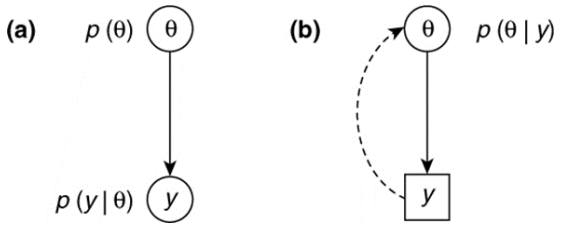


Figure 4 : Graphe orienté acyclique représentant le processus de génération des données (a) et le processus de mise à jour (b). A gauche (a), les données observées sont conditionnelles au paramètre θ. D'un point de vue inférentiel (b), la variable y est observée et contient toute l'information (elle est alors représentée par un carré car elle est considérée comme « certaine »). La distribution a posteriori est obtenue en remontant le processus de génération des données (flèche en traits discontinus) grâce au théorème de bayes. (Osborne 2008) – Sage journals Pre-Approved Permission Requests.

I.C Typologie des concepts regroupés sous le terme d'incertitude

La distribution de probabilité $p(\theta)$ représente l'incertitude dite « paramétrique ». La compréhension des méthodes fondées sur la valeur de l'information nécessite de définir au préalable la grande variété de concepts habituellement regroupés sous le terme « incertitude ». Dans le contexte de l'évaluation économique en santé présenté dans la suite de ce manuscrit, la classification de Briggs et al. (Briggs, Sculpher, and Claxton 2006; Bilcke et al. 2011) constitue la terminologie de référence. Quatre types d'incertitude y sont définis : l'incertitude stochastique, l'incertitude paramétrique, l'hétérogénéité et l'incertitude structurelle.

- L'incertitude stochastique tient au fait qu'un résultat de santé peut être variable entre deux patients, alors même que le paramètre gouvernant ce résultat (probabilité, rapport de cote, etc.) est connu. Ainsi, la durée d'hospitalisation pour appendicectomie peut être différente d'un patient à l'autre, bien que la durée d'hospitalisation moyenne soit connue. L'incertitude stochastique peut être quantifiée par l'écart type de la variable (ou par sa distribution de probabilité), qui reflète la variabilité inter-individuelle entre patients. Le terme « incertitude de premier ordre » est parfois utilisé pour définir l'incertitude stochastique. Il s'agit bien d'un synonyme, mais dont l'utilisation se limite quasi exclusivement à la littérature traitant de la prise de décision médicale (Stinnett and Paltiel 1997).

- L'hétérogénéité correspond quant à elle à la variabilité inter-individuelle qui peut être expliquée à l'aide des caractéristiques des patients. Le résultat de santé moyen peut alors être stratifié sur des caractéristiques tels que l'âge, le sexe, la présence de comorbidités, etc. (ex : table de mortalité stratifiée sur le sexe et l'âge). L'hétérogénéité n'est donc pas une source d'incertitude en tant que telle, mais à l'ère de la médecine personnalisée, sa modélisation est fondamentale pour permettre une décision différenciée selon les caractéristiques de chaque individu (Briggs, Sculpher, and Claxton 2006).

- L'incertitude paramétrique tient au fait que les paramètres gouvernant les résultats de santé ne sont jamais connus avec certitude. Cette incertitude, parfois dénommée « incertitude de second ordre », est bien connue en santé puisque l'intervalle de confiance est une façon de la représenter. Elle peut également être décrite par une distribution de probabilité. La paramétrisation de ses distributions fait l'objet d'une littérature abondante dans le champ de l'analyse médico-économique (Drummond et al. 2015; Briggs, Sculpher, and Claxton 2006). Quelle que soit la métrique utilisée, l'incertitude paramétrique traduit la précision avec laquelle le paramètre a été estimé dans les études disponibles.

- L'incertitude structurelle est introduite par le méthodologiste lorsqu'il construit un modèle. En effet, ces choix en termes de structure (ex : type de modèle de survie) sont susceptibles d'influencer les résultats. Plusieurs outils tels que l'analyse de scénario, l'ajout de paramètres reflétant l'incertitude structurelle, ou le moyennage de modèle, permettent de tenir compte de ce type d'incertitude. Cependant, leur utilisation reste extrêmement marginale en pratique.

L'analogie avec les analyses de régression proposée dans certains ouvrages est utile pour illustrer cette typologie, notamment dans le monde médical (Briggs, Sculpher, and Claxton 2006; Briggs et al. 2012). Dans un modèle de régression, la précision de l'estimation de chacun des paramètres (coefficients) est contenue dans leur erreur standard (incertitude paramétrique) et la variabilité inexpliquée est reflétée par le terme d'erreur de la régression (incertitude stochastique). De même, chaque variable explicative permet d'expliquer une partie de la variabilité du résultat

(hétérogénéité). Enfin, chaque choix de modèle sous-tend une hypothèse sur la relation entre les variables explicatives et la variable à expliquer (incertitude structurelle).

II LES METHODES FONDEES SUR LA VALEUR DE L'INFORMATION

Les méthodes fondées sur la valeur de l'information s'inscrivent dans le prolongement de la théorie de la décision moderne de Raiffa et Schlaifer (Raiffa and Schlaifer 1961). Pour rappel, dans un contexte d'incertitude, la décision est prise en vue de maximiser une fonction-objectif reflétant les préférences du décideur. Cependant, compte tenu de l'incertitude sur les paramètres gouvernant la décision, il existe une probabilité que le choix soit inapproprié. Dès lors, il peut s'avérer pertinent de collecter de nouvelles données pour réduire cette incertitude. Pour répondre à cette problématique, les méthodes fondées sur la valeur de l'information proposent un cadre permettant de quantifier l'incertitude décisionnelle et de valoriser le bénéfice lié à la réduction de cette dernière. Dans une première partie, nous envisageons le cadre (théorique) de la résolution de l'ensemble de l'incertitude entourant les paramètres gouvernant la fonction-objectif. Dans une seconde partie, nous considérons le cadre (pratique) de la résolution de l'incertitude permise par la mise en œuvre d'une vraie étude, dont le design et la taille d'échantillon sont connus.

II.A Valeur attendue d'une information parfaite

Dans l'exemple précédent (Figure 19), nous avions retenu la décision « A » qui maximisait la fonction-objectif du décideur. Pour calculer la valeur attendue d'une information parfaite sur les paramètres dont dépend la fonction-objectif, il convient d'envisager la décision qui serait prise si l'ensemble de l'incertitude était résolu. Cette analyse dite « a posteriori » est purement hypothétique dans la mesure où l'on ne dispose pas de l'information à ce stade.[5] Nous la représentons de nouveau sous la forme d'un arbre de décision (Figure 5). Cette fois, l'utilité est connue conditionnellement à chacun des états du monde et la (meilleure) décision maximisant l'utilité peut être prise. On peut ensuite calculer l'espérance des utilités maximales conditionnellement à chacune des réalisations états du monde.

[5] Elle correspond à « l'analyse terminale » de Raiffa (Raiffa and Schlaifer 1961).

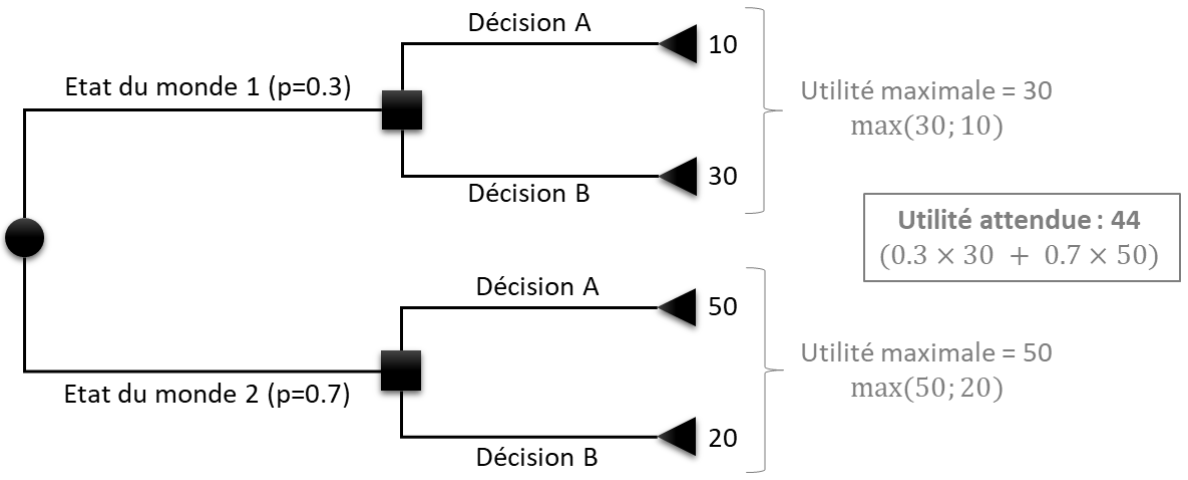


Figure 5 : Arbre de décision représentant l'analyse postérieure du problème décisionnel précédent (Figure 19). L'utilité espérée est maximisée conditionnellement à chacun des états du monde. Si l'on disposait d'une information parfaite concernant la réalisation des états du monde, l'utilité attendue serait alors de 44.

L'analyse a posteriori nous permet de calculer la valeur de l'information parfaite, comme la différence entre l'utilité attendue avec une information sur la réalisation des états du monde et l'utilité espérée en l'absence d'information. Dans notre exemple, la valeur attendue de l'information (parfaite) est donc $44 - 38 = 6$. Cette quantité représente le montant maximal que le décideur est prêt à payer pour disposer d'une information parfaite sur la réalisation des états du monde.

En considérant l'utilité U comme fonction-objectif du décideur, la valeur attendue de l'information parfaite (EVPI, *expected value of perfect information*) se calcule formellement de la manière suivante :

$$\text{EVPI} = E_\theta \left[\max_{\text{d}}\{U(d,\theta)\}\right] - \max_{\text{d}}\{E_\theta[U(d,\theta)]\}$$

Avec $U(d,\theta)$, l'utilité associée à la décision d, compte tenu du (des) paramètre(s) incertain(s) θ. Dans cette formule, on retrouve les deux quantités calculées précédemment : à droite, la maximisation de l'utilité espérée en l'absence d'information, et à gauche, l'utilité attendue en présence d'une information parfaite. Dans notre exemple, la quantité θ correspond à un

paramètre unique (la probabilité de réalisation des états du monde) dont l'incertitude est caractérisée de manière discrète.

L'exemple proposé peut paraître trivial mais son adaptation aux situations complexes, habituellement rencontrées en pratique, est relativement aisée. Tout d'abord, il est courant que le choix se limite à deux alternatives. Dans ce cas, l'analyse de la valeur de l'information est souvent réalisée sur une quantité unique telle que le delta d'utilité. Lors de l'analyse postérieure (Figure 5), nous devons alors tenir compte du fait que la décision A a été prise. Si l'état du monde 2 se réalise, alors la décision A est bien celle maximisant la fonction-objectif. La valeur de l'information nouvelle est donc nulle. A l'inverse, si l'état du monde 1 se réalise, c'est la décision B qui maximise la fonction-objectif. La valeur de l'information correspond ici à un delta de 20, qui une fois pondéré par la probabilité de réalisation de l'état du monde 1 (p=0.3) nous donne une valeur de l'information de 6. Dans ce cadre, la valeur de l'information peut être directement calculée sur l'analyse postérieure grâce à la fonction de perte (moyenne des pertes pondérées par les probabilités de réalisation des états du monde correspondant). Ensuite, l'incertitude entourant le(s) paramètres(s) θ est habituellement représentée par une (ou plusieurs) distribution(s) de probabilité continue(s). Le nombre d'états du monde possibles est donc infini, tout comme les résultats en termes d'utilité. Cette dernière est alors représentée par une distribution de probabilité reflétant l'incertitude sur le(s) paramètre(s) θ. La Figure 6 illustre de manière didactique ce passage d'une distribution discrète à une distribution continue du résultat (ici le bénéfice net incrémental d'une analyse coût-utilité).[6] Sur cette figure, chacune des barres représente la réalisation d'un état du monde, à la manière d'une feuille de l'arbre de décision. Les barres rouges correspondent au coût d'opportunité lié à une décision suboptimale et la hauteur des barres correspond à la probabilité de chacune des issues.

[6] Cette illustration est issue de l'article de Wilson publié en 2015 intitulé « A Practical Guide to Value of Information Analysis » qui décrit les aspects calculatoires des méthodes fondées sur la valeur de l'information.

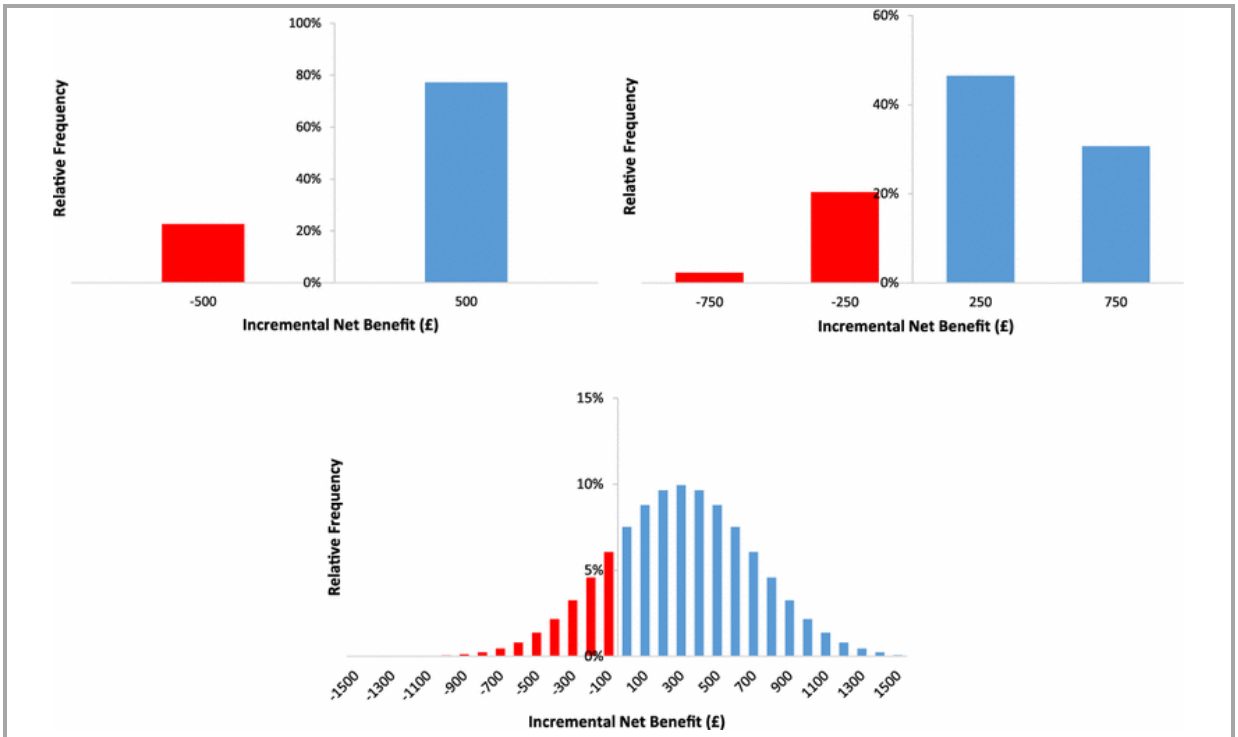


Figure 6 : Evolution de la paramétrisation de l'incertitude entourant le bénéfice net incrémental du cas discret (en haut à droite) vers le cas continu (en bas) - Reproduit avec la permission de Wilson, Copyright Springer Link.

La Figure 6 nous permet ainsi de faire le lien avec le cas continu : la fonction de perte (Figure 7). A partir de cette dernière, on peut directement estimer la valeur de l'information parfaite comme le produit de : (i) la probabilité de prendre une mauvaise décision (en rouge), et (ii) les conséquences associées à cette mauvaise décision (coût d'opportunité, fonction de perte).

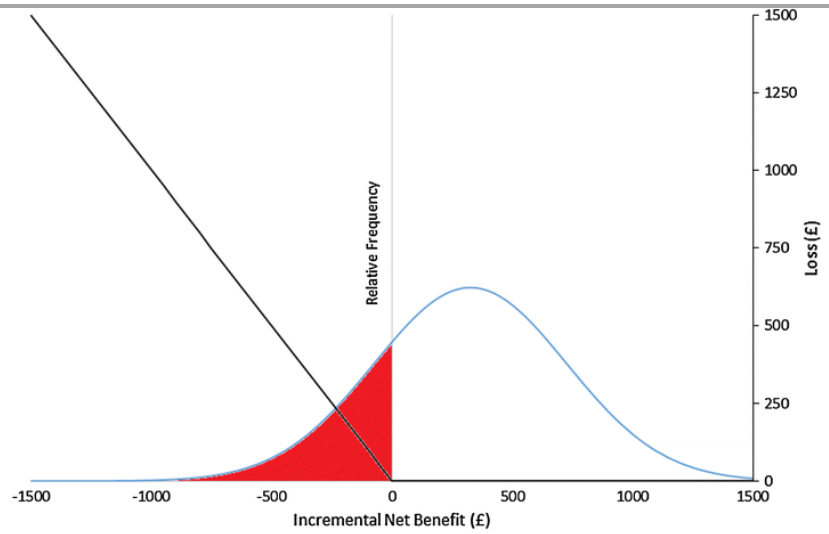


Figure 7 : Distribution du bénéfice net incrémental (BNI, fréquence en abscisse, courbe bleue) et fonction de perte (Loss en abscisse, courbe noire). Lorsque le BNI est positif, la décision prise est optimale et la fonction de perte est nulle. Lorsque le BNI est négatif, la décision prise est suboptimale et la valeur de l'information correspond au produit de la fonction de perte et de sa probabilité (densité de probabilité en rouge). Dans ce cas, la fonction de perte est une droite de coefficient directeur négatif égal au BNI. La valeur de l'information est égale au produit de la fonction de perte par la densité de probabilité. - Reproduit avec la permission de Wilson, Copyright Springer Link.

Enfin, l'exemple ne permet pas d'apprécier l'un des éléments les plus importants : le modèle d'aide à la décision. En effet, bien que la règle de décision soit simple (maximisation de la fonction-objectif du décideur telle que le bénéfice net ou le profit), le résultat n'est pas directement observable. Il doit être construit à l'aide d'un modèle d'aide à la décision, souvent complexe, gouverné par de nombreux paramètres (efficacité, qualité de vie, effets indésirables, coûts, etc.). Ce modèle peut exister indépendamment du calcul de la valeur de l'information, pour la prise de décision en tant que telle. Cependant, la caractérisation de l'incertitude y est généralement insuffisante pour satisfaire à l'exigence méthodologique nécessaire à la réalisation des analyses fondées sur la valeur de l'information. Ce point essentiel sera traité en détail dans le chapitre suivant, dans la mesure où il fait l'objet d'une littérature abondante dans le champ de l'analyse médico-économique.

II.B Valeur attendue de l'information d'échantillonnage

La valeur de l'information parfaite quantifie la valeur associée à la suppression complète de l'incertitude sur l'ensemble des paramètres θ du modèle. Une EVPI non nulle est donc un critère nécessaire mais non suffisant pour justifier la collecte de données supplémentaires. La valeur attendue de l'information d'échantillonnage (EVSI, *expected value of sample information*) permet quant à elle d'estimer la valeur associée à la réduction de l'incertitude grâce à la mise en œuvre d'un design d'étude particulier avec une taille d'échantillon déterminée. Pour ce faire, il est nécessaire d'anticiper les résultats possibles qui seraient obtenus si cette étude était effectivement mise en œuvre, puis de recalculer l'incertitude entourant l'utilité espérée si de nouvelles données étaient disponibles. Le calcul de cette distribution a posteriori fait appel à l'analyse pré-postérieure, une méthode proche de l'analyse postérieure décrite précédemment, mais où la distribution des données $p(y|\theta)$ n'est pas encore disponible et doit donc être prédite ($p(y^*|y)$, Figure 8). De la même manière que précédemment, on peut maximiser la fonction-objectif compte tenu de l'incertitude résiduelle a posteriori de la collecte des données (Figure 9), puis calculer l'espérance sur l'ensemble des échantillons prédits, notés y^* : $E_{y^*}\left[Max_d E_{\theta|y^*} U(d,\theta|y^*)\right]$. Cette quantité peut être comparée à l'utilité espérée en l'absence d'information $\max_d\{E_\theta[U(d,\theta|y^*)]\}$ pour obtenir l'EVSI :

$$\text{EVSI} = E_{y^*}\left[\max_d\{E_{\theta|y^*} U(d,\theta|y^*)\}\right] - \max_d\{E_\theta[U(d,\theta)]\}$$

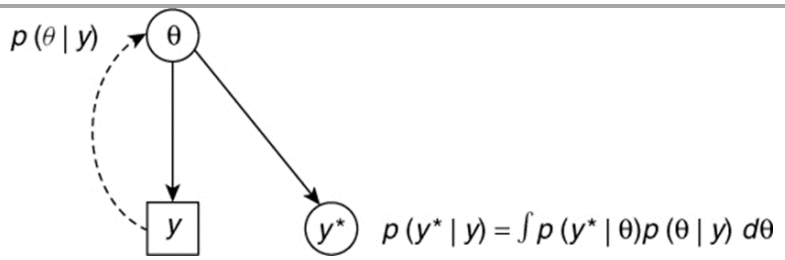


Figure 8 : Graphe orienté acyclique représentant le concept de distribution prédite. Ici, les variables y et y^*sont issues du même processus de génération des données gouverné par le paramètre θ. Une fois que la réalisation $Y = y$ est observée, l'incertitude $p(\theta|y)$ peut être utilisée pour inférer une réalisation possible y^*. (Osborne 2008) – Sage journals Pre-Approved Permission Requests.

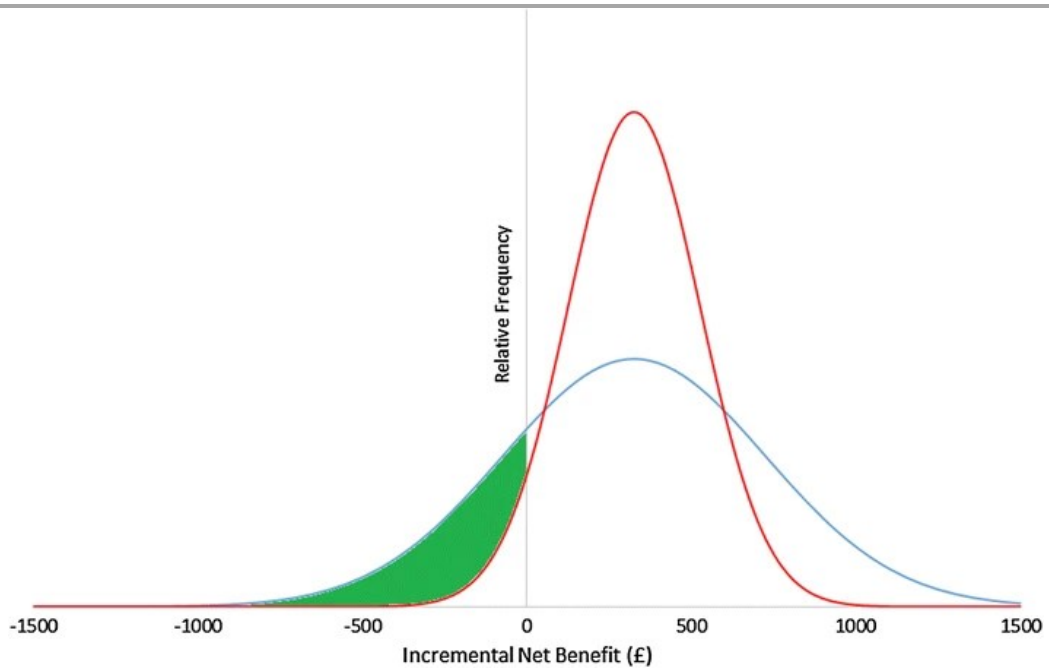


Figure 9 : Réduction de la probabilité de prendre une mauvaise décision (densité de probabilité en vert) comme la différence entre la probabilité a posteriori (courbe rouge, calculée sur la base de la prédiction de possibles données d'échantillonnage) et la probabilité a priori (courbe bleue). La valeur attendue de l'information d'échantillonnage (EVSI) est calculée en multipliant la fonction de perte par la probabilité a posteriori de prendre une mauvaise décision (différence entre la densité de probabilité en vert sur la figure et la densité de probabilité en rouge de la Figure 7)

Dans l'immense majorité des cas, l'EVSI n'est pas calculable analytiquement du fait de la complexité du modèle ou de la caractérisation multivariée de l'incertitude. Les méthodes de calculs disponibles seront détaillées dans la section III.D (page 45).

Enfin, l'EVSI peut être comparée au coût de l'étude pour calculer le bénéfice attendu de l'échantillonnage (ENBS ou ENGS, Expected Net Benefit/Gain of Sampling).

II.C Valeur attendue de l'information partielle

Qu'il s'agisse de l'information parfaite ou de l'information d'échantillonnage, il est souvent utile d'envisager l'acquisition d'une information limitée à certains paramètres du modèle (et non à l'ensemble). La quantité correspondante est la valeur attendue de l'information partielle, parfaite (EVPPI) ou d'échantillonnage (EVPSI). D'un point de vue calculatoire, il faut tenir compte du fait que certains paramètres notés θ_{-i} resteront incertains malgré l'acquisition d'une

information nouvelle. Cette incertitude « résiduelle » est décrite par sa distribution, conditionnelle aux paramètres pour lesquels on disposerait d'une information parfaite, notés θ_i (Brennan et al. 2007). La meilleure décision est celle maximisant la fonction-objectif compte tenu de cette incertitude, et il est nécessaire de modifier le terme de maximisation de l'espérance conditionnelle de l'EVPI et de l'EVSI de la manière suivante :

$$\text{EVPPI} = E_\theta \left[\max_{\text{d}}\left\{E_{\theta_{-i}|\theta_i}\left(U(d,\theta)\right)\right\}\right] - \max_{\text{d}}\{E_\theta[U(d,\theta)]\}$$

$$\text{EVPSI} = E_{y^*} \left[\max_{\text{d}}\left\{E_{\theta_{-i}|y^*}\left(U(d,\theta|y^*)\right)\right\}\right] - \max_{\text{d}}\{E_\theta[U(d,\theta)]\}$$

Pour la valeur d'échantillonnage, cette étape est logique dans la mesure où envisager une nouvelle étude nécessite de restreindre les paramètres étudiés à ceux que l'on peut effectivement collecter. Ainsi, les calculs de la valeur attendue de l'information d'échantillonnage (EVSI) et de l'information partielle d'échantillonnage (EVPSI) sont en général confondus et seul le terme EVSI est rencontré dans la littérature. A l'inverse, le calcul de la valeur de l'information partielle parfaite (EVPPI) est individualisé beaucoup plus clairement. Il constitue une étape préalable (et moins coûteuse d'un point de vue calculatoire) permettant de cibler les paramètres pour lesquels le calcul de l'EVSI est le plus pertinent. Nonobstant, de nombreuses publications se limitent au calcul de l'EVPPI, notamment en raison de la complexité du calcul de l'EVSI (Steuten et al. 2013).

III LA VALEUR DE L'INFORMATION DANS LE CONTEXTE DE L'EVALUATION EN SANTE

Dans ce chapitre, nous envisagerons tout d'abord les modalités d'utilisation des analyses de la valeur de l'information pour orienter les efforts de recherche en santé. Dans une seconde section, nous nous intéresserons aux supports disponibles pour la mise en œuvre du calcul. Dans une troisième section, nous détaillerons l'appropriation des méthodes par deux des principaux acteurs impliqués dans la prise de décision en santé : l'industriel et le régulateur. Enfin, nous nous intéresserons plus particulièrement aux défis posés par le dispositif médical.

III.A Les questions auxquelles les méthodes fondées sur la valeur de l'information sont susceptibles de répondre.

Les méthodes fondées sur la valeur de l'information permettent de répondre aux questions suivantes (Tuffaha, Gordon, and Scuffham 2014a; Eckermann, Karnon, and Willan 2010) :

- L'acquisition de données supplémentaires doit-elle être envisagée ? Et si oui,
- Sur quel(s) paramètre(s) du modèle d'aide à la décision est-il le plus opportun de réduire l'incertitude ?
- Quel est le design d'étude optimal ?

III.A.1 L'acquisition de données supplémentaires doit-elle être envisagée ?

La réponse à cette question est apportée par l'EVPI. Son calcul permet tout d'abord de quantifier le coût de l'incertitude, exprimé en termes d'utilité du décideur (Claxton 2008). L'intérêt d'une étude ne peut être envisagé qu'une fois la population pouvant bénéficier de cette recherche a été déterminée (Box 1). L'importance de cette étape est bien souvent négligée, alors même que son effet de levier est considérable sur l'estimation de la valeur de l'information (Ramos, Maureen, and Al 2015).

Box 1 : Calcul de la valeur de l'information populationnelle

Les quantités telles que l'EVP(P)I et l'EVSI sont calculées à l'échelle individuelle. La valeur de l'information doit ensuite être étendue à l'ensemble de la population susceptible de bénéficier des conclusions d'une nouvelle étude. Plusieurs éléments doivent être pris en compte pour déterminer la taille de cette population « cible » (Fenwick et al. 2020; Koffijberg et al. 2018; Eckermann, Karnon, and Willan 2010) :

- Les données épidémiologiques de prévalence et/ou d'incidence concernant la problématique de santé étudiée. Ces valeurs dépendent de la population couverte par le décideur. A titre d'exemple, les résultats d'une étude financée au niveau national sont susceptibles d'être utiles et utilisés par d'autres pays (Kent et al. 2013). L'utilisation de la population du financeur conduira probablement à une sous-

estimation de la valeur de l'information « globale ». Un modèle de calcul de l'ENBS reflétant la valeur globale de l'information et permettant un échantillonnage transversal a été proposé en réponse à ce constat (Eckermann and Willan 2009). Il s'agit cependant d'une solution avant tout théorique mais dont les applications pratiques sont peu évidentes.

- L'horizon temporel à l'issue duquel les conclusions ne seront plus applicables, à cause d'une innovation technologique, de l'évolution du prix, ou de l'émergence d'une nouvelle information. Une modélisation explicite de cet horizon est souhaitable mais de nombreuses interrogations méthodologiques persistent (Philips, Claxton, and Palmer 2008). Dans ce contexte, certains auteurs proposent une approche historique pour aider à déterminer l'horizon pertinent (Hoyle 2010). En réalité, la détermination d'un horizon temporel est un problème plus large qui concerne également l'analyse sous-jacente (existence de coûts irrécouvrables, question de l'implémentation, etc.) (Salomon, Weinstein, and Goldie 2004).
- Le niveau d'implémentation de la stratégie, présent et futur, est de nature à modifier la taille de la population cible. En effet, il existe toujours un temps de latence entre l'émergence d'une preuve scientifique, sa recommandation et son adoption en pratique courante (Shekelle et al. 2001). La plupart des analyses de la valeur de l'information considèrent implicitement une implémentation parfaite, immédiate et complète, surestimant par la même la population cible (Andronis and Barton 2016a). De nombreux travaux explorent l'articulation entre la valeur de l'information et la problématique de l'implémentation, notamment car le niveau de preuve peut être de nature à impacter cette dernière (Willan and Eckermann 2010; Andronis and Barton 2016b; Fenwick, Claxton, and Sculpher 2008; Grimm, Dixon, and Stevens 2017; Hoomans et al. 2009).

- La valeur de l'information peut se limiter à un sous-groupe de patients pour lesquels l'incertitude est plus importante (Espinoza et al. 2014). Il convient alors de restreindre la population cible à ce sous-groupe de patients.

En parallèle de ces éléments indispensables au calcul de la population cible, certains ajustements peuvent être discutés (Koffijberg et al. 2018) :

- Tout comme les bénéfices et les coûts d'un modèle d'aide à la décision, il peut être pertinent d'actualiser l'EVPI lorsque l'horizon temporel est lointain.
- La prise en compte de l'évolution du prix et de la diffusion du produit peuvent être modélisés, lorsqu'un remboursement est envisagé (Grimm, Dixon, and Stevens 2016).
- Il peut être nécessaire de déduire les patients inclus dans l'étude et ne bénéficiant pas de la stratégie optimale (ex : le bras contrôle d'un essai), dans la mesure où ces derniers ne bénéficieront pas de ses conclusions. Cet ajustement est d'autant plus important que la population cible est petite et/ou que le nombre de patients à inclure est important.
- Si la durée de l'étude envisagée est importante, il faudra tenir compte du délai durant lequel les conclusions de l'étude ne seront pas disponibles et ajuster la population cible en fonction.

Si l'EVPI est nulle ou proche de zéro, le décideur n'aura pas intérêt à rechercher une information supplémentaire puisque, quel que soit le résultat d'une nouvelle étude, sa décision restera identique. A l'inverse, une EVPI non nulle informera le décideur sur la probabilité et les conséquences associées à une éventuelle mauvaise décision. Il pourra alors envisager l'acquisition de données supplémentaires pour réduire cette incertitude en utilisant la borne haute du retour sur investissement comme valeur de référence (Steuten et al. 2013). Néanmoins, certains auteurs jugent que l'EVPI ne constitue pas un critère nécessaire pour envisager la mise en place d'une étude. En effet, il n'est pas possible de la comparer avec le coût d'une étude dont le design est par définition inconnu (Eckermann and Willan 2007a). Nonobstant, les

recommandations proposent de comparer la valeur de l'information parfaite avec les coûts estimés de la mise en œuvre d'une nouvelle étude pour déterminer si l'acquisition de données supplémentaires doit être envisagée (Fenwick et al. 2020).

III.A.2 Sur quel(s) paramètre(s) du modèle d'aide à la décision est-il le plus opportun de réduire l'incertitude ?

Si l'acquisition de données supplémentaires est envisagée, il faut se poser la question des paramètres à étudier en priorité. Cette étape est particulièrement utile lorsqu'un modèle d'aide à la décision synthétise des données de différente nature : efficacité, complications à long terme, coûts, qualité de vie, etc. En effet, une étude unique n'apportera pas une information nouvelle sur l'ensemble de ces paramètres. Le calcul de l'EVPPI permet alors de cibler les paramètres porteurs de la plus grande incertitude décisionnelle. Dans ce cadre, l'approche la plus pertinente est de regrouper les paramètres susceptibles d'être recueillis au sein d'une même étude (Fenwick et al. 2020). D'un point de vue pratique, il s'agit également de ne pas multiplier les calculs d'EVSI, souvent extrêmement complexes et chronophages. De nouveau, rappelons que l'EVPPI reflète la valeur d'une information parfaite et ne constitue pas une condition suffisante à la mise en place de l'étude.

III.A.3 Quel est le design d'étude optimal ?

S'il s'avère qu'un (groupe de) paramètre(s) possède une EVPPI anticipée supérieure au coût d'une (de) potentielle(s) étude(s) complémentaire(s), alors le calcul de l'EVSI peut être envisagé pour quantifier la valeur attendue de l'information, compte tenu du design de l'étude envisagé : type, durée de suivi, taille d'échantillon, etc. L'EVSI pourra ensuite être directement comparée au coût de l'étude proposée grâce à l'ENBS, en tenant compte : des coûts fixes de mise en place de l'étude, des coûts variables liés au nombre de participants inclus, et du coût d'opportunité que supportent les patients recevant une stratégie suboptimale, soit dans le cadre de l'étude, soit dans l'attente des résultats (McKenna and Claxton 2011). A cet effet, un outil permettant de formaliser les coûts d'étude a été développé dans le contexte du financement public des efforts de recherche aux Pays-Bas (van Asselt et al. 2018). Le choix du design le plus efficient se basera ensuite sur la maximisation de l'ENBS parmi un éventail de designs possibles (Fenwick et al. 2020;

McKenna and Claxton 2011). Pour un design donné, on peut par exemple déterminer la taille d'échantillon optimale en comparant l'ENBS pour différentes tailles (Figure 10).

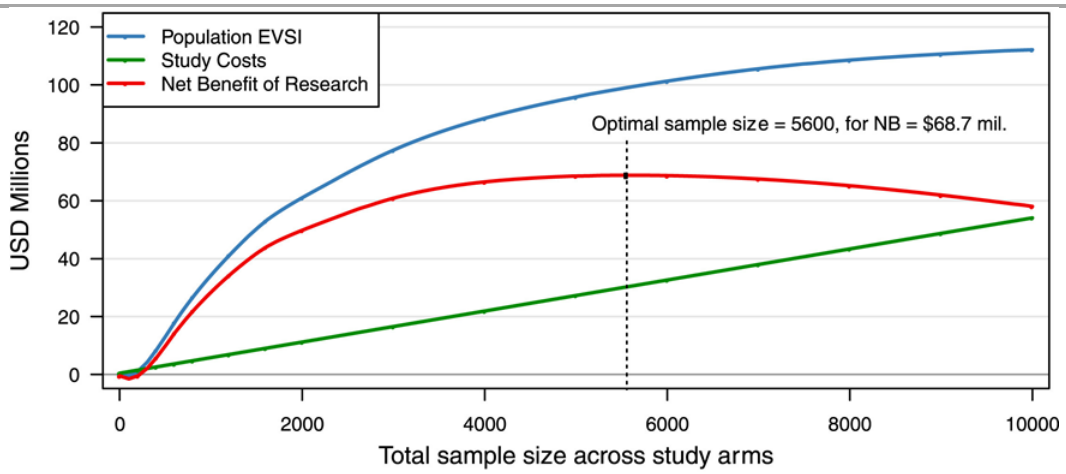


Figure 10 : Exemple d'analyse de la taille optimale d'échantillon pour un essai randomisé comparant la thrombolyse avec le traitement existant de l'accident vasculaire cérébral ischémique dans l'intervalle de 4,5 à 6 heures suivant le début des symptômes. Le bénéfice net attendu de l'échantillonnage (ENBS, en rouge) est calculé comme la différence entre la valeur attendue de l'information d'échantillonnage (EVSI, en bleu) et les coûts de l'étude (en vert) pour une taille d'échantillon comprise entre 0 et 10000 patients. La taille optimale de l'échantillon est estimée à 5600 patients pour un bénéfice attendu de 68,7 millions de dollars US. (Soeteman, Menzies, and Pandya 2017)

III.B Quels supports permettent la réalisation des analyses de la valeur de l'information ?

III.B.1 Un modèle d'aide à la décision

La prise de décision repose sur les résultats fournis par un modèle d'aide à la décision synthétisant l'ensemble des connaissances disponibles sur une question de santé (Briggs, Sculpher, and Claxton 2006; Lumley 2002). Plusieurs alternatives (thérapeutiques, diagnostiques, etc.) sont à la disposition du décideur, qui base son choix sur la maximisation d'une fonction-objectif, tel que le bénéfice net. Les paramètres gouvernant la décision sont néanmoins incertains et la caractérisation de cette incertitude est une étape fondamentale. Elle permet au décideur d'évaluer : l'incertitude entourant sa décision, l'intérêt d'acquérir des données supplémentaires, voire d'envisager le report de sa décision (Drummond et al. 2015). D'un point de vue méthodologique, la caractérisation de l'incertitude garantit une estimation correcte de la fonction-objectif lorsque la relation entre les paramètres et le résultat n'est pas linéaire (ex : modèle de Markov, arbre de décision avec paramètres corrélés, etc.). Dans ce contexte, l'évaluation médico-économique s'apparente conceptuellement à l'analyse décisionnelle telle que modélisée par la théorie de la décision.

L'analyse de sensibilité probabiliste est le gold standard en termes de caractérisation de l'incertitude paramétrique (Saltelli 2002). Pour chacun des paramètres du modèle d'aide à la décision, l'incertitude est calibrée conformément aux données disponibles dans la littérature, puis embarquée dans le modèle sous la forme d'une distribution de probabilité. La paramétrisation de ces distributions fait l'objet de publications et guidelines de référence (Briggs, Sculpher, and Claxton 2006; Briggs et al. 2012) : une probabilité pourra être modélisée par une loi béta, une durée de séjour par une loi gamma, un rapport de risque instantané par une loi log-normale, etc. Dans ce cadre, il est recommandé de modéliser les paramètres corrélés par leur distribution conjointe (par exemple grâce à la matrice de covariance des coefficients d'un modèle de régression multivarié) afin d'éviter de surestimer l'incertitude. Une fois caractérisée, l'incertitude est propagée dans le modèle par simulations de Monte Carlo (répétition d'un tirage aléatoire dans la distribution de chacun des paramètres du modèle). A l'issue de l'analyse de sensibilité probabiliste, on dispose de la fonction-objectif pour chacune des stratégies (RDCR ou

du bénéfice net) que l'on peut moyenner pour obtenir un estimateur non biaisé du résultat (Figure 11).

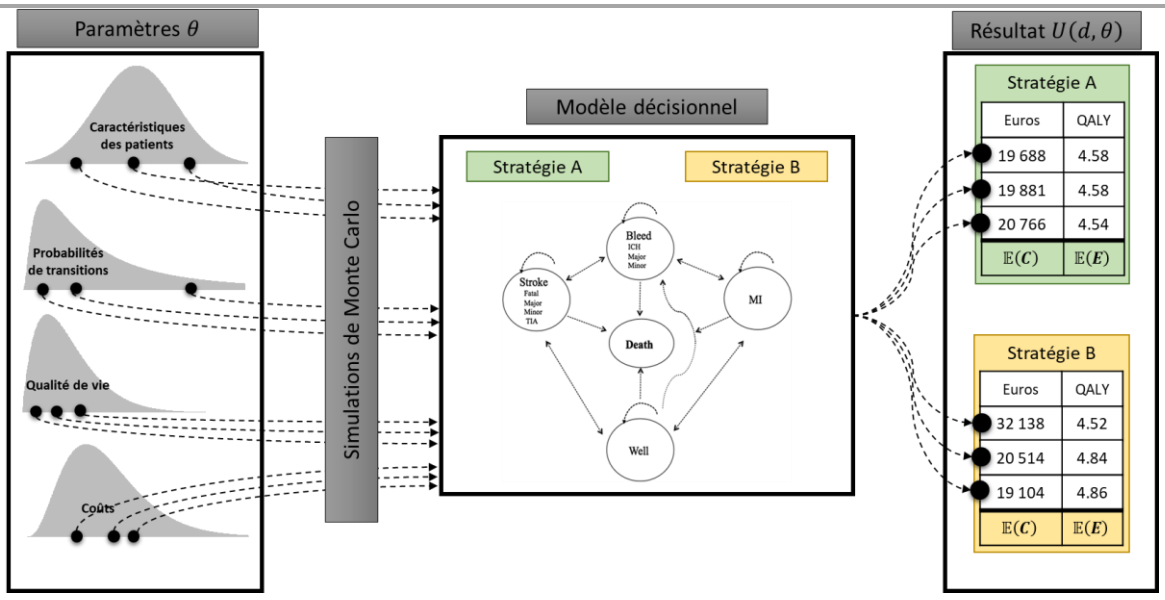


Figure 11 : Analyse de sensibilité probabiliste

En pratique, cette analyse est équivalente à l'analyse terminale utilisée pour le calcul de la valeur de l'information parfaite. En effet, chaque tirage aléatoire de l'analyse de sensibilité probabiliste correspond à la réalisation possible d'un « état du monde », conditionnellement auquel il est possible de maximiser la fonction-objectif. Il est alors aisé de calculer la valeur attendue de l'information parfaite à partir des résultats (Tableau 1). Bien qu'il soit autrement plus complexe en pratique – l'analyse pré-postérieure faisant appel aux méthodes de Monte-Carlo par chaînes de Markov (MCMC) – le principe du calcul de l'EVSI peut être en réalité envisagé de manière similaire (Figure 12).

Tableau 1 : Exemple de calcul de la valeur attendue de l'information parfaite sur les résultats de l'analyse de sensibilité probabiliste (20 simulations de Monte Carlo).

Itération	Stratégie A			Stratégie B			Meilleure stratégie (€)		Coût d'opportunité (€)	Bénéfice net incrémental (€) (réf $d = B$)	
θ	Euros	QALYs	Bénéfice net (€) $U(d = A, \theta)$	Euros	QALYs	Bénéfice net (€) $U(d = B, \theta)$	d	$\max_d\{U(d, \theta)\}$		BNI $(d = A)$	$\max_d$
1	24076	7.67	407576	30391	7.94	427391	B	427391	19815	-19815	0
2	25803	8.07	429303	32443	8.15	439943	B	439943	10640	-10640	0
3	23398	8.24	435398	31396	7.9	426396	A	435398	0	9002	9002
4	23117	8.16	431117	22968	7.71	408468	A	431117	0	22649	22649
5	28044	8.1	433044	21364	7.75	408864	A	433044	0	24180	24180
6	29027	8.3	444027	23088	7.43	394588	A	444027	0	49439	49439
7	28839	7.64	410839	31323	7.95	428823	B	428823	17984	-17984	0
8	22703	8.05	425203	26639	8.04	428639	B	428639	3436	-3436	0
9	26423	7.77	414923	26345	7.76	414345	A	414923	0	578	578
10	29444	7.82	420444	33683	8.1	438683	B	438683	18239	-18239	0
11	22124	8.18	431124	28223	7.86	421223	A	431124	0	9901	9901
12	20387	7.98	419387	34117	7.85	426617	B	426617	7230	-7230	0
13	28348	7.77	416848	26514	7.54	403514	A	416848	0	13334	13334
14	24255	7.97	422755	21304	7.64	403304	A	422755	0	19451	19451
15	20879	8.27	434379	24380	8.05	426880	A	434379	0	7499	7499
16	21142	8.21	431642	22838	7.76	410838	A	431642	0	20804	20804
17	28481	8.3	443481	30792	8.15	438292	A	443481	0	5189	5189
18	25331	8.07	428831	30668	8.19	440168	B	440168	11337	-11337	0
19	28722	8.3	443722	30038	8.15	437538	A	443722	0	6184	6184
20	23729	7.73	410229	21765	7.82	412765	B	412765	2536	-2536	0
E_θ	**25214**	**8.03**	**426714**	**27514**	**7.89**	**421864**	E_θ	**431275**	**4561**	**4850**	**9411**

Chaque ligne correspond à une itération de l'échantillonnage de Monte Carlo dans la distribution de chacun des paramètres du modèle. Pour chacune des stratégies, le résultat en Euros, en QALYs et le bénéfice net (pour une propension à payer de 50000 €/QALY) est calculé. En retenant le bénéfice net comme critère d'utilité, la stratégie A est supérieure à la stratégie B : $\max_d\{E_\theta[U(d, \theta)]\} = 426714€$. *Pour chacune des simulations, l'utilité de la meilleure stratégie conditionnellement au tirage est retenue pour calculer l'utilité attendue en présence d'une information parfaite :* $E_\theta\left[\max_d\{U(d, \theta)\}\right] = 431274€$. *Cette quantité est ensuite comparée à l'utilité espérée de la meilleure stratégie pour calculer la valeur attendue de l'information parfaite :* $\text{EVPI} = E_\theta\left[\max_d\{U(d, \theta)\}\right] - \max_d\{E_\theta[U(d, \theta)]\} = 4561$. *Cette même quantité est obtenue en moyennant les coûts d'opportunité (non nuls lorsque la meilleure stratégie conditionnellement à l'itération est la stratégie B) ou en utilisant l'analyse incrémentale (car il n'y a que deux stratégies).*

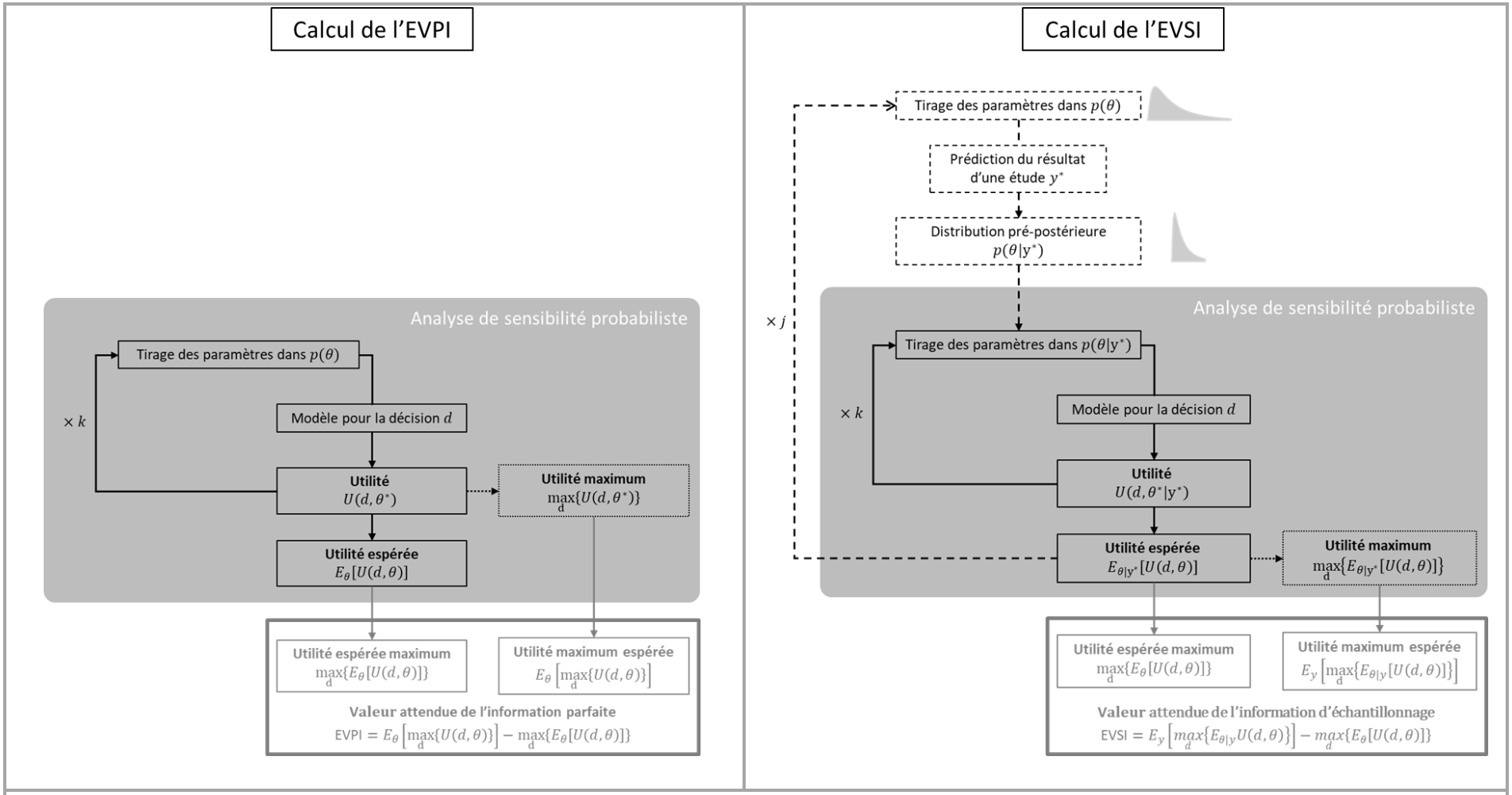


Figure 12 : Calcul de la valeur attendue de l'information parfaire (EVPI, à gauche) et de la valeur attendue de l'information d'échantillonnage (EVSI, à droite) dans le contexte de l'analyse de sensibilité probabiliste. L'analyse de sensibilité probabiliste suffit à calculer l'EVPI moyennant le recueil de l'utilité maximale de chaque simulation de Monte Carlo. Le calcul de l'EVSI implique l'ajout d'une boucle « externe » (en traits discontinus), en supplément de l'analyse probabiliste, qui permet d'évaluer la réduction de l'incertitude paramétrique permise par l'acquisition d'information à travers la réalisation d'une nouvelle étude. Cette étape nécessite l'utilisation des méthodes de Monte-Carlo par chaînes de Markov (MCMC) pour prédire la distribution pré-postérieure des paramètres. Sur le schéma de gauche, on remarquera qu'une connaissance parfaite sur les paramètres θ conduira au calcul de l'EVPI, l'analyse de sensibilité probabiliste étant alors réalisée par l'intermédiaire de la boucle externe.

III.B.2 Des données individuelles

Le calcul de la valeur de l'information peut s'effectuer sur les données individuelles issues notamment d'essais randomisés. Cette approche, parfois qualifiée de modélisation a minima (*minimal modelling*) (Meltzer et al. 2011), partage le rationnel des évaluations économiques adossées aux essais randomisés (Drummond and Davies 1991). Cette modalité de calcul de la valeur de l'information est d'ailleurs souvent retrouvée en complément de tels essais. D'un point de vue méthodologique, deux techniques sont disponibles pour exploiter les données d'une étude :

- L'approche analytique consiste à utiliser les données de l'essai pour modéliser les relations entre le critère de jugement (résultat de santé) et les déterminants mesurés (coûts, qualité de vie, etc.) à l'aide d'une équation uni ou multivariée.
- L'approche par simulation consiste quant à elle à reconstituer l'incertitude grâce aux méthodes de rééchantillonnage (*bootstrap* non paramétrique) ou en réalisant des hypothèses paramétriques sur les distributions d'intérêt (*bootstrap* paramétrique) (Meltzer et al. 2011).

Ces approches présentent l'intérêt de faciliter l'implémentation des méthodes fondées sur la valeur de l'information en minimisant le coût en termes de temps de calcul, notamment pour la valeur de l'information d'échantillonnage. Elles sont donc particulièrement intéressantes lorsque le temps et les ressources nécessaires à l'élaboration d'un modèle d'aide à la décision complet ne sont pas disponibles (Fenwick et al. 2020). Elles présentent également une alternative au calcul du nombre de sujets nécessaires des essais randomisés (Willan 2007a), tout en profitant de la flexibilité de la théorie de la décision pour adopter une perspective sociétale ou industrielle (Willan 2008; Willan and Kowgier 2008).

Cependant, plusieurs limites de ces approches doivent être soulignées. Tout d'abord, cette approche ne peut s'appliquer que dans des champs relativement restreints (Meltzer et al. 2011) : (i) lorsque le suivi des patients de l'étude s'effectue jusqu'à ce qu'aucune différence d'effet ne persiste dans les différents bras (*no modelling*), ou (ii), lorsqu'un traitement n'affecte que la qualité de vie, et non la survie qui peut être modélisée par un modèle simple (*limited modelling*).

Ensuite, à l'inverse des modèles d'aide à la décision, il est rarement possible d'identifier les paramètres responsables de l'incertitude. Pour ce faire, il est nécessaire de recourir à l'approche analytique qui seule permet le calcul de l'EVPI à partir de la distribution multivariée de l'incertitude entourant les paramètres (Koerkamp et al. 2008; Koerkamp et al. 2010). De plus, seuls les paramètres recueillis peuvent être analysés ce qui revient bien souvent à se focaliser sur la distinction entre paramètres de coûts et de qualité de vie. Enfin, l'approche analytique s'appuie souvent sur des hypothèses complexes difficiles à vérifier en pratique.[7] Des connaissances mathématiques avancées sont donc indispensables, au risque de produire des résultats biaisés.

Une alternative adoptée par certains auteurs consiste à construire un modèle d'aide à la décision simple, alimenté grâce aux données d'un essai pilote de petite taille. Le calcul de l'EVSI permet ensuite d'évaluer la pertinence d'une nouvelle étude (Palmer et al. 2016).

III.C Applications des méthodes fondées sur la valeur de l'information

L'évaluation médico-économique commandée par le décideur public répond aux principes de la théorie de la décision. C'est l'une des principales raisons pour lesquelles les méthodes fondées sur la valeur de l'information ont pénétré le domaine de la santé par le biais de l'évaluation médico-économique commanditée par le décideur public pour la hiérarchisation des efforts de recherche, notamment au Royaume Uni (Claxton 1999; Claxton et al. 2001; Claxton et al. 2004). Depuis 2004, plusieurs pays ont incorporé de manière plus ou moins formelle ces méthodes dans leur arsenal d'évaluation (Corro Ramos, Rutten-van Molken, and Al 2013; Goeree and Levin 2006; Goeree et al. 2009; Tuffaha, Gordon, and Scuffham 2016; Tuffaha and Scuffham 2018; Fleurence and Meltzer 2013). Cet effort a donné lieu à la publication de nombreux articles méthodologiques, contrastant avec le faible nombre d'applications pratiques retrouvées dans la littérature. Ainsi, en 2012, une revue systématique de la littérature identifiait que la moitié des 118 publications utilisant les méthodes fondées sur la valeur de l'information étaient de nature méthodologique (Steuten et al. 2013). Cependant, un nombre non négligeable de travaux

[7] Multi-normalité de l'incertitude, relation linéaire entre l'utilité et les paramètres étudiés, etc.

appliqués ont eu recours à ces méthodes pour évaluer l'incertitude et les besoins de recherche, dans le cadre de l'évaluation médico-économique.

III.C.1 Priorisation et le financement des efforts de recherche

Les méthodes fondées sur la valeur de l'information peuvent être utilisées en vue de hiérarchiser les projets de recherche. C'est dès le début des années 2000 que Karl Claxton suggère d'implémenter ces méthodes dans le cadre du processus d'évaluation mené par le régulateur britannique. Claxton mettait alors en exergue le décalage entre l'effort méthodologique et la transparence entourant les décisions d'adoption et de remboursement, et l'absence de tout formalisme pour la priorisation des recherches, alors même que près de 650 millions de livres y étaient allouées chaque année (Claxton and Sculpher 2006). En réponse, deux études pilotes ont été menées et ont permis de démontrer la pertinence des méthodes fondées sur la valeur de l'information (EVPI et EVPPI) comme outil d'aide à la décision pour le financement de la recherche dans le cadre du programme national d'évaluation des technologies de santé du NHS (Claxton et al. 2004) et pour l'élaboration des recommandations de recherche émises par le NICE (Claxton et al. 2005). Dès lors, de nombreux travaux d'évaluation, menées notamment par les équipes universitaires du NICE, ont utilisé ces méthodes pour évaluer l'intérêt de financer l'acquisition de données supplémentaires, en particulier dans le domaine du dispositif médical (Brush et al. 2011; McKenna et al. 2009; Wade et al. 2015; Fortnum et al. 2014; Chen et al. 2012; Stein et al. 2016; Garside et al. 2006; Fox et al. 2007). A ce jour, l'utilisation de ces méthodes ne fait pas l'objet d'une recommandation explicite de la part de la HAS. La dernière version du guide méthodologique « Choix méthodologiques pour l'évaluation économique à la HAS », publié en 2020, en pose néanmoins les bases avec l'obligation pour toute soumission d'explorer et de quantifier « l'incertitude associée à l'estimation de la valeur des paramètres du modèle [en utilisant] une analyse de sensibilité probabiliste, fondée sur une simulation de Monte Carlo de second ordre ». L'évaluation de l'intérêt des méthodes fondées sur la valeur de l'information pour la priorisation des demandes d'étude post-inscription est un objectif de cette thèse qui s'inscrit dans cette perspective.

III.C.2 Choix du design

Les méthodes fondées sur la valeur de l'information sont utilisées ici comme alternative à l'approche classique (fréquentiste) de détermination de la taille d'échantillon, notamment des essais randomisés (Willan and Pinto 2005; Willan 2007b; Tuffaha et al. 2014). Cette dimension est probablement la plus attrayante d'un point de vue théorique dans la mesure où l'ENBS peut être considéré comme un critère nécessaire et suffisant à la mise en place d'une nouvelle étude. Les applications de cette nature restent cependant relativement rares, notamment du fait de la complexité des méthodes mobilisées, de l'absence de recommandation antérieures à 2020, et de la faible pénétration des techniques d'optimisation du calcul de l'EVSI.

III.C.3 Pour le remboursement de l'innovation

La valeur de l'information s'avère un outil utile pour définir l'équilibre entre la valeur d'une innovation en santé et le timing de son accès au remboursement. En effet, les systèmes de santé sont confrontés au défi de l'accès précoce à l'innovation, dans un contexte où le degré de maturité des preuves scientifiques disponibles est par définition plus faible. Les schémas d'accès au marché doivent ainsi tenir compte de deux aspects antagonistes :

- Il peut être risqué de permettre un accès prématuré à une innovation dont les bénéfices de santé sont incertains, ce d'autant que son retrait s'avère bien souvent complexe voire impossible (Chalkidou et al. 2008). Par ailleurs, l'accès précoce au marché est de nature à dissuader l'industriel d'investir dans des études complémentaires concernant l'efficacité ou la sécurité de l'innovation.

- A l'inverse, un schéma d'accès trop restrictif peut faire peser un coût d'opportunité sur les patients en restreignant l'accès à une réelle innovation.

En plus de l'adoption ou du rejet d'une innovation, le régulateur dispose de deux autres modalités d'accès au remboursement : (i) adopter l'innovation conditionnellement à l'acquisition de données complémentaires, ou (ii) rejeter l'innovation mais financer l'acquisition de données (Claxton et al. 2012). En France, ces deux modalités s'apparentent respectivement aux demandes d'études post-inscription et au forfait innovation. Le choix de l'une de ces quatre voies (*Approve, Reject, AWR – Approwe with research, OIR – Only in reseasch*) s'appuie sur une série de questions

clés résumées par la checklist de McKenna, dans le contexte de l'évaluation par le NICE (McKenna et al. 2015) :

1. La technologie est-elle coût-efficace ?
2. L'adoption implique-t-elle des coûts irrécouvrables ?
3. L'acquisition de données supplémentaires a-t-elle un intérêt ?
4. Cette dernière peut-elle s'effectuer sans adoption ?
5. D'autres sources d'incertitudes sont-elles susceptibles de se résoudre dans le temps ?
6. Les bénéfices attendus de l'acquisition de données supplémentaires sont-ils supérieurs au coût d'acquisition ?
7. Le bénéfice de l'adoption est-il supérieur au coût d'opportunité ?

Dans ce cadre, les réponses aux questions (3) et (6) sont apportées par les méthodes fondées sur la valeur de l'information. En 2016, Claxton et al. publient un algorithme décisionnel extrêmement détaillé articulant ces différentes questions (Claxton et al. 2016). Nous y référons le lecteur pour une présentation exhaustive du sujet.

III.C.4 Applications originales

Plusieurs applications et mesures dérivées de la théorie de la valeur de l'information sont retrouvées dans la littérature :

- Dans un contexte de médecine de précision, la valeur attendue d'une prise en charge individualisée (EVIC, Expecte Value of Individualized Care) a été proposée comme mesure permettant d'affiner le choix du traitement en fonction des sous-groupes qui en bénéficient le plus (Basu and Meltzer 2007). L'idée derrière l'EVIC est de modéliser la valeur de l'information comme la fonction de perte liée à l'absence de prise en compte de l'hétérogénéité (Tuffaha, Gordon, and Scuffham 2014b). Elle peut ensuite être comparée au coût d'un test diagnostic susceptible d'apporter une information concernant le traitement le plus adapté pour le patient (Basu, Carlson, and Veenstra 2016; van Gestel et al. 2012). Dans la même veine, Espinoza détaille les aspects

méthodologiques à prendre en compte afin de réaliser un calcul de la valeur de l'information dans un contexte d'hétérogénéité (Espinoza et al. 2014)

- En 2015, Claxton démontre qu'un calcul de la valeur attendue de l'information parfaite peut être adossé à une méta-analyse afin de quantifier l'intérêt de la collecte de données supplémentaires (Claxton et al. 2015; McKenna et al. 2016). Sur la base de la méta-analyse de l'intérêt de la Streptokinase en prévention secondaire de la mortalité liée à l'infarctus, il détaille une méthodologie simple permettant de quantifier la valeur de l'implémentation et la valeur de l'information (Figure 13). Cette étude illustre par ailleurs la possibilité de calculer la valeur de l'information en utilisant d'autres critères que le bénéfice net incrémental. En effet, la valeur de l'information était exprimée en mortalité évitée.

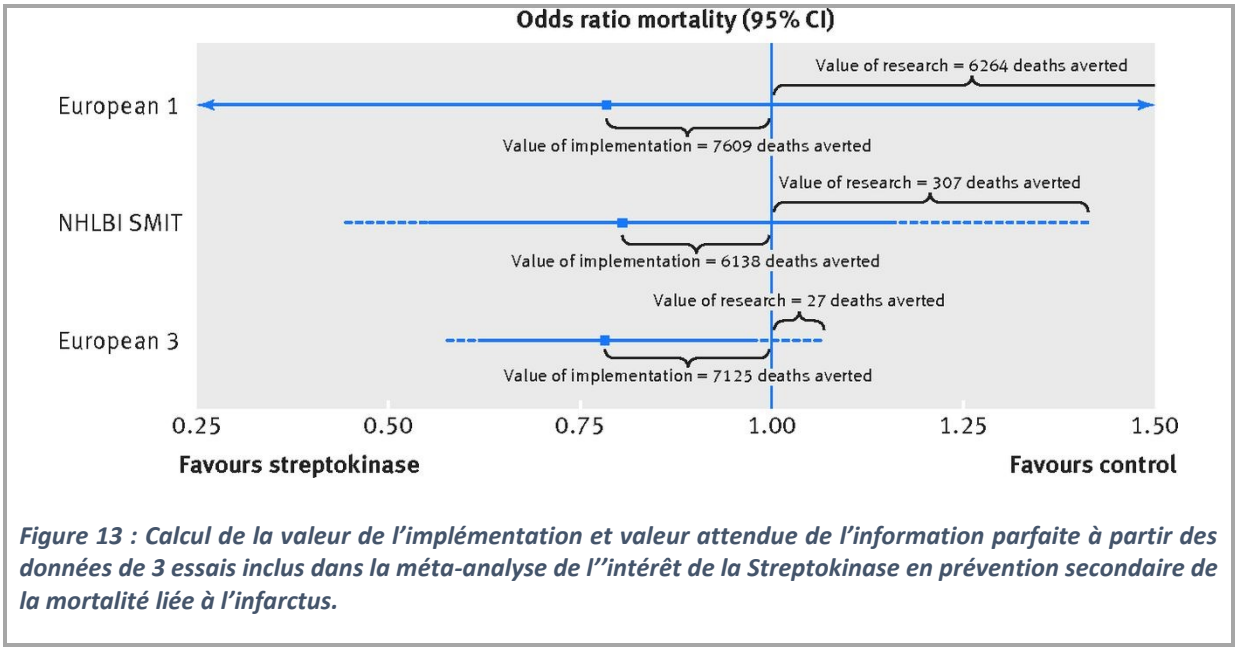


Figure 13 : Calcul de la valeur de l'implémentation et valeur attendue de l'information parfaite à partir des données de 3 essais inclus dans la méta-analyse de l''intérêt de la Streptokinase en prévention secondaire de la mortalité liée à l'infarctus.

- Pour répondre à l'absence d'observation des distributions a priori, une approche fréquentiste du calcul de la taille d'échantillon des études coût-efficacité fondée sur le concept de valeur de l'information est proposée (Bader et al. 2018). Dans ce cadre, l'échantillon optimal est celui maximisant la différence entre la valeur attendue de l'information parfaite avant et après la réalisation d'un essai (et le coût d'inclusion), cette

dernière étant calculée à partir de la distribution d'échantillonnage du bénéfice net incrémental.

- Les dynamiques d'implémentation et l'évolution des prix sont de nature à moduler la valeur associée à la collecte de nouvelles données. Plusieurs tentatives de modélisation sont proposées dans la littérature pour articuler ces éléments avec le calcul de la valeur de l'information et en tenir compte dans le cadre d'accords négociés de mise sur le marché tels que le paiement à la performance ou les règles de partage de risque (Grimm, Dixon, and Stevens 2016; Grimm et al. 2017; Garrison et al. 2013).
- L'application itérative des méthodes fondées sur la valeur de l'information peut permettre de guider des décisions de nature variée. Ainsi, elles ont été proposées pour évaluer la probabilité de remboursement tout au long du développement d'un dispositif médical (Vallejo-Torres et al. 2010).

III.D Aspects calculatoires

L'utilisation des méthodes fondées sur la valeur de l'information se trouve souvent limitée par la difficulté de mettre en œuvre les calculs impliquant deux boucles de simulations imbriquées (EVPPI) ou des méthodes MCMC (EVSI). En réponse à ce problème récurrent, plusieurs méthodes ont été proposées pour diminuer le temps de calcul. Dans la mesure où un recensement exhaustif a été réalisé récemment, dans le cadre de la publication des recommandations de l'ISPOR sur l'utilisation des méthodes fondées sur la valeur de l'information (Rothery et al. 2020), nous nous limitons ici à un résumé des principales techniques disponibles :

- Si les paramètres non concernés par le calcul de l'EVPPI peuvent être modélisés comme une fonction linéaire (ou multilinéaires) du critère de jugement, et qu'aucune corrélation n'existe entre eux, alors la boucle interne du calcul de l'EVPPI (calcul de la quantité $E_{\theta_{-i}|\theta_i}\big(U(d,\theta)\big)$) n'est pas nécessaire car les valeurs moyennes peuvent être utilisées pour la remplacer (Madan et al. 2014). De la même manière, une solution analytique peut être proposée sous certaines conditions (de multi-linéarité) pour le calcul de l'espérance de la boucle interne de l'EVSI, $E_{\theta_{-i}|y^*}\big(U(d,\theta|y^*)\big)$ (Ades, Lu, and Claxton 2004). L'utilisation de telles approximations nécessite néanmoins une connaissance approfondie

du modèle d'aide à la décision sous-jacent ainsi que des relations entre les différents paramètres, ce qui limite ses applications pratiques (Heath, Manolopoulou, and Baio 2017b). A l'inverse, lorsque les conditions d'application sont réunies, l'utilisation des valeurs moyennes est préférable pour diminuer l'erreur de Monte Carlo (Oakley et al. 2010).

- Des méthodes de régression ont été développées pour prédire l'espérance de la boucle interne (de l'EVPPI, $E_{\theta_{-i}|\theta_i}\big(U(d,\theta)\big)$, et de l'EVSI, $E_{\theta_{-i}|y^*}\big(U(d,\theta|y^*)\big)$) sur la base des résultats de l'analyse de sensibilité probabiliste. A l'inverse des régressions classiques, l'objectif est ici de « sur-ajuster » pour prédire le critère de jugement le plus précisément possible, puis d'utiliser le résultat pour économiser la boucle interne de l'EVPPI. Pour ce faire, les auteurs ont recours à des méthodes non-paramétriques telles que les modèles additifs généralisés ou les process gaussiens (Strong, Oakley, and Brennan 2014b; Heath, Manolopoulou, and Baio 2016; Strong et al. 2015; Tuffaha et al. 2016).
- Enfin, le calcul de l'EVSI peut être facilité lorsque la loi a priori du paramètre et la vraisemblance des données sont conjuguées[8] (Willan and Pinto 2005; Ades, Lu, and Claxton 2004).

IV VALEUR D'OPTION REELLE ET VALEUR DE L'INFORMATION

La valorisation par la méthode des options réelles est une méthode d'évaluation des projets de recherche et développement qui permet également la valorisation de l'incertitude. Elle occupe une place de choix dans la littérature économique et constitue une alternative aux méthodes fondées sur la valeur de l'information, notamment pour les industriels, en amont de la mise sur le marché de leur produit. Les deux approches sont distinctes mais étroitement liées. La méthode des options réelles estime la valeur de la flexibilité décisionnelle. En présence d'incertitude, lorsque la décision a un caractère irréversible, la possibilité de reporter la décision (dans l'espoir

[8] La distribution a priori peut être conjuguée avec la vraisemblance pour permettre une inférence exacte et analytique sur la distribution a posteriori. Prenons l'exemple d'un processus de génération des données binomial $p(y|\theta)\sim\mathrm{Bin}(\theta,n)$. La situation conjuguée correspond à une distribution béta du paramètre a priori, $\theta\sim\mathrm{Beta}(a,b)$. Dans ces conditions, la distribution a posteriori $p(\theta|y)$ est connue et suit également une loi beta de paramètres $a+y$ et $b+n-y$, dont les caractéristiques sont parfaitement connues et exploitables analytiquement.

qu'une partie de l'incertitude sera levée au cours du temps) a une valeur économique. De ce point de vue, le découpage d'un projet de R&D en différentes phases en augmente la valeur puisqu'il est possible d'arrêter le développement du produit au regard des informations produites au cours des phases de développement successives. De même, les schémas d'accès à l'innovation s'appuyant sur les méthodes fondées sur la valeur de l'information soulignent l'importance de prendre en compte l'irréversibilité des décisions de remboursement par le régulateur. Ainsi, l'une des questions posées est celle de l'existence de coûts irrécouvrables (McKenna et al. 2015), pouvant amener le régulateur à limiter la diffusion de l'innovation durant l'acquisition de données complémentaire (accès « *Only in reseasch* »).

L'objectif de ce chapitre est de présenter les principes des méthodes de valorisation par les options réelles, puis d'exposer les liens théoriques et mathématiques établis dans la littérature avec les méthodes fondées sur la valeur de l'information. Enfin, nous exposons les modalités de prise en compte de la flexibilité de la décision dans le calcul de la valeur de l'information. Nous concluons en positionnant cette synthèse par rapport aux travaux de Willan et Eckermann sur le sujet (Eckermann & Willan, 2007, 2008).

IV.A Théorie de la valeur d'option

IV.A.1 Une théorie issue de la finance

En finance, une option est un dérivé (*derivative)* conférant à son détenteur la possibilité de vendre ou d'acheter un actif sous-jacent à un prix prédéterminé (*strike price*) à une date spécifiée (*maturity*).[9] Comme son nom l'indique, il s'agit d'une option et non d'une obligation. Ce droit est acheté à un prix (le *premium*). Le détenteur exercera son option d'achat ou de vente en fonction de la valeur à terme de l'actif sous-jacent. Les compagnies aériennes, dont le profit est significativement impacté par les variations du cours du pétrole, auront recours aux options d'achat (*call option*) pour réduire leur exposition. Si au moment d'acheter du pétrole sur le marché le cours a subi une hausse importante, alors la compagnie aérienne exercera son option afin d'acheter le pétrole à un prix inférieur, celui défini au moment de l'achat de l'option. Dans

[9] L'option de vendre est un « put » et l'option d'acheter est un « call ».

le cas contraire, l'option n'aura aucune valeur car le prix du pétrole sera inférieur à ce prix. Au-delà de ce mécanisme d'assurance, les options sont également utilisées comme outil spéculatif. L'option de vendre un actif (*put option*) permet ainsi de parier sur la baisse d'un cours. Si la valeur du cours est inférieure à celle de l'option, alors le spéculateur empoche la différence.

IV.A.2 L'approche par les Options réelles

L'option correspond donc à une décision contingente (Amram and Kulatilaka 1998). C'est l'opportunité de prendre la (bonne) décision une fois l'incertitude levée qui constitue la valeur d'une option. L'analyse des options réelles (*Real Options Analysis*) est la transposition de la théorie de la valeur d'option aux décisions d'investissement concernant des actifs non-financiers. Trois aspects sont pris en compte par l'analyse des options réelles : l'irréversibilité de la décision, l'incertitude entourant ses conséquences et la possibilité de reporter la décision (Dixit and Pindyck 2012).

Le concept d'irréversibilité occupe une place centrale dans l'analyse des options réelles. Henry qualifie une décision d'irréversible si elle réduit significativement et pour longtemps la variété des choix possibles dans le futur (Henry 1974a). Dans le cadre d'un investissement, une quantité significative de ressources est investie immédiatement dans l'espoir d'un bénéfice futur. Le bénéfice est incertain mais une partie plus ou moins importante des ressources ne pourra pas être récupérée si la décision d'investir s'avère mauvaise. De plus des coûts de retournement devront être mis en jeu pour revenir sur la décision. L'importance de ces coûts irrécouvrables caractérise l'irréversibilité de la décision. Dans ce contexte, il peut être intéressant pour le décideur de postposer sa décision, dans l'attente de voir se résoudre l'incertitude. Ce choix n'est possible qu'en présence de flexibilité dans le timing de la décision. Le report de la décision implique néanmoins un coût d'opportunité qui devra être mis en balance avec le bénéfice associé à la résolution de l'incertitude. Le parallèle peut ainsi être ici fait avec l'option d'achat en finance (*call*) qui devient ici le droit d'investir. Le décideur a le choix d'investir immédiatement dans un contexte d'incertitude et de renoncer à son option. S'il n'investit pas, il conserve son option et choisit de l'exercer ou non à terme, c'est-à-dire une fois l'incertitude levée. Si l'utilité à terme est positive, il exerce son option et investit, sinon il renonce simplement à son option et ne perd que le prix d'achat de l'option.

IV.A.3 Développement dans deux champs d'application : l'environnement et les affaires

L'analyse des options réelles a initialement été développée dans le cadre des travaux d'Arrow-Fisher et Henry portant sur l'économie environnementale (Arrow and Fisher 1974; Henry 1974b). Les auteurs définissaient le concept de *Quasi-Option Value* (QOV), comme la valeur de l'acquisition de l'information en cas de report d'une décision irréversible. Ainsi, le bénéfice associé à la décision d'aménagement de l'environnement devait être ajusté pour refléter l'impossibilité d'inverser la décision. Au début des années 1990, Dixit et Pindyck appliquaient la théorie des options aux décisions d'investissement prises dans le monde des affaires (Dixit and Pindyck 2012). Ils proposaient alors de définir la *Real Options Value* (ROV) comme la valeur du report en cas d'opportunité d'acquisition d'information.

Bien que décrites comme équivalentes par Fisher (Fisher 2000), la QOV et la ROV sont en réalité distinctes conceptuellement (Mensink and Requate 2005). Les deux mesures prennent en compte le bénéfice associé à l'obtention d'une information nouvelle mais seule la ROV tient compte du coût d'opportunité associé à l'option. La décision initiale devra donc se baser sur la ROV. Néanmoins, le parallèle entre valeur de l'information et valeur d'option nécessite de s'intéresser au concept de QOV. En effet, la QOV peut être interprétée comme la valeur associée à l'obtention d'information, conditionnellement au report de la décision. La proximité entre ces concepts est illustrée par les travaux d'Hanneman établissant le lien entre QOV et VoI dans le cadre de l'écologie environnementale (Conrad 1980; Hanemann 1989).

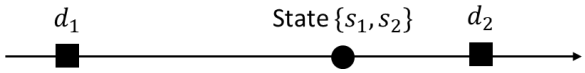
IV.B Aspects mathématiques

IV.B.1 Positionnement du problème

On considère l'alternative décisionnelle suivante : adopter ou non. Il peut s'agir d'investir dans un nouveau business, aménager un terrain, lancer un produit de santé sur le marché, etc. La décision (d_t) est flexible et peut être prise à deux temps ($t \in \{0; 1\}$) : maintenant (d_1) ou à une date ultérieure (d_2). Dans un premier temps, nous considérons la décision comme irréversible. Si l'adoption est décidée maintenant ($d_1 = 1$), alors la décision ne pourra pas être inversée dans le futur et $d_2 = d_1 = 1$. Seule la décision de $d_1 = 0$ préserve l'option d'adopter ou non dans le futur ($d_2 \in \{0; 1\}$). Enfin, une contrainte d'indivisibilité de la décision est imposée. Cette

contrainte concerne surtout le domaine de l'environnement pour lequel un aménagement partiel pourrait être envisagé.

Le bénéfice est incertain et dépend de la réalisation « d'états du monde » dans le futur. Deux états du monde sont possibles $\{s_1; s_2\}$. En cas d'adoption, ils correspondent respectivement à une issue favorable d'utilité b ou à une issue défavorable d'utilité $-c$. Les probabilités de réalisation des états du monde étant π pour s_1 et $1-\pi$ pour s_2. L'adoption immédiate permet l'obtention d'une utilité supplémentaire entre d_1 et d_2 noté a.

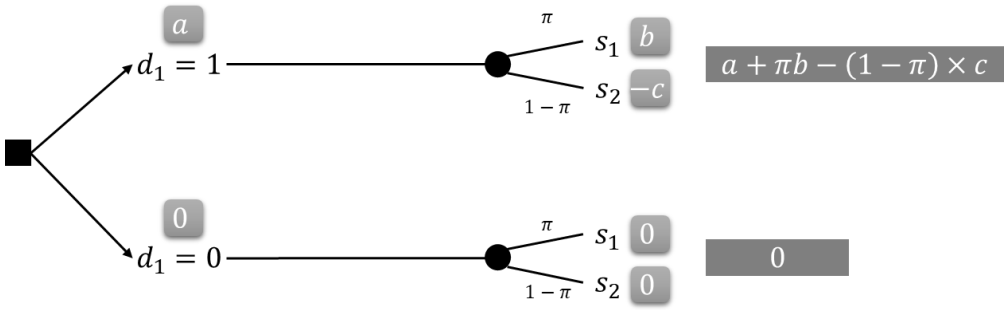


IV.B.2 Les trois types de décideurs de Hannemann

Le premier décideur (N, now) est celui qui prend une décision en ignorant la possibilité d'acquérir une information nouvelle ou de reporter sa décision. Sa décision est donc celle maximisant l'utilité $U^N_{d_1}$.

$$U^N_{d_1=1} = u_{d_1=1} + \mathbb{E}\left(u_{d_2=1}\right) = a + \pi b - (1-\pi) \times c$$

$$U^N_{d_1=0} = u_{d_1=0} + \mathbb{E}\left(u_{d_2=0}\right) = 0$$

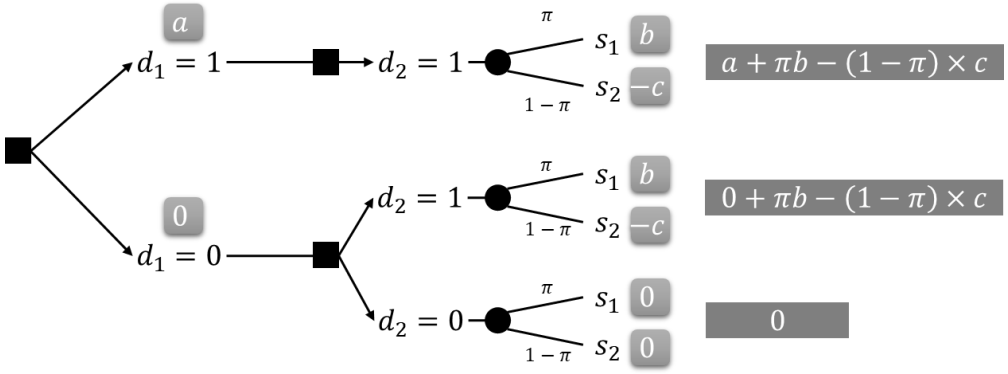


Le second décideur (D, delay) prend une décision en ignorant la possibilité d'acquérir une information nouvelle mais il tient compte de la possibilité de reporter sa décision. Dans un contexte d'irréversibilité, ses possibilités dépendront de d_1. Si $d_1 = 1$, la décision est irréversible

et $d_2 = 1$. Il n'y a alors aucune différence entre l'utilité associée aux décideurs N et D. Si la décision $d_1 = 0$, le décideur D reporte sa décision au temps 2. Il prendra alors une décision fondée sur l'utilité attendue de la décision d'investissement optimale *ex-ante* : $\max_{d_2} \left[\mathbb{E}\left(u_{d_2}\right)\right]$.

$$U^D_{d_1=1} = u_{d_1=1} + \max_{d_2} \left[\mathbb{E}\left(u_{d_2=1}\right)\right] = a + \pi b - (1-\pi) \times c$$

$$U^D_{d_1=0} = u_{d_1=0} + \max_{d_2} \left[\mathbb{E}\left(u_{d_2}\right)\right] = \max[\pi b - (1-\pi) \times c; 0]$$



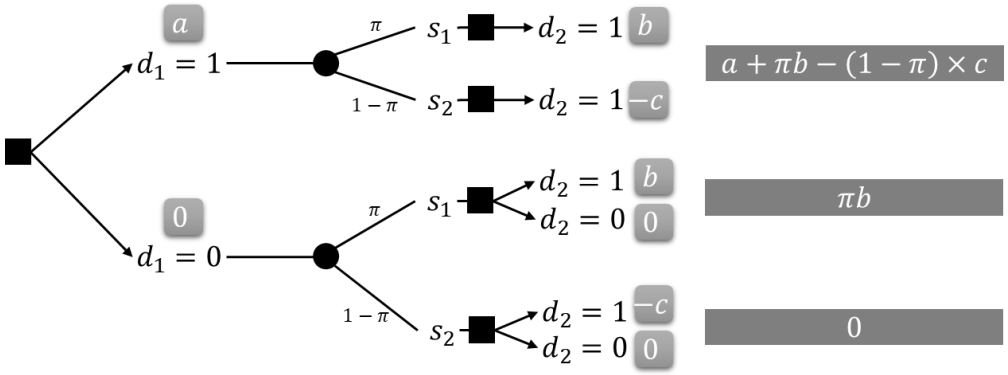
Les hypothèses actuelles rendent peu pertinentes un tel décideur dans la mesure où il reporte sa décision mais ne tient pas compte de l'information. Ce décideur théorique est néanmoins utile à plusieurs égards. Il permet tout d'abord de comprendre le concept de QOV car Hanemann utilise un décideur D comme référence pour le calcul. Il permet également de dissocier le timing de la décision et l'acquisition d'information, ce qui est une hypothèse fondamentale des méthodes fondées sur la valeur de l'information. Enfin, bien que non envisagé ici, le cas d'une utilité négative liée à l'adoption immédiate (a) pourrait amener le décideur à inverser une décision $d_1 = 0$.

Le troisième décideur (L, learning) prend une décision en anticipant l'obtention d'information permise par le report de la décision. Dans un contexte d'irréversibilité, ses possibilités dépendront de nouveau de d_1. Si $d_1 = 1$, la décision est irréversible et $d_2 = 1$. Il n'y a alors aucune différence entre l'utilité associée aux décideurs N, D et L. Si la décision $d_1 = 0$, le

décideur L reporte sa décision au temps 2. Il prendra alors une décision fondée sur l'utilité attendue de la décision d'investissement optimale *ex-post* : $\mathbb{E}\left(\max_{d_2}[u_{d_2}]\right)$.

$$U^L_{d_1=1} = u_{d_1=1} + \mathbb{E}\left(\max_{d_2}[u_{d_2=1}]\right) = a + \pi b - (1-\pi) \times c$$

$$U^L_{d_1=0} = u_{d_1=0} + \mathbb{E}\left(\max_{d_2}[u_{d_2}]\right) = \pi b$$



IV.B.3 Alternatives décisionnelles

La décision est donc composée par (i) la comparaison de l'utilité des alternatives et (ii) le moment de la décision, qui peut être analysé en comparant les décideurs N et L. Il existe donc 4 alternatives qui peuvent d'emblée être rapprochées de la terminologie formalisée dans le cadre des schémas d'accès au remboursement des innovations en santé (Karl Claxton et al., 2016; Eckermann & Willan, 2008) :

- **AN** = Adopter définitivement (*adopt now*, *approve*)
- **RN** = Rejeter définitivement (*reject now*)
- **AL** = Adopter et attendre l'information (*adopt and learn*, *approve with research*)
- **RL** = Rejeter et attendre l'information (*reject and learn*, *only in research*)

L'hypothèse d'irréversibilité de la décision, inhérente au concept de valeur d'option, conduit les alternatives **AN** ($U^N_{d_1=1}$) et **AL** ($U^L_{d_1=1}$) à être équivalentes. La dimension du temps est donc

neutralisée en cas de décision d'adoption car il n'existe aucune flexibilité. L'intérêt de reporter la décision et d'obtenir une information supplémentaire n'est donc possible que pour $d_1 = 0$.

IV.B.4 Valeur d'option réelle

La valeur d'option réelle (ROV) telle que définie par Dixit et Pindyck est la différence entre l'utilité de la décision maximisant l'utilité espérée avec l'information actuelle et l'utilité de la décision maximisant l'utilité espérée avec l'information acquise grâce au report de la décision.

$$ROV = \underbrace{\max\left(U^L_{d_1=0}, U^L_{d_1=1}\right)}_{\text{report et apprentissage}} - \underbrace{\max\left(U^N_{d_1=0}, U^N_{d_1=1}\right)}_{\text{ni report, ni apprentissage}} = \max(RL, AL) - \max(RN, AN)$$

Plusieurs cas de figures doivent être envisagés :

- Si $AL > RL$, la valeur d'option est nulle car $AN = AL$ et $RN \leq RL$.

$$ROV = 0$$

- Si $RL > AL$ et $RN > AN$, le report l'emporte systématiquement face à l'adoption et la valeur d'option est la différence entre $U^L_{d_1=0}$ et $U^N_{d_1=0}$.

$$ROV = U^L_{d_1=0} - U^N_{d_1=0} = RL - RN$$

- Si $RL > AL$ et $AN > RN$, alors

$$ROV = U^L_{d_1=0} - U^N_{d_1=1} = RL - AN$$

IV.B.5 Valeur de quasi-option

La QOV définie par Arrow-Fisher, Henry et Hanemann est équivalente à une taxe virtuelle pénalisant la décision non-optimale d'un développement immédiat. Elle est définie en comparant les décideurs $\boldsymbol{L}$ et $\boldsymbol{D}$ et est calculée conditionnellement au report de la décision. C'est la différence entre deux décideurs utilisant ou non l'information disponible. Elle est donc proche des concepts de valeurs de l'information dans la mesure où la composante temps de la décision semble neutralisée. Elle se définit comme suit :

$$QOV = \underbrace{\left(U^L_{d_1=0} - U^L_{d_1=1}\right)}_{\text{report et apprentissage}} - \underbrace{\left(U^D_{d_1=0} - U^D_{d_1=1}\right)}_{\text{report sans apprentissage}}$$

Pour lier les concepts de QOV et de ROV, il faut déterminer le rapport entre les utilités des décideurs $\boldsymbol{N}$ et $\boldsymbol{D}$. On remarque d'emblée que $U^D_{d_1=1} = U^N_{d_1=1}$ car l'adoption est irréversible. En cas de report, l'utilité est $U^D_{d_1=0} = \mathbb{E}\left(\max_{d_2}[u_{d_2=1}, 0]\right)$. En remarquant que $u_{d_2=1} = U^N_{d_1=1} - u_{d_1=1}$, on peut écrire $U^D_{d_1=0} = \max_{d_2}[U^N_{d_1=1} - u_{d_1=1}, 0] = \max[AN - a\ ,0]$. Le cas $AN - a < 0$ étant celui d'un coût d'opportunité du report supérieur à l'utilité d'une adoption à terme.

On peut donc écrire :

$$QOV = (RL - AL) - (\max[AN - a\ , RN] - AN)$$

Et simplifier grâce à l'hypothèse d'irréversibilité ($AL = AN$) :

$$QOV = RL - \max[AN - \mathrm{a}, RN]$$

La décision de report et d'obtention d'information est comparée à la stratégie optimale en cas de report (i.e. en assumant le coût d'opportunité du report).

IV.C Lien entre la valeur d'option réelle, la valeur de quasi-option et la valeur de l'information

Les quantités suivantes peuvent être calculées :

- $AN = AL = a + \pi b - (1 - \pi) \times c$
- $RN = 0$
- $RL = \pi b$

IV.C.1 En cas de bénéfice lié à l'adoption

On envisage tout d'abord le cas de figure d'un bénéfice lié à l'adoption $\pi b > (1 - \pi) \times c$. Dans le cadre des méthodes fondées sur la valeur de l'information, la valeur maximale que l'on peut envisager d'obtenir correspond à $(1 - \pi) \times c$. Les situations possibles sont résumées dans le tableau ci-dessous. On observe ici que la valeur de l'information maximale que l'on peut obtenir est équivalente à la QOV.

Hypothèse	$\boldsymbol{AL > RL}$	$\boldsymbol{RL > AL}$

Implication de l'hypothèse	$a > (1-\pi) \times c$	$a < (1-\pi) \times c$
QOV	$(1-\pi) \times c$	$(1-\pi) \times c$
ROV	0	$QOV - a$

Si $AL > RL$, le coût d'opportunité est supérieur à la valeur que l'on peut obtenir de l'information. Dans ce cas, l'adoption est l'alternative optimale. L'option de report et donc l'information n'a aucune valeur. Si $RL > AL$, la valeur de l'information est supérieure au coût d'opportunité a et l'option permet un gain équivalent à la QOV déduite du coût d'opportunité.

IV.C.2 Absence de bénéfice lié à l'adoption

On envisage maintenant le cas de figure d'une absence de bénéfice lié à l'adoption $\pi b < (1-\pi) \times c$. Dans le cadre des méthodes fondées sur la valeur de l'information, la valeur maximale que l'on peut envisager d'obtenir correspond à πb.

Hypothèse	$\boldsymbol{AL > RL}$	$\boldsymbol{RL > AL > RN}$	$\boldsymbol{RL > RN > AL}$
Implication de l'hypothèse	$a + (1-\pi) \times c > 0$	$a + (1-\pi) \times c < 0$	$a + (1-\pi) \times c > 0$
QOV	πb	πb	πb
ROV	0	$QOV - (1-\pi) \times c - a - \pi b$	QOV

Si $AL > RL$, l'alternative optimale est d'adopter. L'option de report et donc l'information n'a aucune valeur. Si $RL > AL$, il faut distinguer $\boldsymbol{RN} > AL$ où la valeur de l'information est la QOV et $AL > \boldsymbol{RN}$, où l'information vient compenser le choix initial d'un bras flexible mais non optimal.

On remarque que le cas $RL > AL > \boldsymbol{RN}$ correspond à l'inversion de la décision d'une adoption immédiate. Dans ce cas, la valeur issue de la résolution de l'incertitude n'est pas complètement accessible car elle doit être réduite du coût de cette inversion : le coût d'opportunité en cas de bénéfice positif, la différence initiale en cas de bénéfice négatif. Ce coût de l'inversion est à

rapprocher de la valeur du report pur (Mensink and Requate 2005) et de la valeur d'option simple (Traeger 2014). Nous retrouvons également les résultats d'Hanemann qui établissaient les liens entre QOV et EVPI dans les sciences environnementales (Hanemann 1989).

IV.D Adaptation au contexte de la prise de décision en santé

IV.D.1 Incertitude à terme

L'une des principales différences entre les méthodes fondées sur la valeur de l'information et sur la valeur d'option tient au rôle de l'incertitude. L'incertitude telle qu'envisagée dans les options concerne le futur rendement de l'actif sous-jacent. Ainsi, le prix de l'actif au moment de la décision est connu. Il est par conséquent certain. Ce n'est pas le cas du prix à terme qui dépend de l'évolution du marché. L'évolution de l'incertitude entourant le prix est alors modélisée comme variant de manière aléatoire au fil du temps. Plus on s'éloigne dans le temps, plus l'incertitude augmente. La résolution de l'incertitude à terme est passive, sans intervention du décideur. Pour les produits de santé, l'efficacité immédiate est incertaine et aucune variation de l'efficacité (réelle) n'est attendue à terme. L'approche par les options réelles présente l'avantage d'embarquer ces deux contraintes. Néanmoins, l'incertitude reste complètement résolutive à terme. Pour un produit de santé, l'incertitude à terme restera inchangée en l'absence d'investissement en recherche et seule la réalisation d'une étude est de nature à diminuer l'incertitude (mais pas à la résoudre). En reprenant le principe de l'EVSI, on intègre une résolution partielle de l'incertitude. En pratique, cela signifie que l'information maximale que l'on peut obtenir est inférieure à $(1-\pi)\times c$ ou πb, respectivement pour un bénéfice positif ou négatif. Dans ce cas, la QOV se verra amputée d'une partie de sa valeur pour obtenir QOV_{sample}. La valeur de cette QOV_{sample} correspond à l'EVSI calculée par les méthodes classiques de valeur de l'information.

IV.D.2 Flexibilité de la décision

Le concept fondateur de l'analyse décisionnelle est que l'information a une valeur si elle est de nature à modifier la décision (Raiffa and Schlaifer 1961). En d'autres termes, dans un contexte d'incertitude, il existe une probabilité que la décision prise soit la mauvaise. La théorie de la valeur de l'information implique donc une dissociation entre la décision et le choix de poursuivre

la recherche. La décision est celle maximisant l'utilité pour le décideur puis, dans un deuxième temps, la valeur attendue de l'information (EVPI – EVSI) est calculée indépendamment de la décision. Les méthodes fondées sur la valeur de l'information font ainsi l'hypothèse implicite que la décision est complétement réversible. Cette hypothèse peut être satisfaisante pour un régulateur qui considèrerait sa décision comme complètement réversible. C'est l'hypothèse faite initialement lors de l'implémentation des méthodes fondées sur la valeur de l'information dans le processus de remboursement (Claxton et al. 2004).

Cette hypothèse peut cependant être jugée comme trop forte. Ainsi, le décideur doit prendre en compte la pression des patients, l'absence d'alternatives, etc. Par ailleurs, un dispositif médical implantable ne pourra pas être retiré, ou moyennant un coût très élevé (irréversibilité individuelle). Du point de vue de l'industriel, la mise sur le marché implique des coûts irrécouvrables qui ne seront couverts qu'en cas de décision pérenne : investissement dans une chaîne de production ad hoc, marketing de son produit, etc. Claxton propose ainsi un algorithme décisionnel intégrant à la fois la valeur de l'information mais également l'éventualité de coûts irrécouvrables (Claxton et al. 2016). Bien qu'il confirme l'importance d'intégrer l'irréversibilité dans la décision, Claxton ne fixe pas de seuil au-delà duquel les coûts irrécouvrables peuvent être considérés comme « significatifs ». L'utilisation de l'approche par la valeur d'option nous permet d'intégrer l'irréversibilité dans le processus décisionnel. Il nous faut néanmoins proposer une adaptation afin de permettre une irréversibilité partielle.

IV.E Modèle intégrant la flexibilité de la décision dans le calcul de la valeur de l'information

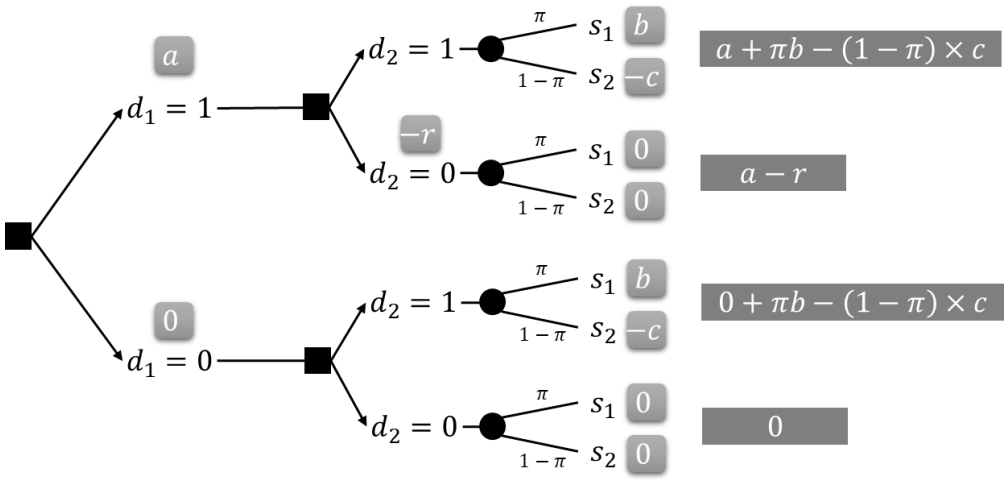
IV.E.1 Le modèle

On intègre ici les coûts de retournement r, représentant l'irréversibilité pour les décideurs. La décision étant maintenant flexible, $d_2 = 0$ est envisagée pour les décideurs $\boldsymbol{D}$ et $\boldsymbol{L}$.

Le second décideur ($\boldsymbol{D}$) prend une décision en ignorant la possibilité d'acquérir une information nouvelle mais il tient compte de la possibilité de reporter sa décision. Si le décideur D reporte sa décision au temps 2. Il prendra alors une décision fondée sur l'utilité attendue de la décision d'investissement optimale *ex-ante* : $\max_{d_2} \left[\mathbb{E}\left(u_{d_2}\right)\right]$.

$$U^D_{d_1=1} = u_{d_1=1} + \max_{d_2}\left[\mathbb{E}\left(u_{d_2}\right)\right] = \max[a + \pi b - (1-\pi) \times c; a - r]$$

$$U^D_{d_1=0} = u_{d_1=0} + \max_{d_2}\left[\mathbb{E}\left(u_{d_2}\right)\right] = \max[\pi b - (1-\pi) \times c; 0]$$

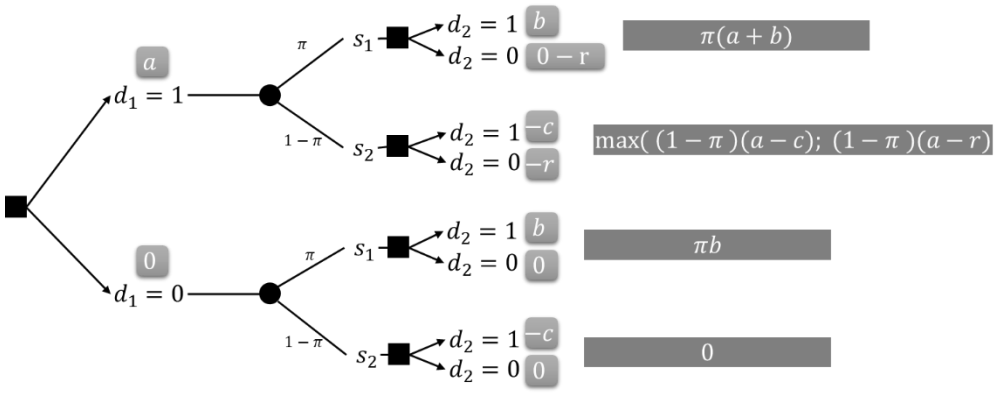


L'ajout de coût de retournement ne modifiera jamais la décision initiale d'adopter du décideur D. En effet, cette dernière se retrouvera confirmée dans la mesure où les coûts de retournement rendent l'utilité de $d_2 = 0$ (sachant $d_1 = 1$) moins attractive que $d_1 = 0$.

Le troisième décideur (L) prend une décision en anticipant l'obtention d'information permise par le report de la décision. Dans ce nouveau contexte, la décision $d_1 = 1$ n'est plus irréversible. S'il reporte sa décision au temps 2, le décideur L prendra une décision basée sur l'utilité attendue de la décision d'investissement optimale *ex-post* : $\mathbb{E}\left(\max_{d_2}\left[u_{d_2}\right]\right)$.

$$U^L_{d_1=1} = u_{d_1=1} + \mathbb{E}\left(\max_{d_2}\left[u_{d_2}\right]\right) = a + \pi b - (1-\pi) \times \min[c; r]$$

$$U^L_{d_1=0} = u_{d_1=0} + \mathbb{E}\left(\max_{d_2}\left[u_{d_2}\right]\right) = \pi b$$



Pour le décideur L, l'ajout des coûts de retournement a pour effet de modifier la valeur de l'utilité de $U^L_{d_1=1}$. Certes, la flexibilité permet de tenir compte de l'information, mais avec un coût de retournement. C'est donc le minimum entre coût de retournement et c qui constituera la valeur de l'information une fois l'adoption décidée. On généralise ainsi les résultats précédents à la situation d'irréversibilité partielle.

La flexibilité de la décision d'adoption permet de définir également une $QOV_{d_1=1}$. En reprenant les cas de figure précédents et en distinguant AL et AN. On définit :

$$QOV_{d_1=1} = AL - \max[AN, RN + a]$$

$$QOV_{d_1=0} = RL - \max[AN - \mathrm{a}, RN]$$

Les quantités suivantes peuvent être recalculées :

- $AN = a + \pi b - (1-\pi) \times c$
- $AL = a + \pi b - (1-\pi) \times \min[c; r]$
- $RN = 0$
- $RL = \pi b$

Si la décision est réversible, $\min[c; r] = r$

Si $\pi b > (1-\pi) \times c$

Hypothèse	$\boldsymbol{AL > AN > RL}$	$\boldsymbol{AL > RL > AN}$	$\boldsymbol{RL > AL > AN}$
Implication de l'hypothèse	$\frac{1}{(1-\pi)}a > c > r$	$c > \frac{1}{(1-\pi)}a > r$	$c > r > \frac{1}{(1-\pi)}a$
$QOV_{d_1=0}$	$(1-\pi) \times c$	$(1-\pi) \times c$	$(1-\pi) \times c$
$QOV_{d_1=1}$	$(1-\pi) \times (c-r)$	$(1-\pi) \times (c-r)$	$(1-\pi) \times (c-r)$
ROV	$QOV_{d_1=1}$	$QOV_{d_1=1}$	$QOV_{d_1=0} - a$

IV.E.2 Comparaison avec l'Option of delay de Willan et Eckermann

Nous retrouvons des résultats proche de ceux de Willan et Eckermann (Eckermann and Willan 2008a, 2008b). Il faut noter que ces derniers ne considèrent que le cas d'un bénéfice net positif (i.e. $\pi b - (1-\pi) \times c > 0$). Dans le cas d'une décision irréversible, $c < r$ et $AL = AN$. Dans ce cas on ne compare que AL et RL pour connaître la valeur et le coût d'opportunité du report. C'est le résultat présenté page 54.

Dans le cas d'une décision réversible $c > r$, $AL > AN$ et plusieurs cas peuvent être envisagés.

- Si le coût d'opportunité est supérieur à l'information maximale que l'on peut obtenir ($\frac{1}{(1-\pi)}a > c$), alors le choix sera systématiquement l'adoption et donc AL. Dans ce cas $\mathrm{ROV} = QOV_{d_1=1}$.

- Si le coût d'opportunité est inférieur à l'information maximale que l'on peut obtenir ($\frac{1}{(1-\pi)}a < c$), alors, il existe un arbitrage entre le coût d'opportunité et le coût de report. Tant que $QOV_{d_1=0} < QOV_{d_1=1} + a$, le cout d'opportunité est supérieur au gain lié à l'absence de coût de retournement. Le choix est donc AL et $\mathrm{ROV} = QOV_{d_1=1}$. Lorsque $QOV_{d_1=0} > QOV_{d_1=1} + a$, le coût d'opportunité ne compense plus le coût de retournement et le choix devient RL avec une valeur de $QOV_{d_1=1} + a$. Le passage de AL à RL s'opère lorsque $r = \frac{1}{(1-\pi)}a$.

En pratique, cela signifie que l'information maximale que l'on peut obtenir est inférieure à $(1-\pi)c$ (pour un bénéfice positif). Comme précédemment la QOV se verra amputée d'une

partie de sa valeur pour obtenir QOV_{sample}. On doit ici calculer deux QOV_{sample} distinctes car elles ne seront pas forcément fondées sur le même nombre de sujets (Eckermann and Willan 2008a, 2008b). De même, le coût d'opportunité sera calculé sur la base du nombre de sujets optimisant $QOV_{d_1=0}$.

V CONCLUSION

La littérature des méthodes fondées sur la valeur de l'information est riche, tant sur les aspects théoriques qu'appliqués. Aujourd'hui, les méthodes sont matures et de nombreux outils sont disponibles pour faciliter leur mise en œuvre. Elles restent cependant très peu utilisées en France. Du fait de ses spécificités, le dispositif médical constitue un domaine de choix pour ces méthodes. L'objectif de cette thèse était donc d'illustrer dans quelle mesure les méthodes fondées sur la valeur de l'information pouvaient être utiles lors des décisions prises dans la vie du dispositif médical, dans le contexte décisionnel français.

Pour ce faire, deux cadres d'utilisation décrits dans la littérature ont été envisagés : la priorisation des efforts de recherche et l'optimisation des designs d'études. Ces deux cadres ont été utilisés à deux temps distincts de l'évaluation du dispositif médical : la détermination des études post-inscription (EPI) demandées à l'occasion de la demande de remboursement et la détermination de la taille d'échantillon lors de l'évaluation précoce de l'utilisabilité réalisée par l'industriel en vue de l'obtention du marquage CE. Ces deux applications ont nécessité de préciser le contexte décisionnel (ensemble des choix, fonction-objectif des parties prenantes, etc.) de développer le modèle d'aide à la décision correspondant à la perspective décisionnelle adoptée et enfin, de caractériser l'incertitude. Un exemple concret a été retenu pour chacun des temps :

- La détermination des demandes d'études post-inscription a été illustré par l'exemple des endoprothèses aortiques pour le traitement de l'anévrisme de l'aorte abdominale ;
- La détermination de la taille d'échantillon optimale a été illustrée sur l'évaluation de l'utilisabilité d'un dispositif innovant d'auto-injection d'adrénaline.

Ces exemples ont permis d'identifier les conditions de mise en œuvre des analyses en termes de données requises, de délai, et de complexité.

Partie 2. ANALYSE DE LA VALEUR DE L'INFORMATION POUR LA PRIORISATION DES EFFORTS DE RECHERCHE : INTERET POUR LA DETERMINATION DES ETUDES POST-INSCRIPTION DANS LE CONTEXTE DECISIONNEL FRANÇAIS

I INTRODUCTION : L'EVALUATION DES DISPOSITIFS MEDICAUX

En 2019, le Syndicat National de l'Industrie des Technologies Médicales (SNITEM) recensait plus de 1500 entreprises composant le tissu industriel du dispositif médical français, dont une large majorité de PME (93%). Son chiffre d'affaires était estimé à 30 milliards d'euros annuel et il représentait près de 90 000 emplois directs. Le secteur du dispositif médical se caractérisait notamment par une forte dynamique avec une croissance moyenne annuelle de 5,3% sur la période 2009-2015 (Lesteven et al. 2015).

Du point de vue de l'évaluation, la tentation est forte d'utiliser le médicament comme référence pour l'analyse du dispositif médical. Cette idée est renforcée par l'existence de nombreux acteurs et outils de régulation communs aux deux champs. Il est néanmoins utile de les distinguer pour appréhender les défis auxquels le secteur du dispositif médical se trouve confronté. Il faut tout d'abord rappeler que la réglementation des dispositifs concernait des acteurs issus d'horizons très variés tels que l'informatique, l'électronique, le textile, l'industrie plastique, etc. A ce titre, elle a longtemps été l'apanage du commissariat européen à l'industrie. Par conséquent, le marquage CE répond à une logique de certification et non d'autorisation, comme pour les médicaments. Le rôle de guichet unique assumé par l'ANSM, à travers la délivrance de l'autorisation de mise sur le marché, n'existe pas pour le dispositif. Dès lors, l'évaluation de l'efficacité et de la sécurité des dispositifs repose essentiellement sur le marquage CE, posant une problématique structurelle de sécurité sanitaire (Lesteven et al. 2015). L'exemple le plus médiatisé fut probablement celui des prothèses mammaires PIP, retirées du marché par l'ANSM (Afssaps à l'époque) en 2010. La faiblesse historique de l'évaluation clinique requise par le marquage CE a permis la commercialisation de dispositifs inefficaces, voire dangereux. Il s'agit par exemple d'endoprothèses aortiques à l'origine de graves thromboses, de ruptures de

prothèses de coudes, d'hyperglycémies non mesurées par des glucomètres défaillants, etc. Ces dispositifs ont finalement été retirés suites aux alertes de matériovigilances mais il est remarquable que la plupart d'entre eux n'avaient pas été autorisés à accéder au marché américain, entre autres raisons grâce à une réglementation plus stricte. En réponse, une nouvelle réglementation européenne a été développée pour mieux encadrer la mise sur le marché des dispositifs médicaux (Encadré 1). Ce règlement européen, entré en application le 26 mai 2021, renforce considérablement les exigences en termes d'évaluations et investigations cliniques, et améliore la transparence des données par l'intermédiaire du registre européen Eudamed.

Encadré 1 : Règlements 2017/745 et 2017/746 du parlement européen et du conseil du 5 avril 2017.

Le marché du dispositif médical est régi par les règlements 2017/745 et 2017/746 du parlement européen et du conseil du 5 avril 2017. Le terme *dispositif médical* y est défini comme : « tout instrument, appareil, équipement, logiciel, implant, réactif, matière ou autre article, destiné par le fabricant à être utilisé, seul ou en association, chez l'homme pour l'une ou plusieurs des fins médicales […] et dont l'action principale voulue dans ou sur le corps humain n'est pas obtenue par des moyens pharmacologiques ou immunologiques ni par métabolisme, mais dont la fonction peut être assistée par de tels moyens ». Cette définition recouvre donc un large éventail de produits de santé qui peuvent être classés en trois catégories :

(i) Dispositif médical à usage individuel : textiles techniques, prothèses, pacemakers, etc.

(ii) Equipement : appareil d'imagerie par résonnance magnétique, logiciel e-santé, etc.

(iii) Dispositifs médicaux de diagnostic in vitro.

La mise sur le marché d'un dispositif est soumise à un processus de certification aux exigences essentielles en matière de sécurité et de performances donnant lieu à l'obtention du marquage CE. Une classification en quatre classes, I, IIa, IIb, et III permet de distinguer les exigences applicables en fonction de la destination et du niveau de risque

inhérent à chaque dispositif. L'évaluation de la conformité est réalisée par un organisme notifié sur la base du dossier technique rédigé par le fabricant (à l'exception de la majorité des dispositifs de classe I pour lesquels le fabricant établit lui-même une déclaration de conformité). La désignation des organismes notifiés est la responsabilité des autorités de santé de chacun des états membres (l'Agence nationale de sécurité du médicament et des produits de santé en France).

Les difficultés d'évaluation du secteur ne sont néanmoins pas le seul fait du cadre réglementaire. En effet, un certain nombre de spécificités liées au secteur des dispositifs médicaux en complexifie l'évaluation. Ces caractéristiques sont bien décrites dans la littérature (Lesteven et al. 2015; Levesque et al. 2014; Rousset 2013). D'un point de vue méthodologique, certains éléments nécessaires à l'obtention d'un niveau de preuve élevé sont souvent plus complexes à obtenir (Campbell 2008). On citera notamment la clause d'ambivalence, le double aveugle, mais également la difficulté de distinguer le facteur humain du facteur technique (courbe d'apprentissage, acte de pose…) lors de l'évaluation clinique. Enfin, le contexte d'évolution incrémentale et les cycles de renouvellement courts rendent souvent difficile la réalisation d'essais cliniques dans des conditions raisonnables. Cette dernière limite est d'autant plus vraie que les populations cibles sont parfois très réduites. Des méthodes flexibles telles que les approches bayésiennes se sont montrées plus adaptées aux contraintes imposées par le dispositif médical. Elles ont ainsi été implémentées avec succès par la FDA (Campbell 2011), notamment dans le cadre des designs d'essais dit « adaptatifs ». Elles permettent par ailleurs la prise en compte de l'information disponible sur les versions antérieures du dispositif (Pibouleau and Chevret 2011).

La taille de la population cible fait également émerger une plus forte contrainte économique. Cette contrainte, particulièrement prégnante dans le cas des PME, est présente tout au long du développement du dispositif. La balance entre les coûts de réalisation d'une étude et le retour sur investissement associé à l'accès au marché devrait en effet conditionner la décision du fabricant d'arrêter ou poursuivre le processus de commercialisation. Bien qu'il existe un corpus

très riche autour de l'évaluation précoce des technologies de santé (*early technology assessment*), la pénétration de ces méthodes reste limitée. De même, la contrainte économique reste présente une fois le dispositif sur le marché. En effet, le suivi du dispositif dans le cadre d'études de phase IV doit se faire à un coût raisonnable. La réutilisation des bases de données médico-administratives (SNDS) dans le cadre d'approches pharamaco-épidémiologiques permet, dans de nombreux cas, d'évaluer l'efficacité, la sécurité et de renseigner sur l'utilisation des dispositifs en vie réelle.

II CONTEXTE ET RATIONNEL

L'accès précoce à l'innovation en santé pose la question de la nature des données disponibles lors de son évaluation initiale. C'est notamment le cas des dispositifs médicaux pour lesquels la décision de primo-inscription s'effectue dans un contexte de substantielle incertitude. En témoigne l'un des principes d'évaluation édictés par la CNEDiMTS qui est de « permettre aux patients d'avoir un accès à des technologies nouvelles, remboursées avec une prise de risque maîtrisée » (Haute Autorité de Santé 2018).

Nous nous positionnons ici dans le cas général d'un dispositif médical à usage individuel candidat au remboursement via la LPPR. L'inscription du dispositif est subordonnée à une évaluation médico-technique, et le cas échéant à une évaluation médico-économique, [10] par les commissions de la HAS. Cette évaluation en primo-inscription permet d'éclairer la décision de remboursement et la négociation du prix par le Comité économique des produits de santé sur la base d'une première estimation de l'efficacité et de l'efficience du dispositif. Celle-ci est cependant entachée de fortes incertitudes compte tenu de la nature des données disponibles à ce stade précoce de maturité. Dans ce contexte, une ou plusieurs étude(s) post-inscription peuvent être demandées[11] en vue de la réinscription du dispositif à une échéance comprise entre un et cinq ans. L'objectif des études post-inscription est d'améliorer la connaissance de l'efficacité et de la tolérance du dispositif, des effets à long terme, des conditions d'utilisation et/ou de prescription, etc. Elles visent ainsi « à collecter des informations pragmatiques, essentielles pour réduire l'incertitude initiale et permettre une réévaluation pertinente des technologies concernées, tant sur les aspects cliniques (bénéfices et risques pour les patients) que sur les aspects collectifs (paramètres économiques, sociétaux…) » (Stamenkovic et al. 2012). La responsabilité de la mise en œuvre des études post-inscriptions repose sur l'industriel. Les types d'études demandées peuvent être très variés : essais pragmatiques, études

[10] Pour toute revendication d'une ASA I, II, ou III avec un impact significatif sur les dépenses de l'Assurance maladie (article R. 165-71-3 du Code de la sécurité sociale).

[11] Par la CT ou la CNEDIMTS, la CEESP et/ou le CEPS.

épidémiologiques observationnelles prospectives, études sur bases de données médico-administratives, etc.

Le fondement rationnel qui sous-tend les demandes d'études post-inscription est largement discuté. Les critères de jugement sont souvent multiples (rapport bénéfice/risque, observance, utilisation des ressources, description des conditions d'utilisation en vie réelle, etc.), complexifiant la mise en œuvre des études par les industriels. À ce problème de hiérarchisation vient s'ajouter l'absence de modèle de référence qui permettrait au décideur de justifier ses choix de manière explicite et transparente (Levesque et al. 2014). Ce manque de prédictibilité et de visibilité quant aux éléments amenant le régulateur français à établir les demandes d'étude post-inscription est régulièrement dénoncé par les fabricants de dispositifs médicaux et de nombreuses études ne sont finalement pas réalisées. Dans ce contexte, l'approche décisionnelle bayésienne et les méthodes fondées sur la valeur de l'information constituent un cadre permettant de considérer explicitement l'incertitude décisionnelle entourant l'adoption d'une technologie de santé et de valoriser l'acquisition d'une information nouvelle en termes de réduction de l'incertitude initiale (Claxton 1999). La valeur de l'information peut être confrontée au coût de mise en œuvre de l'étude proposée, permettant ainsi la comparaison de différentes modalités de recueil de l'information (type d'étude, choix de design et taille d'échantillon, etc.). Ce cadre décisionnel appliqué au contexte français pourrait permettre de prioriser les études post-inscription permettant ainsi de mieux articuler la primo-inscription et la réinscription des produits de santé.

Plusieurs expériences internationales ont été menées afin d'évaluer l'apport des méthodes fondées sur la valeur de l'information dans le processus décisionnel. Le Royaume-Uni a été le premier pays à conduire deux études pilotes évaluant l'intérêt d'incorporer ces méthodes comme outil d'aide à la décision pour le financement de la recherche dans le cadre du programme national d'évaluation des technologies de santé du NHS (Claxton et al. 2004) et pour l'élaboration des recommandations de recherche émises par le NICE (Claxton et al. 2005). A partir de neuf exemples, les études anglaises montrent que l'estimation de l'EVPI et l'EVPPI est possible, moyennant un surcoût raisonnable, à partir des modèles utilisés pour conduire les analyses coût-efficacité. Les paramètres ainsi calculés permettent d'identifier les priorités

de recherche (paramètres, sous-groupes de patients, critères de jugement). Ces travaux, qui demeurent à ce jour les plus complets réalisés sur le sujet, concluaient à l'intérêt de l'approche décisionnelle bayésienne et des méthodes fondées sur la valeur de l'information pour évaluer la pertinence de l'adoption d'un produit de santé, puis de l'articuler avec la question de l'opportunité et du design d'études complémentaires. Le principal obstacle identifié par ces deux études pilotes n'était pas de nature technique ou méthodologique mais plutôt d'ordre contextuel (Claxton and Sculpher 2006). En effet, l'opacité de la prise de décision et l'asymétrie d'information entre les acteurs industriels et institutionnels constituaient autant d'obstacles à l'implémentation de ces méthodes. En 2011, le *Duke Evidence-based Practice Center* évaluait l'utilité pour l'AHRQ américaine d'utiliser les méthodes fondées sur la valeur de l'information pour la priorisation des efforts de recherche (Myers et al. 2011). Bien que ces méthodes aient été jugées « potentiellement utiles » par les membres du panel interrogé, plusieurs limites empêchaient leur implémentation en routine et notamment : l'articulation avec les méthodes existantes, les ressources nécessaires pour mener ces analyses, le timing et le niveau de fidélité de la modélisation. En conclusion, les auteurs jugeaient que ces analyses ne pouvaient se substituer au processus délibératif sous-tendant la détermination des priorités de recherche. Malgré plusieurs publications ultérieures en faveur de l'implémentation (Carlson et al. 2013; Sanders et al. 2016), les recommandations américaines apparaissent en retrait par rapport au NICE qui préconise d'utiliser les méthodes fondées sur la valeur de l'information pour les évaluations menées par les équipes universitaires et, plus récemment encore, pour les soumissions portées par les industriels. A la suite des études pilotes menées par le NICE en 2004 et 2005, plusieurs pays conseillent à des degrés divers d'incorporer une analyse de la valeur de l'information en complément des évaluations « habituelles ». Ainsi les agences d'évaluation des Pays-Bas[12] (Corro Ramos, Rutten-van Molken, and Al 2013), du Canada[13] (Goeree and Levin 2006; Goeree et al. 2009) et dernièrement d'Australie (Tuffaha, Gordon, and Scuffham 2016; Tuffaha and Scuffham 2018),

[12] *Nederlandse Vereniging voor Technology Assessment in de Gezondheidszorg* (Institut d'évaluation des technologies de santé des Pays-Bas)

[13] *Ontario Health Technology Advisory Committee*

ont évalué et implémenté les méthodes fondées sur la valeur de l'information dans leurs guidelines.

Le dénominateur commun de ces pays est l'existence d'un système évaluation des technologies de santé utilisant l'analyse coût-efficacité en routine. La caractérisation et la prise en compte de l'incertitude décisionnelle font d'emblée partie intégrante du processus d'évaluation, notamment à travers la réalisation systématique d'analyses de sensibilité probabilistes et l'existence d'une valeur de référence pour la propension à payer. Les méthodes fondées sur la valeur de l'information ne constituent alors qu'une étape supplémentaire permettant d'évaluer les conséquences d'une mauvaise décision, étape qu'il apparait difficilement justifiable de ne pas mener (Tuffaha 2020). La maîtrise des paradigmes de l'analyse décisionnelle bayésienne est une condition nécessaire à l'appropriation de ces méthodes par les acteurs du processus décisionnel. Dans le contexte français, les membres de la CEESP fondent leurs recommandations sur le rapport coût-efficacité. Ils sont donc familiarisés, au moins implicitement, avec ce type d'approche. Ce n'est pas le cas des membres de la CNEDiMTS pour lesquels le critère de jugement est l'efficacité, ce qui rend le paradigme résolument fréquentiste (intervalle de confiance ou significativité). Cette dualité des critères de jugement, propre au système décisionnel français, est probablement l'un des principaux obstacles à l'implémentation des méthodes fondées sur la valeur de l'information dans le processus d'évaluation. Pour lever ce frein, le critère de jugement habituellement utilisé dans la littérature devra donc être adapté. En effet, le bénéfice net incrémental semble approprié pour la CEESP mais devra être modifié pour la CNEDIMTS qui ne considère (en théorie) que les bienfaits sur la santé.

III OBJECTIF ET STRATEGIE DE RECHERCHE

L'objectif de ce travail était d'évaluer si les analyses de la valeur de l'information pouvaient être utiles à la détermination des études post-inscription dans le contexte décisionnel français. Pour ce faire, nous nous sommes fondés sur un exemple réel de dispositif médical ayant fait l'objet d'une évaluation par la HAS et pour lequel des études post-inscriptions ont été demandées. A

partir de cet exemple, nous avons comparé les études demandées lors de l'évaluation avec celles issues des analyses de la valeur de l'information.

Le choix du dispositif étudié a été réalisé sur la base de trois critères :

(i) Il devait être issu de la liste des dispositifs évalué par la CNEDiMTS et avoir fait l'objet d'une demande d'étude(s) post-inscription ;

(ii) Plusieurs produits devaient être concernés pour ne pas « cibler » un industriel en particulier ;

(iii) L'évaluation du dispositif par la HAS devait avoir été réalisée à distance afin d'éviter le risque d'inférences entre ce travail et l'évaluation ;

(iv) Un modèle de référence devait être disponible pour servir de base à la réalisation des analyses de la valeur de l'information.

Notre choix s'est porté sur les endoprothèses aortiques qui remplissaient ces quatre conditions et pour lesquelles une évaluation a été réalisée en 2009 dans l'indication « anévrisme de l'aorte abdominale sous-rénale » (Lesquelen, Thevenet, and Javerliat 2009a). Il s'agit de l'un des rares exemples pour lequel un modèle d'aide à la décision de référence était disponible. Ce dernier a été développé par le NICE dans le cadre de l'évaluation des endoprothèses menée de manière concomitante au Royaume-Uni. Il a permis à la HAS de disposer des données les plus récentes disponibles à l'époque et a également fait l'objet d'un rapport complémentaire dédié (Lesquelen, Thevenet, and Javerliat 2009b).

Ce travail a été conduit en trois temps :

(i) La première étape a consisté à adapter la structure et les données du modèle d'aide à la décision publié par le NICE au contexte français. De même, la caractérisation de l'incertitude a été revue pour répondre aux prérequis des analyses fondées sur la valeur de l'information ;

(ii) A l'issue de cette première phase, nous avons procédé aux analyses fondées sur la valeur de l'information pour estimer la valeur des informations générées par les d'éventuelles études post-inscription, compte tenu de l'incertitude concernant les différents paramètres du modèle ;

(iii) Enfin nous avons confronté les conclusions de ces analyses aux demandes d'études post-inscription réellement formulées lors de l'évaluation par la HAS.

IV METHODES

Dans une première partie, nous présentons brièvement le modèle initial publié par le NICE. Nous développons ensuite en détail les différentes étapes ayant été nécessaires à l'adaptation du modèle au contexte français. La troisième section est dédiée à la problématique spécifique de la caractérisation de l'incertitude. Enfin, la dernière section présente la mise en œuvre pratique des analyses.

IV.A Modèle d'aide à la décision publié par le NICE

IV.A.1 Prise en charge d'un anévrisme de l'aorte abdominale sous-rénale

L'anévrisme de l'aorte abdominale est une dilatation focale et permanente avec perte du parallélisme des bords de l'artère en aval de l'abouchement des artères rénales (Figure 14). La dilatation est un processus irréversible. L'anévrisme étant le plus souvent asymptomatique, il est découvert à l'occasion d'un dépistage ciblé opportuniste unique des personnes à risque par examen clinique et échographie doppler. En l'absence de dépistage, la rupture de l'anévrisme est le mode le plus fréquent de révélation. Le pronostic est alors sombre avec une mortalité comprise entre 65 et 85% et plus de la moitié des patients qui n'atteignent pas l'hôpital (Sakalihasan, Limet, and Defawe 2005).

Figure 14 : Anévrisme de l'aorte abdominale sous-rénale en CT-angiographie (copyright UpToDate, Inc.)

En 2009, deux techniques étaient disponibles pour le traitement des anévrismes de l'aorte abdominale sous-rénale non rompus. Le traitement de référence, était chirurgical : mise à plat du sac anévrismal puis greffe d'une prothèse synthétique. Depuis le milieu des années 1990, le traitement endovasculaire avec pose d'une endoprothèse par voie fémorale est développé comme alternative thérapeutique à la chirurgie ouverte (Figure 15).

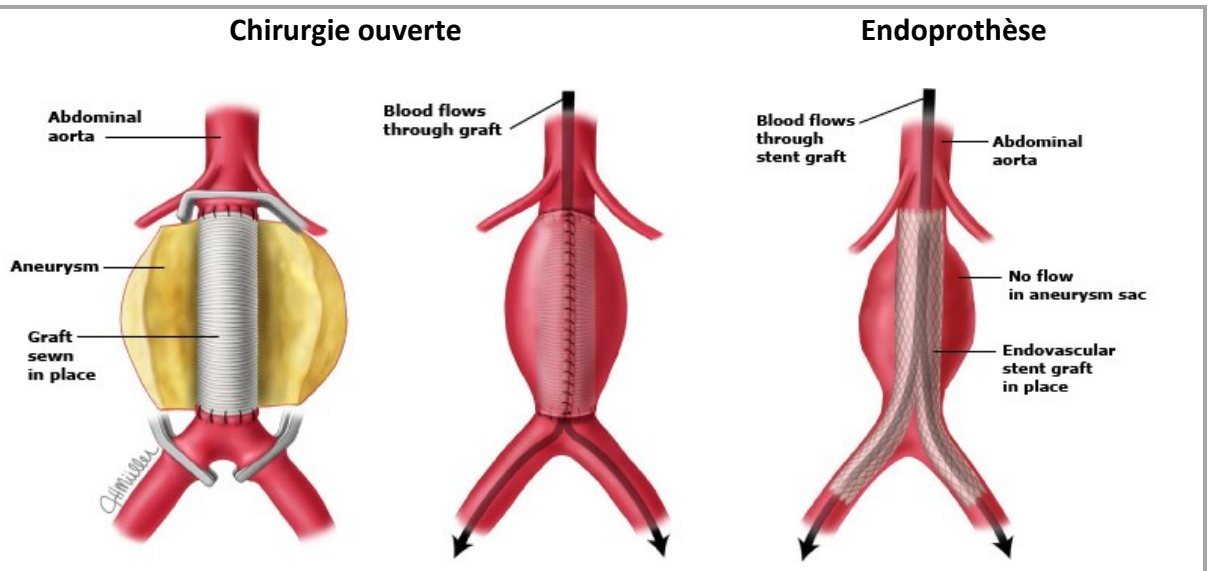


Figure 15 : Alternatives thérapeutiques pour le traitement des anévrismes de l'aorte abdominale sous-rénale non rompus : la chirurgie ouverte à gauche et le traitement endovasculaire à droite (copyright UpToDate, Inc.)

En 2001, l'Afssaps conduisait la première évaluation des bénéfices et risques liés à l'utilisation des endoprothèses dites de 1ère génération. A l'issue de cette évaluation, les endoprothèses avaient vu leur indication restreinte aux patients à risque chirurgical élevé (Encadré 2) ayant un anévrisme de l'aorte abdominale sous-rénale d'un diamètre supérieur à 5 cm ou ayant augmenté de 1 cm en un an.

Encadré 2 : Définition d'un patient à risque chirurgical élevé (Lesquelen, Thevenet, and Javerliat 2009a)

Le patient devait présenter au moins un des facteurs suivants :

- Âge supérieur ou égal à 80 ans.
- Coronaropathie (antécédent[s] d'infarctus de myocarde ou angor) avec test fonctionnel positif et lésions coronariennes pour lesquelles un geste de revascularisation est impossible ou non indiqué.
- Insuffisance cardiaque avec manifestations cliniques patentes.
- Rétrécissement aortique serré non opérable.
- FEVG < 40 %.
- Insuffisance respiratoire chronique objectivée par un des critères suivants : VEMS < 1,2 l/sec ; CV < 50 % de la valeur prédite en fonction de l'âge, du sexe et du poids ; Gazométrie artérielle en l'absence d'oxygène : $PaCO^2$> 45 mm Hg ou PaO^2 < 60 mm Hg ; oxygénothérapie à domicile.
- Insuffisance rénale si créatininémie > 200 µmol/l avant l'injection de produit de contraste.
- Abdomen hostile, y compris présence d'une ascite ou autre signe d'hypertension portale.

Entre 2001 et 2006, des progrès techniques ont permis d'améliorer notablement la sécurité des endoprothèses et un élargissement de la gamme a rendu possible le traitement d'anévrismes de morphologies plus variées. A la suite du renouvellement de l'inscription de 2006, la HAS et l'Afssaps ont décidé conjointement de s'autosaisir pour procéder à la réévaluation des endoprothèses. Leur objectif consistait à réviser les indications de 2001 et à déterminer les études post-inscriptions nécessaires. Cette évaluation s'est achevée en 2009 par la production de deux rapports couvrant les données disponibles sur la période 2001-2009 (Lesquelen, Thevenet, and Javerliat 2009a, 2009b). Dans ce travail, nous avons utilisé les données disponibles dans ces deux rapports pour alimenter le modèle d'aide à la décision sur lequel se fondent les conclusions de l'analyse de la valeur de l'information.

IV.A.2 Le modèle d'aide à la décision développé par le NICE

Le modèle développé par le NICE comparait la chirurgie ouverte (OPEN) et la pose d'endoprothèse par voie endovasculaire (EVAR) chez les patients atteints d'un anévrisme de l'aorte abdominale sous-rénale de 5.5 cm et plus. La perspective retenue était celle du NHS. L'efficacité était mesurée en années de vie ajustée sur la qualité (QALYs, *Quality Adjusted Life Years*) et les coûts étaient mesurés en livres sterlings (2007). La population atteinte d'anévrisme étant à 93% masculine, les femmes étaient exclues de l'analyse.

Le modèle proposé par les Britanniques est un modèle de Markov de cohorte illustré par la Figure 16. Les patients entraient dans le modèle lorsque la décision d'opérer était prise (*Primary Admission*). L''intervention se déroulait soit normalement (transition vers l'état *No Event*) soit les patients décédaient (transition vers l'état *Death*), ou, dans le cas des endoprothèses, subissaient une conversion sur table (transition vers l'état *Conversion to open repair*). Dès lors, les patients parcouraient le modèle par cycles de 6 mois à l'issue desquels ils pouvaient : (i) décéder, (ii) présenter une complication nécessitant une réintervention,[14] (iii) ou rester dans l'état « Absence d'évènement ». L'horizon temporel du modèle était vie entière, c'est-à-dire jusqu'au décès de l'ensemble des patients.

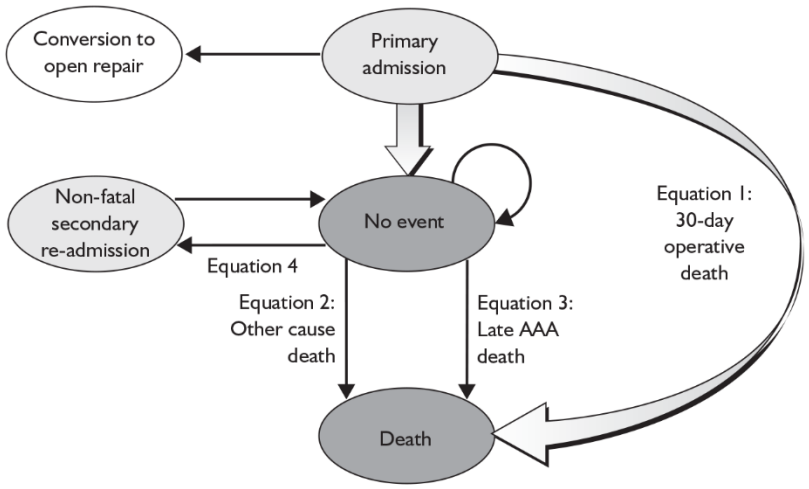


Figure 16 : Structure du modèle d'aide à la décision (reproduit avec la permission de Chambers et al.), UK government's non-commercial license for public sector information v2.0

Deux types de paramètres sont estimés pour calculer les probabilités de transition du modèle : court et long terme. Les paramètres de court terme reflètent le résultat de l'intervention : probabilité de décès opératoire [15] et probabilité de conversion de EVAR vers OPEN. Les paramètres de long terme reflètent la probabilité de réintervention pour complication et celle de

[14] Complications de la laparotomie, endofuites, migrations de prothèses.

[15] Le critère de jugement retenu dans les essais de référence est le décès lors de l'hospitalisation initiale ou jusqu'à 30 jours post-opératoires en extrahospitalier.

décès. Pour cette dernière, la probabilité de décès toutes causes et celle liée à l'anévrisme de l'aorte abdominale sont distinguées afin de modéliser la surmortalité toute cause observée dans le bras EVAR. En effet, le principal avantage de l'abord endovasculaire réside dans le fait que les patients les plus fragiles ont une mortalité opératoire moindre. En contrepartie, une surmortalité non liée à l'anévrisme est observée lors des premières années post-opératoires. Enfin, le calcul du bénéfice net incrémental nécessite l'évaluation du décrément d'utilité et des coûts associés aux différents états du modèle.

Nous avons adapté ce modèle au contexte français en l'alimentant avec les données issues des publications citées dans le rapport de la HAS de 2009. La structure du modèle a également été modifiée. En effet, les recommandations de la HAS tenaient compte de l'efficacité différentielle entre plusieurs sous-groupes, notamment en fonction de l'âge et des comorbidités des patients. Les recommandations publiées en 2001 reflétaient déjà l'importance de cette analyse des caractéristiques des patients. Nous avons donc tenu compte des principales variables définissant les sous-groupes.

IV.B Adaptation du modèle au contexte français

IV.B.1 Matériel

Sur la base d'une revue non systématique de la littérature et des rapports de la HAS, nous avons recensé les données disponibles lors de l'élaboration des recommandations de 2009.

Quatre essais randomisés comparant les deux stratégies pour le traitement des patients avec anévrisme de l'aorte abdominale non rompu et sans contre-indication à la chirurgie ouverte ont été retrouvés. Parmi ces quatre essais, seuls EVAR-1 (Greenhalgh 2004) et DREAM (Prinssen et

al. 2004) remplissaient les exigences de qualité requises.[16,17] Les principales caractéristiques des essais EVAR-1 et DREAM sont synthétisées dans le Tableau 2.[18]

Tableau 2 : Principales caractéristiques des essais DREAM et EVAR-1

	Inclusion	Patients	Critère de jugement		Année
EVAR-1 (Greenhalgh 2004)	2000 à 2003	1082	Mortalité à 30j	↗	2004
EVAR-1 (Greenhalgh 2005)	2000 à 2003	1082	Suivi à 4 ans	NS	2005
EVAR-1 (Brown et al. 2007)	2000 à 2004	1252	Suivi à 4 ans	NS	2007
DREAM (Prinssen et al. 2004)	2000 à 2003	351	Mortalité à 30j	NS	2004
DREAM (Blankensteijn, de Jong, Prinssen, van der Ham, Buth, van Sterkenburg, et al. 2005)	2000 à 2003	351	Suivi à 2 ans	NS	2005

Parmi les trois registres disponibles à l'époque, seul le registre européen EUROSTAR (*European Collaborators on Stent Graft Techniques for Abdominal Aortic Aneurysm Repair*) a été retenu (Tableau 3). En effet, c'était à la fois le plus important en termes de recrutement et le plus pertinent dans le contexte français puisqu'il incluait de nombreux centres de référence du territoire parmi lesquels Grenoble, Lille, Lyon, Montpellier, Nancy, Nîmes, et Paris. Les données issues des deux autres registres, RETA et NVD, n'ont pas été prises en compte dans la mesure où seuls des patients britanniques étaient inclus.

[16] Les essais sont randomisés, les caractéristiques à baselines sont identiques dans les deux groupes, les critères d'éligibilité sont détaillés, les exclus sont pris en compte, un calcul de puissance a priori a été réalisé et l'analyse est en intention de traiter.

[17] Les données de deux autres essais ne sont pas prises en compte pour les raisons suivantes : (i) manque de puissance (pas de calcul de puissance a priori pour Cuypers), (ii) qualité méthodologique faible, (iii) design ne permettant pas de conclure sur la mortalité et (iv) suivi limité (ex : 1 mois pour Cuypers). (Cuypers et al. 2001a) (Cuypers et al. 2001b; Soulez et al. 2005)

[18] Le calcul de puissance de l'essai DREAM se basait sur un critère de jugement à court terme (composite 30 jours – sortie d'hospitalisation). Seuls les résultats d'EVAR-1 ont donc été utilisés pour estimer la mortalité au-delà de 30 jours.

Tableau 3 : Principales caractéristiques du registre EUROSTAR

	Inclusion	Pays	Patients	Suivi
EUROSTAR (EVAR)	1996 à 2003	Europe	5466	6 ans

IV.B.2 Influence des caractéristiques des patients sur les paramètres du modèle

Trente-quatre études ont exploré l'influence des caractéristiques à *baseline* sur le devenir des patients EVAR (Chambers et al. 2009). La majorité de ses études se basaient sur les données du registre EUROSTAR. Un chevauchement important existait entre les patients analysés dans les différentes publications. Ces études étaient pour la plupart de qualité méthodologique moyenne et pouvaient concerner d'anciennes générations d'endoprothèses, surtout lorsque les suivis étaient longs. Les principaux résultats des études concernant l'influence des caractéristiques des patients sur les paramètres du modèle sont détaillés ci-dessous et résumés dans le Tableau 4 :

- L'âge est associé à une augmentation de la mortalité à 30 jours et toutes causes à long terme, et semble être lié à la mortalité causée par l'anévrisme de l'aorte abdominale (Peppelenbosch et al. 2004; Lange et al. 2005; Timaran et al. 2007).[19] L'âge est également un facteur d' augmentation du risque d'endofuites (van Marrewijk et al. 2004) mais pas du risque de réintervention (Hobo and Buth 2006).

- L'insuffisance rénale augmente la mortalité à 30 jours (van Eps et al. 2007). Elle est également associée à la mortalité toutes causes et semble être associée à la mortalité par anévrisme (Peppelenbosch et al. 2004).

- Le score de l'American Society of Anaesthesiologists (ASA) reflète l'état général préopératoire du patient. Un score ASA élevé (>II) est un facteur prédictif de mortalité à

[19] Certaines études traitent l'âge à l'aide d'un seuil (70 ou 80 ans) quand les autres l'analysent comme une variable quantitative.

30 jours et semble être associé à une mortalité toutes causes et par anévrisme plus élevée (van Eps et al. 2007; Leurs et al. 2006).[20]

- Les contre-indications à la chirurgie ouverte ont été définies a posteriori sur les données du registre EUROSTAR (Buth et al. 2002). Elles sont classées en sept catégories : (i) maladies cardiovasculaires, (ii) affections pulmonaires, (iii) maladies cardio-pulmonaires, (iv) affections malignes, (v) contre-indications anatomiques ou à l'abord abdominal, (vi) affection systémique (hémodialyse, myasthénie, schizophrénie, etc.), (vii) mauvais état général. Il semble exister une association entre l'existence d'une contre-indication à la chirurgie ouverte et la mortalité à 30 jours (Buth et al. 2002). De la même manière, l'existence d'une contre-indication à la chirurgie ouverte est également corrélée à la mortalité à long terme liée à l'anévrisme de l'aorte abdominale (Torella 2004). Les preuves du lien avec la mortalité toutes causes sont quant à elles limitées (Leurs, Hobo, and Buth 2004).
- La taille de l'anévrisme est un facteur pronostic de mortalité à 30 jours et à long terme, toutes causes et par anévrisme (Peppelenbosch et al. 2004; van Eps et al. 2007; Buth et al. 2002; Torella 2004). Mais la littérature ne permet pas de conclure au sujet de l'association entre la taille de l'anévrisme et les réinterventions/endofuites.
- Les anciennes générations d'endoprothèses semblent associées à une mortalité à long terme par anévrisme plus élevée (Torella 2004).

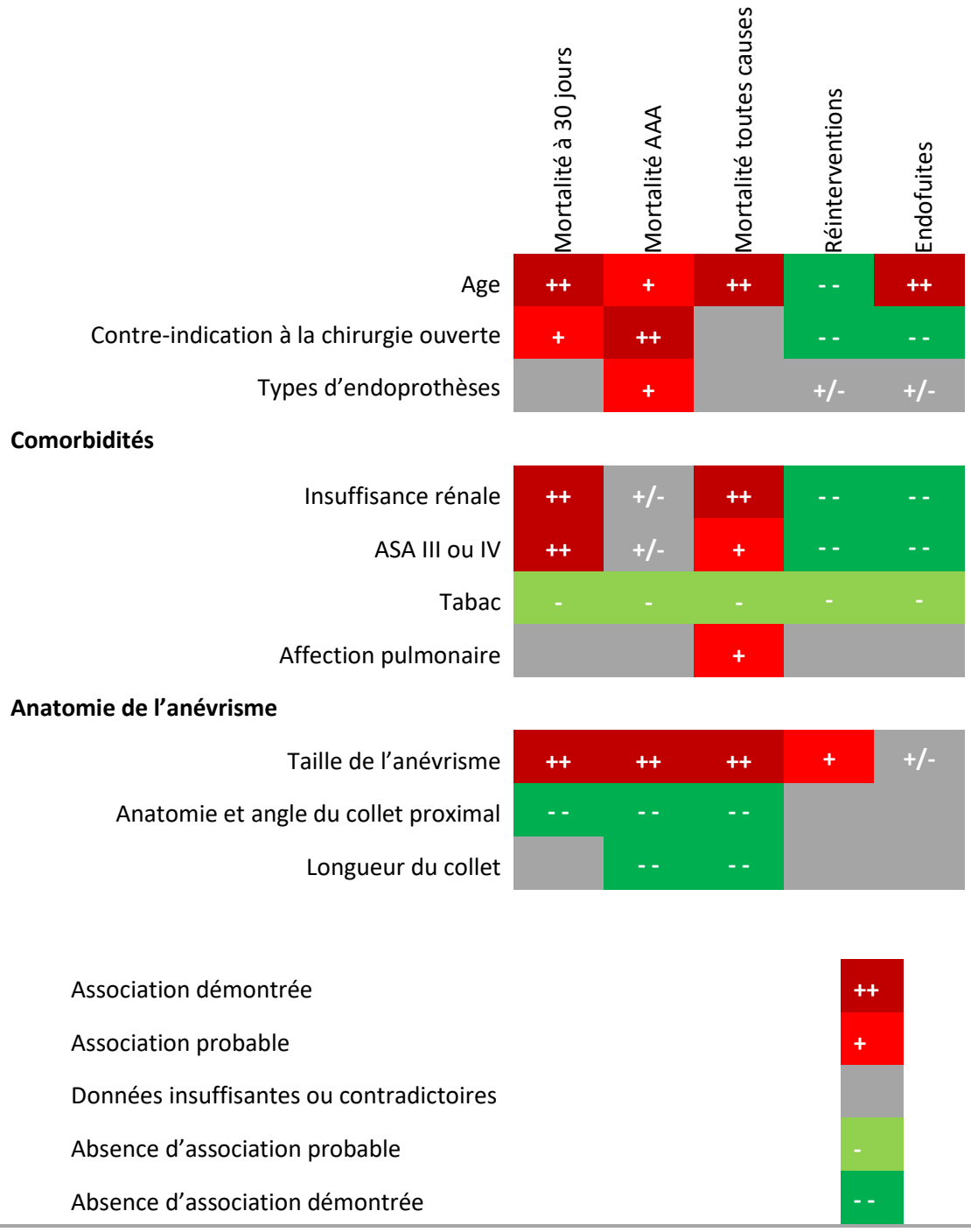
Trois caractéristiques ont finalement été retenues pour stratifier la population en 24 sous-groupes de patients « homogènes » en termes de résultats de santé :

- L'âge : 75 ou 85 ans
- L'état général du patient établi sur la base des comorbidités présentes (insuffisance rénale, score ASA et affection pulmonaire) : bon, moyen, mauvais.

[20] Le rôle de plusieurs comorbidités a été individuellement évalué dans le pronostic d'EVAR : insuffisance pulmonaire, diabète, insuffisance cardiaque, obésité, anémie et hypertension.

- La taille de l'anévrisme : de 5 à 5.4 cm, de 5.5 à 5.9 cm, de 6.0 à 6.4 cm, plus de 6.5 cm.

Tableau 4 : Principaux facteurs de risques modélisés par les études observationnelles



IV.B.3 Estimation de la mortalité opératoire (jusqu'à 30 jours)

La mortalité opératoire a été estimée à l'aide de deux paramètres : la probabilité de décès dans le groupe EVAR et l'odds ratio de décès entre les groupes EVAR et OPEN.

Le risque de décès opératoire des patients du groupe EVAR a été estimé sur les données de registre EUROSTAR (Chambers et al. 2009), à l'aide d'un modèle de régression logistique multivarié (Tableau 5). Les principaux déterminants du décès opératoire étaient l'âge, la clause d'ambivalence (possibilité d'une chirurgie ouverte), la taille de l'anévrisme, la génération de l'endoprothèse et la présence de comorbidités (atteinte rénale, score ASA). La taille de l'anévrisme utilisée comme référence était de 5.5cm. Un coefficient négatif permettait de calculer la mortalité pour un anévrisme de 5cm.

Tableau 5: Risque de décès dans les 30 jours suivant la pose d'une endoprothèse aortique : modèle de régression logistique multivarié sur les données du registre EUROSTAR (1994-2006)

	Coefficient	SE	OR
Par année au-delà de 74 ans	0.071	0.010	1.074
Contre-indication à la chirurgie ouverte	0.631	0.143	1.879
Par cm d'AAA > 5.5 cm	0.298	0.045	1.347
Ancienne génération d'endoprothèse	0.430	0.157	1.537
Atteinte rénale	0.680	0.142	1.974
ASA 3 or 4	0.704	0.165	2.023
Constante	-4.885	0.155	

Le modèle de régression logistique nous permet de prédire la probabilité de décès ($Y = 1$) sachant les covariables ($X = x_1, \dots, x_n$) en utilisant :

$$\text{logit}(Y) = ln\frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Le $\text{logit}(Y)$ est reformuler sous forme de probabilité :

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n+\varepsilon}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n+\varepsilon}} = \frac{e^{logit}}{1+e^{logit}}$$

Pour un patient de 75 ans avec un AAA de 6,5 cm et une atteinte rénale on obtient : $\text{logit}(Y) = 0.071 + 0.298 + 0.680 - 4.88 = -3.836$. Et on en déduit la probabilité de décès $P(Y=1|X) = \frac{e^{-3.836}}{1+e^{-3.836}} = \frac{0.0216}{1+0.0216} = 0.0211.$

L'odds ratio de décès a été calculé par méta-analyse des essais DREAM et EVAR-1 (Chambers et al. 2009). Dans cette méta-analyse, le traitement par EVAR était associé à une réduction significative de la mortalité à 30 jours, OR = 0.35 [IC95% 0.19 à 0.64] (Figure 17). Ce lien était maintenu quel que soit l'état général ou l'âge du patient (Brown et al. 2007; Schermerhorn, O'Malley, et al. 2008).

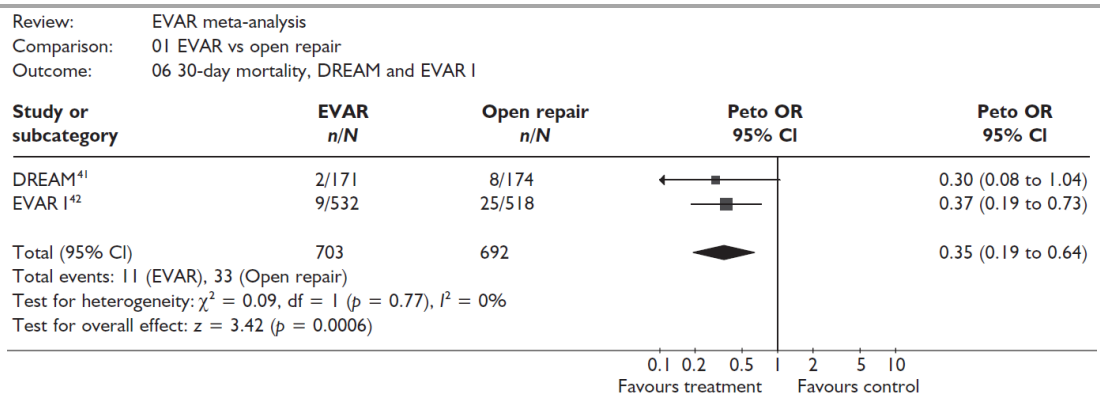


Figure 17 : Métanalyse de l'odds ratio de mortalité à 30 jours comparant chirurgie ouverte (OPEN) et endoprothèse (EVAR). Reproduit avec la permission de Chambers et al. (Chambers et al. 2009), UK government's non-commercial license for public sector information v2.0.

IV.B.4 Estimation de la mortalité à long terme (au-delà de 30 jours)

L'étude de Schermerhorn menée sur une cohorte de 45,560 assurés à Medicare comparait la survie à 5 ans des patients EVAR et OPEN, appariés par la méthode du score de propension (Schermerhorn, O'Malley, et al. 2008). Malgré une diminution de la mortalité précoce, le bénéfice initial du traitement par endoprothèse n'était pas maintenu à moyen, voire long terme. L'importance du bénéfice initial associé à EVAR, ainsi que le délai jusqu'à convergence des deux courbes de survie, augmentaient avec l'âge du patient au moment de l'intervention (Figure 18). Pour un patient de moins de 75 ans, la réduction absolue du risque de mortalité opératoire était inférieure à 2.5% et la convergence était observée avant 24 mois. Au-delà de 85 ans, la réduction absolue du risque de mortalité opératoire était de 8.5% et le bénéfice initial du traitement par endoprothèse était maintenu environ 4 ans. Les données des essais EVAR-1 et DREAM indiquaient quant à eux une convergence des courbes de survie à deux ans (Greenhalgh 2005; Blankensteijn, de Jong, Prinssen, van der Ham, Buth, and van Sterkenburg 2005).

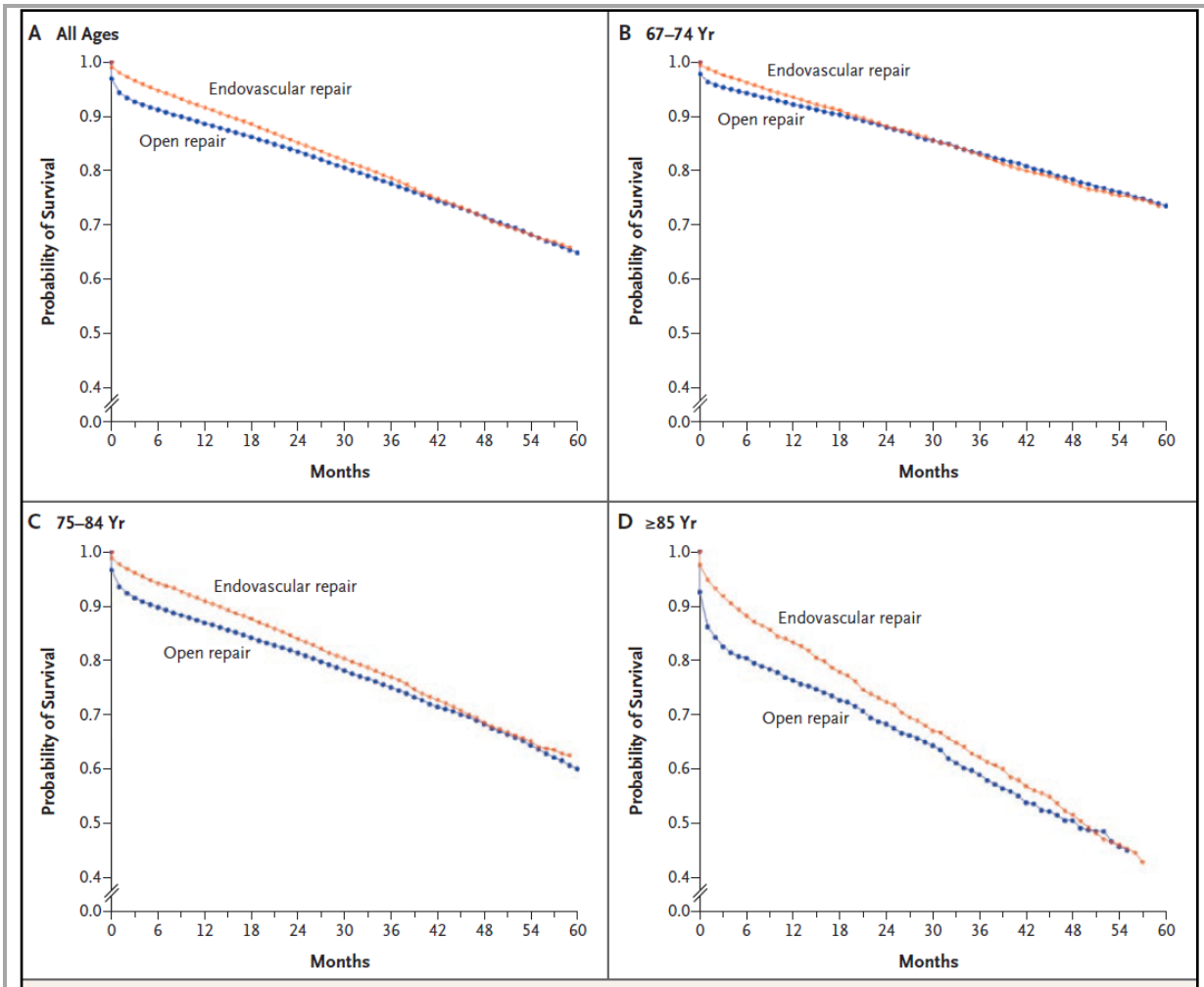


Figure 2. Survival of Patients Undergoing Endovascular Repair or Open Repair of Abdominal Aortic Aneurysms, Overall and According to Age.
Data are shown for all patients (Panel A), for those 67 to 74 years of age (Panel B), those 75 to 84 years of age (Panel C), and those 85 years of age or older (Panel D).

Figure 18 : Survie après réparation endovasculaire ou chirurgie ouverte d'un anévrisme aortique abdominal, globale et en fonction de l'âge. Reproduit avec la permission de Schermerhorn et al., Copyright Massachusetts Medical Society.

Plusieurs hypothèses étaient proposées pour expliquer l'absence de maintien de l'avantage initial du traitement par endoprothèse :

- Le risque de décès lié à l'anévrisme serait plus élevé dans le bras endoprothèse. Dans l'essai EVAR-1, le nombre de décès liés à l'anévrisme est de 2 dans le groupe chirurgie contre 7 dans le groupe endoprothèse avec un suivi d'environ 1250 patients-année. Cette différence n'était pas observée dans l'essai DREAM avec un décès lié à l'anévrisme dans chaque bras.

- Le gain initial associé à EVAR pourrait s'expliquer par la lourdeur de la chirurgie ouverte. Face à des patients par nature à haut risque chirurgical, la chirurgie ouverte précipiterait plus souvent le décès qu'EVAR. Sur le moyen terme, ces patients décéderaient inexorablement conduisant à un rééquilibrage entre les deux groupes.

Afin de modéliser ces deux hypothèses, le modèle proposé par le NICE distinguait le risque de mortalité toutes causes (h_{other}) et le risque de mortalité spécifique liée à l'anévrisme (h_{AAA}). Le risque instantané de mortalité globale était donc :

$$h(t) = h_{\text{other}}(t) + h_{\text{AAA}}(t)$$

La mortalité non liée à l'anévrisme (h_{other}) a été calculée à partir du risque instantané de décès en population générale (h_0) multiplié par le surrisque associé à un anévrisme de grande taille (HR_{large}) et, dans le groupe endoprothèse, par un surrisque supplémentaire (HR_{EVAR}) permettant la convergence des courbes de survie. Le risque de décès à « baseline » (h_0) a été déterminé à partir des tables de mortalité de 2008 fournies par le Centre d'épidémiologie sur les causes médicales de décès (CépiDc). La surmortalité associée à un patient ayant survécu à la chirurgie (HR_{large}) a été calculée sur les données du registre EUROSTAR. A cet effet, une population de référence a été constituée au sein même du registre, à partir des patients ayant un anévrisme de petite taille (<5cm) dont le risque de décès était considéré comme identique à la population générale (h_0). La survie des patients présentant un anévrisme de grande taille a été modélisée par un modèle de Cox ajusté sur les facteurs de risques retrouvés dans les études observationnelles (cf. Tableau 4, page 80). Le modèle est présenté dans le Tableau 6. Dans ce dernier, l'évènement d'intérêt était le décès non lié à l'anévrisme (les décès liés à l'anévrisme étaient traités comme des censures). Le surrisque de mortalité non liée à l'anévrisme dans le bras endoprothèse (HR_{EVAR}) a été calculé sur les données en intention de traiter de l'essai EVAR-1. Pour des effectifs et des durées de suivies comparables, le nombre de décès observés au-delà de 30 jours était de 74 dans le groupe endoprothèses et de 69 dans le groupe chirurgie. Le surrisque de décès dans le groupe EVAR était donc estimé à $HR_{\text{EVAR}} = 1.07$. Ce surrisque était appliqué tant que la survie dans le groupe endoprothèse était supérieure à la survie dans le groupe chirurgie.

Tableau 6: Surrisque de décès non lié à l'anévrisme au-delà de 30 jours comparativement à la population générale : modèle de Cox multivarié sur les données EUROSTAR 1994-2006.

	Coefficient	SE	HR
Par année au-delà de 74 ans	0.043	0.004	Ajustement
Contre-indication à la chirurgie ouverte	0.396	0.076	1.485
AAA 5.1-5.4 cm	0.185	0.112	1.204
AAA 5.5-5.9 cm	0.290	0.113	1.336
AAA 6 -6.4 cm	0.429	0.116	1.536
AAA 6.5+	0.565	0.108	1.759
Ancienne génération d'endoprothèse	0.141	0.070	1.152
Atteinte pulmonaire	0.250	0.067	1.284
ASA 3 or 4	0.334	0.070	1.397
Atteinte rénale	0.332	0.076	1.394

La seconde composante de la mortalité à long terme, la mortalité liée à l'anévrisme (h_{AAA}) a été estimée à partir du risque de décès pour cause d'anévrisme dans le groupe endoprothèse ($h_{\mathrm{AAA|EVAR}}$) multiplié le cas échéant par un facteur protecteur associé au groupe chirurgie ouverte ($HR_{\mathrm{AAA|OPEN}}$). La littérature retrouvait une association forte entre la taille de l'anévrisme et le risque de décès lié à l'anévrisme au-delà de 30 jours, ce risque étant croissant dans le temps (Peppelenbosch et al. 2004; Buth et al. 2002; Torella 2004). La nature de l'évolution du risque dans le temps était néanmoins très incertaine (croissance, stabilité ou décroissance), notamment à cause du faible nombre d'évènements et de l'utilisation d'endoprothèses plus anciennes pour les suivis les plus longs.[21] Le risque de décès a donc été considéré comme constant et modélisé

[21] L'analyse de survie permet de modéliser l'effet du temps sur la probabilité de décès. La fonction de survie (S(t)) est ainsi définie comme la probabilité qu'un individu soit vivant à l'instant t. De même, le hazard (noté h(t)) traduit le risque instantané de décès sur un intervalle Δt sachant que l'individu était vivant au temps t : $h(t) = \lim_{dt\to\infty} \frac{\Pr\,(t\leq T\leq t+\Delta t)}{dt.S(t)}$. Le hazard correspond donc à l'intensité de la transition entre un état « vivant » et l'état « décédé ». A intervalle de temps constant, il peut être transformé en probabilité de transition à l'aide de la relation suivante : $Pr = 1 - e^{-h}$.

Le hazard peut être constant ou varier au cours du temps. La forme de la relation entre le hazard et le temps dépend du modèle de survie utilisé. Le modèle de survie exponentielle correspond à un hazard constant. D'autres modèles comme les modèles de survie Weibull et Gamma permettent de définir une variation monotone du hazard (i.e. un risque de décès croissant ou décroissant au cours du temps). Enfin, les modèles log-normal, log-logistique, etc. permettent de modéliser une variation non monotone (i.e relation en U). Chacune de ces fonctions de survie comporte un ou plusieurs paramètre(s) permettant sa définition. Ce(s) paramètre(s) est (sont) calculé(s) à l'aide de

sur les données du registre EUROSTAR par un modèle paramétrique exponentiel ajusté sur les principaux facteurs de risques retrouvés dans les études observationnelles (cf. Tableau 4, page 80). Deux modèles de survie paramétriques, respectivement de Weibull et Log-normal, ont été estimés sur les même données mais n'ont finalement pas été retenus (moindre vraisemblance statistique du modèle). Le modèle de survie exponentielle est détaillé dans le Tableau 7.

Tableau 7: Risque de décès lié à l'anévrisme au-delà 30 jours suivant la pose d'une endoprothèse aortique : modèle de régression de survie exponentielle multivarié sur les données EUROSTAR 1994-2006.

	Coefficient	**SE**	**HR**
Par année au-delà de 74 ans	0.041	0.010	1.074
Contre-indication à la chirurgie ouverte	0.518	0.143	1.879
Par cm d'AAA > 5.5 cm	0.684	0.045	1.347
Ancienne génération d'endoprothèse	1.324	0.157	1.537
Atteinte rénale	0.895	0.142	1.974
ASA 3 or 4	0.619	0.165	2.023
Constante	-8.525	0.155	

Le modèle de régression de survie exponentielle nous permet de prédire le risque de décès $h(t)$ à partir des différentes covariables x et des coefficients de la régression β : $h(t) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n+\varepsilon}$. Ce risque peut être reformulé sous forme de probabilité de survie jusqu'au temps t : $P(T < t) = 1 - e^{-t*h}$. Ainsi, pour un patient de 75 ans avec un AAA de 6.5 cm, on obtient une probabilité de décès à 6 mois de : $P(T < 6) = 1 - e^{-6*exp(0.041+1.324-8.525)} = 0.00465$.

Le surrisque de décès au-delà de 30 jours lié à l'anévrisme observé dans le groupe EVAR a conduit à l'utilisation d'un Hazard Ratio protecteur en faveur de la chirurgie ($HR_{\mathrm{AAA|OPEN}}$). Comme pour la mortalité non liée à l'anévrisme, les données de l'essai EVAR-1 ont été utilisées. Le nombre de décès liés à l'anévrisme au-delà de 30 jours était de 7 dans le groupe endoprothèse et de 2 dans

méthodes de régression « classique » sur les données de survie disponibles dans la littérature. Le choix entre les différents modèles de survie est réalisé sur un critère statistique, la vraisemblance du modèle. Le modèle le plus vraisemblable est retenu. Lorsque des modèles comportent un nombre différent de paramètres, ils sont comparés à l'aide de critères de parcimonie (Akaike Information Criteria, Bayesian Information Criteria) qui valorisent la vraisemblance tout en pénalisant le nombre de paramètre utilisés.

le groupe chirurgie ouverte. Ces chiffres sont à comparer à deux estimations existantes : HR=2.46 [IC95% 0.48 à 12.7] (Epstein et al. 2008) et HR=1.15 [IC95% 0.39 à 3.41] sur les données brutes. Nous avons retenu un hazard ratio de 1.5 tel que proposé par le NICE (Chambers et al. 2009).

IV.B.5 Réadmission au-delà de 30 jours pour complications

Seul l'essai randomisé EVAR-1 traitait de la question de la réadmission de manière détaillée. Les taux de réintervention étaient de 6.9/100 personnes-années et de 2.4/100 personnes-années respectivement dans les groupes endoprothèse et chirurgie ouverte (Greenhalgh 2005). Le taux de réintervention était plus élevé dans le groupe endoprothèse avec un HR=2.7 [IC95% 1.8 à 4.1]. Néanmoins, les taux incluaient à la fois les réinterventions lors de l'hospitalisation initiale et des réadmissions ultérieures. Le *hazard ratio* des seules réadmissions pour réintervention était estimé par une régression de Weibull (Tableau 8). Le hazard ratio estimé était cohérent avec les données de la revue de la littérature de la HAS (Lesquelen, Thevenet, and Javerliat 2009a; Verzini et al. 2002; Biebl et al. 2005; Brewster et al. 2006; Sampram et al. 2003; Hobo and Buth 2006; Hinchliffe et al. 2006; Schouten et al. 2005).

Tableau 8: Risque de réadmission au-delà 30 jours suivant la pose d'une endoprothèse aortique : modèle de régression de survie de Weibull sur les données EVAR-1.

	Coefficient	**SE**	**HR**
$\boldsymbol{HR}$	1.91	0.38	6.75
$\boldsymbol{ln(\lambda)}$	-6.12	0.43	
$\boldsymbol{ln(\gamma)}$	-0.513	0.001	

Contrairement au modèle de survie exponentielle, le risque dépend du temps. Le paramètre de forme γ détermine la nature de l'évolution du risque au cours du temps et le paramètre d'échelle λ détermine le délai d'occurrence de l'évènement. Ainsi, le modèle de régression de survie de Weibull nous permet de prédire le risque de réadmission pour complications : $h(t,\gamma,\lambda) = \lambda * \gamma * t^{\gamma-1}$. Ce risque peut être reformulé sous forme de probabilité de survie jusqu'au temps t : $P(T < t) = 1 - e^{-t*h}$.

IV.B.6 Ressources utilisées et coûts des procédures EVAR et OPEN

L'objectif était de déterminer un coût moyen pour chacune des procédures, un coût moyen pour une réintervention et le coût du suivi. Pour obtenir ce résultat, nous avons utilisé la base

nationale du Programme de Médicalisation des Systèmes d'Information (PMSI), base de données médico-administratives qui enregistre l'activité médicale de l'ensemble des établissements de santé en France. Cette base regroupe notamment les diagnostics (codés selon la 10e Classification Internationale des Maladies CIM10) et les actes réalisés au cours des séjours (codés selon la Classification Commune des Actes Médicaux CCAM). Les séjours sont catégorisés en Groupes Homogènes de Malades (GHM) puis facturés en Groupes Homogènes de Séjour (GHS).

Les actes associés à EVAR et OPEN ont été identifiés dans la CCAM en vigueur en 2008 (version 10). Après exclusion des gestes qui concernaient les anévrysmes supra-rénaux et ceux qui décrivaient un clampage supra-rénal, nous avons retenu 3 actes pour EVAR et 6 actes pour OPEN. Le code et le libellé de chacun de ces actes CCAM sont détaillés dans le Tableau 9.

Tableau 9 : Actes CCAM caractérisant les procédures OPEN et EVAR.

Acte CCAM	Libellé
	EVAR
DGLF0010	Pose d'endoprothèse couverte bifurquée aortobisiliaque, par voie artérielle transcutanée
DGLF0020	Pose d'endoprothèse couverte aorto-uniiliaque, par voie artérielle transcutanée
DGLF0050	Pose d'endoprothèse couverte rectiligne dans l'aorte abdominale infrarénale, par voie artérielle transcutanée
	OPEN
DGPA0050	Mise à plat d'un anévrisme aortique infrarénal non rompu avec remplacement prothétique aortoaortique infrarénal, par laparotomie avec clampage infrarénal
DGPA0100	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque non rompu avec remplacement prothétique aortobifémoral, par laparotomie avec clampage infrarénal
DGPA0120	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque non rompu avec remplacement prothétique aortobisiliaque, par laparotomie avec clampage infrarénal
DGPA0160	Mise à plat d'un anévrisme aorto-ilio-fémoral avec remplacement prothétique bifurqué aorto-ilio-fémoral, par laparotomie avec clampage infrarénal
DGPA0180	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque rompu avec remplacement prothétique, par laparotomie
EDPA0050	Mise à plat d'un anévrisme iliaque avec remplacement prothétique aorto-iliaque ou aortofémoral unilatéral, par laparotomie

Les statistiques publiques de la base nationale du PMSI (disponibles sur scansante.fr) nous ont ensuite fourni les répartitions de ces actes par GHM en 2008. Nous n'avons retenu que les GHM relatifs aux affections de l'appareil circulatoire (CMD 05), compatibles avec une intervention sur un anévrisme de l'aorte abdominale, afin de pallier la non-spécificité des actes sélectionnés et les erreurs de codage.[22,23] Le code et le libellé de chacun des GHM sont détaillés dans le Tableau 10.

Tableau 10 : GHM caractérisant les procédures OPEN et EVAR.

GHM	Libellé
	EVAR
05C101	Chirurgie majeure de revascularisation, niveau 1
05C102	Chirurgie majeure de revascularisation, niveau 2
05C103	Chirurgie majeure de revascularisation, niveau 3
05C104	Chirurgie majeure de revascularisation, niveau 4
05C111	Autres interventions de chirurgie vasculaire, niveau 1
05C112	Autres interventions de chirurgie vasculaire, niveau 2
05C113	Autres interventions de chirurgie vasculaire, niveau 3
05C114	Autres interventions de chirurgie vasculaire, niveau 4
05K061	Endoprothèses vasculaires sans infarctus du myocarde, niveau 1
05K062	Endoprothèses vasculaires sans infarctus du myocarde, niveau 2
05K063	Endoprothèses vasculaires sans infarctus du myocarde, niveau 3
05K064	Endoprothèses vasculaires sans infarctus du myocarde, niveau 4
05K06T	Endoprothèses vasculaires sans infarctus du myocarde, très courte durée
	OPEN
05C101	Chirurgie majeure de revascularisation, niveau 1
05C102	Chirurgie majeure de revascularisation, niveau 2
05C103	Chirurgie majeure de revascularisation, niveau 3
05C104	Chirurgie majeure de revascularisation, niveau 4

[22] Exemple : exclusion du GHM 05K05V « Endoprothèses vasculaires et infarctus du myocarde sans CMA »

[23] Exemple : pour l'acte OPEN DGPA005, 1213 hospitalisations ont été enregistrées dans la base nationale du PMSI en 2008 sur plus de 20 GHM différents. 7 GHM relevaient de la CMD 05 mais on trouvait parmi eux 2 GHM de pontage aortocoronariens (05C05W et 05C04W). Finalement, les 5 GHM restants (05C06V, 05C06W, 05C08W, 05C10V, 05C10W) représentaient 1167 hospitalisations avec l'acte DGPA005.

Pour le calcul des coûts associés à chacune des deux procédures, nous avons utilisé le référentiel national de coûts qui fournit une estimation des coûts moyens par GHM à partir d'un échantillon d'établissements. Ce dernier se base sur des outils de comptabilité analytique qui permettent de ventiler les coûts sur différents postes de dépenses. Les coûts utilisés excluent les charges de structure (immobilier et financier). Les données 2008 ont été publiées pour des GHM au format v11b (i.e. en intégrant les niveaux de sévérité), permettant le traitement individuel des GHM.

Pour chacune des deux procédures, nous avons extrait l'ensemble des GHM d'intérêt entre le 1er mars 2009 et le 28 février 2010 (année de facturation), puis nous avons exclu ceux ne contenant pas un acte CCAM caractéristique de la procédure. L'agrégation des nombres d'hospitalisations par GHM a abouti à un case-mix pour OPEN et EVAR. La répartition du nombre de séjours sélectionnés entre ces GHM définissait leur pondération.[24] La procédure EVAR était répartie sur trois racines (05C10, 05C11 et 05K06), pour un total de 13 GHM. La procédure OPEN est répartie sur une racine de GHM (05C10) composée de 4 GHM (Chirurgie majeure de revascularisation, niveaux 1 à 4). Les données fournies par le référentiel national de coûts et correspondant à ces GHM sont résumées dans le Tableau 11 et le Tableau 12, respectivement pour EVAR et OPEN.

[24] Exemple : 3489 séjours ont été identifiés comme des hospitalisations OPEN sur les critères actes/GHM précédents. Le case-mix en racines était le suivant : 05C06 (Autres interventions cardiothoraciques, âge supérieur à 1 an, ou vasculaires quel que soit l'âge, avec circulation extracorporelle) 14 séjours (0.40%), 05C08 (Autres interventions cardiothoraciques, âge supérieur à 1 an, ou vasculaires quel que soit l'âge, sans circulation extracorporelle) 2 séjours (0.06%), 05C10 (Chirurgie majeure de revascularisation) 3473 séjours (99.54%).

Tableau 11 : Référentiel ENCC v11b (2008) pour les GHM 05C10 (Chirurgie majeure de revascularisation, niveaux 1 à 4), 05C11 (Autres interventions de chirurgie vasculaire, niveaux 1 à 4) et 05K06 (Endoprothèses vasculaires sans infarctus du myocarde, niveaux 1 à 4 et très courte durée). Le nombre de chirurgies EVAR (n) n'est pas présent dans le référentiel mais est calculé à partir du nombre d'actes de pose d'endoprothèses aortiques dans une extraction de la base nationale du PMSI. Les valeurs sont arrondies pour clarifier la présentation.

GHM	Nb France	Nb ENCC	Nb EVAR	Coût	σ
05C101	6065	1500	207	8322	605
05C102	4336	1249	427	11294	920
05C103	2478	677	107	17132	1402
05C104	1182	310	37	28191	1319
05C111	3634	882	222	5873	414
05C112	1787	481	200	9331	626
05C113	979	223	54	15479	744
05C114	354	82	17	25431	1181
05K061	36011	8710	745	4078	267
05K062	9599	2513	853	6651	487
05K063	1530	395	360	11457	975
05K064	261	76	79	18352	1552
05K06T	6539	1991	4	3028	413

Tableau 12 : Référentiel ENCC v11b (2008) pour les GHM 05C10 (Chirurgie majeure de revascularisation, niveaux 1 à 4). Le nombre de chirurgies OPEN (n) n'est pas présent dans le référentiel mais est calculé à partir du nombre d'actes de chirurgie ouverte d'anévrisme de l'aorte abdominale sous-rénale dans une extraction de la base nationale du PMSI. Les valeurs sont arrondies pour clarifier la présentation.

GHM	Nb France	Nb ENCC	Nb OPEN	Coût	σ
05C101	6065	1500	985	8322	605
05C102	4336	1249	1030	11293	920
05C103	2478	677	473	17131	1402
05C104	1182	310	238	28191	1319

L'endoprothèse fait partie de la liste des dispositifs facturés en sus. Nous avons établi le coût moyen d'une endoprothèse à 5447€ sur la base des trois actes de pose et de leur proportion respective dans le case-mix EVAR. Le détail du calcul de coût est présenté dans le Tableau 13. Pour le calcul du coût moyen d'EVAR, nous avons déduit les charges directes pour les DMI facturables en sus puis ajouté le tarif moyen d'un DMI EVAR.

Tableau 13 : Coût d'une endoprothèse aortique. Le coût du corps, d'une extension/jambage est issu de la LPP 2008.

Acte CCAM	Libellé	%	Corps	Ext./Jamb.	Total
DGLF001	Pose d'endoprothèse couverte bifurquée aortobisiliaque, par voie artérielle transcutanée	72	3577€	2254€	5831€
DGLF002	Pose d'endoprothèse couverte aorto-uniiliaque, par voie artérielle transcutanée	22	3577€	1127€	4704€
DGLF005	Pose d'endoprothèse couverte rectiligne dans l'aorte abdominale infrarénale, par voie artérielle transcutanée	6	3577€	0€	3577€

Pour les conversions sur table (abord endovasculaire vers chirurgie ouverte), le PMSI ne permet pas de les détecter puisqu'un seul acte est pris en compte pour le groupage et qu'il n'existe ni acte ni diagnostic spécifique à ce cas de figure. Nous avons donc utilisé la somme des coûts d'OPEN et d'EVAR comme approximation du coût d'une conversion sur table. Le coût moyen d'une réintervention a été déterminé à partir de l'extraction des séjours des patients pour les 4 années suivant l'intervention initiale. Seuls les GHM en rapport avec une intervention de l'aorte étaient considérés. Le coût moyen et l'erreur standard étaient calculés selon la même méthode que précédemment. Enfin le coût moyen d'une visite de suivi correspondait à la somme des coûts (en 2008) d'un scanner abdomino-pelvien (96.27€) et d'un acte de consultation de spécialiste (25€), à savoir 121.27€.

Au total, l'exploitation des données du référentiel national des coûts de l'ENCC a abouti aux résultats présentés dans le Tableau 14. La moyenne pondérée des coûts issus de l'ENC par racine (après déduction des charges directes pour les DMI facturables en sus et ajout du tarif moyen d'un DMI EVAR moyen) conduit à une estimation du **coût moyen de 10648€ pour EVAR et de 12497€ pour OPEN**. Une note contenant une version détaillée du calcul des coûts et des modalités de caractérisation de l'incertitude est proposée à l'annexe Partie 6.I, page 179.

Tableau 14 : Coûts moyens et Erreur Standard des procédures EVAR et OPEN issus du référentiel national de coûts 2008. Les valeurs sont arrondies pour clarifier la présentation.

Procédure	Cout moyen	Erreur Standard
EVAR	12 748€	824€
OPEN	12 708€	1 091€
Conversion	25 456€	1 915€
Réintervention	8 228€	2 114€

IV.C Caractérisation de l'incertitude

La caractérisation de l'incertitude paramétrique est un élément fondamental de la réalisation des analyses fondées sur la valeur de l'information. Pour chacun des paramètres du modèle, nous avons défini une distribution de probabilité reflétant l'incertitude de l'estimation dans la littérature. La forme de chacune des distributions variait en fonction du type de paramètres (probabilité, odds ratio, coût, utilité, etc.) et se basait sur les recommandations de référence en termes de modélisation de l'incertitude paramétrique dans les analyses médico-économique (Drummond et al. 2015; Briggs, Sculpher, and Claxton 2006; Briggs et al. 2012). A titre d'exemple, l'incertitude entourant l'odds ratio était modélisée par une loi log-normale et la probabilité de conversion par une loi béta. Pour chacun des paramètres du modèle d'aide à la décision, la distribution de probabilité, ses paramètres et la source bibliographique sont résumés dans le Tableau 15.

Tableau 15 : Caractérisation de l'incertitude paramétrique.

	Distribution	Paramètres	Source
Mortalité opératoire (30 jours)			
Probabilité de décès (EVAR)	Normale conjointe	Tableau 5	EUROSTAR
Odds ratio EVAR vs OPEN	Log-normal	Moy=-1.05 ; SE=0.3	Meta-analyse
Probabilité de conversion	Béta	Alpha=4 ; Beta=496	EVAR trial 1
Mortalité à long terme non liée à l'anévrisme			
HR anévrisme	Normale conjointe	Tableau 6	EUROSTAR
HR EVAR vs OPEN	Log-normal	Moy=0.07 ; SE=0.16	EVAR trial 1
Mortalité à long terme liée à l'anévrisme			
Hazard mortalité EVAR	Normale conjointe	Tableau 7	EUROSTAR
HR EVAR vs OPEN	Log-normal	Moy=0.41 ; SE=0.54	NICE
Réhospitalisation pour complication			
HR EVAR vs OPEN	Log-normal	Tableau 8	Evar trial 1
Coûts (€, référence 2008)			
Procédure OPEN	Gamma	SE=480	ENCC
Procédure EVAR	Gamma	SE=310	ENCC
Conversion	Gamma	SE=23695	ENCC
Réhospitalisation	Gamma	SE=73	ENCC
Utilités			
Décrément OPEN	Gamma	SE=0.016	Kind et al.
Décrément EVAR	Gamma	SE=0.014	Kind et al.

Plusieurs paramètres dépendent des caractéristiques du patient et sont modélisés par l'intermédiaire de modèles statistiques multivariés :

- La probabilité de décès est estimée par une régression logistique multivariée (Tableau 5).
- Le surrisque de décès non lié à l'anévrisme de la population porteuse d'un anévrisme aortique abdominal est estimé par une régression de Cox multivariée (Tableau 6).
- Le surrisque de décès lié à l'anévrisme dans le groupe EVAR est estimé par un modèle multivarié de survie paramétrique exponentielle (Tableau 7).

- La probabilité de ré-hospitalisation pour complication est estimée par une régression de survie de Weibull (Tableau 8).

Ces modèles statistiques ont l'avantage de prendre en compte les corrélations entre les différents paramètres et de modéliser plus fidèlement le résultat de santé. L'incertitude paramétrique associée aux paramètres d'un même modèle statistique (par exemple l'erreur standard des coefficients de la régression logistique) ne peut être considérée de manière indépendante. L'utilisation d'une distribution de probabilité normale conjointe a permis de propager simultanément l'incertitude à l'ensemble du modèle statistique et de préserver les corrélations entre ces paramètres. Pour ce faire, une transformation de Cholesky a été appliquée à la matrice de variance-covariance du modèle pour servir de support aux simulations de Monte Carlo (Briggs et al. 2012; Briggs, Sculpher, and Claxton 2006).

La propagation de l'incertitude a été réalisée par l'intermédiaire de 10,000 simulations de Monte Carlo. Une simulation consistait en un tirage aléatoire dans chacune des distributions de probabilité des paramètres, puis par le calcul du modèle de Markov pour obtenir les résultats de santé associés aux deux alternatives, EVAR et OPEN. En pratique, cette étape consistait en une boucle interne de 48 itérations (inner loop) correspondant aux 24 sous-groupes de patients pour chacune des deux alternatives (Groot Koerkamp et al. 2010). Les résultats de santé étaient ensuite pondérés en fonction de la répartition des patients dans chacun des groupes. Au total, 24,000 itérations du modèle de Markov étaient nécessaires pour obtenir l'analyse de sensibilité probabiliste correspondant à une des deux alternatives. Les résultats de santé moyens étaient issus de l'espérance des 10,000 simulations, pour refléter toute asymétrie dans la distribution des résultats de santé (Briggs et al. 2012).

L'EVPI était calculée comme suit, grâce à l'analyse de sensibilité probabiliste :

$$\mathrm{EVPI} = E_\theta \left[\max_{\mathrm{d}}\{U(d,\theta)\}\right] - \max_{\mathrm{d}}\{E_\theta[U(d,\theta)]\}$$

Avec $U(d,\theta)$ représentant l'utilité du critère de jugement pour la décision d sur l'ensemble des paramètres du modèle notés θ.

Le calcul de l'EVPPI impliquait une boucle extérieure supplémentaire (outer loop). Cette dernière permettait de tirer aléatoirement une réalisation possible du (des) paramètre(s) d'intérêt. Ce(s) paramètre(s) étai(en)t ensuite maintenu(s) constant(s) pour le calcul du modèle de Markov tel que précédemment décrit. L'EVPPI pour du sous-ensemble de paramètre(s) θ_i était calculé de la manière suivante :

$$\text{EVPPI} = E_\theta \left[\max_{\text{d}} \{ E_{\theta_{-i}|\theta_i} (U(d,\theta)) \} \right] - \max_{\text{d}} \{ E_\theta [U(d,\theta)] \}$$

Au total, le calcul de l'EVPPI pour un sous-ensemble de paramètres impliquait un total de 48 millions de simulations de Monte Carlo.

IV.D Analyses

Une cohorte hétérogène de 10,000 patients candidats à un traitement de leur anévrisme aortique abdominal sous-rénal a été introduite dans le modèle de Markov. La cohorte était composée des 24 sous-groupes de patients identifiés grâce à la revue de la littérature (voir chapitre IV.B.2, page 78). La répartition des patients dans chacun des sous-groupes était issue des données du registre EUROSTAR. Les transitions entre chacun des états du modèle étaient réalisées par cycle de 6 mois après un premier cycle initial de 30 jours correspondant à la période post-opératoire. Les patients parcouraient le modèle jusqu'au décès (horizon temporel vie entière). L'incertitude était propagée à travers le modèle par simulation de Monte Carlo, grâce à un tirage aléatoire dans les distributions de probabilité de chacun des paramètres du modèle. Les coûts et les utilités étaient actualisés au taux annuel de 3.5%. Le modèle a été programmé sous le logiciel « R statistical programming language » version 3.6.3 (R Development Core Team 2010).

Pour chacune des deux stratégies évaluées, les critères de jugement suivants ont été estimés :

- Années de vie : la survie moyenne d'un patient en nombre d'années.
- QALYs : la survie moyenne d'un patient pondérée par l'utilité associée aux différents états de santé et reflétant la qualité de vie (0 pour la mort, 1 pour la santé parfaite).

- Coûts : coût moyen « vie entière » exprimé en euros (€, année de référence 2008). La perspective retenue était celle du système de santé.
- Ratio différentiel coût-résultat.

Concernant les décideurs, les critères de jugement retenus étaient :

- Du point de vue de la CNEDiMTS : les années de vie.
- Du point de vue de la CEESP : le bénéfice net incrémental exprimé en euros et calculé sur la base d'une hypothèse de propension à payer égale à 100 000€/QALY. En l'absence de consensus concernant le seuil de propension à payer en France, nous avons systématiquement présenté le ratio différentiel coût-résultat.

La population cible se basait sur une incidence annuelle de 5900 patients traités pour un anévrisme de l'aorte abdominal sous-rénal non rompu en France. Les conclusions des analyses fondées sur la valeur de l'information étaient supposées s'appliquer durant 5 ans, jusqu'à échéance de la demande d'étude post-inscription.

Dans un premier temps, nous avons évalué l'adéquation entre la décision prise par la HAS en 2009 de généraliser la recommandation d'utilisation des endoprothèses à l'ensemble des patients et les conclusions obtenues grâce au modèle d'aide à la décision. Dans un second temps, nous avons caractérisé l'incertitude entourant la décision par l'intermédiaire d'une analyse de sensibilité de probabiliste et des analyses fondées sur la valeur de l'information. Les résultats obtenus ont été comparés aux demandes d'études post-inscriptions demandées en pratique. Enfin, nous avons évalué l'intérêt de la prise en compte des résultats stratifiés en fonction des sous-groupes sur la décision d'adoption et sur les conclusions des analyses fondées sur la valeur de l'information.

V RESULTATS

V.A Efficacité et efficience des endoprothèses aortiques pour le traitement de l'anévrisme de l'aorte abdominale sous rénal.

Dans le contexte français de 2009, les résultats calculés sur l'ensemble de la population étaient favorables au traitement par endoprothèse et ce quel que soit le critère de jugement retenu (Tableau 16).

Tableau 16 : Résultats de l'analyse coût-efficacité.

	EVAR	OPEN
Années de vie	7,23	7,17
QALYs	4,49	4,43
Coûts	15 375€	13 054€
Bénéfice net*	434 007€	430 000€

* La propension à payer est fixée à 100 000€/QALY

Lorsque les résultats sont analysés du point de vue de l'efficacité et de la sécurité (critères de jugement de la CNEDiMTS), le traitement endovasculaire permet un gain moyen par patient de 0.06 années de vie avec des survies moyennes de 7,17 et 7,23 années. Dans notre modèle, le bénéfice observé en termes d'efficacité s'explique par une mortalité opératoire plus faible dans le groupe EVAR. La survie globale de ces patients converge puis rejoint celle des patients du groupe OPEN dans les deux années suivant l'intervention, sous l'effet de la surmortalité « long terme » observée dans le groupe EVAR.

Lorsque les résultats sont analysés du point de vue de l'efficience (critère de jugement CEESP), le traitement est coût-efficace pour une propension à payer de 100 000€/QALY. Le ratio différentiel coût-résultat (RDCR) du traitement par endoprothèse aortique est de 36 676€/QALY. Le coût « vie entière » du traitement par endoprothèse est plus élevé que pour la chirurgie ouverte (coûts respectifs de 15 375€ et 13 054€). Dans la mesure où les coûts des procédures initiales sont très proches, la différence de coûts « vie entière » observée en faveur de la chirurgie ouverte est essentiellement liée au nombre plus important de réinterventions dans le groupe EVAR et,

marginalement, au coût des conversions de l'abord endovasculaire vers la chirurgie ouverte. Le bénéfice net incrémental est de 4 007€ en faveur du traitement par endoprothèse.

Au total, les résultats obtenus grâce à la modélisation sont concordants avec les recommandations d'adoption des endoprothèses telles que formulées en 2009 par la HAS (Lesquelen, Thevenet, and Javerliat 2009a).

V.B Analyse de sensibilité probabiliste

L'analyse de sensibilité probabiliste est présentée graphiquement sur un plan coût-efficacité (Figure 19). On observe qu'il existe une probabilité non nulle que la décision d'adopter les endoprothèses aortique soit la mauvaise, indépendamment du critère de jugement retenu :

- Du point de vue de l'efficacité, la probabilité que le traitement par endoprothèse améliore la durée de vie du patient est estimée à 73%. En ajustant sur la qualité de vie, la probabilité augmente pour atteindre 87% (i.e. environ 13% des simulations sont situées à gauche de l'axe des ordonnées du plan coût-efficacité).
- Du point de vue de l'efficience, la probabilité que le traitement soit coût-efficace est de 72% au seuil de 100 000€/QALYs.

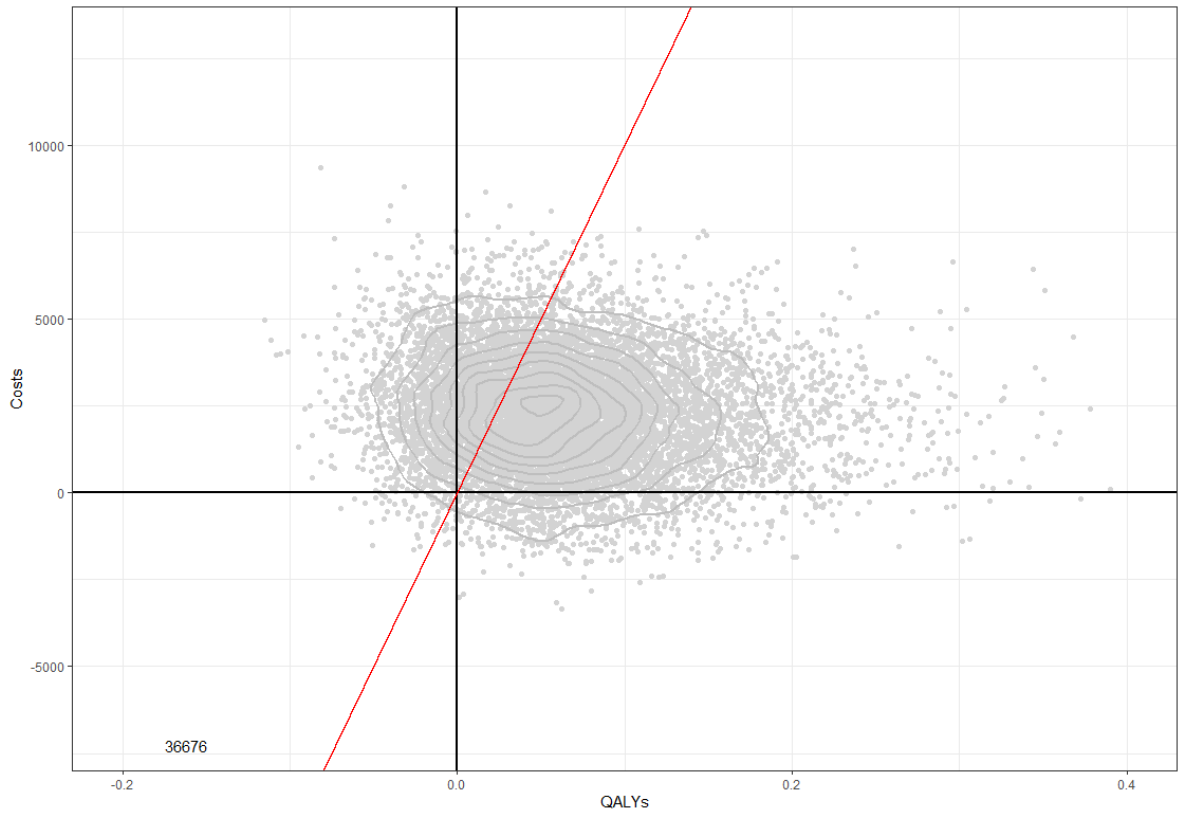


Figure 19 : Plan coût-efficacité représentant l'analyse de sensibilité probabiliste. Le traitement par endoprothèse est comparé à l'alternative de référence (chirurgie ouverte). Le nuage de points correspond au résultat obtenu pour chacune des 10 000 simulations de Monte Carlo. La droite rouge correspond à une propension à payer de 100 000€/QALY.

V.C Analyses fondées sur la valeur de l'information

Suivant le critère de jugement retenu, l'EVPI populationnelle[25] était estimée à 327 années de vie, 88 QALYs où 26 259 576€. Dans la mesure où l'EVPI n'est pas nulle, il y a un intérêt à poursuivre les efforts de recherche, et ce malgré la recommandation d'adoption. D'un point de vue théorique, il s'agit de la valeur maximale que le décideur devrait être prêt à payer pour disposer d'une connaissance parfaite sur l'ensemble des paramètres du modèle d'aide à la décision (absence d'incertitude décisionnelle). Du point de vue du décideur, le résultat confirme par ailleurs la nécessité d'études complémentaires afin de résoudre toute ou partie de l'incertitude

[25] Pour une population cible annuelle de 5900 patients et un horizon des conclusions des analyses fondées sur la valeur de l'information de 5 ans.

décisionnelle. Cependant, l'EVPI ne constitue qu'un ordre de grandeur et il est nécessaire de calculer l'EVPPI pour évaluer les paramètres les plus pertinents à investiguer, principaux « drivers » de l'incertitude décisionnelle.

Les paramètres ont été regroupés de manière à refléter les études pouvant être demandées et mises en œuvre en pratique. Cinq groupes (non exclusifs) ont été identifiés : les paramètres d'efficacité, les paramètres d'efficacité à court terme, les paramètres d'efficacité à long terme, les paramètres d'utilité et les paramètres de coûts. Les résultats sont présentés dans le Tableau 17.

Tableau 17 : Valeur de l'information parfaite par groupes de paramètres (EVPPI).

	Années de vie	QALYs	Bénéfice net
Efficacité	327	63	21 565 395 €
Efficacité court terme	94	2	5 931 914 €
Efficacité long terme	95	11	9 010 052 €
Utilité	0	0	406 416 €
Coûts	0	0	26 974 €
Total (EVPI)	**327**	**88**	**26 259 576 €**

Les principaux paramètres à l'origine de l'incertitude décisionnelle sont les paramètres d'efficacité. Si les paramètres d'efficacité à court et long terme étaient étudiés séparément, l'EVPPI serait significativement diminuée. D'autre part, il ne semble pas pertinent de financer une étude portant sur l'utilité ou les coûts des procédures, dans la mesure où l'incertitude paramétrique entourant ces paramètres semble conduire à une incertitude décisionnelle limitée.

V.D Comparaison avec les études post-inscriptions demandées par la HAS

A ce stade, nous pouvons comparer les résultats obtenus grâce au modèle d'aide à la décision avec les études post-inscriptions effectivement demandées. En effet, le modèle d'aide à la décision n'a utilisé que des données disponibles en 2009 et a conduit à la même décision d'adopter les endoprothèses. En 2009, la HAS subordonnait le renouvellement d'inscription à la réalisation « *d'une étude de suivi mise en place sur une cohorte de patients représentative de la*

population traitée. L'objectif de cette étude est d'évaluer l'intérêt de la technique à long terme, c'est-à-dire au-delà de 5 ans. Cette étude de cohorte devra concerner les 150 premiers patients implantés après inscription sur la LPPR. Des critères simples et précis sont ainsi proposés par le groupe de travail : mortalité globale, complications (endofuite, migration), conversion chirurgicale, évolution et rupture de l'anévrisme» (Lesquelen, Thevenet, and Javerliat 2009a).

Les résultats obtenus grâce au modèle d'aide à la décision formalisent et confirment la pertinence de centrer les investigations sur les paramètres d'efficacité (mortalité liée ou non à l'anévrisme et réintervention). Bien que cette dernière soit porteuse d'incertitude sur le long terme (EVPI de 95 années ou 9 010 052 €), il semble pertinent d'étudier simultanément les paramètres d'efficacité à court terme (mortalité opératoire et conversion) pour maximiser la valeur de l'information (EVPI = 327 années ou 21 565 395 €). L'étude isolée des paramètres du modèle ne présente quant à elle aucun intérêt. Enfin, l'absence de demande d'étude post-inscription s'intéressant aux coûts ou à l'utilité apparaît justifiée, mais doit être interprétée à la lumière de l'absence d'avis d'efficience publié à l'époque.

V.E Prise en compte des caractéristiques des patients

La revue de la littérature a montré que les résultats de santé étaient significativement liés à l'âge du patient, à ses comorbidités, et la taille de l'anévrisme au moment de la chirurgie. Les résultats stratifiés sur les 24 groupes de patients sont présentés dans le Tableau 18. La pose d'endoprothèse est associée à un meilleur résultat de santé lorsque l'âge du patient augmente, lorsque son état général se dégrade et, dans une moindre mesure, lorsque la taille de l'anévrisme est élevée. Ainsi, le nombre d'années de vie supplémentaires est quasi nul pour les patients jeunes (75 ans) et en bon état général. A l'inverse, les patients en mauvais état général bénéficient particulièrement du traitement endovasculaire avec un gain compris entre 0.16 et 0.20 années de vie. Du point de vue de l'efficience, le RDCR est inférieur à 40 000€/QALY quel que soit le sous-groupe, hormis chez les patients de 75 ans en bon état général pour lesquels il est supérieur à 100 000€/QALY. Les meilleurs résultats observés pour les patients plus âgés avec comorbidités confortent ainsi l'approche initiale de la HAS qui, jusqu'à 2009, restreignait l'indication des endoprothèses aux patients à haut risque chirurgical. Sur la base des données disponibles en 2009, il apparaît cohérent de lever cette restriction. A l'inverse, les résultats ne

permettent pas de trancher en faveur de l'une ou l'autre des techniques pour les patients de 75 ans en bon état général, tant du point de vue efficacité (delta quasi nul) que de celui de l'efficience (absence de seuil consensuel de propension à payer).

Tableau 18 : Résultats de l'analyse coût-efficacité stratifiée par sous-groupes en fonction de l'âge, de l'état général et de la taille de l'anévrisme du patient.

		Années de vie			QALYs			Coûts			Bénéfice net			RDCR
		EVAR	OPEN	DELTA	EVAR	OPEN	DELTA	EVAR	OPEN	DELTA	EVAR	OPEN	DELTA	€/QALYs
75 ans														
Etat général bon														
5,0 à 5,4 cm	15%	10,33	10,32	0,005	6,20	6,17	0,027	16 010 €	13 116 €	2 893 €	603 956 €	604 162 €	-206 €	107 657
5,5 à 5,9 cm	8%	9,71	9,72	-0,007	5,88	5,86	0,021	15 896 €	13 107 €	2 790 €	571 832 €	572 499 €	-667 €	131 409
6,0 à 6,4 cm	6%	8,99	9,00	-0,003	5,50	5,48	0,025	15 763 €	13 095 €	2 668 €	534 320 €	534 538 €	-218 €	108 879
Plus de 6,5 cm	8%	8,15	8,18	-0,033	5,04	5,04	0,009	15 598 €	13 081 €	2 517 €	488 797 €	490 458 €	-1 660 €	293 823
Etat général moyen														
5,0 à 5,4 cm	9%	8,66	8,60	0,066	5,32	5,26	0,066	15 697 €	13 086 €	2 611 €	516 658 €	512 650 €	4 008 €	39 451
5,5 à 5,9 cm	6%	8,09	8,02	0,068	5,01	4,94	0,068	15 584 €	13 075 €	2 509 €	485 662 €	481 326 €	4 336 €	36 650
6,0 à 6,4 cm	5%	7,42	7,34	0,079	4,64	4,57	0,076	15 448 €	13 062 €	2 386 €	448 941 €	443 737 €	5 204 €	31 432
Plus de 6,5 cm	6%	6,67	6,61	0,068	4,23	4,16	0,071	15 291 €	13 047 €	2 244 €	407 462 €	402 630 €	4 832 €	31 709
Etat général mauvais														
5,0 à 5,4 cm	5%	7,03	6,86	0,171	4,43	4,29	0,135	15 361 €	13 048 €	2 313 €	427 326 €	416 146 €	11 180 €	17 142
5,5 à 5,9 cm	3%	6,52	6,33	0,186	4,14	3,99	0,145	15 249 €	13 036 €	2 213 €	398 309 €	385 995 €	12 314 €	15 236
6,0 à 6,4 cm	2%	5,91	5,71	0,203	3,79	3,63	0,157	15 114 €	13 021 €	2 093 €	363 725 €	350 138 €	13 587 €	13 347
Plus de 6,5 cm	3%	5,28	5,07	0,204	3,42	3,26	0,159	14 966 €	13 005 €	1 961 €	326 803 €	312 877 €	13 926 €	12 346
85 ans														
Etat général bon														
5,0 à 5,4 cm	3%	4,92	4,88	0,031	3,23	3,19	0,043	14 923 €	13 019 €	1 904 €	308 048 €	305 664 €	2 384 €	44 407
5,5 à 5,9 cm	3%	4,54	4,51	0,031	3,00	2,96	0,043	14 829 €	13 009 €	1 820 €	285 169 €	282 664 €	2 506 €	42 073
6,0 à 6,4 cm	2%	4,10	4,06	0,036	2,73	2,68	0,047	14 714 €	12 996 €	1 718 €	257 873 €	254 915 €	2 958 €	36 746
Plus de 6,5 cm	3%	3,63	3,60	0,030	2,43	2,39	0,043	14 589 €	12 981 €	1 607 €	228 718 €	226 032 €	2 685 €	37 445
Etat général moyen														
5,0 à 5,4 cm	3%	3,85	3,77	0,083	2,57	2,49	0,078	14 643 €	12 983 €	1 660 €	242 335 €	236 228 €	6 107 €	21 370
5,5 à 5,9 cm	2%	3,53	3,44	0,090	2,36	2,28	0,082	14 553 €	12 972 €	1 581 €	221 823 €	215 191 €	6 632 €	19 251
6,0 à 6,4 cm	2%	3,14	3,04	0,097	2,12	2,03	0,087	14 443 €	12 958 €	1 485 €	197 311 €	190 119 €	7 191 €	17 119
Plus de 6,5 cm	2%	2,76	2,66	0,097	1,87	1,78	0,086	14 329 €	12 943 €	1 386 €	172 521 €	165 273 €	7 248 €	16 056
Etat général mauvais														
5,0 à 5,4 cm	1%	2,89	2,73	0,161	1,95	1,82	0,129	14 360 €	12 941 €	1 419 €	180 983 €	169 473 €	11 510 €	10 976
5,5 à 5,9 cm	1%	2,62	2,44	0,170	1,78	1,64	0,135	14 275 €	12 928 €	1 347 €	163 305 €	151 105 €	12 199 €	9 943
6,0 à 6,4 cm	1%	2,30	2,12	0,175	1,57	1,43	0,138	14 173 €	12 913 €	1 259 €	142 376 €	129 812 €	12 565 €	9 111
Plus de 6,5 cm	1%	1,99	1,81	0,174	1,36	1,23	0,136	14 070 €	12 898 €	1 172 €	122 118 €	109 655 €	12 463 €	8 597
Total														
	100%	7,23	7,17	0,059	4,49	4,43	0,063	15 375 €	13 054 €	2 321 €	434 007 €	430 000 €	4 007 €	36 676

De la même manière que précédemment, une analyse de sensibilité probabiliste a été réalisée par sous-groupe de patients. Les plans coût-efficacité sont représentés sur la Figure 20. L'incertitude varie selon le sous-groupe et se révèle maximum pour les patients de 75 ans en bon état général. Chez ces patients, la probabilité que le traitement soit efficace (i.e. que le nombre d'années de vie supplémentaires soit positif) est de 50% pour ceux dont la taille de l'anévrisme est comprise entre 5 et 5.5cm, et de 38% pour ceux dont la taille de l'anévrisme est supérieure à 6.5 cm. Des résultats semblables sont obtenus du point de vue de l'efficience, puisque les probabilités que le traitement soit coût-efficace à un seuil de 100 000€/QALY sont respectivement de 46% et 39%. L'incertitude est considérablement réduite à mesure que l'état général se dégrade et/ou que l'âge augmente. Les probabilités que le traitement soit efficace et efficient pour les patients de 85 ans en mauvais état général sont respectivement supérieures à 99% et 97%, quelle que soit la taille de l'anévrisme.

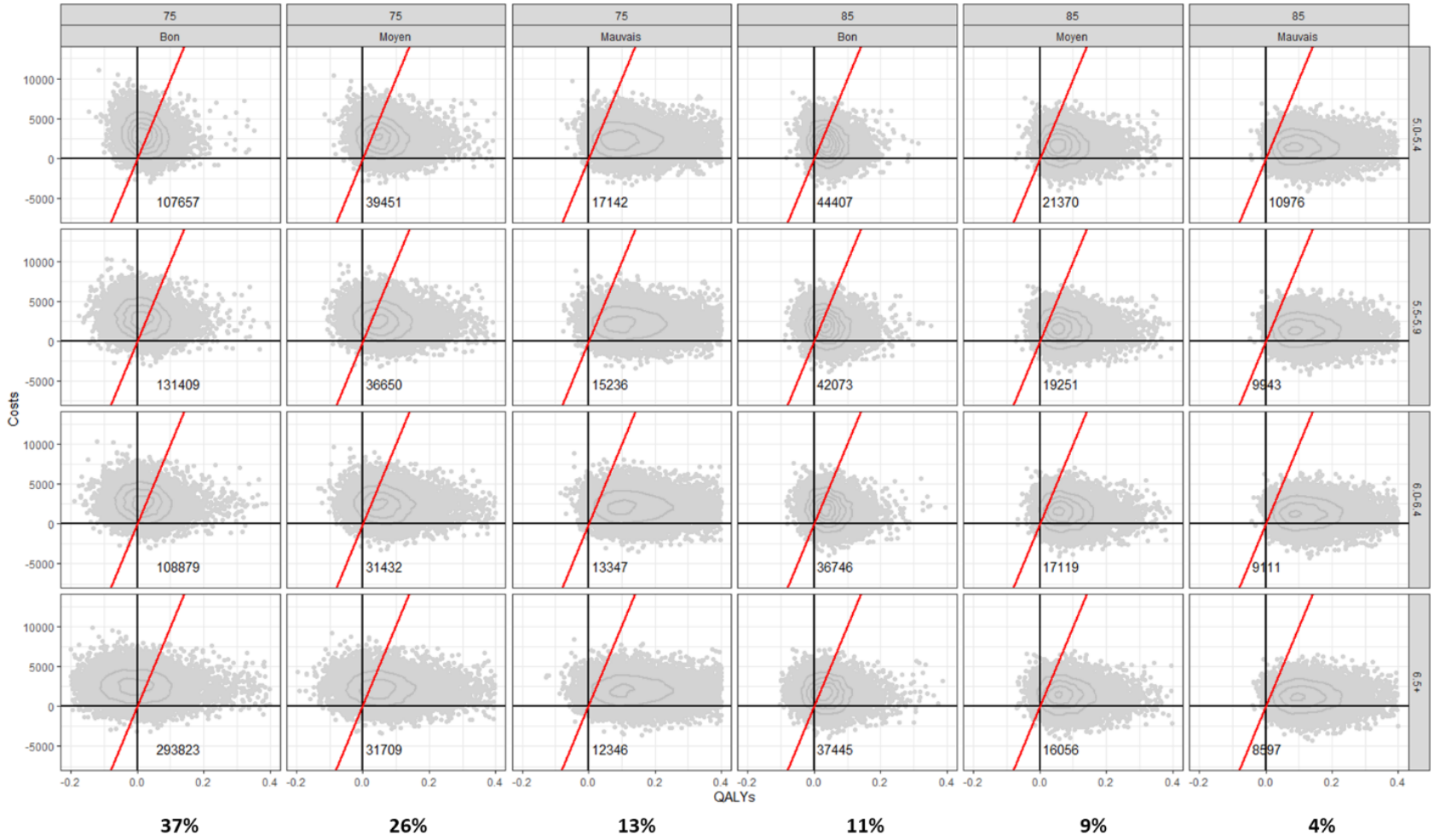


Figure 20 : Analyse de sensibilité probabiliste menée sur les 24 strates de patients. Les trois colonnes de gauche correspondent aux patients de 75 ans et les trois colonnes de droite aux patients de 85 ans. L'état général se dégrade de gauche à droite et la taille de l'anévrisme augmente de haut en bas.

L'analyse de l'incertitude paramétrique dans chacun des sous-groupes soulève donc la question de la pertinence de financer des efforts de recherche de manière indifférenciée. Les analyses fondées sur la valeur de l'information nous renseignent sur l'opportunité d'effectuer ces investigations supplémentaires dans chacun des sous-groupes. Les résultats de ces analyses (Tableau 19) suggèrent que les efforts de recherche pourraient se concentrer sur les patients jeunes et en bon état général. En effet, l'EVPI des patients de 75 ans en bon état général atteint à elle seule 344 années de vie et plus de 24 millions d'euros, soit des valeurs comparables voire supérieures à l'EVPI de l'ensemble de la population. Chez ces patients, une étude portant sur les paramètres d'efficacité à long terme semble particulièrement pertinente. Chez les patient âgés et/ou en mauvais état général, l'EVPI est faible ou nulle dans la plupart des sous-groupes et l'acquisition d'information supplémentaire ne semble pas de nature à remettre en cause la supériorité des endoprothèses en termes d'efficacité et d'efficience.

En conclusion, il aurait été pertinent de restreindre les demandes d'études post-inscriptions à certains sous-groupes de patients pour lesquels l'incertitude est maximale. Il faut cependant préciser qu'en 2009, seules les données brutes étaient disponibles en sous-groupes et le niveau de formalisation proposé ici est bien supérieur à celui de l'époque. Par ailleurs, la faisabilité d'une telle étude en termes de recrutement devrait être envisagée.

Tableau 19 : EVP(P)I populationnelle stratifiée par sous-groupes en fonction de l'âge, de l'état général et de la taille de l'anévrisme du patient.

		Années de vie						Bénéfice net					
		EVPI	Efficacité	Court terme	Long terme	QALYs	Coûts	EVPI	Efficacité	Court terme	Long terme	QALYs	Coûts
75 ans													
Etat général bon													
5,0 à 5,4 cm	15%	98	100	46	80	0	0	8 078 631 €	6 836 992 €	3 403 099 €	5 613 478 €	2 768 735 €	1 975 596 €
5,5 à 5,9 cm	8%	74	72	31	61	0	0	4 825 445 €	4 356 496 €	1 758 676 €	3 668 257 €	971 922 €	631 177 €
6,0 à 6,4 cm	6%	63	62	29	53	0	0	4 224 656 €	3 880 114 €	1 800 110 €	3 280 316 €	1 059 410 €	751 357 €
Plus de 6,5 cm	8%	109	102	25	92	0	0	7 058 344 €	6 462 092 €	1 834 631 €	5 707 155 €	379 155 €	200 861 €
Etat général moyen													
5,0 à 5,4 cm	9%	14	15	4	1	0	0	1 924 792 €	1 453 439 €	428 292 €	523 057 €	42 797 €	3 319 €
5,5 à 5,9 cm	6%	20	20	6	6	0	0	1 668 091 €	1 396 061 €	429 326 €	545 349 €	17 667 €	33 €
6,0 à 6,4 cm	5%	13	13	4	2	0	0	1 111 615 €	887 304 €	256 046 €	280 080 €	2 460 €	0 €
Plus de 6,5 cm	6%	44	44	14	22	0	0	2 715 059 €	2 515 794 €	754 515 €	1 216 495 €	6 826 €	0 €
Etat général mauvais													
5,0 à 5,4 cm	5%	1	1	0	0	0	0	212 133 €	148 454 €	22 802 €	913 €	0 €	0 €
5,5 à 5,9 cm	3%	1	1	0	0	0	0	149 293 €	103 342 €	20 863 €	0 €	0 €	0 €
6,0 à 6,4 cm	2%	1	1	0	0	0	0	91 266 €	62 365 €	11 277 €	0 €	0 €	0 €
Plus de 6,5 cm	3%	3	3	1	0	0	0	204 587 €	154 999 €	35 923 €	0 €	0 €	0 €
85 ans													
Etat général bon													
5,0 à 5,4 cm	3%	4	4	1	1	0	0	610 920 €	377 874 €	41 151 €	217 834 €	97 630 €	32 903 €
5,5 à 5,9 cm	3%	6	6	2	2	0	0	603 423 €	421 835 €	69 783 €	236 682 €	70 467 €	22 223 €
6,0 à 6,4 cm	2%	4	4	1	1	0	0	399 107 €	263 191 €	37 233 €	131 244 €	30 347 €	7 051 €
Plus de 6,5 cm	3%	12	12	4	6	0	0	823 928 €	670 538 €	146 681 €	386 288 €	53 220 €	15 974 €
Etat général moyen													
5,0 à 5,4 cm	3%	0	0	0	0	0	0	148 366 €	65 063 €	1 669 €	9 308 €	0 €	0 €
5,5 à 5,9 cm	2%	1	1	0	0	0	0	128 227 €	56 802 €	2 919 €	5 263 €	0 €	0 €
6,0 à 6,4 cm	2%	0	0	0	0	0	0	78 372 €	33 233 €	1 704 €	2 480 €	0 €	0 €
Plus de 6,5 cm	2%	1	1	0	0	0	0	139 295 €	68 585 €	6 940 €	2 367 €	0 €	0 €
Etat général mauvais													
5,0 à 5,4 cm	1%	0	0	0	0	0	0	16 099 €	8 310 €	593 €	0 €	0 €	0 €
5,5 à 5,9 cm	1%	0	0	0	0	0	0	12 330 €	7 336 €	572 €	0 €	0 €	0 €
6,0 à 6,4 cm	1%	0	0	0	0	0	0	7 431 €	4 118 €	415 €	0 €	0 €	0 €
Plus de 6,5 cm	1%	0	0	0	0	0	0	10 832 €	7 958 €	683 €	0 €	0 €	0 €
Total													
Ensemble	**100%**	**327**	**328**	**94**	**95**	**0**	**0**	**26 259 576 €**	**21 565 395 €**	**5 931 914 €**	**9 010 052 €**	**406 416 €**	**26 974 €**
24 sous-groupes	**100%**	**469**	**462**	**168**	**327**	**0**	**0**	**35 242 242 €**	**30 242 295 €**	**11 065 903 €**	**21 826 566 €**	**5 500 636 €**	**3 640 494 €**

VI DISCUSSION

Nous avons donc montré que les analyses de la valeur de l'information auraient pu être utiles à la détermination des études post-inscription dans le contexte décisionnel français. Sur la base de l'exemple de l'évaluation de l'intérêt des endoprothèses aortiques dans le traitement de l'anévrisme de l'aorte abdominale sous-rénale non rompu par la HAS (Lesquelen, Thevenet, and Javerliat 2009b, 2009a), nous avons formalisé un modèle d'aide à la décision qui a permis d'estimer la valeur de l'information, conformément aux recommandations méthodologiques récemment publiées (Fenwick et al. 2020; Rothery et al. 2020). Les études post-inscriptions recommandées par les méthodes fondées sur la valeur de l'information ont ensuite été comparées à celles qui avaient réellement été demandées par la HAS.

Les résultats moyens obtenus grâce au modèle d'aide à la décision ont confirmé la décision prise à l'époque d'adopter les endoprothèses aortiques. En effet, les endoprothèses apportaient un bénéfice au patient en termes d'années de vie, de qualité de vie, et étaient coût-efficaces avec un bénéfice net estimé à environ 4000€/patient pour une propension à payer de 100,000€/QALY. Ce résultat est peu surprenant dans la mesure où le modèle sur lequel nous nous basions avait été élaboré à la demande du NICE à la même période (Epstein et al. 2008; Chambers et al. 2009). Les analyses fondées sur la valeur de l'information confirmaient l'intérêt d'une étude sur les paramètres d'efficacité, particulièrement sur le long terme, telle que demandée par la HAS. L'acquisition de données supplémentaires sur les paramètres de coût et d'utilité ne présentait que peu d'intérêt. L'interprétation de ce dernier constat était limitée en regard de l'absence d'évaluation formelle de l'efficience à l'époque. A l'inverse, il aurait été pertinent de restreindre les études post-inscription aux patients jeunes et en bon état général. En effet, au-delà du questionnement éthique indispensable lié à l'âge et l'espérance de vie des patients, d'éventuelles études menées sur des patients âgés et/ou en mauvais état général ne semblaient pas en mesure de modifier la décision.

L'utilisation des méthodes fondées sur la valeur de l'information nécessite de formaliser le modèle d'aide à la décision. Ce socle commun permet d'agréger les connaissances issues de multiples sources au sein d'un modèle unique pouvant servir de base aux échanges entre les

industriels et le décideur, mais également entre les différentes commissions (CNEDiMTS et CEESP) dont les critères de jugement sont différents. Cependant l'étape d'élaboration d'un modèle s'avère bien souvent complexe, tant d'un point de vue technique, que pour établir un consensus sur la structure et les données pertinentes. Par ailleurs, sa qualité influencera directement la pertinence des résultats de l'analyse de la valeur de l'information. A ce titre, la réalisation des modèles par des équipes universitaires, comme au Royaume-Uni, constitue un avantage pour garantir la « neutralité » du modèle sous-jacent. Ainsi plusieurs modèles d'aide à la décision élaborés par des industriels et qui ont été soumis de manière concomitante ont montré un avantage bien plus marqué que dans le modèle finalement produit par l'équipe universitaire de York (Medtronic 2007). Dans le contexte français, la réalisation de ce type de modèle est confiée aux industriels qui adaptent bien souvent un modèle de référence développé par la maison mère. Ce modèle est ensuite fourni à la CEESP qui l'expertise et en tire ses recommandations. Des modèles de qualité moyenne (validation insuffisante, structure non adaptée au contexte français, etc.) pourraient constituer une limite à la mise en œuvre des analyses fondées sur la valeur de l'information.

La caractérisation de l'incertitude paramétrique entourant les paramètres du modèle requiert un travail supplémentaire à l'occasion de la sélection puis de la synthèse des études disponibles. Cet exercice n'est pas nécessaire en l'absence de modèle d'aide à la décision. Habituellement, une revue détaillée de la littérature est réalisée et les données disponibles sont présentées sous forme de tableau. L'étape consistant à synthétiser les données par un estimateur unique, en sélectionnant les études les plus pertinentes, voire en réalisant une méta-analyse, n'est que rarement réalisée. Le décideur doit alors « piocher » dans les estimateurs à sa disposition, ouvrant la porte à une série de biais liés à la subjectivité du décideur et à l'absence de systématisation de la synthèse. De nouveau, ce prérequis est celui de l'analyse décisionnelle en général et n'est pas propre aux analyses fondées sur la valeur de l'information. Cependant, la qualité de la caractérisation de l'incertitude nécessaire à la réalisation de ces analyses vient renforcer cette exigence.

Les choix de modélisation du risque en fonction du temps étaient fondamentaux dans la mesure où la durée maximale de suivi permettant de calibrer la fonction de survie est de 6 ans et que les

recommandations d'études post-inscriptions mentionnaient un horizon au-delà de 5 ans. La question de la qualité de la caractérisation de l'incertitude associée à l'extrapolation des courbes de survie à long terme s'est alors posée. Il s'avère que les estimateurs de la valeur de l'information se sont montrés particulièrement performants dans ce contexte et ont capturé la nécessité d'acquisition de données concernant l'efficacité des endoprothèses à long terme. L'explication semble tenir à l'effet de levier exercé par les estimateurs des modèles de survie. En effet, alors même que l'incertitude est « fixée », son impact augmente avec le temps. Cette propriété est utile car elle permet d'éviter d'ajouter de l'incertitude structurelle aux méthodes d'extrapolation.

En termes de conditions d'application, ce travail pose la question de l'expertise matérielle et technique nécessaire à la mise en œuvre des analyses fondées sur la valeur de l'information. La revue de la littérature réalisée afin d'intégrer les données pertinentes dans le contexte français ne nécessite pas de travail supplémentaire par rapport à une analyse classique. A l'inverse, le niveau de complexité du modèle d'aide à la décision est bien plus important. Le modèle de référence fournit par le NICE a dû être largement réécrit pour permettre une caractérisation de l'incertitude répondant aux exigences de calcul de la valeur de l'information. Ces étapes complémentaires requièrent un savoir-faire technique dont la majorité des modélisateurs ne disposent pas forcément et qui nécessite une formation spécifique. Ainsi, la prise en compte des corrélations a considérablement complexifié la propagation de l'incertitude et requis l'utilisation de méthodes de simulations multivariées. De plus, la nécessité d'estimer les résultats de santé et la valeur de l'information pour les sous-groupes de patients a imposé une modification de la structure du modèle avec l'adjonction d'une boucle interne de simulation. Enfin, la caractérisation de l'incertitude a dû être adaptée pour de multiples paramètres. A titre d'exemple, le hazard ratio traduisant le surrisque de décès non lié à l'anévrisme était supérieur à 1, puis égal à 1 lorsque les mortalités convergeaient. L'incertitude telle que modélisée initialement pouvait aboutir à un hazard ratio (de surrisque) inférieur à 1, ce qui conduisait les courbes à diverger. L'impact était certes limité sur la probabilité d'efficience mais devenait considérable sur les estimateurs de la valeur de l'information. Nous avons donc reparamétré l'incertitude afin de corriger ce résultat aberrant. Ce cas d'école illustre l'effort supplémentaire

requis en termes de qualité de la caractérisation de l'incertitude pour garantir la pertinence des résultats. Enfin, la mise en œuvre des simulations de Monte Carlo nécessite rapidement une puissance calculatoire que seuls des serveurs sont en mesure de fournir. A titre d'exemple, plus de 24h sont nécessaires pour le calcul des 48 millions de modèles permettant l'estimation de l'EVPPI pour un sous-groupe de paramètres.

Les métamodèles permettent d'apporter une alternative aux simulations imbriquées nécessaires au calcul de l'EVPPI et de l'EVSI (Menzies 2016a; Tuffaha et al. 2016; Menzies 2016b; Strong et al. 2015; Strong, Oakley, and Brennan 2014a; Heath, Manolopoulou, and Baio 2019; Heath, Manolopoulou, and Baio 2018a, 2017a). L'objectif de ces modèles est de lever le verrou calculatoire ayant longtemps limité l'utilisation de ces méthodes. Cependant, ces techniques ont été développées sur des modèle d'aide à la décision relativement simples (essais randomisés, arbre de décision ou Markov homogène) mais s'avèrent inutilisables dès lors qu'ils se complexifient. A notre connaissance, aucune publication disponible à ce jour ne traite de la gestion des corrélations et de l'hétérogénéité de la population dans les métamodèles. En l'absence de travaux sur le sujet, nous avons eu recours à l'approche par simulation, la seule qui garantisse la validité des résultats de notre modèle. Ce faisant, nous n'avons pu réaliser le calcul de l'EVSI pour des raisons calculatoires, ce qui constitue une limite majeure de ce travail et plus généralement de la mise en œuvre de ces analyses en routine.

Antérieurement à 2009, le remboursement des endoprothèses ne concernait qu'une partie des patients présentant un anévrisme de l'aorte abdominale. Les critères de prise en charge étaient fondés sur la taille de l'anévrisme, l'âge du patient et ses comorbidités. Il a donc été nécessaire d'adapter le modèle afin de tenir compte de ce raisonnement en sous-groupe. Le NICE prenait en compte l'hétérogénéité au travers de scénarii qui permettaient un raisonnement stratifié sur la valeur moyenne de l'utilité dans chaque groupe, sans toutefois se montrer satisfaisants pour calculer la valeur de l'information à l'échelle de l'ensemble de la population. Dans notre étude, les analyses stratifiées sur les sous-groupes de population mettent en exergue la différence d'efficacité et d'efficience de la stratégie EVAR selon les caractéristiques des patients. Ainsi, les patients plus âgés et présentant des comorbidités étaient les principaux bénéficiaires de l'abord endovasculaire. A l'inverse, le bénéfice était incertain pour les patients jeunes et avec peu de

comorbidités. Pour ces derniers, l'analyse fondée sur la valeur de l'information retrouvait une incertitude décisionnelle maximale qui aurait pu justifier une demande d'étude ciblée sur cette sous-population.

La décision de restreindre l'indication d'une technologie de santé à des sous-groupes de population est de plus en plus fréquente, notamment dans un contexte de personnalisation de la médecine. Le problème de la prise en compte de l'hétérogénéité dans le calcul de la valeur de l'information est donc une question fondamentale du fait de sa portée générale. L'un des résultats remarquables est l'observation de valeurs plus élevées d'EVP(P)I lors de l'analyse en sous-groupes que lors de l'analyse de l'ensemble de la population. Plusieurs éléments peuvent expliquer ce résultat. Dans le paradigme des analyses fondées sur la valeur de l'information, la décision prise est celle maximisant l'utilité. Lorsque l'analyse est réalisée sur l'ensemble de la population, les endoprothèses aortiques l'emportent, et ce quel que soit le critère de jugement. A l'inverse, lorsque l'analyse est réalisée en sous-groupes, la chirurgie ouverte est l'alternative de choix dans certains sous-groupes. Par conséquent, la décision de référence n'est plus la même[26] et les mesures ne sont alors plus comparables. L'impact est d'autant plus important que les sous-groupes concernés représentaient plus d'un tiers de la population cible. Un second élément d'explication est lié à la transformation de la variabilité (par définition inexpliquée) en hétérogénéité. Ce faisant, la structure de l'incertitude paramétrique se trouve modifiée, ce qui a généralement pour effet de diminuer la valeur de l'information. Des travaux théoriques définissant le concept de « valeur de l'information dynamique » visent à quantifier deux sources de valeur : (i) la valeur associée à la réduction de l'incertitude sur les estimateurs des paramètres conditionnels et (ii) la valeur associée à l'estimation des paramètres de stratification (Espinoza et al. 2014). Cette deuxième composante est la « valeur de l'hétérogénéité » et n'est pas étudiée dans ce travail. L'impact respectif de ces deux phénomènes n'est pas quantifiable mais la décision différenciée entre les sous-groupes semble clairement l'emporter.

[26] Dans l'analyse de l'ensemble de la population, l'information avait une valeur si elle amenait le décideur à modifier sa décision, c'est-à-dire de recommander OPEN plutôt qu'EVAR. Dans l'analyse des sous-groupes de patients jeunes et sans comorbidités, l'information a une valeur si le décideur modifie sa décision en faveur d'EVAR.

En conclusion, ce travail nous a permis de montrer l'apport des analyses de la valeur de l'information pour la détermination des études post-inscriptions dans le contexte décisionnel français et d'envisager les conditions de mise en œuvre

Partie 3. APPLICATION DES METHODES FONDEES SUR LA VALEUR DE L'INFORMATION A L'EVALUATION PRECOCE DES DISPOSITIFS MEDICAUX : L'EXEMPLE DES ETUDES D'UTILISABILITE

I ESTIMATING THE NUMBER OF USABILITY PROBLEMS AFFECTING MEDICAL DEVICES: MODELLING THE DISCOVERY MATRIX – PUBLIÉ DANS BMC MEDICAL RESEARCH METHODOLOGY

TECHNICAL ADVANCE Open Access

Estimating the number of usability problems affecting medical devices: modelling the discovery matrix

Vincent Vandewalle[1,2†], Alexandre Caron[1*†], Coralie Delettrez[3], Renaud Périchon[1], Sylvia Pelayo[1,4], Alain Duhamel[1,3] and Benoit Dervaux[1,3]

* Correspondence: alexandre.caron2@univ-lille.fr
†Vincent Vandewalle and Alexandre Caron contributed equally to this work.
[1]Univ. Lille, CHU Lille, ULR 2694 Evaluations des technologies de santé et des pratiques médicales, F-59000 Lille, France
Full list of author information is available at the end of the article

I.A Abstract

I.A.1 Background

Usability testing of medical devices are mandatory for market access. The testings' goal is to identify usability problems that could cause harm to the user or limit the device's effectiveness. In practice, human factor engineers study participants under actual conditions of use and list the problems encountered. This results in a binary discovery matrix in which each row corresponds to a participant, and each column corresponds to a usability problem. One of the main challenges in usability testing is estimating the total number of problems, in order to assess the completeness of the discovery process. Today's margin-based methods fit the column sums to a binomial model of problem detection. However, the discovery matrix actually observed is truncated because of undiscovered problems, which corresponds to fitting the marginal sums

without the zeros. Margin-based methods fail to overcome the bias related to truncation of the matrix. The objective of the present study was to develop and test a matrix-based method for estimating the total number of usability problems.

I.A.2 Methods

The matrix-based model was based on the full discovery matrix (including unobserved columns) and not solely on a summary of the data (e.g. the margins). This model also circumvents a drawback of margin-based methods by simultaneously estimating the model's parameters and the total number of problems. Furthermore, the matrix-based method takes account of a heterogeneous probability of detection, which reflects a real-life setting. As suggested in the usability literature, we assumed that the probability of detection had a logit-normal distribution.

I.A.3 Results

We assessed the matrix-based method's performance in a range of settings reflecting real-life usability testing and with heterogeneous probabilities of problem detection. In our simulations, the matrix-based method improved the estimation of the number of problems (in terms of bias, consistency, and coverage probability) in a wide range of settings. We also applied our method to five real datasets from usability testing.

I.A.4 Conclusions

Estimation models (and particularly matrix-based models) are of value in estimating and monitoring the detection process during usability testing. Matrix-based models have a solid mathematical grounding and, with a view to facilitating the decision-making process for both regulators and device manufacturers, should be incorporated into current standards.

I.B Background

I.B.1 Introduction

The usability testing is a cornerstone of medical device development, and proof of usability is mandatory for market access in both the European Union and the United States (Food and Drug Administration 2016). The overall objective of a usability assessment is to ensure that a medical device is designed and optimized for use by the intended users in the environment in which the device is likely to be used (UK-MHRA 2017). The goal is to identify problems (called "use errors") that could cause harm to the user or impair medical treatment (e.g. an inappropriate number of inhalations, finger injection with an adrenaline pen, etc.) (Food and Drug Administration 2012). The detection of usability problems must be as comprehensive as possible because medical devices are safety-critical systems (Borsci et al. 2013). However, the total number of usability problems is never known in advance. The main challenge during the usability testing is thus to estimate this number, in order to assess the completeness of the problem discovery process (Borsci et al. 2014).

In practice, participants are placed under actual conditions of use (real or simulated), and usability problems are observed and listed by human factor engineers. The experimental conditions are defined in a risk analysis that gathers together possible usability problems. Throughout the usability testing, problems are discovered and added to a discovery matrix - a binary matrix with the participants as the rows and the problems as the columns. The current approach involves estimating the total number of problems as the usability testing progresses, starting from the first sessions. The number is estimated iteratively as the sample size increases, until the objective of completeness has been achieved (Lewis 1994).

From a statistical perspective, the current estimation procedure is based on a model of how the usability problems are detected; this is considered to be a binomial process. The literature suggests that the total number of usability problems can be estimated from the discovery matrix's problem margin (the sum of the columns) (Kanis 2011; Lewis 2001; Hertzum and Jacobsen 2003; Schmettow 2012; Borsci, Londei, and Federici 2011). However, this estimation is complicated by (i) the small sample size usually encountered in usability testing of medical

devices (Faulkner 2003) and (ii) as-yet unobserved problems that truncate the margin and bias estimates (Lewis 2000; Sauro and Lewis 2016; Thomas and Gart 1971).

The objective of the present study was to develop a matrix-based estimation of the number of usability problems affecting a medical device. This new method is based on the likelihood of the discovery matrix (rather than the matrix's margins alone), so as to avoid a reduction in the level of information prior to modeling. The method's main targets are (i) regulatory agencies and notified bodies involved in the pre-market evaluation of medical devices, and (ii) medical device manufacturers (more specifically, the human factors engineers in charge of ensuring that the devices are usable).

I.B.2 Data collected during the usability testing: the discovery matrix

The human factor engineer collects the results of the usability testing in a problem-discovery matrix $\mathbb{d}$. Each row corresponds to a participant, and each column corresponds to a usability problem. The result is 1 if the participant discovered the problem and 0 if not. Considering that after the inclusion of n participants, j problems have been discovered, a $n \times j$ matrix is built. By way of an example, the discovery matrix obtained after $n = 8$ participants (in rows) might be the one presented below:

$$\mathbb{d} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

In this example, $j = 10$ different problems (in columns) have been detected so far. The first participant discovered only one problem (column 1), whereas the second discovered two new problems (columns 2 and 3), etc.

At this stage, some problems might not have been detected, and the total number of usability problems (m) is unknown. It should be noted that by definition, $m \geq j$ and $m - j$ problems remain undetected. Indeed, $\mathbb{d}$ comes from a complete but unobserved matrix of dimensions

$n \times m$. This matrix is denoted as $\mathbb{x}$. Thus, the "observed" matrix $\mathbb{d}$ is a truncated version of the "complete" matrix $\mathbb{x}$; it lacks the columns corresponding to the as-yet undetected problems. Hereafter, we use the following notation: $\mathbb{x} = (x_{il})_{1 \leq i \leq n, 1 \leq l \leq m}$ where $x_{il} = 1$ if the participant i experiences the problem l, and $x_{il} = 0$ otherwise.

$$\mathbb{x} = \begin{pmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{il} & \cdots & x_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nl} & \cdots & x_{nm} \end{pmatrix}$$

The human factor engineer's goal is to estimate the total number of problems m from the discovery matrix $\mathbb{d}$ and thus deduce $m - j$ - the number of problems that have not been detected. The new method presented below addresses this goal.

I.B.3 Conventional estimation of $\boldsymbol{m}$ using a margin-based probabilistic model

In this section, we describe the margin-based methods currently employed to estimate the number of usability problems. As mentioned above, m is currently estimated by fitting a probabilistic (binomial) model to the discovery matrix's problems margin. More specifically, the probability with which a given usability problem is discovered by a participant is modelled by a Bernoulli trial with a probability of success (i.e. detection) p. For a given problem, the Bernoulli trial is considered to apply independently to each of the n participants in the usability testing. Thus, the problem margin sums can be considered as an independent, identically distributed sequence of Bernoulli trials, in which the number of times a given usability problem (a random variable X) has been observed after n participants follows a binomial distribution, $X \sim \mathcal{B}in(n, p)$. Considering the binomial distribution of the margin sums, the proportion of problems that has been discovered at least once after n participants is given by the cumulative function of the shifted geometric distribution (Lewis 1994; Virzi 1992; Nielsen and Landauer 1993):

$$P(X > 0) = 1 - (1 - p)^n \quad (1)$$

The total number of problems m is then deduced from the following relationship:

$$j = (1 - (1 - p)^n) \times m \quad (2)$$

The discovery progress is thus assessed in two steps: the probability of detection p is first estimated and then plugged into Equation (2) to estimate the number of problems m. A wide range of literature methods are available for estimating the probability of problem detection. The simplest way involves computing the naive estimate (denoted as $\hat{p}$) using the observed discovery matrix $\mathbb{d}$, considering that only j problems have been detected so far:

$$\hat{p} = \frac{\sum_{i=1}^{n}\sum_{l=1}^{j} x_{il}}{n * j} \qquad (3)$$

As mentioned above, the naïve estimate is systematically biased - especially for small samples. Indeed, unobserved problems result in zero columns that shrink the probability space and lead to overestimation of p, particularly at the beginning of the process when $j \ll m$. Consequently, m is systematically underestimated, which generates safety concerns in the medical device field. In response, several strategies have been employed to overcome the truncated matrix problem.

In 2001, Hertzum and Jacobsen suggested normalizing the value of $\hat{p}$ (Hertzum and Jacobsen 2003). This procedure considers that the lower boundary of the probability of detection estimated with n participants is $1/n$. For example, in a sample of 5 participants, $\hat{p} \in [0.2\,;1]$. Conversely, the normalized estimator $\hat{p}_{Norm} \in [0;1]$, and is computed as follows:

$$\hat{p}_{\text{Norm}} = \frac{\hat{p} - \frac{1}{n}}{1 - \frac{1}{n}} \qquad (4)$$

However, the normalized approach suffers from a major limitation when estimating the total number of problems with Equation (4). In fact, if each participant has discovered only one problem and if each problem was discovered only once, $\hat{p} = \frac{1}{n}$, $\hat{p}_{Norm} = 0$, and the estimated number of problems $\hat{m}$ is infinite. We will not discuss this estimation method further.

Turing and Good developed a discounting method for estimating the probability of unseen species on the basis of observed data (Good 1953). Lewis suggested that the Good-Turing (GT) adjustment could be used to reduce the magnitude of the overestimation of p by increasing the probability space and thus accounting for unobserved usability problems (Lewis 2001). The GT adjustment is computed as the proportion of singletons relative to the total number of events

(i.e. the proportion of problems discovered only once, $x_{il} = 1$), and is incorporated in the estimation as follows:

$$\hat{\mathrm{p}}_{\mathrm{GT}} = \frac{\hat{p}}{1 + GT} \tag{5}$$

However, Lewis observed that use of the GT estimator overestimated p. He empirically assessed the best adjustment for a small sample size by carrying out Monte Carlo simulations on a range of usability testing databases involving web or software user interfaces with known true values. Based on these simulations, Lewis concluded that the best method was to average the GT adjustment and a "double-deflation" term:

$$\hat{p}_{\text{double-deflation}} = \frac{1}{2}\left[\frac{\hat{p}}{1 + GT_{adj}}\right] + \frac{1}{2}\left[\left(\hat{p} - \frac{1}{n}\right) \times \left(1 - \frac{1}{n}\right)\right] \tag{6}$$

Nevertheless, the degree of adjustment of the probability space for unobserved problems is essentially empirical. The residual bias is not known to trend towards over- or underestimation.

In 2009, Schmettow considered the problem margin sums in a zero-truncation framework (Schmettow 2009). Indeed, the distribution of the problems so far observed follows a binomial distribution with only a positive integer as support (i.e. a positive or conditional distribution). The distribution is zero-truncated because problems only appear in the discovery matrix once they have been discovered. The probability is then estimated using standard mathematical techniques, such as the maximum likelihood or moment estimator (Finney 1947; Rider 1955; Shah 1961). The probability mass function is:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{7}$$

and zero truncation is achieved as follows:

$$P(X = k)_{\mathrm{zt}} = \begin{cases} 0 & if\ k = 0 \\ \dfrac{P(X = k)}{1 - P(X = 0)} & if\ k > 0 \end{cases} \tag{8}$$

The probability of problem discovery is then estimated by using maximum likelihood techniques to fit the marginal sums to the zero-truncated binomial distribution. It should be noted that the

expected probability of unobserved problems, $\Pr(X = 0)$, is deduced from the non-truncated function (Schmettow 2009).

I.B.4 Methods taking account of a heterogeneous problem detection probability

All the methods presented above assume that the probability of detection is the same for all usability problems (i.e., the same p). However, this assumption is unrealistic and does not hold true in real-life usability testing. Schmettow showed that overdispersion was frequent in the problem margin sums, reflecting heterogeneity in the probability of detection (Schmettow 2008). Furthermore, erroneously ignoring the presence of heterogeneity by using a single, average value of p leads to overestimation of the completeness of the discovery process (Jensen's inequality) (Caulton 2001). Schmettow tackled this problem by developing a model that incorporated heterogeneity. The probability of detection was considered to be a random variable, which enabled each problem to have its own probability of detection. Schmettow used the logit-normal distribution as a plugin distribution for the probability of detection. Formally, the logit of the probability of detection follows a normal distribution $\mathcal{N}(\mu, \sigma)$. In this model, the problem margin sums follows a logit-normal binomial distribution and the probability mass function is:

$$P(X = k) = \binom{n}{k} \frac{1}{\sqrt{2\pi}\sigma} \int_0^1 (1-p)^{n-k-1} p^{k-1} \exp\left(-\frac{(logit(p) - \mu)^2}{2\sigma^2}\right) dp \qquad (9)$$

Using the zero truncation technique presented in equation (8), Schmettow developed the logit-normal binomial zero truncated (LNBzt) model and applied it to the usability of medical infusion pumps (Schmettow, Vos, and Schraagen 2013a). To the best of our knowledge, this model is the only one that accounts for both heterogeneity and unobserved problems.

I.B.5 Statistical limitations of margin-based methods

The primary limitation of the margin-based methods presented above is that they estimate the probability of detection only. The number of problems m is deduced but not estimated *per se*. It would be possible to estimate both m and p by summarizing the discovery matrix on the basis of the participants' margin. In such a case, each sum follows a binomial $\mathcal{B}in(m, p)$, thus enabling estimation of both the number of attempts and the probability of success in a binomial setting.

However, DasGupta and Rubin established that there were no unbiased estimates for essentially any functions of either the number of attempts or the probability of success (DasGupta and Rubin 2005). This problem was initially considered by Fisher and Haldane for estimating species abundance (Fisher 1941; Haldane 1941). It has also been considered by Olkin, Petkau, and Zidek, who developed both a moment and a maximum likelihood estimator, and by Carroll and Lombard, who proposed an estimator in a Bayesian setting (leading to a beta-binomial distribution) (Carroll and Lombard 1985; Olkin, Petkau, and Zidek 1981). Hall also considered this problem in an asymptotic framework (Hall 1994).

The second limitation of margin-based methods is information loss, relative to the initially available data. For example, *j* and the number of singletons were the only data used in the GT estimates. In the same way, the zero-truncated method considered only the column sums for the problems and omitted the pattern of detection (i.e., the users).

Here, we tackle these problems by directly modelling the full discovery matrix (including unobserved columns) and not only a summary of the data (e.g. the margins). In the Methods section, we describe the statistical basis of the matrix-based method and detail a Bayesian approach for estimating the number of problems. In the Results section, we compare the matrix-based method's statistical properties with those of existing models in a simulation study and then in actual usability studies. Lastly, we discuss the implications of our results with regard to estimation of the number of problems in usability testing.

I.C Methods

We first specify the statistical basis underpinning the matrix-based method, and the principle of column permutation in particular. Next, we present our estimation of the number of problems in a Bayesian setting. The last part is dedicated to the methods used to assess the matrix-based model's performance.

I.C.1 The matrix-based method

We first present the matrix-based method. For the sake of clarity, we simplified the problem by considering that the probability of problem detection was homogeneous. The concept of

heterogeneous probability will be introduced in the second part of this section, along with the Bayesian estimation.

I.C.1.a Presentation of the method

Consider the complete discovery matrix $\mathbb{x}$. The probability of $\mathbb{x}$ can be written as follows:

$$P(\mathbb{x}|p,m) = p^{\mathbb{x}_{\bullet\bullet}}(1-p)^{nm-\mathbb{x}_{\bullet\bullet}} \quad (10)$$

where $\mathbb{x}_{\bullet\bullet} = \sum_{i=1}^{n}\sum_{l=1}^{m} x_{il}$ is the total number of problems observed by n participants.

An example of a possible matrix $\mathbb{x}$ obtained from two participants during a usability testing of a medical device with $m = 3$ problems is given below (with users in rows and problems in columns):

$$\mathbb{x} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad (11)$$

As seen above, the complete discovery matrix $\mathbb{x}$ is never observed, and the discovery matrix $\mathbb{d}$ is the only one available. It is similar to the matrix $\mathbb{x}$, except that unobserved problems are missing. Considering the above example, neither of the users observed the second problem, and the resulting observed discovery matrix $\mathbb{d}$ would be:

$$\mathbb{d} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (12)$$

It should be noted that if the total number of problems m is known, then the complete matrix $\mathbb{x}$ could be reconstituted (with permutation), based on the matrix $\mathbb{d}$. For instance, if we take the matrix $\mathbb{x}$ and consider (wrongly, in this case) that the number of problems $m = 5$, then the reconstituted complete matrix denoted by $\hat{\mathbb{x}}^m$ would be obtained by padding the matrix $\mathbb{d}$ with columns of zeros (corresponding to as-yet unobserved problems):

$$\hat{\mathbb{x}}^{m=5} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \quad (13)$$

Thus, noting that $\mathbb{x}_{\bullet\bullet} = \mathbb{d}_{\bullet\bullet}$, it is possible to compute the likelihood of the complete matrix $\hat{\mathbb{x}}^m$ on the basis of the discovery matrix $\mathbb{d}$. This likelihood is given by the following equation:

$$P(\hat{\mathbb{x}}^m|p,m) = p^{\mathbb{x}_{\bullet\bullet}}(1-p)^{nm-\mathbb{x}_{\bullet\bullet}} \quad (14)$$

Note that the definition of $\hat{\mathbb{x}}^m$ depends on the value m, which is unknown. Thus, any inference based on $\hat{\mathbb{x}}^m$ will induce some bias. For instance, a maximum likelihood estimation of (p, m) based on $\hat{\mathbb{x}}^m$ (consisting in maximizing $p(\hat{\mathbb{x}}^m|p, m)$ with respect to m and p) leads to $\hat{m} = j$ (where j is the number of problems observed so far) and $p = \frac{\mathbb{x}_{\bullet\bullet}}{nj}$, which are known to be biased. We tackled this issue by modeling the distribution of the observed discovery matrix $p(\mathbb{d}|p, m)$.

It should be noted that the matrix $\mathbb{d}$ is defined in a lexicographic order, which simply means that the problems are ordered in the order of detection. For instance, the six possible complete matrices $\mathbb{x}$ leading to the previous matrix $\mathbb{d}$ if $m = 3$ are presented in Table 1.

Table 1: Six possible complete matrices $\hat{\mathbb{x}}^{m=3}$ leading to the observed discovery matrix $\mathbb{d} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$

Possibility 1	Possibility 2	Possibility 3
$\hat{\mathbb{x}}_1^{m=3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\hat{\mathbb{x}}_2^{m=3} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\hat{\mathbb{x}}_3^{m=3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
Possibility 4	**Possibility 5**	**Possibility 6**
$\hat{\mathbb{x}}_4^{m=3} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$	$\hat{\mathbb{x}}_5^{m=3} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$	$\hat{\mathbb{x}}_6^{m=3} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

In fact, if we could consider the label (the name of the usability problem) associated with each column, only one matrix $\mathbb{x}$ could lead to the matrix $\mathbb{d}$. However, since we have no means of finding the names of the columns in the initial matrix $\mathbb{x}$, we will consider that the matrix $\mathbb{d}$ has unnamed columns. Removing these column names allows us to consider the matrix $\mathbb{d}$ for the observed data (for which the definition does not vary as a function of the model's definition of the model – in contrast to $\hat{\mathbb{x}}^m$). Thus:

$$P(\mathbb{d}|m = 3, p) = \sum_{h=1}^{6} P(\hat{\mathbb{x}}_h^{m=3}|m = 3, p) \qquad (15)$$

and more generally

$$P(\mathbb{d}|m,p) = \sum_{h=1}^{H(\mathbb{d},m)} P(\mathbb{X}_h^m|m,p) \quad (16)$$

where $H(\mathbb{d},m)$ is the number of different matrices $\mathbb{X}_h^m$ with m columns leading to the same discovery matrix $\mathbb{d}$.

In the simple example presented above (Table 1), $H(\mathbb{d},m) = 6$ and each matrix $\mathbb{X}_h^m$ has the same probability, i.e. $p^2(1-p)^4$. It follows that:

$$P(\mathbb{d}|m=3,p) = \mathrm{H}(\mathbb{d},\mathrm{m}=3) \times P(\mathbb{X}_h^{m=3}|m=3,p) = 6 \times p^2(1-p)^4 = A_3^2 \times p^2(1-p)^4 \quad (17)$$

More generally, the number of matrices $\mathbb{x}$ with m columns associated with an observed discovery matrix $\mathbb{d}$ is:

$$\mathrm{H}(\mathbb{d},\mathrm{m}) = \frac{m!}{(m-j)!\,j_1!\ldots j_r!} = \frac{1}{j_1!\ldots j_r!} \times A_m^j \quad (18)$$

where r is the number of different columns of $\mathbb{d}$, and j_h $(1 \le h \le r)$ is the number of repetitions of the column of type h. Of course, $j = j_1 + \cdots + j_r$. Here, we recognize a familiar equation: that associated with the number of anagrams of a word in which each type of column corresponds to a different letter, including the null column (repeated $m-j$ times).

Lastly, since each matrix $\mathbb{X}_h^m$ has the same probability, we obtain the likelihood of $\mathbb{d}$ as follows:

$$P(\mathbb{d}|p,m) = \frac{1}{j_1!\ldots j_r!} \times A_m^j \times P(\mathbb{X}_h^m|m,p) \quad (19)$$

In practice, the computation of $\frac{1}{j_1!\ldots j_r!}$ has no impact on the estimation, since it is the same for all values of m and p. This result is not limited to the homogenous setting and would remain valid for any probability of $\mathbb{x}$ with a column-wise exchangeability property.

In the particular case of the homogeneous setting, we obtain:

$$P(\mathbb{d}|p,m) = \frac{1}{j_1!\ldots j_r!} \times A_m^j \times p^{\mathbb{x}_{\bullet\bullet}}(1-p)^{nm-\mathbb{x}_{\bullet\bullet}}. \quad (20)$$

In the homogeneous setting, our matrix-based approach could be extended to perform maximum likelihood inference or Bayesian inference on the parameters. However, as explained above, this setting is unrealistic in practice and so a heterogeneous probability of detection should be considered in the following section.

I.C.1.b Heterogeneity and Bayesian estimation

We considered a heterogeneous probability of detection; i.e. each problem l has its own probability of detection p_l. In line with Schmettow's method, we assume that the probabilities of detection are independent and follow a logit-normal distribution, i.e. $\text{logit}(p_l) \sim \mathcal{N}(\mu, \sigma)$. The model's parameters are m, μ and σ. Note that $p_1, \dots, p_m$ are considered as latent random variables - like random effects in the mixed model.

Given these parameters, the likelihood of the discovery matrix $\mathbb{d}$ can be written as

$$P(\mathbb{d}|\mu, \sigma, m) = \int_0^1 \dots \int_0^1 P(\mathbb{d}|p_1, \dots, p_m, m) f(p_1, \dots, p_m|\mu, \sigma) dp_1 \dots dp_m \qquad (21)$$

where $f(p_1, p_2, \dots, p_m|\mu, \sigma)$ is the probability density function of $p_1, p_2, \dots, p_m$. Given that the columns are exchangeable, we can also write

$$P(\mathbb{d}|\mu, \sigma, m) = \frac{1}{j_1! \dots j_r!} \times A_m^j \times P(\hat{\mathbb{x}}_h^m|\mu, \sigma, m) \qquad (22)$$

which will be useful for subsequent computations.

We now consider a Bayesian framework (Robert 2007) for estimation of the parameters. This framework has good theoretical properties and can include prior knowledge about the problem's parameters. Indeed, the distribution of the parameters $P(\mu, \sigma, m)$ must first be defined. Moreover, assuming the prior independence of μ, σ and m, $P(\mu, \sigma, m) = P(\mu)P(\sigma)P(m)$. We assume a prior uniform distribution for m:

$$P(m) = \frac{1}{M} \forall m \in \{1, \dots, M\} \qquad (23)$$

The value of M is the pre-determined upper boundary for m, and should be chosen by the human factor engineer according to the expected maximum possible number of problems. To prevent

underestimation, a high value should be used. However, if M is unnecessarily high, it will lead to an increase in the computing time.

Since our goal here is to estimate the number of problems, our main interest is $P(m|\mathbb{d})$, which is obtained using Bayes' theorem:

$$P(m|\mathbb{d}) = \frac{\mathrm{P}(m) \times P(\mathbb{d}|m)}{\sum_{m'=1}^{M} \mathrm{P}(m') \times P(\mathbb{d}|m')} \quad (24)$$

Thus, we need to compute $P(\mathbb{d}|m)$ for each possible value of m in $\{1, \dots, M\}$. This computation requires computation of the integrated likelihood $P(\mathbb{d}|m)$, as follows

$$P(\mathbb{d}|m) = \int_{0}^{+\infty} \int_{-\infty}^{+\infty} P(\mathbb{d}|\mu, \sigma, m) P(\mu) P(\sigma) d\mu \, d\sigma \quad (25)$$

The choice of prior distributions for $P(\mu)$ and $P(\sigma)$ is discussed below. $P(\mathbb{d}|m)$ can be computed by approximating this integral with Markov chain Monte Carlo (MCMC) techniques.

Even though $P(m|\mathbb{d})$ is the main quantity of interest, $P(\mu|\mathbb{d})$ and $P(\sigma|\mathbb{d})$ are also of interest because they can be used as prior distributions for future studies; this will decrease the sample size and improve early estimates as part of an early control strategy.

I.C.1.c Computational aspects

From a computational perspective, and since $P(\mathbb{d}|\mu, \sigma, m) = \frac{1}{j_1! \dots j_r!} \times A_m^j \times P(\hat{\mathbb{x}}_h^m|\mu, \sigma, m)$, we will first focus on the computation based on $\hat{\mathbb{x}}_h^m$ and will then deduce the results for $\mathbb{d}$.

Let now consider the choice of a prior distribution for μ and σ. Since μ and σ are Gaussian distribution parameters and in the absence of additional information (e.g. from previous usability studies), we chose the following flat priors:

- $\mu \sim \mathcal{N}(0; \mathcal{A})$: a Gaussian distribution with a high variance $\mathcal{A}$, (e.g. $\mathcal{A} = 10^8$), mimicking a uniform distribution on $\mathbb{R}$,
- $\sigma^2 \sim \mathrm{inv} - \chi_\nu^2$: an inverse chi-squared distribution with ν degrees of freedom (typically $\nu = 1$).

When the data has a Gaussian distribution, choosing the above priors leads to a conjugated posterior distribution. However, a logistic-normal distribution of the probabilities of detection means that conjugacy cannot be obtained. Thus, estimation of the posterior distribution required the use of MCMC methods. This consisted in drawing μ and σ for each possible value of m, $m \in 1, \ldots, M$ according to their posterior distribution $P(\mu, \sigma | m, \mathbb{d})$, and deducing a numerical approximation of $P(\mathbb{d}|m)$ from the Monte-Carlo sample. Lastly, $P(m|\mathbb{d})$ was computed using Bayes' theorem.

For a fixed value of m, we consider sampling from $P(\mu, \sigma | \mathbb{x}_h^m, m)$, computing the integrated likelihood $P(\mathbb{x}_h^m | m)$ with bridge sampling (Meng and Wong 1996), and deducing $P(\mathbb{d}|m)$.

The parameters μ and σ (given $\mathbb{x}_h^m$ and m) are sampled using the parameter space augmented by $p_1, \ldots, p_m$, i.e. the discovery probabilities associated with each column of $\mathbb{x}_h^m$. Thus, we will now sample from $\mu, \sigma, p_1, \ldots, p_m | \mathbb{x}_h^m$, using stan software (adaptative Hamiltonian Monte Carlo algorithm).

I.C.2 Assessment of the performance of the matrix-based method

We compared the performance of five methods (naïve, GT, double-deflation, LNBzt, and matrix-based methods) first in a simulation study and then using literature data from actual usability studies.

I.C.2.a Simulation study

Each simulation consisted in generating an observed discovery matrix $\mathbb{d}$ from the usability testing of a hypothetical medical device with a known total number of usability problems m and a sample size n. The probability of detection was normally distributed ($\mathcal{N}(\mu, \sigma)$) on a logit scale. The combinations of parameters used in the simulations are specified in Table 2. The values were chosen to reflect a wide range of parameters encountered in usability testing of medical devices.

Table 2: Combinations of parameters for the simulation testing with homogeneous and heterogeneous probabilities of detection.

Parameter	Values
Total number of usability problems	$m = 20,50,100$
Sample size	$n = 15,20,30,40,50$
Probability of problem detection	$\mu = logit(0.1), logit(0.2)$ $\sigma = 0.5, 1, 2$
Number of combinations tested	**90**

In each setting (i.e. for each combination of m, μ, σ and n), we simulated $S = 2 \times 10^4$ complete discovery matrices, $\mathbb{x}_{m,\mu,\sigma,n,i}$, $i \in \{1,2, \dots, S\}$.. The matrices $\mathbb{d}$ were obtained by truncation of the zero columns (problems not yet discovered). We averaged the estimates of m over the S simulations and computed the 95% fluctuation interval (0.025 and 0.975 quantiles). We also calculated the prediction's root mean square error (RMSE) as the square root of the mean square difference between the predicted and true values of m:

$$RMSE(m) = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(m - \hat{m}_i)^2} \qquad (26)$$

When the sample is small, little information is available; a tight credible interval might reflect overconfidence rather than a good estimation. Thus, to gauge the level of confidence that human factor engineers can place in each method, we computed the coverage probability. In each setting, this is the proportion of 95% confidence intervals for the simulated $\hat{m}_i$ that include the true value of m. The confidence intervals for $\hat{m}_i$ were computed using 1000 parametric bootstrap repetitions with the parameters $(\hat{m}_i, \hat{\mu}_i, \hat{\sigma}_i, n)$. For the matrix-based method, we were able to directly compute the 95% confidence interval of the posterior distribution of each simulation, which saved substantial computation time.

I.C.2.b Application to actual usability studies

We applied the above-described methods to the discovery matrices of five published usability studies. Four did not involve a medical device: the EDU3D dataset encompassed 119 problems discovered by 20 participants during the evaluation of virtual environments (Bach and Scapin 2010), the MACERR dataset encompassed 145 problems discovered by 15 participants during a scenario-driven usability testing of an integrated office system (Lewis, Henry, and Mack 1990), the MANTEL dataset encompassed 30 problems submitted by 76 expert participants evaluating the specifications of a computer program, and the SAVINGS dataset encompassed 48 usability problems discovered by 34 participants on voice response systems MANTEL and SAVINGS comes from the same experiment on heuristic evaluations (Nielsen and Molich 1990). These four studies were included because they have been used in important publications in this field (Lewis 2001) and they enabled us to address heterogeneity in the probability of discovery, in particular (Schmettow 2008). The fifth usability testing involved a medical device: INFPUMP encompassed 107 usability problems discovered by 34 participants (intensive care unit nurses and anesthesiologist) evaluating a prototype medical infusion pump (Schmettow, Vos, and Schraagen 2013a).

For each of the five datasets, we computed the estimates and the 95% confidence intervals for the final data. When a sufficient number of participants had been included (i.e. for MANTEL, SAVINGS, and INFPUMP), we addressed the change in the estimates as a function of the sample size.

All the analyses were carried out running R software (version 3.6.1) on several servers equipped with 12-core Intel® Xeon® E5-2650 v4 processors (http://hpc.univ-lille.fr/cluster-hpc-htc). The MCMC was performed using the *Stan* library (http://mc-stan.org) via the *rstan* package (Team 2018). The integrated likelihood was obtained using the *bridge_sampler* function of the *bridgesampling* package (Gronau, Singmann, and Wagenmakers 2017). In order to facilitate the matrix-based method's application in practice, a short step-by-step tutorial [see Additional file 1] and the code [see Additional file 2] is provided as supplementary material. A reproducible R code with the data and the simulation study performed in this manuscript is available on GitHub

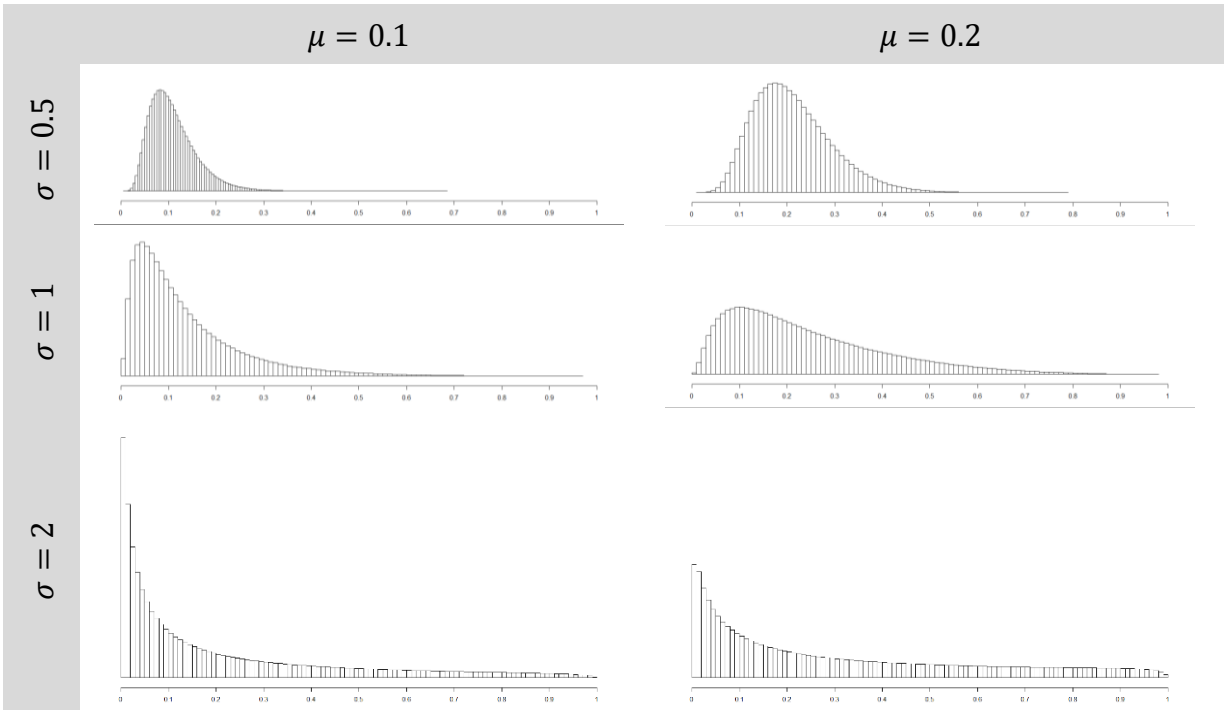
(https://github.com/alexandre-caron/matrix_based-usability). The link to the archived version referenced in this manuscript is available in the "Availability of data and materials" section.

I.D Results

I.D.1 The simulation study

The distributions of the probability of detection for each setting are summarized in Table 3. The distribution shifted to a highest average probability of detection when μ increased. It is noteworthy that a higher dispersion (σ) not only flattened the distribution but also led to an increase in probability of very rare problems.

Table 3: Distribution of the probability of detection as a function of μ and σ. The probability of detection followed a logit-normal distribution: $logit(p_i) \sim \mathcal{N}(\mu, \sigma)$.



The results of the simulation are presented for the five methods (naïve, GT, double-deflation, LNBzt, and matrix-based). The prediction error of m as a function of the sample size n are presented in Figure 1. The RMSE is presented in Figure 2. A tabulated version of these data is also provided as supplementary material [Appendix 3]. As mentioned by Schmettow, extreme estimates of m can be obtained with the LNBzt method when the number of singletons is high.

We decided to discard any results with $\widehat{m}_{LNBzt} > 500$, to avoid penalizing the method with estimates that would not be realistic in real life (Schmettow 2009).

As expected, the accuracy of the estimation of the number of problems increased with the sample size for all estimates, with less bias and greater consistency (i.e. the RMSE tended towards zero as the sample size increased). Likewise, the estimates were better as the number of problems to discover m increased. For all methods, the bias was higher as the number of "rare" problems increased (i.e. for a higher σ).

Methods accounting for heterogeneity: the matrix-based and LNBzt estimates

The matrix-based method showed less bias overall; the bias ranged from -8.5% to +14.7% for the 90 simulated combinations. This range was narrower (from -5.1 to +1.2%) when the participant sample size was 30 or more. In contrast, the LNBzt method displayed systematic upward bias; although the lower boundary was -0.1%, the upper boundary was 54.7%. This bias was still observed for 30 participants, with an upper boundary of 23.8%.

When $\sigma = 2$, the matrix-based method underestimated the number of problems. However, this underestimation was less than -5.1% for $n \geq 30$. For lower values of σ, the matrix-based method's bias ranged from -2.6% to +1.2% for $n \geq 30$. The bias associated with the LNBzt method was high for $\sigma = 2$. Although the bias decreased with n, it was still +11.8% for $n = 50$. For a lower value of σ, the bias associated with the LNBzt method ranged from -2.6% to +1.2% for $n \geq 30$.

The matrix-based method gave the lowest RMSE in all settings. This was particularly true when the number of "rare" problems was high ($\sigma > 0.5$). The LNBzt gave the highest average RMSE. As mentioned in the Methods, this bias resulted from a few very high estimates of m, which increased the average RMSE dramatically. This was true for the lowest average probability of detection (i.e. $\mu = \text{logit}(0.1)$) and the highest variance (i.e. $\sigma = 2$).

Methods not accounting for heterogeneity: the naïve, GT, and double-deflation estimates

The estimates that did not take account of heterogeneity showed the strongest bias. The naïve estimate was the worst; it systematically underestimated the true value of m (range: -33.2% to -

0.2%). This underestimation was slightly lower for the GT estimate, especially when σ was low. However, the range was still broad: from -32.2% to -0.2%. The double-deflation method compensated even more for underestimation but sometimes led to overestimation (range: -32.0% to +8.6%).

When σ was lower (i.e. 0.5 or 1), the trend towards underestimation was less pronounced for the double-deflation and the GT methods (with lower boundaries of -14.1% and -17.2, respectively) than for the naïve method (lower boundary: -22.8%. The bias persisted for larger sample sizes: it was still as high as -6.4% for the three methods for $n = 50$.

The naïve RMSE estimate was again the worst of the methods that did not take account of heterogeneity. Although the GT and the double-deflation methods gave acceptable RMSEs, this feature must be interpreted with caution. In fact, the acceptable RMSEs resulted essentially from systematic underestimation, which in turn limited the range of possible $\widehat{m}$ (which can never be lower than j). Hence, the interpretation of the RMSE was limited for these methods.

Coverage probability.

As explained in the Methods, human factor engineers do not know the variables for the usability testing they are carrying out. The coverage probability enables them to study the reliability of the estimate (and its 95% confidence interval). A tabulated version of the data is provided as supplementary material [see **Error! Reference source not found.** in Appendix 3].

For the matrix-based method, the coverage probability was always over 80% (except for $m = 100, n = 15, \mu = logit(0.1)$, and $\sigma = 0.5$, where the probability of coverage dropped to 72%) with an average of 94% over the range of settings tested in the simulations study. The probability was at least 81% for $n \geq 20$ and at least 88% for $n \geq 30$. The LNBzt method's coverage probability was always over 80%, with an average of 92%. The LNBzt performed particularly well for small sample sizes, with a minimum coverage of 89% for $n = 15$, of 86% for $n = 20$, and of 82% for $n = 30$. Indeed, the LNBzt method provided the broadest confidence intervals of the five methods studied here. It is noteworthy that the LNBzt method was the only one that sometimes failed to fit the data (in 33% of cases). However, it was impossible to adjust the method's

parameter for each individual simulation. In practice, changing the optimization function's starting values would avoid most of the fitting failures.

The methods not taking account of heterogeneity provided a low, erratic coverage probability in most settings. On average, the coverage probabilities were 17.9%, 31.5% and 33.7% for the naïve, GT, and double-deflation methods, respectively. Furthermore, the three methods frequently yielded excessively high estimated levels of confidence - especially for high values of m.

Lessons learned from the simulation study

From the human factor engineer's point of view, the matrix-based and LNBzt methods are the only reliable ones; they gave a good coverage probability in almost any setting and for almost any sample size. Conversely, the methods not taking account of heterogeneity were unreliable and so could not be trusted.

I.D.2 Application to real data from published usability studies

The estimated number of problems computed from the discovery matrices of five published usability studies are presented in Table 4. Although the real number of problems is not known, we can compare the matrix-based method's predictions with those of the other methods (and especially the LNBzt method).

Table 4: The estimated number of problems for five real datasets from published usability studies.

	n*	j**	naïve	Good-Turing	double deflation	LNBzt	matrix-based
EDU3D	**20**	**119**					
$\hat{m}$			120	121	122	155	152
95%CI			117 – 121	118 – 125	120 – 129	132 – 195	135 – 167
MACERR	**15**	**145**					
$\hat{m}$			156	178	184	449	382
95%CI			146 – 160	171 – 207	192 – 245	256 – 1301	346 – 440
MANTEL	**76**	**30**					
$\hat{m}$			30	30	30	31	30
95%CI			30 – 30	30 – 30	30 – 30	31 – 35	30 – 37
SAVINGS	**34**	**44**					
$\hat{m}$			44	44	44	46	45
95%CI			44 – 45	44 – 45	44 – 45	42 – 50	44 – 51
INFPUMP	**34**	**107**					
$\hat{m}$			107	107	107	122	120
95%CI			107 – 108	106 – 108	106 – 108	110 – 136	112 – 143

* n is the number of participants in the study
** j is the number of problems discovered after analyses by n participants

In these five datasets, the number of participants ranged from 15 to 76. Previous studies of these datasets (Schmettow 2008, 2009; Schmettow, Vos, and Schraagen 2013a; Lewis 2001) demonstrated that the probability of problem detection was heterogeneous. As suggested by the results of the simulation study, the methods not taking account of heterogeneity considered that the discovery process was complete or very close to being complete for all datasets (except MACERR: see below). Thus, we compared the results of the methods that do account for

heterogeneity. It is noteworthy that the estimates of μ and σ^2 by both the LNBzt and the matrix-based methods fell within the range observed in our simulation study for all datasets other than MACERR.

All five methods considered that the SAVINGS and MANTEL datasets were complete after 34 and 76 participants had been included, respectively. However, the confidence intervals produced by the matrix-based and the LNBzt methods suggest that few problems had yet to be discovered.

The matrix-based and the LNBzt methods estimated similar number of problems for EDU3D ($\widehat{m}_{\text{matrix-based}} = 152$ and $\widehat{m}_{LNBzt} = 155$). The 95% confidence interval was broader for the LNBzt method (132 to 195) than for the matrix-based method (135 to 167).

The infusion pumps in the INFPUMP study were in early-stage development, and an additional re-design phase (for fixing the usability problems discovered) was planned; this explains why n= 107 unique problems were detected by the 34 participants in the usability testing. The LNBzt and matrix-based methods gave similar estimates and confidence intervals: $\widehat{m}_{\text{LNBzt}} = 122$ (i.e. 15 undiscovered problems), with a 95% confidence interval from 115 to 131, whereas $\widehat{m}_{\text{matrix-based}} = 120$, with a 95% confidence interval from 112 to 143. The parameters computed by the matrix-based method predicted an average probability of detection $\hat{\mu}_{\text{matrix-based}} = logit(0.136)$ and a dispersion of $\hat{\sigma}_{\text{matrix-based}} = 1.52$. For the LNBzt method, the probability $\hat{\mu}_{\text{LNBzt}} = logit(0.136)$ was the same, and the dispersion was slightly higher ($\hat{\sigma}_{\text{LNBzt}} = 1.50$). The confidence interval (from 110 to 136) was narrower. The true number of problems with the pump was not known because it was redesigned after 34 participants had tested the device. However, if we accept the parameters $\hat{\mu}$ and σ as true and apply the results of our simulation study, the INFPUMP data suggest that the LNBzt and matrix-based methods are both reliable. Nevertheless, the breadth of the respective confidence intervals emphasizes the remaining uncertainty for these two methods.

Using the MACERR data, the LNBzt predicted a very low average probability of detection ($\hat{\mu}_{\text{LNBzt}} = logit(0.014)$) and a high level of heterogeneity ($\hat{\sigma}_{\text{LNBzt}} = 1.90$). These values were out of the range of the settings tested in the simulation study, and suggested that the number of "rare" problems was high. This might explain the high number of problems predicted by the LNBzt

method ($\widehat{m}_{\text{LNBzt}} = 449$), and the very large 95% confidence interval (from 256 to 1301). The matrix-based method's estimate was lower ($\widehat{m}_{Matrix-based} = 382$), and the 95% confidence interval was narrower (346 to 440). However, the number of participants included in MACERR was low (n=15); a larger number of participants would have been necessary to discover new problems and improve the estimates.

On average, computation of the estimate and its confidence interval took less than ten minutes for the matrix-based method, less than one minute for the LNBzt method, and only a few seconds for the three other methods.

I.E Discussion

We decided to model the full discovery matrix (including unobserved columns) and not just a summary of the data (e.g. the margins). The estimation problem was considered simultaneously in terms of the (heterogeneous) probability of problem detection and the number of problems. Although the experimental conditions in real-life usability testing are unknown, the matrix-based method outperformed the other methods and appeared to be the most reliable in a broad range of settings.

Most of the currently available methods assume that the probability of detection is the same for all problems. This assumption is likely to be wrong, since real data show that the probability of detection varies (Schmettow 2008, 2009). Furthermore, ignoring heterogeneity is known to strongly bias the results (Caulton 2001; Woolrych and Cockton 2001). We therefore developed a method that accounted for heterogeneity in the probability of problem discovery p; we used a logit-normal distribution as a plugin to model this uncertainty. The choice of this distribution was convenient in that it allowed us to compare our method with the only published model that accounts for heterogeneity. However, there are no data for confirming the validity of this choice. Nevertheless, this limitation could be easily overcome by replacing the logit-normal by another distribution (such as beta or gamma) if it proves to be more appropriate. This choice could be made using model choice criteria (e.g. the Akaike information criterion or the Bayesian information criterion). However, it should be borne in mind that for a small sample size, fitting

for both incompleteness and heterogeneity is complex and inevitably leads to a high degree of uncertainty.

Here, we sampled μ and σ for fixed values of m. This turned out to be a rather time-consuming strategy because we had to run as many chains as there were values of m. We chose not to sample directly from the joint distribution $P(\mu, \sigma, m|\mathbb{d})$ because the dimension of the latent parameters $p_1, p_2, \ldots, p_m$ varied as a function of m - making it impossible to use a standard MCMC algorithm. In this particular situation, use of the reversible jump algorithm (Green 1995) might be a solution but would considerably complicate our algorithm.

There are two key moments in medical device development for assessing the best method. Early in the development cycle, the device is not mature; usability testing is referred to as "formative" because many usability problems are being discovered and corrected in an iterative design improvement process. Just before market access, usability testing is referred to as "validation" testing; they are performed on the final version of the device to ensure that no critical usability problems remain (Food and Drug Administration 2016; UK-MHRA 2017).

The number of participants in the validation testing is an important parameter for both the regulatory authorities and the device manufacturer. Indeed, a sufficient sample size will (i) guarantee the medical device's compliance with the safety standards required for market authorization, and (ii) avoid a "black swan" effect that would strongly affect the manufacturer's credibility and profitability (Bias and Mayhew 2005). The validation testing focuses on the detection of infrequent usability problems. The US Food and Drug Administration requires a minimum of 15 participants (Food and Drug Administration 2016). This minimum is based on a naïve estimate, which has been proven to dramatically underestimate the true number of usability problems for this number of participants (Faulkner 2003). Indeed, the average coverage probability observed in our simulation study for $n = 20$ was as low as 12% and did not exceed 51%. Furthermore, this threshold does not consider heterogeneity in the probability of problem detection. Our findings suggest that to produce a relevant estimate with the matrix-based method, at least 20 participants are required in the validation step. In fact, the matrix-based method displayed good statistical properties with as few as 20 participants.

Since the validation testing only concerned problems that are probably less frequent, one could question the need to use methods that account for a heterogeneous probability of problem detection. In fact, problems are expected to be "homogeneously rare". To the best of our knowledge, however, the assumption of homogeneity for rare problems has no theoretical or experimental basis. Furthermore, human factor engineers will define the usability testing's experimental conditions according to the risk analysis, in order to facilitate the detection of problems previously described in the literature. If an engineer suspects the existence of problem removing the cap from an adrenaline pen, he/she might choose to evaluate the device in a more realistic test environment (e.g. with an actor pretending to go into anaphylactic shock); the problem is more likely to occur there than in a quiet, low-fidelity environment. By making some problems more detectable, the human factor engineer might introduce a degree of heterogeneity into the discovery process.

The choice of method was even more obvious for "formative" testing. In our simulations, the "formative" testing corresponds to a setting in which usability problems are frequent and numerous. Schmettow's usability testing of a medical infusion pump is also an example of a formative assessment because it was followed by a redesign. Here, we proved that matrix-based methods are more reliable and have low bias and high consistency. As in the case of the infusion pump, a reliable estimate from a small number of participants is an economic advantage for the manufacturer, who can shorten redesign cycles, accelerate device development, and hasten market access. The matrix-based method met this requirement because it required the fewest participants to guarantee good statistical properties. Another strength of the matrix-based method is its ability to embed previous knowledge through the prior parameters. Indeed, we used weakly informative priors for μ and σ to avoid introducing information that we did not have about the medical device in question. However, one could take advantage of prior knowledge from earlier stages in device development or from a formative usability assessment to increase the accuracy of the estimate, especially when the sample size is small (i.e. an early control strategy). This approach is actually encouraged by regulatory bodies for medical device clinical trials (Food and Drug Administration 2010) and helps to reduce the overall sample size.

Although we have suggested a threshold of 20 participants as the minimum sample size for obtaining a reliable estimate with the matrix-based method, we do not consider this to be the final threshold or a "magic number". Indeed, as suggested by various researchers, the estimation models should be run iteratively as the sample size increases (Borsci et al. 2013). Thus, estimation models constitute a means of controlling and ensuring quality in formative testing and should not solely be considered as a checkpoint for validation testing. Although the matrix-based method was more reliable, the LNBzt method could be used to double check the estimates - especially when high dispersion and/or the presence of very rare problems is suspected. Indeed, the LNBzt method's coverage probability is high, and the overestimation bias makes it a conservative method that could usefully prevent the usability testing from being stopped too early.

I.F Conclusions

Estimation models (and particularly matrix-based models) are of value in estimating and monitoring the detection process during usability testing. Matrix-based models have a solid mathematical grounding and, with a view to facilitating the decision-making process for both regulators and device manufacturers, should be incorporated into current standards. To this end, the step-by-step tutorial provided here should facilitate the practical use of the matrix-based method in the evaluation of medical devices.

II THE OPTIMAL SAMPLE SIZE FOR USABILITY TESTING, FROM THE MANUFACTURER'S PERSPECTIVE: A VALUE-OF-INFORMATION APPROACH – PUBLIÉ DANS VALUE IN HEALTH

ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Methodology

The Optimal Sample Size for Usability Testing, From the Manufacturer's Perspective: A Value-of-Information Approach

Alexandre Caron, MD,* Vincent Vandewalle, PhD,* Romaric Marcilly, PhD, Jessica Rochat, MSc, Benoit Dervaux, PhD

*Alexandre Caron and Vincent Vandewalle contributed equally to the manuscript and should both be considered first author.

II.A Abstract

II.A.1 Introduction

For medical devices, a usability assessment is mandatory for market access; the objective is to detect potentially harmful use errors that stem from the device's design. The manufacturer assesses the final version of the device and determines the risk-benefit ratio for remaining errors. However, the decision rule currently used to determine the sample size for this testing suffers from statistical limitations and the lack of a clear decision-making perspective.

II.A.2 Methods

As an alternative, we developed a value-of-information analysis from the medical device manufacturer's perspective. The consequences of use errors not detected during usability testing and the errors' probability of occurrence were embedded in a loss function. The value of further testing was assessed as a reduction in the expected loss for the manufacturer. The optimal sample size was determined using the expected net benefit of sampling ($ENBS$, the difference between the value provided by new participants and the cost of their inclusion).

II.A.3 Results

The value-of-information approach was applied to a real usability test of a needle-free adrenaline auto-injector. The initial estimate (carried out on the first $n = 20$ participants) gave an optimal sample size of 100 participants and an ENBS of €250,000. This estimation was updated iteratively as new participants were included. After the inclusion of 90 participants, the $ENBS$ was null for any sample size; hence, the cost of adding more participants outweighed the expected value-of-information, and the study could therefore be stopped.

II.A.4 Conclusion

Based on these results, our method appears to be highly suitable for sample size estimation in the usability testing of medical devices before market access.

II.B Introduction

II.B.1 Usability of medical devices

Medical devices' design issues generate errors that have a negative impact on use; these range from inconvenience or difficulty of use to potentially life-threatening errors (e.g. poor usability of a cardiac defibrillator, leading to a delay in defibrillation) (Reeson, Kyeremanteng, and D'Egidio 2018). Indeed, design issues account for more than a third of medical device recalls and thus constitute the leading cause of these incidents (Food and Drug Administration 2012). With the overall goal of making medical devices safer, the regulatory bodies have decided that a usability assessment is mandatory for market access (European Commission 2017). The objective is to ensure that the device is "designed and optimized for use by the intended users in the context in which the device is likely to be used" (UK-MHRA 2017).

The usability assessment aims at detecting potentially harmful use errors that stem from the device's design. Each error's acceptability is assessed according to the trade-off between the hazard associated with decreased effectiveness, the risk of harm (severity), and the benefit still provided by the medical device. If an error is considered to be acceptable by the manufacturer, the associated hazard can be mitigated by mentioning it in the user manual or by training the end user appropriately. Conversely, if the error is unacceptable, the medical device must be redesigned. During the device's development phase, several iterations of "formative" usability studies are performed to mitigate or correct use errors.(Pelayo, Marcilly, and Bellandi 2020) Once the medical device's design is mature, the manufacturer performs "summative" testing in order to assess the risk associated with residual errors. This summative testing is the cornerstone of the premarket submission and is governed by two requisites: that the medical device is free of unacceptable use errors, and that any residual errors are known and deemed acceptable.

Box 1: Usability assessment in the regulatory approval of medical devices.

Regulatory bodies recognize that use errors induced by medical device design issues are a potential cause of patient injury or death. To prevent the occurrence of errors due to usability issues and to optimize the use of the medical device by users, regulatory bodies

around the world require medical device manufacturers to implement a human factor engineering (HFE) process. This process is closely related to the risk management process; it aims to assess and mitigate usability-induced errors that could lead to risks for patients. With this objective in mind, manufacturers must first mitigate the hazards and risks associated with use (e.g., through design changes). To confirm that these mitigation efforts have been successful and that users can use the device safely and effectively, manufacturers must then evaluate the medical device's usability, assess the residual risks, and determine whether or not these risks are acceptable (i.e. validation testing). If the HFE process has been followed, then the usability of a medical device with respect to safety is presumed to be acceptable - unless the evaluation provides evidence to the contrary.

In practice, the participants in the validation study are representative users who are asked to use the medical device in an experimental environment that mirrors the actual conditions of use. Each use error made is added to a discovery matrix - a binary matrix with the participants as the rows and the errors as the columns. The result is 1 if the participant made the error and 0 if not. For instance, let us imagine that after $n = 5$ participants, $j = 10$ errors have been discovered. The corresponding $n \times j$ matrix (denoted as $\mathbb{d}$) might be as follows:

$$\mathbb{d}_{n=5} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$

A new column is added to the matrix each time a participant discovers a new error. Of course, a given error can be discovered several times - leading to several 1s in a given column.

Today's standard approach involves assessing the exhaustiveness of the error discovery process by modelling the probability of detection of use errors and thus estimating the expected number of errors (denoted as m) (Lewis 2001; Virzi 1992; Nielsen and Landauer 1993). An early control strategy (based on the results provided by the first participants) is currently recommended by the regulators (Food and Drug Administration 2016; Faulkner 2003). The proportion of the total number of use errors discovered is estimated iteratively as the sample size increases, until it

reaches a predetermined threshold (i.e. as part of an adaptive strategy) (Lewis 1994). For more information, the reader is referred to a detailed review of current approaches (Vandewalle et al. 2020).

In earlier work, we introduced a matrix-based Bayesian approach for estimation of the number of errors affecting medical devices (Vandewalle et al. 2020). Our model overcomes the statistical limitations of older approaches (Lewis 2000; Sauro and Lewis 2016) by (i) modelling the full discovery matrix, including as-yet unobserved columns that truncate the margin and bias estimates,(Thomas and Gart 1971) (ii) simultaneously estimating the model's parameters and the total number of errors (DasGupta and Rubin 2005), and (iii) accounting for a heterogeneous probability of error detection with a logit-normal binomial distribution (Schmettow 2008, 2009; Caulton 2001). To facilitate understanding of the present work, a summary of the Bayesian approach and a comprehensive overview of the mathematics used in the matrix-based model are provided in Appendix 1.

II.B.2 Value-of-information analysis

Here, we hypothesize that the use of exhaustiveness as the criterion for sample size determination is not consistent with the patient safety objective pursued by the healthcare system's stakeholders in general and by regulators and manufacturers in particular. Firstly, exhaustiveness does not account for the acceptability of undetected errors. Indeed, the persistence of an unacceptable error is far more serious because it will prompt the regulator to order a device recall. Secondly, the use of an arbitrary threshold (e.g. 90%) means that the omission of 1 error in 10 and the omission of 5 errors in 50 are treated equally, which is not appropriate.

Value-of-information analysis has proved to be an effective means of decision support during the health technology life cycle. It provides a formal assessment of the value of research and help to guide research design.(Willan and Pinto 2005) Moreover, it is particularly suitable for the small samples typically employed in usability testing of medical devices.(Rothery et al. 2017) The value-of-information approach is based on Bayesian decision theory,(Raiffa and Schlaifer 1961) and additional research is considered to be valuable if it reduces the likelihood (and consequences) of making a wrong decision.(Fenwick et al. 2020) Here, we carried out a value-of-information

analysis from the manufacturer's perspective. Indeed, the manufacturer is primarily responsible for evaluating its medical device and seeks to bring the product to market after an appropriate risk-benefit assessment and regulatory body approval (e.g. the CE mark or FDA approval). Ideally, the manufacturer would prefer to mitigate usability-induced use errors before the market launch, in order to avoid the economic consequences of failure: re-engineering costs, loss of turnover, legal action, reputational damage, and loss of customer trust.(Bias and Mayhew 2005; Slawomirski, Auraaen, and Klazinga 2017) Furthermore, adopting the manufacturer's perspective allows one to shift the sample size paradigm from risk avoidance to risk management. Instead of eliminating the risk (which is nearly impossible), the goal of the value-of-information analysis is to quantify it and to adjust the sample size accordingly.(Hall et al. 2017; Pezeshk and Gittins 2006) The primary objective of the present study was to present a framework for optimal sample size determination from the manufacturer's perspective via the application of a value-of-information analysis. This framework (detailed in the Methods section) was applied to actual usability testing of a new medical device (a needle-free adrenaline auto-injector).

II.C Methods

II.C.1 The loss function

A value-of-information analysis requires formalization of the decision-maker's preferences. Here, we supposed that a medical device manufacturer behaves as a risk-neutral profit maximizer.(Chen and Willan 2013) From the manufacturer's standpoint, additional research is valuable if it reduces the expected loss (in terms of profit) associated with errors that are left undetected and that might be experienced by end users. Thus, deferring the request for regulatory approval while acquiring additional usability information might be of interest to the manufacturer. If more participants are added to the usability test (leading to the discovery of new use errors), the reduction in the loss corresponds to the expected value of sample information (EVSI).

The loss function was based on the assumption that the discovery of use errors by the medical device's end users mirrors the discovery process observed during the usability testing. Thus, the matrix-based method could be applied on the discovery matrix $\mathbb{d}$ to estimate the probability and

of an end user discovering at least one error. The mathematics underlying these computations are detailed in Appendix 2. To determine the opportunity cost of an error being experienced by an end user, we hypothesized that the economic consequences depended on whether the error was acceptable or not. The method used to estimate the number of undiscovered use errors (denoted as y_+ and y_- for acceptable and unacceptable errors, respectively) is detailed in Appendix 3. For a population of N end users, the expected loss is:

$$\mathcal{L} = \begin{cases} \bar{y}_+ \times c_+, & \bar{y}_- = 0 \\ \bar{y}_+ \times c_+ + c_-, & \bar{y}_- > 0 \end{cases}$$

where c_+ and c_- are respectively the opportunity costs associated with the detection of at least one acceptable or one unacceptable error by an end user. Thus, the expected loss is $y_+ \times c_+$ if all unacceptable errors have been detected or $y_+ \times c_+ + c_-$ otherwise. For the sake of clarity, we considered that any error experienced by an end user was acted upon and led to consequences for the manufacturer. This assumption can be relaxed by either weighting the opportunity cost (e.g., by the likelihood of legal pursuit) or increasing the threshold for triggering the consequences of an unacceptable error. However, this assumption holds true if we consider that most users will not buy the same device again after an error (leading to a loss of custom for the manufacturer).

II.C.2 The optimal sample size, according to a value of information analysis

The optimal sample size is that which maximizes the difference between the value of the information and the cost of obtaining that information. The testing must be continued until the cost of running an extra test exceeds the value of information expected to be provided by adding more participants. This is equivalent to maximizing the expected net benefit of sampling (ENBS, i.e. the difference between the EVSI and the costs of increasing the sample size), which means finding the highest number of participants n for which the marginal gain $ENBS(n) - ENBS(n-1) > 0$.

Since the extended discovery matrix $\mathbb{d}'$ (that obtained after adding n' participants to the initial sample of size n) is not available at this stage, the computation of $\mathcal{L}'$ requires the use of a Bayesian tool called pre-posterior analysis. This technique is based on the classical posterior analysis in which the prior distribution is updated with sampled data to give the parameter's

posterior distribution. However, the pre-posterior analysis deals with the fact that the data (i.e. the extended discovery matrix $\mathbb{d}'$) has not been observed. It usefully solves this issue by predicting the likelihood function (the distribution of the sampled data) that is conditional on the prior data. The mathematics underlying the pre-posterior analysis are detailed in Appendix 4. The value of information that can be expected by adding n' new participants is:

$$EVSI(n') = \mathcal{L} - \mathcal{L}'$$

and the corresponding $ENBS$ is:

$$ENBS(n') = EVSI(n') - n' \times C_v$$

where C_v is the variable cost of including a new participant in the usability test (fixed costs were already incurred at the beginning of the trial and are considered as sunk costs). Indeed, maximizing the $ENBS$ is the same as finding the highest value of n' for which the marginal gain $ENBS(n') - ENBS(n' - 1) > 0$. Beyond this value of n', the cost of including new participants is greater than the size of the reduction in the expected loss. A simulation approach was used to compute $\mathcal{L}'$ and maximize $ENBS(n')$ for a wide range of possible n', and a moving average was used to smooth the simulated $ENBS$. The optimal sample size is denoted as $n^\star = n + n'$.

II.C.3 A case study: the needle-free adrenaline auto-injector

The usability of the needle-free adrenaline auto-injector was tested as part of a wider research project presented in Box 1. In the case study, we modelled the consequences as follows. We considered that an acceptable error would prompt the end user to ask the manufacturer for a reimbursement – meaning that the manufacturer will make no profit. An unacceptable error would generate failure costs in terms of product recall, product reengineering, and temporary loss of market access. Nevertheless, we are well aware that estimating the costs of undetected use error is complex and normally requires a sophisticated risk analysis. This task is the responsibility of the manufacturer, using the information (most of which is confidential) at its disposal. Even so, the scenario described in the case study is based on the recall of a similar medical device, and we consider it to be relevant for our present purposes. In line with the current regulatory process and in order to ensure reliability, $n = 20$ participants were included before the value of information was first estimated. Next, participants were added in blocks of

ten; which number corresponds approximately to a full day of usability testing. We considered that all the errors observed with $n = 20$ participants were acceptable; otherwise, the medical device would have had to return to the redesign/formative testing stage. All the statistical routines in this paper have been implemented in an R package called useval, which is available on GitHub.

Box 2: Context of the usability testing of the needle-free adrenaline auto-injector.

The usability of the Zeneo® needle-free adrenaline auto-injector (Crossject, Dijon) was tested as part of the USEVAL project. The research project was designed to establish the critical methodological choices for the usability validation of medical devices. The project focused on the added value of variations in ecological validity (the fidelity of the testing environment, the group of end users, etc.) and in specific cultural features (different countries). Two European usability centers for health technologies (Lille University Hospital, Lille, France; Geneva University Hospital, Geneva, Switzerland) designed and carried out usability tests in collaboration with device manufacturers. In line with the applicable legislation in each country, the protocols were reviewed and authorized by national investigational review boards (reference for France: CPP Est I 2017-A02847-46; reference for Switzerland: LRH RS 810.30). During the project, we collected micro-costing data that reflected actual expenditure. These included the cost of drafting the protocol by human factors engineers and statisticians, and the costs of usability testing (the fee for the actress simulating an anaphylactic shock, the rental of the premises, the cost of the device, etc.).

The needle-free adrenaline auto-injector was tested by 140 participants (80 in Switzerland and 60 in France). This testing enabled the detection of 39 acceptable errors and 0 unacceptable errors. The sample size of 140 was unusually high for usability testing but was chosen because we included groups that differed with regard to the fidelity of the testing environment (high vs. low fidelity), end user experience (inexperienced vs. highly experienced), and country of residence (France vs. Switzerland). The groups' composition and characteristics will not be considered in the subsequent analysis. All subsamples were

pooled without regard to the participants' characteristics. The observations were then randomly ordered and sequentially integrated into our computations.

II.D Results

II.D.1 Data for the value-of-information analysis

To carry out the value-of-information analysis, we used data from different sources; this included validation testing and micro-costing (Table 1). Two figures were estimated: (i) the profit that would be lost in the event of market suspension, and (ii) the number of times the device will be used by end users (corresponding to the number of times that use errors can potentially be experienced). The manufacturer applies for the CE mark, which grant access to a market of 513 million people in the European Union.

Table 1: Data for the value-of-information analysis

Data	Value	Source
Epidemiology of anaphylactic shock		
Population of the European Union (2019)	513,471,676	Eurostat
Prevalence of at-risk people	3/1000	Panesar et al.(Panesar et al. 2013)
Annual incidence of anaphylactic shock	7.9/100,000	Helbing et al., Abi Khalil et al.(Abi Khalil, Damak, and Décosterd 2014; Helbling et al. 2004)
Medical device		
Unit selling price	€50	Calculated based on assumptions (see text)
Gross margin rate	55%	
Market share	10%	
At-risk persons carrying a device	75%	
Usability testing		
Fixed costs	€30,000	Micro-costing
Variable costs (per participant)	€800	Micro-costing
Opportunity cost of undetected use errors		
Acceptable	€27.5	Calculated based on assumptions (see text)
Unacceptable	€4,865,659	

The profit was based on the expected sales volume. By assuming that the prevalence of at-risk people was 3 per 1000,(Panesar et al. 2013) we expect to have a target population of 1,540,415 people - 75% of whom will buy an adrenaline injector. Given an expected market share of 10% (according to the manufacturer's financial report) and an expiration date of 2 years, we estimated that 57,766 people per year will buy the device (Figure 1). As a general rule that applies to all the currently marketed adrenaline injectors, the health authorities recommend that a person carries

two devices (in case one device malfunctions). Thus, we estimated that 115,531 devices will be sold each year.

The number of times the device will be used by end users was estimated through the incidence rate for anaphylactic shock requiring an adrenaline injection (7.9 per 100000 person-years),(Abi Khalil, Damak, and Décosterd 2014; Helbling et al. 2004) which corresponds to an annual incidence of 40,565. After accounting for the expected market share, we estimated that the auto-injector would be used by 4057 persons with anaphylactic shock per year (Figure 1).

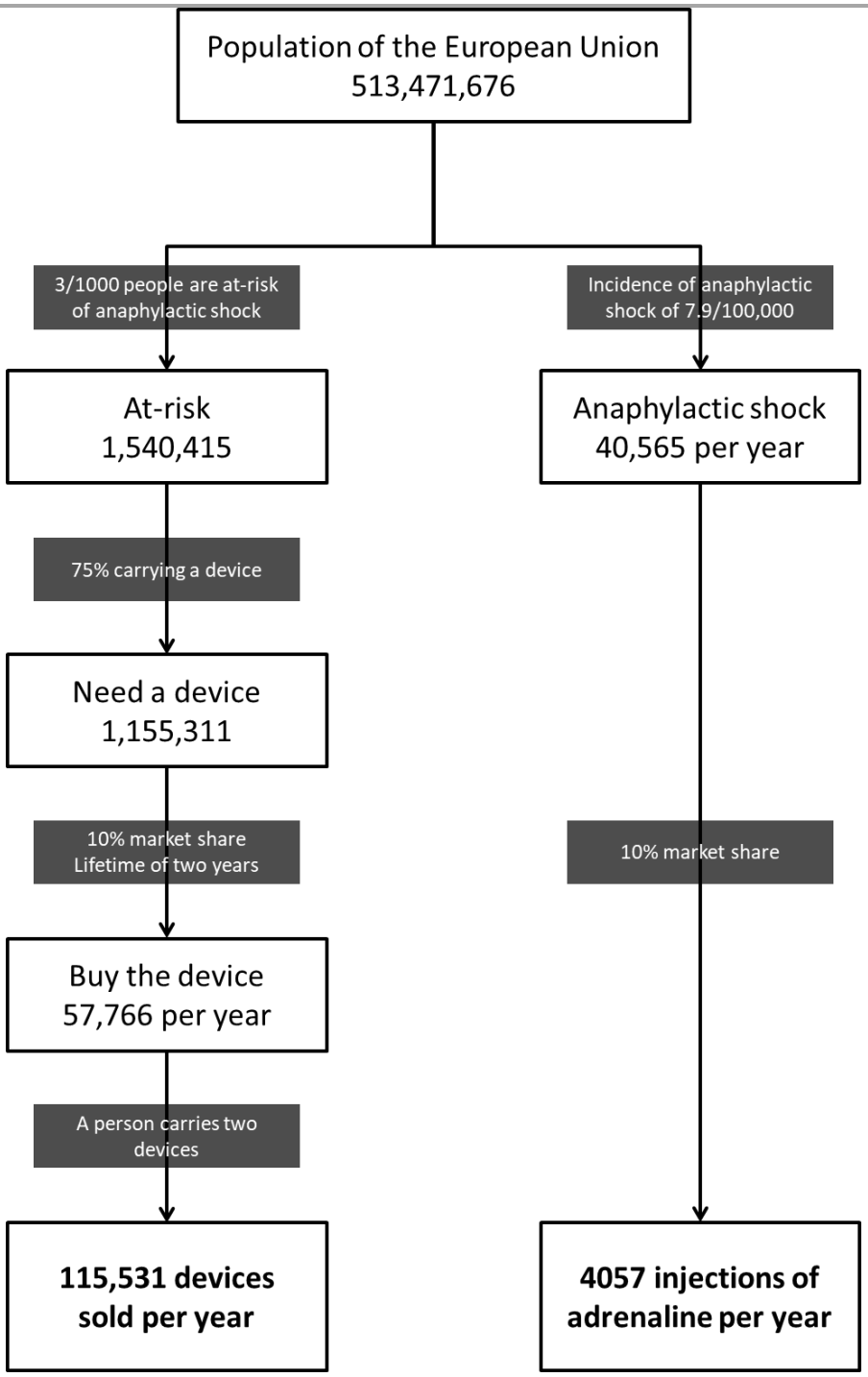


Figure 1: Flow chart illustrating the steps in the estimation of the number of devices to be sold (left column) and the number of times the device is used (right column). The 115,531 devices sold each year will determine the manufacturer's annual profit and (if unacceptable errors lead to market suspension) potential losses. The number of times the device will actually be used (4057 a year) will determine the number of errors experienced by end users.

As detailed in the Methods section, the discovery of a previously undetected acceptable error by an end user will cost the manufacturer its profit. This profit is generated by the expected annual sale of 115,351 devices. Given a market price of €50 and a gross margin rate of 55% - a typical rate in the medical device sector, the opportunity cost was:

$$c_+ = €50 \times 0.55 = €27.50$$

If an unacceptable error is discovered by an end user, we expect to see an opportunity cost equivalent to 18 months of market suspension plus the re-engineering costs (assumed to be €100,000). Our simulated 18-month market suspension was based on the temporary withdrawal of a similar medical device: an auto-injector with needle commercialized by one of Crossject's competitors (a significant proportion of pens required greater than normal force for activation, with the possible life-threatening hindrance of adrenaline injection). The cost of at least one undetected unacceptable error discovered by the end users is therefore:

$$c_- = 115{,}531 \times €27.50 \times 1.5 + 100{,}000 = €4{,}865{,}659$$

We considered that the testing's conclusion is valid for 5 years but that after that time, the manufacturer might have updated and replaced the current device, or the person might have purchased a competing product. With this lifetime of 5 years, the population size is $N = 4057 \times 5 = 20{,}285$ end users.

II.D.2 Sample size estimation with the first 20 participants

In the following section, we describe the two steps in the value-of-information analysis used to estimate the sample size: (i) computation of the probability of undetected use errors and deduction of the expected loss in the absence of further testing, and (ii) the pre-posterior analysis used to estimate the ENBS. In line with the early control strategy recommended by the regulators, we started the sample size estimation after including the first $n = 20$ participants. At this stage, 26 use errors had been discovered; none of these were deemed to be unacceptable.

In the first step, we estimated the probability of undiscovered use errors (see the heatmap in Figure 1) and then deduced what the loss would be in the absence of further testing. On one hand, all the acceptable errors were unlikely to have been detected by the first 20 participants

(<10%). Indeed, we estimated that there was an 88% chance that the total number of acceptable errors (represented on the x axis of the heatmap) was between 27 and 36. On the other hand, the probability of an unacceptable error remaining undetected was low (6.4%). As a result, the expected loss in the absence of further usability testing was estimated to be €390'815.

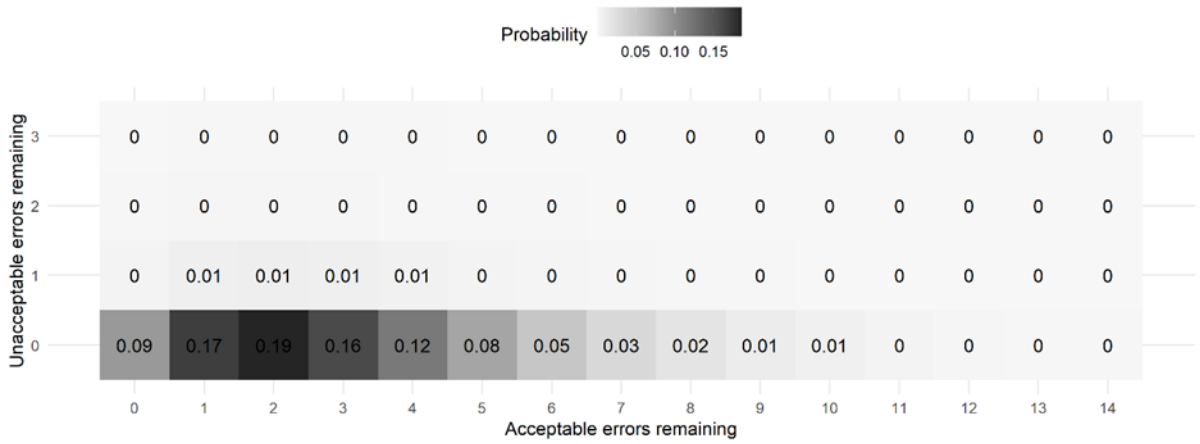


Figure 2: Heatmap of the posterior probability of undetected use errors. Acceptable errors are shown on the x-axis, and unacceptable errors are shown on the y-axis. Each rectangle in the grid corresponds to the probability of observing a given combination of acceptable and unacceptable errors (rounded up to the second decimal). The probability is also represented on a grey scale, with dark grey corresponding to a higher probability. In this example, the darker bottom row indicates that the probability of an unacceptable error remaining undetected is low. Conversely, it is likely that acceptable errors are still to be discovered. As shown in the bottom left corner of the heatmap, the probability whereby all errors have already been detected is low (9%).

In a second step, we used a pre-posterior analysis to estimate the ENBS for sample sizes ranging from 30 to 120 participants. For each sample size, we computed the ENBS as the difference between the EVSI and the cost of recruiting new participants (Figure 2). The maximum ENBS was achieved for a total sample size of $n^{\star} = 100$ participants, corresponding to the recruitment of

n'=80 additional participants. Beyond this sample size, the cost of recruiting new participants is expected to outweigh the gain in the value of information (i.e. decreasing the ENBS).

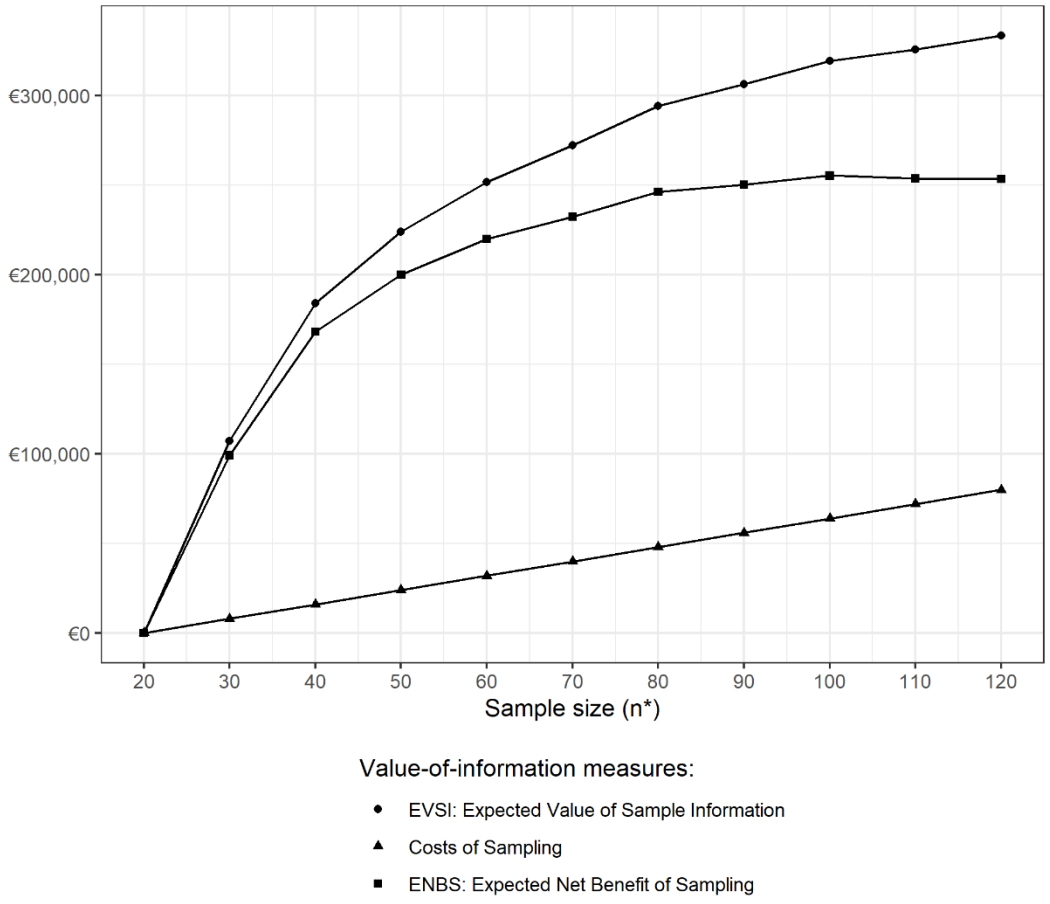


Figure 3: The expected net benefit of sampling (ENBS, ■), defined as the difference between the expected value of sample information (EVSI, ●) and the cost of adding new participants (cost of sampling, ▲). At this stage, $n = 20$ participants have been included, and the total sample size $n^{\star}$ is represented on the x axis. The optimal sample size is estimated to be $n^{\star} = 100$, meaning that another $n' = 80$ participants should be recruited. The corresponding ENBS (€255'453) is obtained by computing the difference between the EVSI (€319'453) and the cost of testing new participants (€64'000). The EVSI is the difference between the expected loss without further information (€390'815), and the expected loss with the additional information provided by further testing (€71'362).

II.D.3 Adaptative estimation of the sample size

Our first estimation of the sample size was uncertain, as illustrated by the similar $ENBS$ for values ranging from $n^{\star} = 80$ to $n^{\star} = 120$. As the number of participants increased, the estimation of the expected loss became more accurate. An adaptive strategy enables the

discovery process to be monitored and the testing to be stopped when appropriate. Indeed, we incremented the sample size by 10 participants until the $ENBS$ became negative for any sample size, i.e. when the cost of adding more participants outweighed the expected value of information. As the number of participants increases, the estimated optimal sample size ranged between $n^\star = 80$ and $n^\star = 100$. These changes are explained by (i) the pattern of detection of new participants, and (ii) the reduction in uncertainty, leading to a convergence in the estimation. Hence, the study should be stopped when the sample size is $n = 90$ participants. At this stage, $j_+ = 36$ acceptable errors and $j_- = 0$ unacceptable errors were discovered. The maximum $ENBS$ is reached for $n' = 0$ because the cost of acquiring more information outweighed the expected value (in term of loss reduction). Indeed, the ENBS values computed for $n^\star \geq 100$ are negative.

A sensitivity analysis was also carried out to assess the effect of decreased/increased costs of unacceptable use errors. The optimal sample size was $n^\star = 80$ when dividing the cost by 2 and $n^\star = 130$ when doubling it. Lastly, we estimated the sample size by using conventional methods: the naïve, Good Turing and double deflation methods (Vandewalle et al. 2020). The three methods estimated that the error discovery process was complete after the inclusion of 20 participants - illustrating the statistical weakness of these approaches. Conversely, the logit-normal binomial zero-truncated model estimated that the total number of errors ranged from 33 to 49 for sample sizes from 20 to 90. However, the use of logit-normal binomial zero-truncated model would always prompt the recruitment of more participants because (i) use errors would always remain and (ii) the costs and benefits of further recruitment are not taken into account.

II.E Discussion

Here, we described a method for estimating the sample size of usability studies carried out before medical devices are granted market access. The framework is based on a value-of-information analysis, and we adopted the manufacturer's perspective. The full Bayesian model and the decision theory approach overcome the two main hurdles observed to date, namely the limitations of conventional statistical methods and the lack of a clear decision-making perspective. We illustrated our method for sample size estimation by application to real usability testing of a new, needle-free adrenaline auto-injector.

II.E.1 Value-of-information analysis in health technology assessment

To the best of our knowledge, the present work is the first to have used a value-of-information approach to determine the optimal sample size for usability testing of medical device. The underlying conceptual framework has been widely used in health technology assessment to formalize decision problems under uncertainty and to inform adoption and research decisions (Eckermann and Willan 2007b; Claxton et al. 2004). The optimal sample size is computed by maximizing the difference between the $EVSI$ and the cost of obtaining that information. In this framework, the decision is based on the consequences of insufficient pre-market research rather than an arbitrary threshold, such as type 1 and 2 errors (Willan and Pinto 2005). Given the higher level of uncertainty observed in the evaluation of medical devices, the use of a value-of-information framework is particularly appropriate (Rothery et al. 2017). It allows the manufacturer to formally characterize decision uncertainty and the trade-off between the value of early market approval and the benefit of further evaluation (Eckermann and Willan 2008a). Furthermore, the manufacturer can refine the analysis by using other parameters at its disposal, such as the expected market share and the time horizon. For example, we used a 5-year horizon (because a different product is likely to emerge after that length of time) and a 10% market share (the target in the manufacturer's business plan, given the existence of well-established competitors). The potential use of these variables illustrates how a value-of-information framework can be integrated into the manufacturer's strategic planning.

II.E.2 Meeting the regulator's objective of patient safety

Here, we argued that the methods currently applied by manufacturers do not guarantee the patient's safety. Indeed, the recommended completeness threshold is often reached with as few as 15 to 20 participants (Schmettow, Vos, and Schraagen 2013b; Schmettow 2008). However, the severity of the still-undiscovered errors are not embedded in the decision rule – even though the data are available. According to human factors guidelines, the acceptability of use error must always be assessed. This acceptability is based on several dimensions: the severity of the use error, the error's probability of occurrence, the trade-off between the potential harm and the benefit provided by the medical device, the underlying condition of use, etc. (ANSI 2009). For

sake of clarity, we presented a simple alternative that reflects the main decision facing the manufacturer: to accept the use error or not. Each type of error was associated with an opportunity cost for the manufacturer, corresponding to the loss incurred if the error is discovered by end users. We were thus able to embed the patient safety objective into the evaluation process in a much more satisfactory way, by linking it to the manufacturer's profit. It is noteworthy that although only two opportunity costs were used to feed the loss function, the probability of occurrence was embedded into the model. Thus, the heterogeneity in the probability of error occurrence meant that the consequences were also variable - leading to a broad range of possible losses for the manufacturer.

II.E.3 Implementation of the value-of-information framework within current usability practices

In the present study, we estimated the optimal sample size for the first usability test of a medical device. This so-called "early control" strategy was chosen because the estimate is based on the best available data. This strategy is also in line with current usability standards (Sauro and Lewis 2016; Borsci et al. 2014), which we hope will facilitate its uptake. Indeed, the widespread adoption of this framework would be beneficial for both manufacturers and regulators. The advantages for the manufacturer are obvious, as long as the utility function accurately reflects its perspective. Moreover, the method is more in line with the patient safety goal pursued by the regulator. This is particularly true for critical, unacceptable errors. The persistence of these errors is largely prevented by the dramatic economic impact on the manufacturer. Although we cannot prove that the implementation of this framework will result in socially optimal allocation of the resources, it certainly gets us closer to that goal in the present regulatory context. The medical device's risk-benefit ratio will be appraised differently by the manufacturer and the regulator. The extent to which our method might improve patient safety is an important area for further research.

II.E.4 The manufacturer's perspective

We chose to adopt the manufacturer's standpoint because the current regulatory paradigm places the responsibility for patient safety on the manufacturer. This paradigm has been carried over from the industrial roots of medical device regulation. Alternatively, we could have adopted

the regulator's perspective, as is common in other fields. In such a case, the net benefit framework is usually assumed to be the most relevant utility function. In our framework, (i) the opportunity loss would be valued in terms of the net benefit forgone for the population of end users, and (ii) the optimal sample size would maximize the expected net benefit (minus the cost of further research), computed as the difference between the benefit in terms of health outcomes (e.g. survival and quality of life) and the costs (e.g. resuscitation care and long-term sequelae). The health benefits could be modeled as a function of the probability of residual errors. For instance, a fatal use error would be valued by converting the forgone lifetime (unadjusted or quality-adjusted life years) into monetary units with an appropriate willingness to pay and by weighting the value by the probability of occurrence given by the pre-posterior analysis. The probability of this loss would decrease as the sample size increases, as observed in this manuscript. Thus, adoption of the regulator's perspective is possible but would require a more complex model and the implementation of several assumptions (such as the relationship between use errors and net benefit) and definitions (such as willingness to pay for the benefit obtained).

II.E.5 Main limitations

The first limitation of the present work relates to the determination of the opportunity cost, which is a cornerstone of our framework. The values presented in the case study are primarily illustrative. We used an example with a real device as a proxy for the opportunity costs and performed a sensitivity analysis for a wide range of possible costs, in order to make the example as realistic as possible. However, the opportunity costs should be computed by the manufacturer on the basis of a more in-depth risk analysis. Furthermore, this assessment requires confidential corporate information and is specific to the device under evaluation. Nevertheless, we consider that given the potential consequences of a recall (re-engineering costs, loss of turnover, legal action, reputational damage, and loss of customer trust), the manufacturer must have this kind of expert assessment at its disposal (Bias and Mayhew 2005; Slawomirski, Auraaen, and Klazinga 2017). The methods used to analyze risk are beyond of the scope of the present work, and the reader is referred to human factors textbooks for more details (ANSI 2009). A second limitation relates to our assumption that the decision maker was risk-neutral. This is a common assumption

in the value-of-information literature because most researchers adopt a societal or healthcare system perspective. Although a risk-neutral perspective might be relevant for manufacturers with a large product portfolio, it is probably not relevant for the SMEs that comprise a high proportion of medical device manufacturers. However, the risk-neutral utility function could be replaced by a constant relative (or absolute) risk-aversion utility function with only marginal changes to our framework. A third potential limitation relates to the decision rule. We assumed that the manufacturer seeks to maximize profit. This objective function is intuitive and is commonly used to describe a producer's economic behavior. However, there are other options: "satisficing", for example, which assumed bounded rationality rather than substantive rationality (Simon 1956).

II.E.6 Conclusion

In conclusion, we developed a framework for optimal sample size determination from the manufacturer's perspective via the application of a value-of-information analysis. We then applied the framework to actual usability testing of a needle-free adrenaline auto-injector. Based on our results, our method appears to be highly suitable for estimating the sample size in usability testing before medical devices are granted market access.

Partie 4. CONCLUSIONS GENERALES

Dans cette thèse, nous avons montré que les méthodes fondées sur la valeur de l'information pouvaient être utiles aux prises de décisions tout au long de la vie du dispositif médical en (i) caractérisant l'incertitude décisionnelle et (ii) déterminant l'opportunité et la valeur attendue de l'acquisition de données supplémentaires. Nous avons appliqué ces méthodes à deux temps de l'évaluation du dispositif. En amont de la mise sur le marché, nous avons montré que l'implémentation d'une évaluation médico-économique précoce, incluant la mesure de la valeur de l'information, était de nature à permettre la justification du nombre de sujets nécessaires dans les études d'utilisabilité imposées par la législation européenne en vue de l'apposition du marquage CE. Lors de l'évaluation pour inscription au remboursement, nous avons montré l'apport des analyses de la valeur de l'information pour la détermination des études post-inscriptions dans le contexte décisionnel français.

Les travaux menés durant cette thèse ont montré le rôle central de la caractérisation de l'incertitude. Ainsi, dans le cadre des études d'utilisabilité, une revue de la littérature préalable concernant les méthodes existantes pour le calcul du nombre de sujets nécessaires a mis en exergue la faiblesse statistique des modèles d'aide à la décision existants. Par conséquent, un calcul de la valeur de l'information utilisant ces modèles s'est rapidement avéré impossible, eu égard à une caractérisation de l'incertitude insatisfaisante. Il a donc été nécessaire de redévelopper le modèle sous-jacent pour répondre aux prérequis de l'analyse de la valeur de l'information. De même, le modèle d'aide à la décision utilisé par le NICE pour comparer les endoprothèses aortiques à la chirurgie ouverte a dû être largement réécrit pour permettre une caractérisation de l'incertitude répondant aux exigences de calcul de la valeur de l'information. Ainsi, la modélisation des corrélations entre les caractéristiques des patients a nécessité l'ajout des matrices de Cholesky et l'étude du coût des interventions a nécessité le calcul de l'incertitude entourant des interventions multi-GHM. Enfin, l'hétérogénéité de la population a dû être intégrée au modèle pour modéliser les résultats distincts dans les sous-groupes. Au total, les exigences méthodologiques pour la mise en œuvre des analyses fondées sur la valeur de l'information sont fortes. La structure du modèle sous-jacent est réinterrogée et une attention

particulière est nécessaire quant à la qualité de la caractérisation de l'incertitude qui constitue un élément fondamental des analyses fondées sur la valeur de l'information.

Durant cette thèse, nous avons eu l'occasion de présenter les méthodes fondées sur la valeur de l'information à de nombreux acteurs issus de champs variés : industriels, académiques, régulateurs, etc. Nous avons ainsi pu constater une appropriation plus délicate par les acteurs non-économistes. En effet, le corpus des méthodes fondées sur la valeur de l'information a été proposé dans le contexte de l'évaluation médico-économique et semble bien appréhendé par les acteurs de ce champ. Pour les autres, l'implémentation dans le processus de décision se confronte à des obstacles de nature variée. Il s'agit tout d'abord de la compréhension des concepts de la théorie de la décision, et notamment l'approche bayésienne, qui se heurte au paradigme de l'analyse fréquentiste prévalent en santé. On observe par exemple des références systématiques au risque de première espèce habituellement utilisé pour l'interprétation des résultats d'essais. De la même manière, la formalisation d'un critère de décision objectif unique synthétisant l'ensemble des perspectives n'est pas chose aisée. Ici, nous avons utilisé la flexibilité des méthodes fondées sur la valeur de l'information pour proposer différents critères de jugement, selon que le décideur soit le fabriquant (profit) ou le régulateur (efficacité et efficience), en nous appuyant sur un modèle d'aide à la décision permettant d'embarquer l'ensemble des perspectives. Bien que nous ayons obtenu des résultats concordants pour les endoprothèse aortiques, des analyses de scenarii menées sur les données anglaises ont révélé l'existence de discordances susceptibles entre les critères d'efficacité et d'efficience. Une telle situation nécessiterait d'établir une méthodologie d'arbitrage non envisagée dans ce manuscrit. Du point de vue des acteurs issus du champ de l'évaluation médico-économique, les méthodes étaient bien appréhendées mais l'interprétation des résultats demeurait difficile dans le contexte français. Cela est principalement lié à l'absence d'une propension à payer pour un résultat de santé qui constitue un des freins majeurs à l'implémentation de ce type de méthodes. En effet, la valeur de l'information peut varier de manière non monotone en fonction de la propension à payer et des valeurs très différentes peuvent être obtenues selon la propension retenue. Ce problème récurrent est de nouveau à envisager dans un environnement plus large qui est celui de la doctrine de l'évaluation médico-économique en France. Enfin, certains décideurs,

principalement du côté des fabricants mais également du côté des autorités de santé, expriment une aversion pour le risque. Lorsqu'une telle attitude vis-à-vis du risque est identifiée, son impact sur les estimations de la valeur de l'information devrait être quantifié. Il est utile de rappeler que ce problème n'est pas spécifique et concerne de façon générale l'estimation de l'utilité espérée. Cependant, et bien que la littérature soit relativement abondante sur le sujet, les applications pratiques restent rares. Dans le cas spécifique du calcul la valeur de l'information, certains travaux proposent de prendre en compte cet aspect, notamment à travers des accords négociés de mise sur le marché tels que le paiement à la performance ou les règles de partage de risque entre le régulateur et le fabricant (Grimm, Dixon, and Stevens 2016; Grimm et al. 2017; Garrison et al. 2013).

Les méthodes fondées sur la valeur de l'information ont connu un développement considérable durant la dernière décennie aboutissant à la publication de recommandations internationales sur leur utilisation (Fenwick et al. 2020; Rothery et al. 2020). On retrouve en leur sein les principales décisions pouvant être appuyées par ce type d'analyses : la priorisation de la recherche, les choix de design de la recherche, le remboursement, et la prise de décision efficace tout au long du cycle de vie d'un produit de santé. Les neuf recommandations de bonnes pratiques contenues dans ce rapport appuient de nombreux enseignements issus de nos travaux. Ainsi, les deux premières portent sur les modalités de caractérisation de l'incertitude, notamment l'importance de considérer l'ensemble des paramètres simultanément. Au cours de cette thèse, un effort important a été déployé pour garantir une caractérisation précise de l'incertitude. Dans ce cadre, les résultats obtenus aux deux temps de l'évaluation des dispositifs démontrent à la fois l'importance, mais également la difficulté d'opérationnaliser une telle exigence. Les recommandations 3 et 4 portent essentiellement sur la manière de justifier certains choix méthodologiques tels que l'horizon temporel des conclusions et la taille de la population. De telles recommandations sont essentielles car elles constituent un cadre de référence permettant d'homogénéiser les pratiques. Enfin, les quatre dernières recommandations détaillent les conclusions pouvant être tirées des principales métriques que sont l'EVPI, l'EVPPI, l'EVSI et l'ENBS. Dans ce cadre, les rôles centraux de l'EVPPI pour la priorisation des efforts de recherche et de l'ENBS pour la détermination du design optimal sont soulignés.

D'un point de vue méthodologique, plusieurs freins à la mise en œuvre ont été levés durant la dernière décennie. Tout d'abord, l'émergence de techniques de simplifications calculatoires facilite considérablement le calcul de l'EVPPI et de l'EVSI (Oakley et al. 2010; Strong and Oakley 2013; Strong, Oakley, and Brennan 2014b; Strong et al. 2015; Heath, Manolopoulou, and Baio 2017a; Heath and Baio 2018; Heath, Manolopoulou, and Baio 2018b; Heath, Manolopoulou, and Baio 2019). Ensuite, les recommandations harmonisent les hypothèses et les choix de modélisation susceptibles de modifier les résultats de l'analyse tels que la détermination de l'horizon temporel, le calcul de la taille de la population cible, l'actualisation, etc. Néanmoins, malgré l'existence d'un corpus théorique d'une densité remarquable, les applications des analyses de la valeur de l'information restent encore limitées, notamment le calcul des quantités les plus complexes telle que l'EVSI et l'ENBS (Steuten et al. 2013). Le niveau de compétence en termes de modélisation statistique, de calcul mathématique et de programmation explique probablement en partie ce constat. La mise à disposition de calculateurs en ligne gratuit (ex : SAVI - Sheffield Accelerated Value of Information), de packages R (BCEA, SAVI, EBASS, etc.) et le développement d'un centre de ressources par le Collaborative Network for Value of Information Analysis (https://www.convoi-group.org/resources/) sont autant d'initiatives de nature à lever les obstacles techniques résiduels et à favoriser une diffusion plus large des méthodes.

En conclusion, les résultats de cette thèse offrent une perspective encourageante quant à l'opportunité d'implémenter les méthodes fondées sur la valeur de l'information dans le contexte français. Sur le modèle anglais ou américain, des groupes de travail associant les parties prenantes du processus décisionnel (CEESP, CNEDiMTS, CEPS, Industriel, etc.) pourraient être mobilisés afin d'évaluer la faisabilité et l'apport des analyses de la valeur de l'information dans le processus décisionnel. Dans ce cadre, des exemples tels que ceux présentés dans ce manuscrit pourraient servir de support de réflexion aux groupes de travail. Le résultat d'une telle expérimentation pourrait aboutir à la publication de recommandations concernant leur utilisation dans le contexte français, encore inexistantes à ce jour.

Partie 5. REFERENCES

Abi Khalil, Muriel, Hassen Damak, and Dumeng Décosterd. 2014. 'Anaphylaxie et état de choc anaphylactique', *Rev Med Suisse*, 10: 1511-5.

Ades, A. E., G. Lu, and K. Claxton. 2004. 'Expected value of sample information calculations in medical decision modeling', *Medical Decision Making*, 24: 207-27.

Amram, Martha, and Nalin Kulatilaka. 1998. 'Real options:: Managing strategic investment in an uncertain world', *OUP Catalogue*.

Andronis, L., and P. M. Barton. 2016a. 'Adjusting Estimates of the Expected Value of Information for Implementation: Theoretical Framework and Practical Application', *Medical Decision Making*, 36: 296-307.

———. 2016b. 'Adjusting Expected Value of Sample Information Using Realistic Expectations around Implementation', *Med Decis Making*, 36: 284.

ANSI. 2009. 'ANSI/AAMI HE75-2009: human factors Engineering—design of medical devices', *Arlington, VA: Association for the Advancement of Medical Instrumentation*.

Arrow, Kenneth J, and Anthony C Fisher. 1974. 'Environmental preservation, uncertainty, and irreversibility.' in, *Classic papers in natural resource economics* (Springer).

Arrow, Kenneth Joseph. 1965. *Aspects of the theory of risk-bearing* (Helsinki: Yrjö Jahnsonian Sä¨tiö).

Bach, Cedric, and Dominique L Scapin. 2010. 'Comparing inspections and user testing for the evaluation of virtual environments', *Intl. Journal of human–computer interaction*, 26: 786-824.

Bader, Clément, Sébastien Cossin, Aline Maillard, and Antoine Bénard. 2018. 'A new approach for sample size calculation in cost-effectiveness studies based on value of information', *BMC Medical Research Methodology*, 18: 113.

Basu, Anirban, Josh J Carlson, and David L Veenstra. 2016. 'A framework for prioritizing research investments in precision medicine', *Medical Decision Making*, 36: 567-80.

Basu, Anirban, and David Meltzer. 2007. 'Value of Information on Preference Heterogeneity and Individualized Care', *Medical Decision Making*, 27: 112-27.

Bernoulli, Daniel. 1738. 'Exposition of a new theory on the measurement', *Econometrica*, 22: 23-36.

Bias, Randolph G, and Deborah J Mayhew. 2005. *Cost-justifying usability: An update for the Internet age* (Elsevier).

Biebl, M., A. G. Hakaim, B. Hugl, W. A. Oldenburg, R. Paz-Fumagalli, and J. M. McKinney. 2005. 'Endovascular aortic aneurysm repair with the Zenith AAA Endovascular Graft: does gender affect procedural success, postoperative morbidity, or early survival?', 71: 1001-08.

Bilcke, Joke, Philippe Beutels, Marc Brisson, and Mark Jit. 2011. 'Accounting for Methodological, Structural, and Parameter Uncertainty in Decision-Analytic Models: A Practical Guide', *Medical Decision Making*, 31: 675-92.

Blankensteijn, J. D., S. E. C. de Jong, M. Prinssen, A. C. van der Ham, J. Buth, and S. M. M. van Sterkenburg. 2005. 'Two-year outcomes after conventional or endovascular repair of abdominal aortic aneurysms', 352: 2398-405.

Blankensteijn, J. D., S. E. de Jong, M. Prinssen, A. C. van der Ham, J. Buth, S. M. van Sterkenburg, H. J. Verhagen, E. Buskens, D. E. Grobbee, and Group Dutch Randomized Endovascular Aneurysm Management Trial. 2005. 'Two-year outcomes after conventional or endovascular repair of abdominal aortic aneurysms', *N Engl J Med*, 352: 2398-405.

Borsci, S., A. Londei, and S. Federici. 2011. 'The Bootstrap Discovery Behaviour (BDB): a new outlook on usability evaluation', *Cogn Process*, 12: 23-31.

Borsci, S., R. D. Macredie, J. L. Martin, and T. Young. 2014. 'How many testers are needed to assure the usability of medical devices?', *Expert Rev Med Devices*, 11: 513-25.

Borsci, Simone, Robert D Macredie, Julie Barnett, Jennifer Martin, Jasna Kuljis, and Terry Young. 2013. 'Reviewing and extending the five-user assumption: a grounded procedure for interaction evaluation', *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20: 1-23.

Brennan, Alan, Samer Kharroubi, Anthony O'hagan, and Jim Chilcott. 2007. 'Calculating partial expected value of perfect information via Monte Carlo sampling algorithms', *Medical Decision Making*, 27: 448-70.

Brewster, D. C., J. E. Jones, T. K. Chung, G. M. Lamuraglia, C. J. Kwolek, and M. T. Watkins. 2006. 'Long-term outcomes after endovascular abdominal aortic aneurysm repair – the first decade', 244: 426-38.

Briggs, Andrew H., Milton C. Weinstein, Elisabeth A. L. Fenwick, Jonathan Karnon, Mark J. Sculpher, and A. David Paltiel. 2012. 'Model Parameter Estimation and Uncertainty Analysis: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–6', *Medical Decision Making*, 32: 722-32.

Briggs, Andrew, Mark Sculpher, and Karl Claxton. 2006. *Decision modelling for health economic evaluation* (Oup Oxford).

Brown, L. C., R. M. Greenhalgh, S. Howell, J. T. Powell, and S. G. Thompson. 2007. 'Patient fitness and survival after abdominal aortic aneurysm repair in patients from the UK EVAR trials', *Br J Surg*, 94: 709-16.

Brush, J., K. Boyd, F. Chappell, F. Crawford, M. Dozier, E. Fenwick, J. Glanville, H. McIntosh, A. Renehan, D. Weller, and M. Dunlop. 2011. 'The value of FDG positron emission tomography/computerised tomography (PET/CT) in pre-operative staging of colorectal cancer: a systematic review and economic evaluation', *Health Technol Assess*, 15: 1-192, iii-iv.

Buth, J., C. J. van Marrewijk, P. L. Harris, W. C. J. Hop, V. Riambau, and R. J. F. Laheij. 2002. 'Outcome of endovascular abdominal aortic aneurysm repair in patients with conditions considered unfit for an open procedure: a report on the EUROSTAR experience', 35: 211-19.

Campbell, G. 2008. 'Statistics in the world of medical devices: the contrast with pharmaceuticals', *J Biopharm Stat*, 18: 4-19.

———. 2011. 'Bayesian statistics in medical devices: innovation sparked by the FDA', *J Biopharm Stat*, 21: 871-87.

Carlson, J. J., R. Thariani, J. Roth, J. Gralow, N. L. Henry, L. Esmail, P. Deverka, S. D. Ramsey, L. Baker, and D. L. Veenstra. 2013. 'Value-of-Information Analysis within a Stakeholder-Driven Research Prioritization Process in a US Setting: An Application in Cancer Genomics', *Medical Decision Making*, 33: 463-71.

Carroll, Raymond J, and F Lombard. 1985. 'A note on N estimators for the binomial distribution', *Journal of the American Statistical Association*, 80: 423-26.

Caulton, David A. 2001. 'Relaxing the homogeneity assumption in usability testing', *Behaviour & Information Technology*, 20: 1-7.

Chalkidou, K., J. Lord, A. Fischer, and P. Littlejohns. 2008. 'Evidence-based decision making: when should we wait for more information?', *Health Aff (Millwood)*, 27: 1642-53.

Chambers, D., D. Epstein, S. Walker, D. Fayter, F. Paton, K. Wright, J. Michaels, S. Thomas, M. Sculpher, and N. Woolacott. 2009. 'Endovascular stents for abdominal aortic aneurysms: a systematic review and economic model', *Health Technol Assess*, 13: 1-189, 215-318, iii.

Chen, M. H., and A. R. Willan. 2013. 'Determining optimal sample sizes for multistage adaptive randomized clinical trials from an industry perspective using value of information methods', *Clinical Trials*, 10: 54-62.

Chen, Y. F., J. Madan, N. Welton, I. Yahaya, P. Aveyard, L. Bauld, D. Wang, A. Fry-Smith, and M. R. Munafo. 2012. 'Effectiveness and cost-effectiveness of computer and other electronic aids for smoking

cessation: a systematic review and network meta-analysis', *Health Technology Assessment*, 16: 1-+.

Claxton, K., J. T. Cohen, and P. J. Neumann. 2005. 'When is evidence sufficient?', *Health Aff (Millwood)*, 24: 93-101.

Claxton, K., L. Ginnelly, M. Sculpher, Z. Philips, and S. Palmer. 2004. 'A pilot study on the use of decision theory and value of information analysis as part of the NHS Health Technology Assessment programme', *Health Technology Assessment*, 8: 1-+.

Claxton, K., S. Griffin, H. Koffijberg, and C. McKenna. 2015. 'How to estimate the health benefits of additional research and changing clinical practice', *BMJ*, 351: h5987.

Claxton, K. P., and M. J. Sculpher. 2006. 'Using value of information analysis to prioritise health research: some lessons from recent UK experience', *Pharmacoeconomics*, 24: 1055-68.

Claxton, K., S. Palmer, L. Longworth, L. Bojke, S. Griffin, M. Soares, E. Spackman, and C. Rothery. 2016. 'A Comprehensive Algorithm for Approval of Health Technologies With, Without, or Only in Research: The Key Principles for Informing Coverage Decisions', *Value Health*, 19: 885-91.

Claxton, Karl. 1999. 'The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies', *Journal of health economics*, 18: 341-64.

———. 2008. 'Exploring Uncertainty in Cost-Effectiveness Analysis', *Pharmacoeconomics*, 26: 781-98.

Claxton, Karl, Simon Eggington, Laura Ginnelly, Susan Griffin, Christopher McCabe, Zoe Philips, Paul Tappenden, and Alan Wailoo. 2005. "A pilot study of value of information analysis to support research recommendations for the National Institute for Health and Clinical Excellence." In.

Claxton, Karl, Peter J. Neumann, Sally Araki, and Milton C. Weinstein. 2001. 'BAYESIAN VALUE-OF-INFORMATION ANALYSIS: An Application to a Policy Model of Alzheimer's Disease', *International Journal of Technology Assessment in Health Care*, 17: 38-55.

Claxton, Karl, S Palmer, L Longworth, Laura Bojke, Susan Griffin, Claire McKenna, M Soares, Eldon Spackman, and J Youn. 2012. 'Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development'.

Conrad, Jon M. 1980. 'Quasi-option value and the expected value of information', *The Quarterly Journal of Economics*, 94: 813-20.

Corro Ramos, I., M. P. Rutten-van Molken, and M. J. Al. 2013. 'The role of value-of-information analysis in a health care research priority setting: a theoretical case study', *Med Decis Making*, 33: 472-89.

Cuypers, P. M. W., M. Gardien, J. Buth, C. H. Peels, J. A. Charbon, and W. Hop. 2001a. 'Randomized study comparing cardiac response in endovascular and open abdominal aortic aneurysm repair', 88: 1059-65.

Cuypers, PWM, Martin Gardien, Jaap Buth, CH Peels, JA Charbon, and WCJ Hop. 2001b. 'Randomized study comparing cardiac response in endovascular and open abdominal aortic aneurysm repair', *British Journal of Surgery*, 88: 1059-65.

DasGupta, A, and Herman Rubin. 2005. 'Estimation of binomial parameters when both n, p are unknown', *Journal of Statistical Planning and Inference*, 130: 391-404.

Dixit, Robert K, and Robert S Pindyck. 2012. *Investment under uncertainty* (Princeton university press).

Drummond, Michael F, and Linda Davies. 1991. 'Economic analysis alongside clinical trials: revisiting the methodological issues', *International journal of technology assessment in health care*, 7: 561-73.

Drummond, Michael F, Mark J Sculpher, Karl Claxton, Greg L Stoddart, and George W Torrance. 2015. *Methods for the economic evaluation of health care programmes* (Oxford university press).

Eckermann, S., J. Karnon, and A. R. Willan. 2010. 'The value of value of information: best informing research design and prioritization using current methods', *Pharmacoeconomics*, 28: 699-709.

Eckermann, S., and A. R. Willan. 2007a. 'Expected value of information and decision making in HTA', *Health Economics*, 16: 195-209.

———. 2007b. 'Expected value of information and decision making in HTA', *Health Econ*, 16: 195-209.
———. 2008a. 'The option value of delay in health technology assessment', *Med Decis Making*, 28: 300-5.
———. 2008b. 'Time and expected value of sample information wait for no patient', *Value Health*, 11: 522-6.
———. 2009. 'Globally optimal trial design for local decision making', *Health economics*, 18: 203 - 16.
Eeckhoudt, Louis, and Philippe Godfroid. 2000. 'Risk aversion and the value of information', *The Journal of Economic Education*, 31: 382-88.
Epstein, D. M., M. J. Sculpher, A. Manca, J. Michaels, S. G. Thompson, L. C. Brown, J. T. Powell, M. J. Buxton, and R. M. Greenhalgh. 2008. 'Modelling the long-term cost-effectiveness of endovascular or open repair for abdominal aortic aneurysm', *Br J Surg*, 95: 183-90.
Espinoza, Manuel A., Andrea Manca, Karl Claxton, and Mark J. Sculpher. 2014. 'The Value of Heterogeneity for Cost-Effectiveness Subgroup Analysis: Conceptual Framework and Application', *Medical Decision Making*, 34: 951-64.
European Commission. 2017. "Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EEC." In.
Faulkner, Laura. 2003. 'Beyond the five-user assumption: Benefits of increased sample sizes in usability testing', *Behavior Research Methods, Instruments, & Computers*, 35: 379-83.
Fenwick, E., K. Claxton, and M. Sculpher. 2008. 'The value of implementation and the value of information: Combined and uneven development', *Medical Decision Making*, 28: 21-32.
Fenwick, Elisabeth, Lotte Steuten, Saskia Knies, Salah Ghabri, Anirban Basu, James F Murray, Hendrik Erik Koffijberg, Mark Strong, Gillian D Sanders Schmidler, and Claire Rothery. 2020. 'Value of information analysis for research decisions—an introduction: report 1 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force', *Value in Health*, 23: 139-50.
Finney, DJ. 1947. 'The truncated binomial distribution', *Annals of Eugenics*, 14: 319-28.
Fisher, Anthony C. 2000. 'Investment under uncertainty and option value in environmental economics', *Resource and Energy Economics*, 22: 197-204.
Fisher, Lloyd D. 1996. 'Comments on Bayesian and frequentist analysis and interpretation of clinical trials', *Controlled clinical trials*, 17: 423-34.
Fisher, Ronald A. 1941. 'The negative binomial distribution', *Annals of Eugenics*, 11: 182-87.
Fleurence, Rachael L, and David O Meltzer. 2013. "Toward a science of research prioritization? The use of value of information by multidisciplinary stakeholder groups." In.: Sage Publications Sage CA: Los Angeles, CA.
Food and Drug Administration. 2010. 'Guidance for the use of Bayesian statistics in medical device clinical trials', *Maryland: US Food and Drug Administration*.
———. 2012. 'Medical device recall report FY2003 to FY2012', *Center for Devices and Radiological Health*.
———. 2016. 'Applying human factors and usability engineering to medical devices: Guidance for industry and Food and Drug Administration staff', *Washington, DC: FDA*.
Fortnum, H., P. Leighton, M. D. Smith, L. Brown, M. Jones, C. Benton, E. Marder, A. Marshall, and K. Sutton. 2014. 'Assessment of the feasibility and clinical value of further research to evaluate the management options for children with Down syndrome and otitis media with effusion: a feasibility study', *Health Technol Assess*, 18: 1-147, v-vi.
Fox, M., S. Mealing, R. Anderson, J. Dean, K. Stein, A. Price, and R. S. Taylor. 2007. 'The clinical effectiveness and cost-effectiveness of cardiac resynchronisation (biventricular pacing) for heart failure: systematic review and economic model', *Health Technol Assess*, 11: iii-iv, ix-248.

Frey, Bruce B. 2018. *The SAGE encyclopedia of educational research, measurement, and evaluation* (Sage Publications).

Garrison, Louis P., Adrian Towse, Andrew Briggs, Gerard de Pouvourville, Jens Grueger, Penny E. Mohr, J. L. Severens, Paolo Siviero, and Miguel Sleeper. 2013. 'Performance-Based Risk-Sharing Arrangements—Good Practices for Design, Implementation, and Evaluation: Report of the ISPOR Good Practices for Performance-Based Risk-Sharing Arrangements Task Force', *Value in Health*, 16: 703-19.

Garside, R., M. Pitt, M. Somerville, K. Stein, A. Price, and N. Gilbert. 2006. 'Surveillance of Barrett's oesophagus: exploring the uncertainty through systematic review, expert workshop and economic modelling', *Health Technology Assessment*, 10: 1-+.

Gilks, Walter R, Sylvia Richardson, and David Spiegelhalter. 1995. *Markov chain Monte Carlo in practice* (CRC press).

Goeree, Ron, Les Levin, Kiran Chandra, James M Bowen, Gord Blackhouse, Jean-Eric Tarride, Natasha Burke, Matthias Bischof, Feng Xie, and Daria O'Reilly. 2009. 'Health technology assessment and primary data collection for reducing uncertainty in decision making', *Journal of the American College of Radiology*, 6: 332-42.

Goeree, Ron, and Leslie Levin. 2006. 'Building Bridges Between Academic Research and Policy Formulation', *Pharmacoeconomics*, 24: 1143-56.

Good, Irving J. 1953. 'The population frequencies of species and the estimation of population parameters', *Biometrika*, 40: 237-64.

Goodman, Steven N. 1999a. 'Toward evidence-based medical statistics. 1: The P value fallacy', *Annals of internal medicine*, 130: 995-1004.

———. 1999b. 'Toward evidence-based medical statistics. 2: The Bayes factor', *Annals of internal medicine*, 130: 1005-13.

Green, Peter J. 1995. 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika*, 82: 711-32.

Greenhalgh, R. M. 2004. 'Comparison of endovascular aneurysm repair with open repair in patients with abdominal aortic aneurysm (EVAR trial 1), 30-day operative mortality results: randomised controlled trial', *The Lancet*, 364: 843-48.

———. 2005. 'Endovascular aneurysm repair versus open repair in patients with abdominal aortic aneurysm (EVAR trial 1): randomised controlled trial', *The Lancet*, 365: 2179-86.

Grimm, S. E., S. Dixon, and J. W. Stevens. 2017. 'Assessing the Expected Value of Research Studies in Reducing Uncertainty and Improving Implementation Dynamics', *Medical Decision Making*, 37: 523-33.

Grimm, Sabine E., Simon Dixon, and John W. Stevens. 2016. 'When Future Change Matters: Modeling Future Price and Diffusion in Health Technology Assessments of Medical Devices', *Value in Health*, 19: 720-26.

Grimm, Sabine Elisabeth, Mark Strong, Alan Brennan, and Allan J Wailoo. 2017. 'The HTA risk analysis chart: visualising the need for and potential value of managed entry agreements in health technology assessment', *Pharmacoeconomics*, 35: 1287-96.

Gronau, Quentin F, Henrik Singmann, and Eric-Jan Wagenmakers. 2017. 'Bridgesampling: An R package for estimating normalizing constants', *arXiv preprint arXiv:1710.08162*.

Groot Koerkamp, B., M. C. Weinstein, T. Stijnen, M. H. Heijenbrok-Kal, and M. G. M. Hunink. 2010. 'Uncertainty and patient heterogeneity in medical decision models', *Medical Decision Making*, 30: 194-205.

Haldane, John BS. 1941. 'The fitting of binomial distributions', *Annals of Eugenics*, 11: 179-81.

Hall, P. S., A. Smith, C. Hulme, A. Vargas-Palacios, A. Makris, L. Hughes-Davies, J. A. Dunn, J. M. S. Bartlett, D. A. Cameron, A. Marshall, A. Campbell, I. R. Macpherson, Rea Dan, A. Francis, H. Earl, A. Morgan,

R. C. Stein, C. McCabe, and Optima Trial Management Group on behalf of the. 2017. 'Value of Information Analysis of Multiparameter Tests for Chemotherapy in Early Breast Cancer: The OPTIMA Prelim Trial', *Value in Health*, 20: 1311-18.

Hall, Peter. 1994. 'On the erratic behavior of estimators of N in the binomial N, p distribution', *Journal of the American Statistical Association*, 89: 344-52.

Hanemann, W Michael. 1989. 'Information and the concept of option value', *Journal of Environmental Economics and management*, 16: 23-37.

Haute Autorité de Santé. 2018. "Principes d'évaluation de la CNEDiMTS relatifs aux dispositifs médicaux à usage individuel en vue de leur accès au remboursement." In *Evaluation des dispositifs médicaux. Mis à jour Juin*, 2017-11.

Heath, A., and G. Baio. 2018. 'Calculating the Expected Value of Sample Information Using Efficient Nested Monte Carlo: A Tutorial', *Value in Health*, 21: 1299-304.

Heath, A., I. Manolopoulou, and G. Baio. 2016. 'Estimating the expected value of partial perfect information in health economic evaluations using integrated nested Laplace approximation', *Statistics in Medicine*, 35: 4264-80.

———. 2017a. 'A Review of Methods for Analysis of the Expected Value of Information', *Med Decis Making*, 37: 747-58.

———. 2017b. 'A Review of Methods for Analysis of the Expected Value of Information', *Medical Decision Making*, 37: 747-58.

———. 2018a. 'Efficient Monte Carlo Estimation of the Expected Value of Sample Information Using Moment Matching', *Med Decis Making*, 38: 163-73.

———. 2018b. 'Efficient Monte Carlo Estimation of the Expected Value of Sample Information Using Moment Matching', *Medical Decision Making*, 38: 163-73.

Heath, Anna, Ioanna Manolopoulou, and Gianluca Baio. 2019. 'Estimating the Expected Value of Sample Information across Different Sample Sizes Using Moment Matching and Nonlinear Regression', *Medical Decision Making*, 39: 347-59.

Helbling, A, T Hurni, UR Mueller, and WJ Pichler. 2004. 'Incidence of anaphylaxis with circulatory symptoms: a study over a 3 ‐ year period comprising 940 000 inhabitants of the Swiss Canton Bern', *Clinical & Experimental Allergy*, 34: 285-90.

Henry, Claude. 1974a. 'Investment decisions under uncertainty: the" irreversibility effect"', *The American Economic Review*, 64: 1006-12.

———. 1974b. 'Option values in the economics of irreplaceable assets', *The Review of Economic Studies*, 41: 89-104.

Hertzum, Morten, and Niels Ebbe Jacobsen. 2003. 'The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods', *International Journal of Human-Computer Interaction*, 15: 183-204.

Hinchliffe, R. J., L. Bruijstens, S. T. MacSweeney, and B. D. Braithwaite. 2006. 'A randomised trial of endovascular and open surgery for ruptured abdominal aortic aneurysm – results of a pilot study and lessons learned for future studies', 32: 506-13.

Hobo, R., and J. Buth. 2006. 'Secondary interventions following endovascular abdominal aortic aneurysm repair using current endografts. A EUROSTAR report', 43: 896-902.

Hoomans, Ties, Elisabeth A. L. Fenwick, Steve Palmer, and Karl Claxton. 2009. 'Value of Information and Value of Implementation: Application of an Analytic Framework to Inform Resource Allocation Decisions in Metastatic Hormone-Refractory Prostate Cancer', *Value in Health*, 12: 315-24.

Howard, Ronald A. 1966. 'Information value theory', *IEEE Transactions on systems science and cybernetics*, 2: 22-26.

Hoyle, Martin. 2010. 'Historical Lifetimes of Drugs in England: Application to Value of Information and Cost-Effectiveness Analyses', *Value in Health*, 13: 885-92.

Huygens, Christiaan. 1657. 'De ratiociniis in ludo aleae (On reckoning at games of chance)', *London: T. Woodward*.
Kanis, H. 2011. 'Estimating the number of usability problems', *Appl Ergon*, 42: 337-47.
Kast, Robert. 1993. *La théorie de la décision* (La Découverte Paris).
Kent, S., A. Briggs, S. Eckermann, and C. Berry. 2013. 'Are value of information methods ready for prime time? An application to alternative treatment strategies for NSTEMI patients', *Int J Technol Assess Health Care*, 29: 435-42.
Knight, Frank Hyneman. 1921. *Risk, uncertainty and profit* (Houghton Mifflin).
Koerkamp, B. G., J. J. Nikken, E. H. Oei, T. Stijnen, A. Z. Ginai, and M. G. Hunink. 2008. 'Value of information analysis used to determine the necessity of additional research: MR imaging in acute knee trauma as an example', *Radiology*, 246: 420‐25.
Koerkamp, Bas Groot, Sandra Spronk, Theo Stijnen, and M. G. Myriam Hunink. 2010. 'Value of Information Analyses of Economic Randomized Controlled Trials: The Treatment of Intermittent Claudication', *Value in Health*, 13: 242-50.
Koffijberg, H., C. Rothery, K. Chalkidou, and J. Grutters. 2018. 'Value of Information Choices that Influence Estimates: A Systematic Review of Prevailing Considerations', *Med Decis Making*, 38: 888-900.
Kolmogorov, Andreĭ Nikolaevich. 1933. *Foundations of the theory of probability*.
Lange, C., L. J. Leurs, J. Buth, H. O. Myhre, and collaborators Eurostar. 2005. 'Endovascular repair of abdominal aortic aneurysm in octogenarians: an analysis based on EUROSTAR data', 42: 624-30.
Lesquelen, A., N. Thevenet, and I. Javerliat. 2009a. "Évaluation des endoprothèses aortiques abdominales utilisées pour le traitement des anévrismes de l'aorte abdominale sous-rénale." In.: Haute Autorité de Santé.
———. 2009b. 'Évaluation des endoprothèses aortiques abdominales utilisées pour le traitement des anévrismes de l'aorte abdominale sous-rénale (rapport complémentaire)'.
Lesteven, P., F. Simon-Delavelle, F. Auvigne, C. Witchitz, and E. Peyrat. 2015. "La régulation des dispositifs médicaux." In *Revue de dépenses*, edited by Inspection générale des finances - Inspection générale des affaires sociales.
Leurs, L. J., R. Hobo, and J. Buth. 2004. 'The multicenter experience with a third-generation endovascular device for abdominal aortic aneurysm repair – a report from the EUROSTAR database', 45: 293-300.
Leurs, L. J., J. Kievit, P. C. Dagnelie, P. J. Nelemans, and J. Buth. 2006. 'Influence of infrarenal neck length on outcome of endovascular abdominal aortic aneurysm repair', 13: 640-48.
Levesque, Karine, Claire Coqueblin, Bernard Guillot, Lucie Aubourg, Bernard Avouac, Cédric Carbonneil, Michel Cucherat, Patricia Descamps-Mandine, Serge Hanoka, and Marcel Goldberg. 2014. 'Les études post-inscriptions en France, quels enjeux pour les dispositifs médicaux', *Therapie*, 69: 303-12.
Lewis, James R. 1994. 'Sample sizes for usability studies: Additional considerations', *Human factors*, 36: 368-78.
———. 2000. "Using discounting methods to reduce overestimation of p in problem discovery usability studies." In.: Citeseer.
Lewis, James R, Suzanne C Henry, and Robert L Mack. 1990. "Integrated office software benchmarks: A case study." In *Interact*, 337-43.
Lewis, James R. 2001. 'Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples', *International Journal of Human-Computer Interaction*, 13: 445-79.
Lumley, Thomas. 2002. 'Network meta‐analysis for indirect treatment comparisons', *Statistics in medicine*, 21: 2313-24.

Madan, Jason, Anthony E. Ades, Malcolm Price, Kathryn Maitland, Julie Jemutai, Paul Revill, and Nicky J. Welton. 2014. 'Strategies for Efficient Computation of the Expected Value of Partial Perfect Information', *Medical Decision Making*, 34: 327-42.

McKenna, C., S. Griffin, H. Koffijberg, and K. Claxton. 2016. 'Methods to place a value on additional evidence are illustrated using a case study of corticosteroids after traumatic brain injury', *J Clin Epidemiol*, 70: 183-90.

McKenna, C., C. McDaid, S. Suekarran, N. Hawkins, K. Claxton, K. Light, M. Chester, J. Cleland, N. Woolacott, and M. Sculpher. 2009. 'Enhanced external counterpulsation for the treatment of stable angina and heart failure: a systematic review and economic analysis', *Health Technol Assess*, 13: iii-iv, ix-xi, 1-90.

McKenna, Claire, and Karl Claxton. 2011. 'Addressing Adoption and Research Design Decisions Simultaneously: The Role of Value of Sample Information Analysis', *Medical Decision Making*, 31: 853-65.

McKenna, Claire, Marta Soares, Karl Claxton, Laura Bojke, Susan Griffin, Stephen Palmer, and Eldon Spackman. 2015. 'Unifying Research and Reimbursement Decisions: Case Studies Demonstrating the Sequence of Assessment and Judgments Required', *Value in Health*, 18: 865-75.

Medtronic. 2007. "Endovascular aneurysm repair (EVAR) for the treatment of infra-renal abdominal aortic aneurysms (AAA)." In.: A submission to the National Institute for Health and Clinical Excellence (NICE) by Medtronic.

Meltzer, David O., Ties Hoomans, Jeanette W. Chung, and Anirban Basu. 2011. 'Minimal Modeling Approaches to Value of Information Analysis for Health Research', *Medical Decision Making*, 31: E1-E22.

Meng, Xiao-Li, and Wing Hung Wong. 1996. 'Simulating ratios of normalizing constants via a simple identity: a theoretical exploration', *Statistica Sinica*: 831-60.

Mensink, Paul, and Till Requate. 2005. 'The Dixit–Pindyck and the Arrow–Fisher–Hanemann–Henry option values are not equivalent: a note on Fisher (2000)', *Resource and Energy Economics*, 27: 83-88.

Menzies, N. A. 2016a. 'An Efficient Estimator for the Expected Value of Sample Information', *Medical Decision Making*, 36: 308-20.

———. 2016b. 'An Efficient Estimator for the Expected Value of Sample Information', *Med Decis Making*, 36: 308-20.

Myers, E., G. D. Sanders, D. Ravi, D. Matchar, L. Havrilesky, G. Samsa, B. Powers, A. McBroom, M. Musty, and R. Gray. 2011. 'AHRQ Methods for Effective Health Care.' in, *Evaluating the Potential Use of Modeling and Value-of-Information Analysis for Future Research Prioritization Within the Evidence-Based Practice Center Program* (Agency for Healthcare Research and Quality (US): Rockville (MD)).

Nielsen, Jakob, and Thomas K Landauer. 1993. "A mathematical model of the finding of usability problems." In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, 206-13.

Nielsen, Jakob, and Rolf Molich. 1990. "Heuristic evaluation of user interfaces." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 249-56.

O'Hagan, Anthony, Christopher McCabe, Ron Akehurst, Alan Brennan, Andrew Briggs, Karl Claxton, Elisabeth Fenwick, Dennis Fryback, Mark Sculpher, David Spiegelhalter, and Andrew Willan. 2005. 'Incorporation of uncertainty in health economic modelling studies', *Pharmacoeconomics*, 23: 529-36.

Oakley, Jeremy E., Alan Brennan, Paul Tappenden, and Jim Chilcott. 2010. 'Simulation sample sizes for Monte Carlo partial EVPI calculations', *Journal of Health Economics*, 29: 468-77.

Olkin, Ingram, A John Petkau, and James V Zidek. 1981. 'A comparison of n estimators for the binomial distribution', *Journal of the American Statistical Association*, 76: 637-42.

Osborne, Jason W. 2008. *Best practices in quantitative methods* (Sage).
Palmer, S., F. Cramp, E. Clark, R. Lewis, S. Brookes, W. Hollingworth, N. Welton, H. Thom, R. Terry, K. A. Rimes, and et al. 2016. 'The feasibility of a randomised controlled trial of physiotherapy for adults with joint hypermobility syndrome', *Health technology assessment*, 20: 1 - 290.
Panesar, SS, S Javad, D De Silva, BI Nwaru, L Hickstein, A Muraro, G Roberts, M Worm, MB Bilò, and V Cardona. 2013. 'The epidemiology of anaphylaxis in Europe: a systematic review', *Allergy*, 68: 1353-61.
Pascal, Blaise. 1654. *Traité du triangle arithmétique, avec quelques autres petits traitez sur la mesme matière. Par Monsieur Pascal* (Guil. Desprez).
Pelayo, Sylvia, Romaric Marcilly, and Tommaso Bellandi. 2020. 'Human factors engineering for medical devices: European regulation and current issues', *International Journal for Quality in Health Care*.
Peppelenbosch, N., J. Buth, P. L. Harris, C. van Marrewijk, and G. Fransen. 2004. 'Diameter of abdominal aortic aneurysm and outcome of endovascular aneurysm repair: does size matter? A report from EUROSTAR', 39: 288-97.
Pezeshk, Hamid, and John Gittins. 2006. 'Bayesian approach to determine the number of subsequent users of a new treatment', *Statistical methods in medical research*, 15: 585-92.
Philips, Z., K. Claxton, and S. Palmer. 2008. 'The half-life of truth: what are appropriate time horizons for research decisions?', *Med Decis Making*, 28: 287-99.
Pibouleau, Leslie, and Sylvie Chevret. 2011. 'Bayesian statistical method was underused despite its advantages in the assessment of implantable medical devices', *Journal of clinical epidemiology*, 64: 270-79.
Pratt, John W. 1978. 'Risk aversion in the small and in the large.' in, *Uncertainty in economics* (Elsevier).
Prinssen, M., E. L. Verhoeven, J. Buth, P. W. Cuypers, M. R. van Sambeek, R. Balm, E. Buskens, D. E. Grobbee, J. D. Blankensteijn, and Group Dutch Randomized Endovascular Aneurysm Management Trial. 2004. 'A randomized trial comparing conventional and endovascular repair of abdominal aortic aneurysms', *N Engl J Med*, 351: 1607-18.
R Development Core Team. 2010. "The R project for statistical computing." In.: R Foundation for Statistical Computing Vienna.
Raiffa, Howard, and Robert Schlaifer. 1961. *Applied statistical decision theory*.
Ramos, Isaac Corro, PMH Maureen, and Maiwenn J Al. 2015. 'Determining the impact of modeling additional sources of uncertainty in value-of-information analysis', *Value in Health*, 18: 100-09.
Reeson, Marc, Kwadwo Kyeremanteng, and Gianni D'Egidio. 2018. 'Defibrillator design and usability may be impeding timely defibrillation', *The Joint Commission Journal on Quality and Patient Safety*, 44: 536-44.
Rider, Paul R. 1955. 'Truncated binomial and negative binomial distributions', *Journal of the American Statistical Association*, 50: 877-83.
Robert, Christian. 2007. *The Bayesian choice: from decision-theoretic foundations to computational implementation* (Springer Science & Business Media).
Rothery, C., K. Claxton, S. Palmer, D. Epstein, R. Tarricone, and M. Sculpher. 2017. 'Characterising Uncertainty in the Assessment of Medical Devices and Determining Future Research Needs', *Health Econ*, 26 Suppl 1: 109-23.
Rothery, C., M. Strong, H. E. Koffijberg, A. Basu, S. Ghabri, S. Knies, J. F. Murray, G. D. Sanders Schmidler, L. Steuten, and E. Fenwick. 2020. 'Value of Information Analytical Methods: Report 2 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force', *Value Health*, 23: 277-86.
Rousset, Guillaume. 2013. "La place croissante des dispositifs médicaux dans le progrès médical: repères et enjeux, E. Couty, E. Vicaut, P. De Puylaroque (Eds.), Éditions de Santé-Presses de Sciences Po (2013), 107 pp." In.: Elsevier.

Sakalihasan, N., R. Limet, and O. D. Defawe. 2005. 'Abdominal aortic aneurysm', *The Lancet*, 365: 1577-89.

Salomon, Joshua A, Milton C Weinstein, and Sue J Goldie. 2004. 'Taking account of future technology in cost effectiveness analysis', *BMJ*, 329: 733-36.

Saltelli, Andrea. 2002. 'Sensitivity analysis for importance assessment', *Risk analysis*, 22: 579-90.

Sampram, E. S. K., M. T. Karafa, E. J. Mascha, D. G. Clair, R. K. Greenberg, and S. P. Lyden. 2003. 'Nature, frequency, and predictors of secondary procedures after endovascular repair of abdominal aortic aneurysm', 37: 930-37.

Sanders, Gillian D, Anirban Basu, Evan Myers, and David Meltzer. 2016. 'Potential value of an aspirin-dose trial for secondary prevention of coronary artery disease: informing PCORI and future trial design', *Circulation*, 134: A20405-A05.

Sauro, Jeff, and James R Lewis. 2016. *Quantifying the user experience: Practical statistics for user research* (Morgan Kaufmann).

Savage, Leonard J. 1954. *The foundations of statistics* (Courier Corporation).

Schermerhorn, M. L., A. J. O'Malley, A. Jhaveri, P. Cotterill, F. Pomposelli, and B. E. Landon. 2008. 'Endovascular vs. open repair of abdominal aortic aneurysms in the Medicare population', *N Engl J Med*, 358: 464-74.

Schermerhorn, M., A. O'Malley, A. Jhaveri, P. Cotterill, F. Pomposelli, and B. Landon. 2008. 'Endovascular vs open repair of abdominal aortic aneurysms in the Medicare population', 358: 464-74.

Schmettow, M., W. Vos, and J. M. Schraagen. 2013a. 'With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems', *J Biomed Inform*, 46: 626-41.

Schmettow, Martin. 2008. 'Heterogeneity in the usability evaluation process', *People and Computers XXII Culture, Creativity, Interaction 22*: 89-98.

———. 2009. "Controlling the usability evaluation process under varying defect visibility." In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, 188-97. British Computer Society.

———. 2012. 'Sample size in usability studies', *Communications of the ACM*, 55: 64-70.

Schmettow, Martin, Wendy Vos, and Jan Maarten Schraagen. 2013b. 'With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems', *Journal of biomedical informatics*, 46: 626-41.

Schouten, O., V. H. van Waning, M. D. Kertai, H. H. H. Feringa, J. J. Bax, and E. Boersma. 2005. 'Perioperative and long-term cardiovascular outcomes in patients undergoing endovascular treatment compared with open vascular surgery for abdominal aortic aneurysm or iliaco-femoro-popliteal bypass', 96: 861-66.

Shah, SM. 1961. 'The asymptotic variances of method of moments estimates of the parameters of the truncated binomial and negative binomial distributions', *Journal of the American Statistical Association*, 56: 990-94.

Shekelle, Paul, Martin P Eccles, Jeremy M Grimshaw, and Steven H Woolf. 2001. 'When should clinical guidelines be updated?', *BMJ*, 323: 155-57.

Simon, Herbert A. 1956. 'Rational choice and the structure of the environment', *Psychological review*, 63: 129.

Slawomirski, Luke, Ane Auraaen, and Nicolaas S Klazinga. 2017. 'The economics of patient safety: Strengthening a value-based approach to reducing patient harm at national level'.

Soeteman, Djøra I., Nicolas A. Menzies, and Ankur Pandya. 2017. 'Would a Large tPA Trial for Those 4.5 to 6.0 Hours from Stroke Onset Be Good Value for Information?', *Value in Health*, 20: 894-901.

Soulez, Gilles, Eric Thérasse, Amir Abbas Tahami Monfared, Jean-Francois Blair, Manon Choiniére, Elkoury Stéphane, Nathalie Beaudoin, Marie-France Giroux, Andrée Cliche, and Jacques Lelorier. 2005.

'Pain and quality of life assessment after endovascular versus open repair of abdominal aortic aneurysms in patients at low risk', *Journal of Vascular and Interventional Radiology*, 16: 1093-100.

Stamenkovic, Sophie, Anne Solesse, Laura Zanetti, Pascale Zagury, and Muriel Vray. 2012. 'Guide de la Haute autorité de santé (HAS): les études post-inscription sur les technologies de santé (médicaments, dispositifs médicaux et actes): principes et méthodes', *Therapie*, 67: 409-21.

Stein, R. C., J. A. Dunn, J. M. Bartlett, A. F. Campbell, A. Marshall, P. Hall, L. Rooshenas, A. Morgan, C. Poole, S. E. Pinder, D. A. Cameron, N. Stallard, J. L. Donovan, C. McCabe, L. Hughes-Davies, and A. Makris. 2016. 'OPTIMA prelim: a randomised feasibility study of personalised care in the treatment of women with early breast cancer', *Health Technol Assess*, 20: xxiii-xxix, 1-201.

Steuten, Lotte, Gijs van de Wetering, Karin Groothuis-Oudshoorn, and Valesca Retèl. 2013. 'A Systematic and Critical Review of the Evolving Methods and Applications of Value of Information in Academia and Practice', *Pharmacoeconomics*, 31: 25-48.

Stinnett, Aaron A, and A David Paltiel. 1997. 'Estimating CE ratios under second-order uncertainty: the mean ratio versus the ratio of means', *Medical decision making*, 17: 483-89.

Strong, M., and J. E. Oakley. 2013. 'An Efficient Method for Computing Single-Parameter Partial Expected Value of Perfect Information', *Medical Decision Making*, 33: 755-66.

Strong, M., J. E. Oakley, and A. Brennan. 2014a. 'Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: a nonparametric regression approach', *Med Decis Making*, 34: 311-26.

———. 2014b. 'Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: A nonparametric regression approach', *Medical Decision Making*, 34: 311-26.

Strong, M., J. E. Oakley, A. Brennan, and P. Breeze. 2015. 'Estimating the Expected Value of Sample Information Using the Probabilistic Sensitivity Analysis Sample: A Fast, Nonparametric Regression-Based Method', *Medical Decision Making*, 35: 570-83.

Team, Stan Development. 2018. "RStan: the R Interface to Stan. R package version 2.17. 3." In.

Thomas, Donald G, and John J Gart. 1971. 'Small sample performance of some estimators of the truncated binomial distribution', *Journal of the American Statistical Association*, 66: 169-77.

Timaran, C. H., F. J. Veith, E. B. Rosero, J. G. Modrall, F. R. Arko, and G. P. Clagett. 2007. 'Endovascular aortic aneurysm repair in patients with the highest risk and in-hospital mortality in the United States', 142: 520-24.

Torella, F. 2004. 'Effect of improved endograft design on outcome of endovascular aneurysm repair', 40: 216-21.

Traeger, Christian P. 2014. 'On option values in environmental and resource economics', *Resource and Energy Economics*, 37: 242-52.

Tuffaha, H. 2020. 'Value of Information Analysis: Are We There Yet?', *Pharmacoecon Open*.

Tuffaha, H. W., L. G. Gordon, and P. A. Scuffham. 2014a. 'Value of information analysis in healthcare: a review of principles and applications', *J Med Econ*, 17: 377-83.

———. 2014b. 'Value of information analysis in oncology: the value of evidence and evidence of value', *J Oncol Pract*, 10: e55-62.

Tuffaha, H. W., and P. A. Scuffham. 2018. 'The Australian Managed Entry Scheme: Are We Getting it Right?', *Pharmacoeconomics*, 36: 555-65.

Tuffaha, H. W., M. Strong, L. G. Gordon, and P. A. Scuffham. 2016. 'Efficient Value of Information Calculation Using a Nonparametric Regression Approach: An Applied Perspective', *Value Health*, 19: 505-9.

Tuffaha, Haitham W., Louisa G. Gordon, and Paul A. Scuffham. 2016. 'Value of Information Analysis Informing Adoption and Research Decisions in a Portfolio of Health Care Interventions', *MDM Policy & Practice*, 1: 2381468316642238.

Tuffaha, Haitham W., Heather Reynolds, Louisa G. Gordon, Claire M. Rickard, and Paul A. Scuffham. 2014. 'Value of information analysis optimizing future trial design from a pilot study on catheter securement devices', *Clinical Trials*, 11: 648-56.

UK-MHRA. 2017. "Human Factors and Usability Engineering – Guidance for Medical Devices Including Drug-Device Combination Products." In, edited by Medicines & Healthcare products Regulatory Agency.

Vallejo-Torres, Laura, Lotte Steuten, Bonny Parkinson, Alan J. Girling, and Martin J. Buxton. 2010. 'Integrating Health Economics Into the Product Development Cycle: A Case Study of Absorbable Pins for Treating Hallux Valgus', *Medical Decision Making*, 31: 596-610.

van Asselt, Thea, Bram Ramaekers, Isaac Corro Ramos, Manuela Joore, Maiwenn Al, Ivonne Lesman-Leegte, Maarten Postma, Pepijn Vemer, and Talitha Feenstra. 2018. 'Research costs investigated: a study into the budgets of Dutch publicly funded drug-related research', *Pharmacoeconomics*, 36: 105-13.

van Eps, R. G. S., L. J. Leurs, R. Hobo, P. L. Harris, and J. Buth. 2007. 'Impact of renal dysfunction on operative mortality following endovascular abdominal aortic aneurysm surgery', 94: 174-78.

van Gestel, Aukje, Janneke Grutters, Jan Schouten, Carroll Webers, Henny Beckers, Manuela Joore, and Johan Severens. 2012. 'The Role of the Expected Value of Individualized Care in Cost-Effectiveness Analyses and Decision Making', *Value in Health*, 15: 13-21.

van Marrewijk, C. J., G. Fransen, R. J. F. Laheij, P. L. Harris, and J. Buth. 2004. 'Is a type II endoleak after EVAR a harbinger of risk? Causes and outcome of open conversion and aneurysm rupture during follow-up', 27: 128-37.

Vandewalle, Vincent, Alexandre Caron, Coralie Delettrez, Renaud Périchon, Sylvia Pelayo, Alain Duhamel, and Benoit Dervaux. 2020. 'Estimating the number of usability problems affecting medical devices: modelling the discovery matrix', *BMC Medical Research Methodology*, 20: 234.

Verzini, F., P. Cao, S. Zannetti, G. Parlani, P. De Rango, and A. Maselli. 2002. 'Outcome of abdominal aortic endografting in high-risk patients: a 4-year single-center study', 9: 736-42.

Virzi, Robert A. 1992. 'Refining the test phase of usability evaluation: How many subjects is enough?', *Human factors*, 34: 457-68.

Von Neumann, John, and Oskar Morgenstern. 1947. 'Theory of games and economic behavior, 2nd rev'.

Wade, R., E. Sideris, F. Paton, S. Rice, S. Palmer, D. Fox, N. Woolacott, and E. Spackman. 2015. 'Graduated compression stockings for the prevention of deep-vein thrombosis in postoperative surgical patients: a systematic review and economic model with a value of information analysis', *Health Technology Assessment*, 19: 1.

Whittaker, J. 1990. 'Graphical Models in Applied Multivariate Statistics Wiley New York 1', *Zbl0732*, 62056.

Willan, A., and M. Kowgier. 2008. 'Determining optimal sample sizes for multi-stage randomized clinical trials using value of information methods', *Clin Trials*, 5: 289-300.

Willan, A. R. 2007a. 'Clinical decision making and the expected value of information', *Clinical Trials*, 4: 279-85.

———. 2007b. 'Clinical decision making and the expected value of information', *Clin Trials*, 4: 279-85.

———. 2008. 'Optimal sample size determinations from an industry perspective based on the expected value of information', *Clinical trials (london, england)*, 5: 587-94.

Willan, A. R., and S. Eckermann. 2010. 'Optimal clinical trial design using value of information methods with imperfect implementation', *Health Econ*, 19: 549-61.

Willan, A. R., and E. M. Pinto. 2005. 'The value of information and optimal clinical trial design', *Stat Med*, 24: 1791-806.

Woolrych, Alan, and Gilbert Cockton. 2001. "Why and when five test users aren't enough." In *Proceedings of IHM-HCI 2001 conference*, 105-08. Eds)(Cépaduès Editions, Toulouse, FR, 2001).

Partie 6. ANNEXES

I ARTICLE VALUE IN HEALTH

I.A Appendices 1: mathematical background

In these appendices, we detail the mathematical background used to estimate the parameters for the value-of-information analysis:

(i) the number of use errors affecting the medical device.

(ii) the number of use errors discovered by end users once the device is on the market.

(iii) the acceptability of the use errors.

(iv) the effect of adding new participants to usability studies.

Each point is detailed in a dedicated sub-section below.

I.A.1 Estimating the number of use errors

As mentioned in the Introduction, the discovery matrix $\mathbb{d}$ observed during usability testing is a truncated version of a larger (unobserved) matrix $\mathbb{x}$. It lacks the columns corresponding to the as-yet undetected errors. In the matrix $\mathbb{d}$, the columns are ordered according to the discovery of use errors. A probabilistic model is assumed on $\mathbb{x}$, and a model on $\mathbb{d}$ is then deduced. The following notation is used: $\mathbb{x} = (x_{il})_{1\leq i\leq n, 1\leq l\leq m}$ where $x_{il} = 1$ if participant i detects error l, and $x_{il} = 0$ otherwise.

$$\mathbb{x} = \begin{pmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{il} & \cdots & x_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nl} & \cdots & x_{nm} \end{pmatrix}$$

Schmettow introduced heterogeneity into the estimation by modeling the probability of detection with a logit-normal binomial distribution.[1] In this model, each error l has its own probability of detection p_l, and the probabilities of detection (are independent and) follow a

logit-normal distribution, i.e. $\text{logit}(p_l) \sim \mathcal{N}(\mu, \sigma)$. Here, $p_1, \ldots, p_m$ are considered to be latent random variables - like random effects in the mixed model.

The likelihood of the complete discovery matrix $\mathbb{X}$ is obtained by integrating over the latent variables $p_1, \ldots, p_m$:

$$P(\mathbb{X}|\mu, \sigma, m) = \int_0^1 \ldots \int_0^1 P(\mathbb{X}|p_1, \ldots, p_m, m) f(p_1, \ldots, p_m|\mu, \sigma) dp_1 \ldots dp_m$$

where $f(p_1, p_2, \ldots, p_m|\mu, \sigma)$ is the probability density function of $p_1, p_2, \ldots, p_m$.

We used a Bayesian framework[2] to estimate the parameters. We assumed the prior independence of μ, σ and m:

$$P(\mu, \sigma, m) = P(\mu)P(\sigma)P(m)$$

The following priors were defined for the Bayesian estimation. They were chosen in order to avoid the introduction of prior knowledge - a common criticism of Bayesian approaches. As a result, p_l and m must be considered as "flat" or "non-informative" priors:

- $\mu \sim \mathcal{N}(0; \mathcal{A})$: a Gaussian distribution with variance $\mathcal{A} = 1.5$,
- $\sigma^2 \sim \text{inv} - \chi^2_\nu$: an inverse chi-squared distribution with $\nu = 1$ degrees of freedom.
- $P(m) = \frac{1}{M} \forall m \in \{1, \ldots, M\}$: a uniform distribution with M being a pre-determined upper boundary for m.

For each possible value of m in a grid $\{1, \ldots, M\}$, it is possible to compute the integrated likelihood $P(\mathbb{X}|m)$:

$$P(\mathbb{X}|m) = \int_0^{+\infty} \int_{-\infty}^{+\infty} P(\mathbb{X}|\mu, \sigma, m) P(\mu) P(\sigma) d\mu \, d\sigma$$

We approximated this integral with Markov chain Monte Carlo (MCMC) techniques. Given the previous equation, we focused on the computation based on $\mathbb{X}$. In practice, μ and σ are drawn for each possible value of m, $m \in \{1, \ldots, M\}$. For a fixed value of m, we sample from

$P(\mu, \sigma|\mathbb{\hat{X}}^m, m)$. The parameters are sampled using the parameter space augmented by $p_1, \ldots, p_m$ (i.e. from $\mu, \sigma, p_1, \ldots, p_m|\mathbb{\hat{X}}^m$) using the adaptative Hamiltonian Monte Carlo algorithm of *stan.*[3] Next, a numerical approximation of the integrated likelihood $P(\mathbb{X}|m)$ is obtained via bridge sampling.[4]

However, $\mathbb{X}$ is not observed, and thus the Bayesian inference is carried out on the observed matrix $\mathbb{d}$. The matrix $\mathbb{d}$ can be linked to $\mathbb{X}$ while noting that the columns are exchangeable, i.e. that the integrated likelihood on $\mathbb{X}$ is the same for any permutation of its columns:

$$P(\mathbb{d}|m) = \frac{1}{j_1! \ldots j_r!} \times A_m^j \times P(\mathbb{\hat{X}}^m|\, m),$$

where $\mathbb{\hat{X}}^m$ is the complete discovery matrix obtained from $\mathbb{d}$ given the value m, i.e. the matrix $\mathbb{d}$ padded with $m - j$ null columns.

To estimate the number of errors, we focused on $P(m|\mathbb{d})$, which is obtained using Bayes' theorem:

$$P(m|\mathbb{d}) = \frac{\mathrm{P}(m) \times P(\mathbb{d}|m)}{\sum_{m'=1}^{M} \mathrm{P}(m') \times P(\mathbb{d}|m')}$$

Likewise, the hyperparameters of the logit normal distribution, $P(\mu|\mathbb{d})$ and $P(\sigma|\mathbb{d})$, can be estimated.

It is noteworthy that various parameters are computed during the MCMC step and can be sampled "backwards":

- The number of errors m can be sampled from $P(m|\mathbb{d})$
- The probability of detection for each of the m errors $p_1, \ldots, p_m$ can be sampled from

  $P(p_1, \ldots, p_m|\mathbb{d}, m) = P(p_1, \ldots, p_m|\mathbb{\hat{X}}^m)$

I.A.2 Modelling the discovery of usability-induced use errors by end users

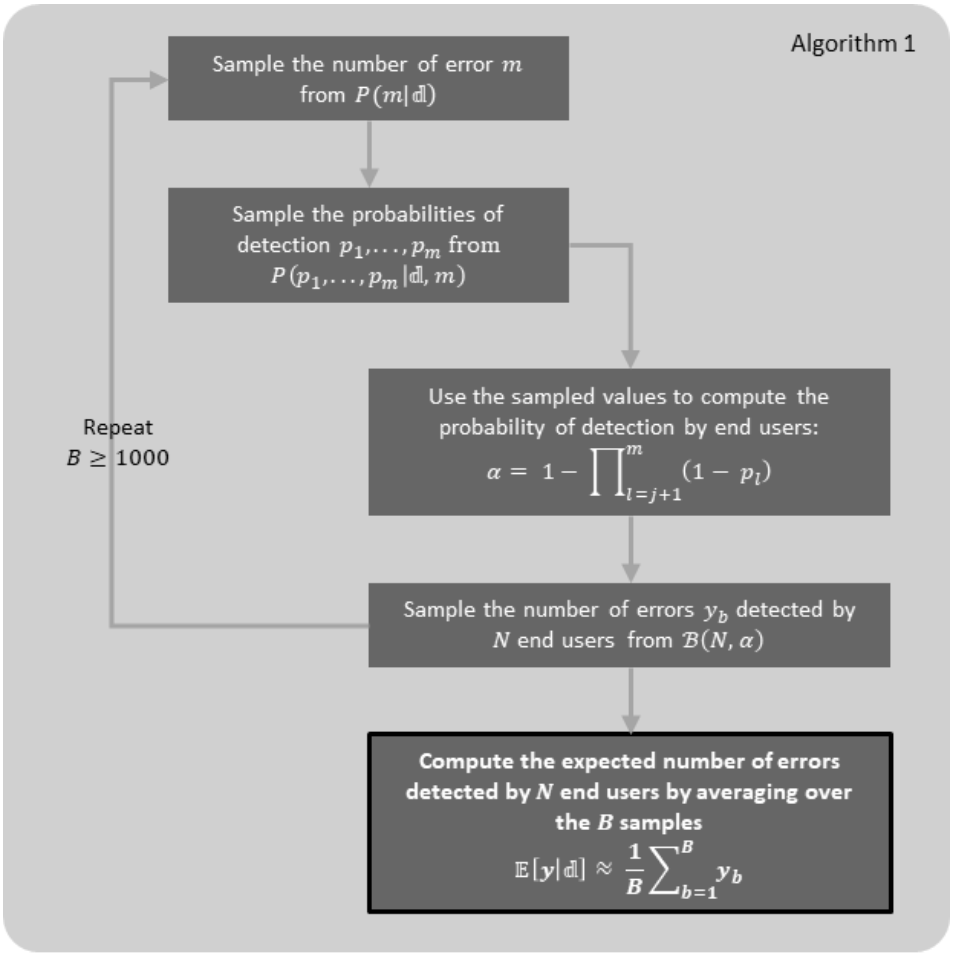
The medical device assessed in the validation study is the final version. The regulator grants market access on the basis of this design and therefore prohibits any modifications. Since the

same medical device is used by both usability testing participants and end users, we assume that the post-marketing error discovery is a continuation of that observed during the validation study. Indeed, the number of errors experimented by end users will follow the same Bernoulli process as that observed during the testing.

Once the device has been commercialized, the end user might discover one or more of the $m - j$ errors not detected during the usability testing. Since each error has its own probability of detection p_l, the probability α whereby an end user detect at least one of the $m - j$ remaining errors is $\alpha = 1 - \prod_{l=j+1}^{m}(1 - p_l)$, i.e. the probability of "failing" in the $m - j$ independent Bernoulli trials. When considering the whole population of size N, the distribution of the number of end users discovering at least one error (denoted as y) is written as follows:

$$y|m, p_1, \ldots, p_m \sim \mathcal{B}\left(N, \alpha = 1 - \prod_{l=j+1}^{m}(1 - p_l)\right)$$

In this equation, it is noteworthy that y is a random variable that depends on m and $p_1, \ldots, p_m$, which are unknown. Furthermore, the distribution of these numbers depends on parameters (e.g. μ, σ) considered to be random in a Bayesian framework. However, these parameters can be sampled backwards, as explained above. The algorithm for computation of $\mathbb{E}[y|\mathbb{d}]$ is shown in Supplementary Figure 1.



Supplementary Figure 1: Sampling algorithm for computing the expected number of errors detected by end users.

I.A.3 The acceptability of the use errors

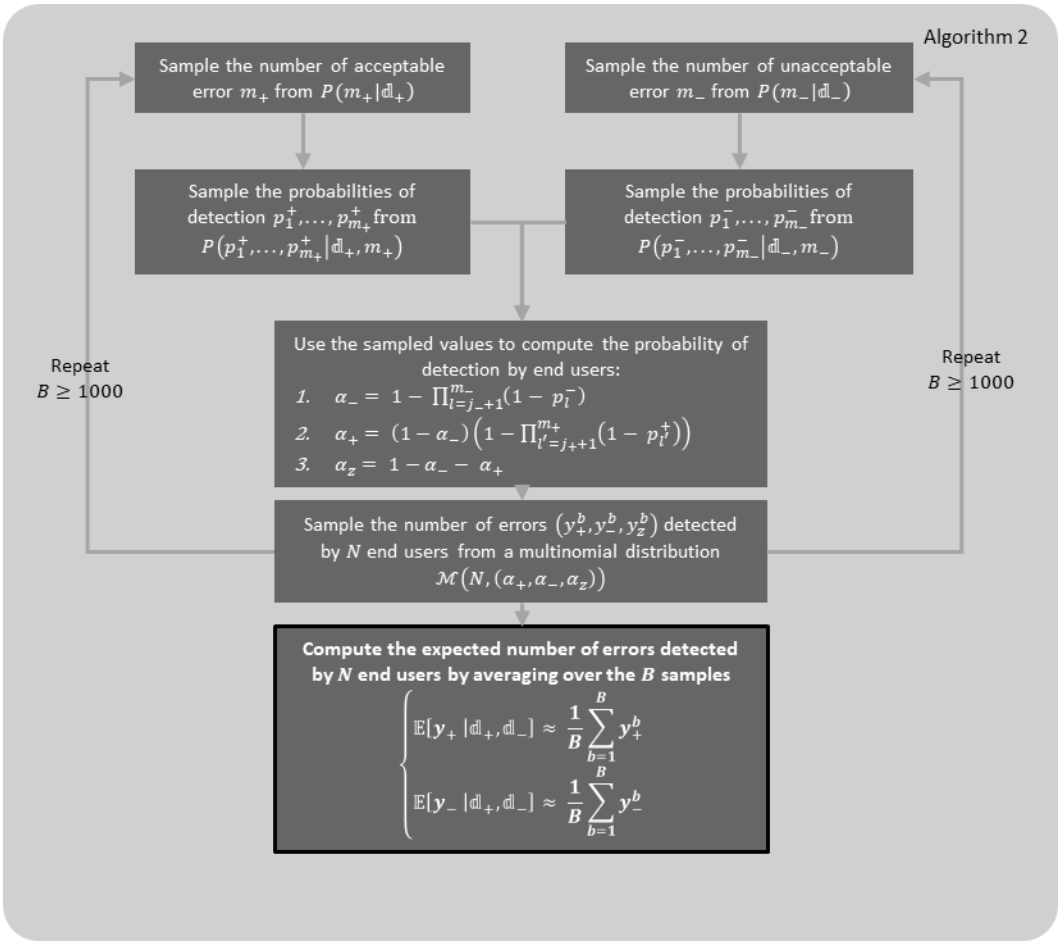
We now distinguish between two levels of acceptability, namely acceptable errors and unacceptable errors denoted with "+" and "−" subscripts, respectively. Let us assume that j_+ and j_- errors were discovered during the evaluation; the corresponding discovery matrices are now $\mathbb{d}_+$ and $\mathbb{d}_-$. It is noteworthy that if an unacceptable error is observed during the usability test, the medical device would need to be redesigned and the entire usability test would have to be repeated from the beginning on the new device. Thus, in our case, $j_- = 0$. However, the mathematics presented here are valid for $j_- \geq 0$ without loss of generalizability.

It is widely acknowledged that for a given error, the probability of detection is not linked to the acceptability. Thus, we used the same prior distribution for the probabilities of detection for both acceptable and unacceptable errors. For instance, non-informative priors were defined for both

acceptability (denoted as m_+ and m_-): $P(m_+) = \frac{1}{M_+} \forall m_+ \in \{1, \dots, M_+\}$ and $P(m_-) = \frac{1}{M_-} \forall m_- \in \{1, \dots, M_-\}$

Since the two discovery matrices are supposed to be independent (with each generated by its own discovery parameters), we obtain $P(m_+, m_-|\mathbb{d}_+, \mathbb{d}_-) = P(m_+|\mathbb{d}_+)P(m_-|\mathbb{d}_-)$. Thus, the number of errors can be estimated out separately. Indeed, a posterior analysis can be performed (as described above) for $\mathbb{d}_+$ on the one hand and $\mathbb{d}_-$ on the other.

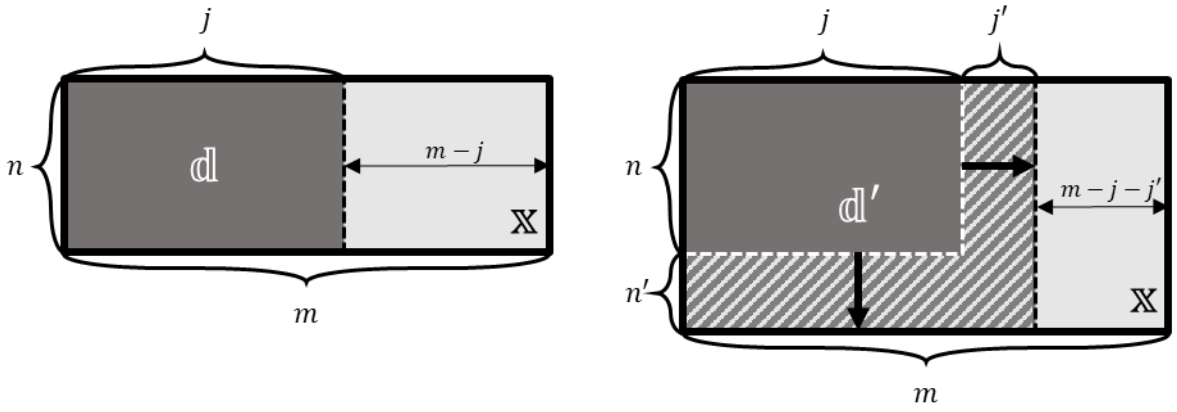
In contrast to $y|m, p_1, \dots, p_m$, the computations of y_+ and y_- involve multinomial sampling because the end user can experience (i) one or more unacceptable errors, (ii) one or more acceptable errors, or (iii) no errors. The probabilities of these three outcomes are denoted as α_+, α_- and α_z, respectively. In multinomial sampling, we considered that an unacceptable error prevailed over an acceptable error; if an end user experiences both types, he/she is considered to be only in the "unacceptable error" category. The computation algorithm for calculating the number of undetected errors while accounting for error acceptability is given in Supplementary Figure 2.



Supplementary Figure 2: Sampling algorithm for computing the expected number of acceptable and unacceptable errors detected by end users. We first need to sample parameters for acceptable and unacceptable errors separately, and then adapt the sampling of the number of users encountering critical and non-critical errors according to a multinomial distribution. α_+ is the probability with which an end user encounters at least one previously undiscovered acceptable error, α_- is the probability with which an end user encounters at least one previously undiscovered unacceptable error but no acceptable errors, and α_z is the probability with which a user encounters no new errors ($\alpha_+ + \alpha_- + \alpha_z = 1$).

I.A.4 The effect of adding new participants to the usability testing

For the sake of clarity, we first examine the effect of adding new participants without accounting for different levels of error acceptability. After including n participants, j errors were discovered. Adding n' new participants ($n' > 0$) reduces the number of undetected errors and confirms the lower probability of occurrence of undetected errors. Eventually, the discovery process continues and j' new errors are discovered. The impact on the discovery matrix is shown in Supplementary Figure 3.



Supplementary Figure 3: Progress of the discovery process and impact on the discovery matrix after n participants (left matrix) and after the addition of n' new participants (right matrix). j' is the number of new errors discovered after adding n' participants.

The discovery matrix after $n + n'$ participants (denoted as $\mathbb{d}'$) has more rows than $\mathbb{d}$ (corresponding to the new participants) and – possibly - more columns than $\mathbb{d}$ (corresponding to possible j' newly detected errors), represented by the hatched area in Supplementary Figure 3. Since the number of undetected errors decreases (the light grey area in Supplementary Figure 3), the expected number discovered by end users y also decreases. In other words, the errors are detected during testing or after the market launch.

In the remainder of this paper, we will focus on (i) assessing the magnitude of the reduction in the number of previously undetected errors and (ii) showing that the recruitment of additional participants confirms the lower probability of occurrence of undetected errors. We are primarily interested in the expected number of errors discovered by the end users. Noting that the matrix $\mathbb{d}'$ remains unobserved so far, it is considered as a random variable; the expected number of errors detected by the end users must be computed over all possible $\mathbb{d}'$, given the matrix $\mathbb{d}$ already observed.

$$\mathbb{E}[\mathbb{E}[y|\mathbb{d}']\,|\mathbb{d}]$$

Computation of the expected number of errors detected by the end users requires the use of a Bayesian tool called pre-posterior analysis. This technique is based on the classical posterior analysis in which the prior distribution is updated with sampled data to give the parameter's

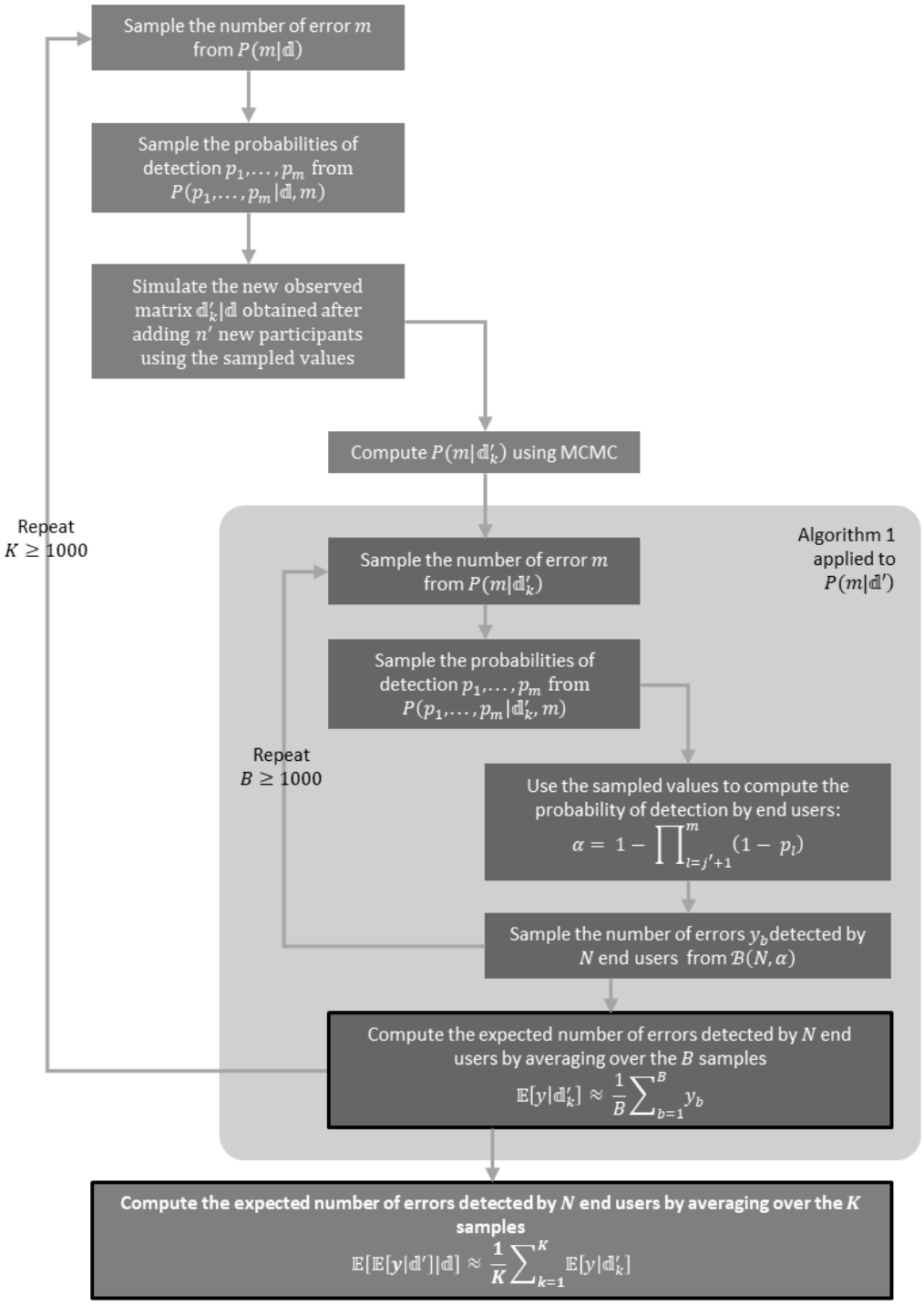
posterior distribution. In our case, the prior distribution of the number of errors (given the observed discovery matrix) is $P(m|\mathbb{d})$, the sampled data is $\mathbb{d}'$, and the corresponding updated posterior distribution is $P(m|\mathbb{d}')$. However, the posterior analysis cannot be carried out in this way because (as mentioned above) $\mathbb{d}'$ has not been observed. The pre-posterior analysis usefully solves this issue by predicting the likelihood function (the distribution of the sampled data) that is conditional on the prior data. Next, the posterior distribution $P(m|\mathbb{d}')$ is generated and the algorithms described above are applied. The pre-posterior algorithm is shown in Supplementary Figure 4 (without taking account of error acceptability) and Supplementary Figure 5 (taking account of error acceptability). It should be noted that the use of the full pre-posterior algorithm can be computationally prohibitive due to the need to sample from $y|\mathbb{d}'$ with an expensive MCMC algorithm for each possible value of $\mathbb{d}'$. In order to make the computation tractable, we suggest:

1. Keeping values of m, μ, σ that had been used to sample $\mathbb{d}'$ (μ and σ are obtained from the MCMC chain while sampling from $p_1, \ldots, p_m|\mathbb{d}, m$).

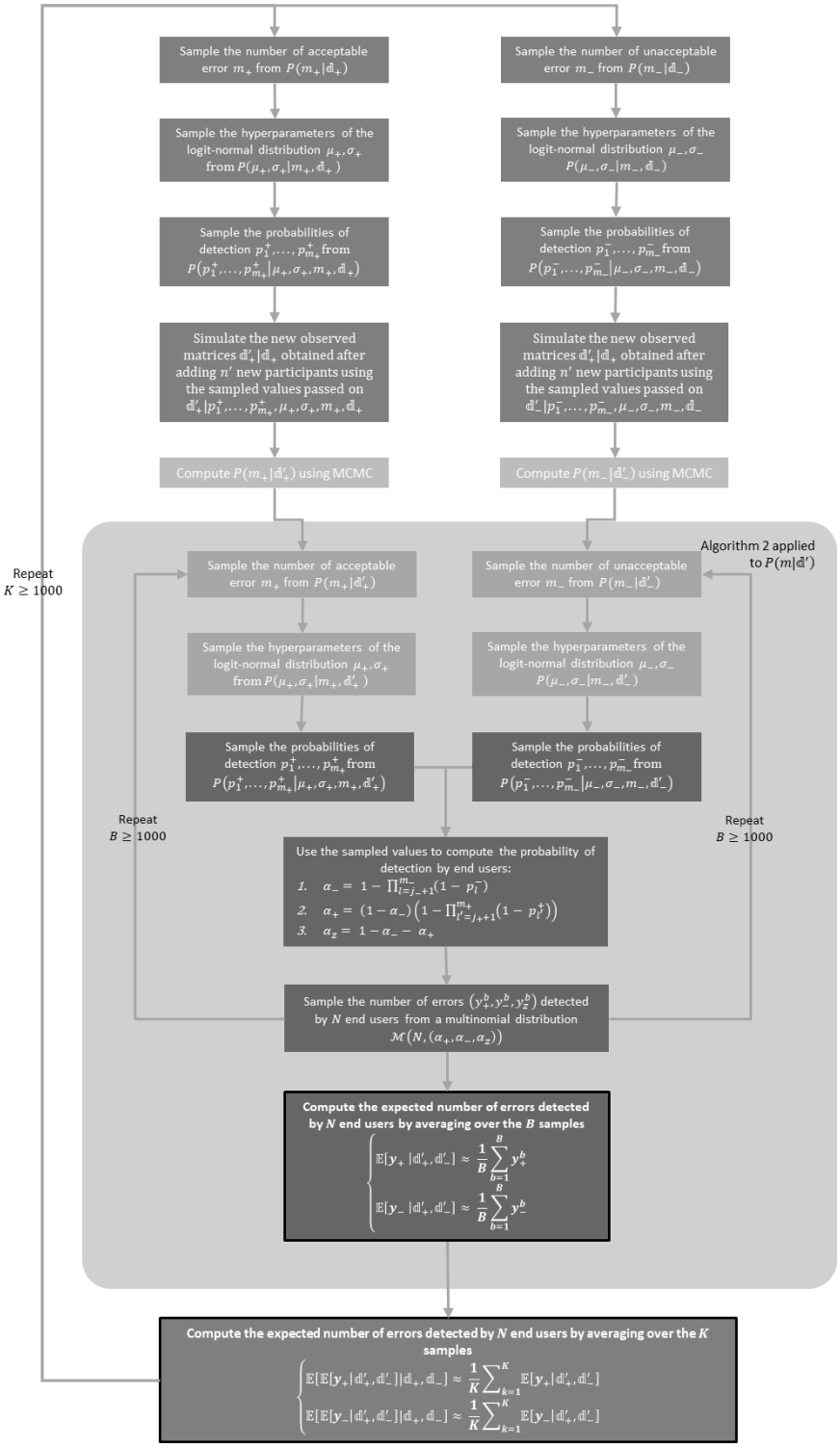
2. Sampling independently each p_j for undetected problems, according to

   $P(p_j|\mu, \sigma, x_{.j}' = 0)$

The rationale of this approach is that the sampling of $\mathbb{d}'$ resulting from m, μ, σ, will not give any significant new information about these parameters. The main update needed is for the probabilities of the non-discovered problem in $\mathbb{d}'$ (involved in the sampling of y). These can be obtained in the same way for each undiscovered problem. This consists in sampling the probability of error j given the generation parameters μ and σ, and knowing that problem j has not yet been observed by the $n + n'$ subjects of the augmented discovered matrix ($x_{.j}' = 0$, denoting that column j of the completed discovery matrix is null). This is particularly tractable since the burdensome multivariate MCMC sampler is replaced with a simple univariate sampler (the same one for each probability of an undiscovered error).



Supplementary Figure 4: Sampling algorithm for the pre-posterior analysis of the expected numbers of acceptable and unacceptable errors newly detected after adding n' participants to the usability test.



Supplementary Figure 5: Sampling algorithm for the pre-posterior analysis of the expected number of acceptable and unacceptable errors detected by end users after adding n' participants to the usability testing. In order to make the computation tractable, we kept values of m, μ, σ that had been used to sample $\mathbb{d}'$ (μ and σ are obtained from the MCMC chain while sampling from $p_1, \dots, p_m|\mathbb{d}, m$) and sampled independently each p_j for undetected problems, according to $P(p_j|\mu, \sigma, x_{.j}' = 0)$. These steps are represented in light grey.

I.A.5 Reference

1. Schmettow M. Controlling the usability evaluation process under varying defect visibility. British Computer Society; 2009:188-197.

2. Robert C. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media; 2007.

3. Carpenter B, Gelman A, Hoffman MD, et al. Stan: A probabilistic programming language. *Journal of statistical software*. 2017;76(1)

4. Meng X-L, Wong WH. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*. 1996:831-860.

I.B Appendices 2: case-study by “classical” methods

We estimated the sample size by using “classical” methods: the naïve, Good Turing and double deflation methods.(Vandewalle et al. 2020) The three methods estimated that the error discovery process was complete after the inclusion of 20 participants - illustrating the statistical weakness of these approaches.

The logit-normal binomial zero-truncated model estimated that the total number of errors ranged from 33 to 49 for sample sizes from 20 to 90. However, the use of logit-normal binomial zero-truncated model would always prompt the recruitment of more participants because (i) use errors would always remain and (ii) the costs and benefits of further recruitment are not taken into account.

II ESTIMATION DES COUTS DES PROCEDURES EVAR ET OPEN

La détermination des coûts des procédures OPEN et EVAR est réalisée à partir des bases de données médico-administratives nationales hospitalières. L'objectif est de déterminer un coût moyen pour chacune des procédures, un coût moyen pour une réintervention et le coût du suivi. Dans une première section, nous présenterons brièvement la base nationale du Programme de Médicalisation des Systèmes d'Information (PMSI) qui contient l'ensemble des séjours hospitaliers en France. Ensuite, dans une deuxième section, nous détaillerons l'étude nationale des coûts (ENC) qui réalise chaque année l'estimation des coûts moyens nationaux par groupes homogènes de malades (GHM) et qui permet la caractérisation de l'incertitude entourant ceux-ci. Enfin, dans une dernière section, nous présenterons l'application à notre cas d'étude et les résultats obtenus. En annexe, nous fournirons les documents et codes nécessaires à la reproductibilité de l'analyse.

II.A Le Programme de Médicalisation des Systèmes d'Information (PMSI)

II.A.1 Description du PMSI

En France, l'activité médicale des établissements de santé est enregistrée dans une base de données médico-administratives appelée Programme de Médicalisation des Systèmes d'Information (PMSI). Les données médicales d'hospitalisation sont traduites en diagnostics (codés selon la 10^{e} Classification Internationale des Maladies CIM10) et en actes (codés selon la Classification Commune des Actes Médicaux CCAM). Les séjours hospitaliers sont regroupés en Groupes Homogènes de Malades (GHM).

Ces GHM sont déterminés par un algorithme de groupage qui prend en compte le motif principal d'hospitalisation (« diagnostic principal »), la présence d'actes spécifiques dits « classants » (ex : les actes opératoires classent les séjours dans des GHM de type chirurgical – un seul acte est pris en compte) et la cohérence entre le diagnostic principal et l'acte classant (cohérence entre la pathologie et l'organe opéré). L'âge du patient peut également être pris en compte. Les GHM sont encodés sur 6 caractères. Schématiquement, les 2 premiers chiffres représentent l'organe ou le groupe d'organes concernés (Catégorie Majeure de Diagnostic, CMD) et le 3^{e} caractère est une lettre qui signe le caractère chirurgical (C), interventionnel (K), médical (M) ou indifférencié

(Z) du GHM. Les 2 chiffres servent de compteur. Les 5 premiers caractères définissent les racines des GHM. Les niveaux de sévérité encodés sur le 6e caractère tels qu'ils existent actuellement ont été mis en place à partir de 2009 (version v11 des GHM).

II.A.2 Case-mix pour les procédures OPEN et EVAR

Nous avons identifié les actes associés à EVAR et OPEN dans la CCAM en vigueur en 2009 (version 15). Après exclusion des gestes qui concernaient les anévrismes supra-rénaux et ceux qui décrivaient un clampage supra-rénal, nous avons retenu 3 actes pour EVAR et 6 actes pour OPEN. Le code et le libellé de chacun de ces actes CCAM sont détaillés dans le Tableau 9.

Tableau 20 : Actes CCAM caractérisant les procédures OPEN et EVAR.

Acte CCAM	Libellé
	EVAR
DGLF0010	Pose d'endoprothèse couverte bifurquée aortobisiliaque, par voie artérielle transcutanée
DGLF0020	Pose d'endoprothèse couverte aorto-uniiliaque, par voie artérielle transcutanée
DGLF0050	Pose d'endoprothèse couverte rectiligne dans l'aorte abdominale infrarénale, par voie artérielle transcutanée
	OPEN
DGPA0050	Mise à plat d'un anévrisme aortique infrarénal non rompu avec remplacement prothétique aortoaortique infrarénal, par laparotomie avec clampage infrarénal
DGPA0100	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque non rompu avec remplacement prothétique aortobifémoral, par laparotomie avec clampage infrarénal
DGPA0120	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque non rompu avec remplacement prothétique aortobisiliaque, par laparotomie avec clampage infrarénal
DGPA0160	Mise à plat d'un anévrisme aorto-ilio-fémoral avec remplacement prothétique bifurqué aorto-ilio-fémoral, par laparotomie avec clampage infrarénal
DGPA0180	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque rompu avec remplacement prothétique, par laparotomie
EDPA0050	Mise à plat d'un anévrisme iliaque avec remplacement prothétique aorto-iliaque ou aortofémoral unilatéral, par laparotomie

Nous avons ensuite déterminé les GHM compatibles avec les procédures EVAR et OPEN à partir d'une liste de GHM contenant au moins l'un des codes CCAM sélectionnés. Nous n'avons retenu que les GHM relatifs aux affections de l'appareil circulatoire (CMD 05) compatibles avec une intervention d'AAA (ex : exclusion du GHM 05K05V « Endoprothèses vasculaires et infarctus du myocarde sans CMA ») afin de pallier la non-spécificité des actes sélectionnés[27] et les erreurs de codage. Le code et le libellé de chacun de ces actes CCAM sont détaillés dans le Tableau 10.

Tableau 21 : GHM caractérisant les procédures OPEN et EVAR.

GHM	Libellé
	EVAR
05C101	Chirurgie majeure de revascularisation, niveau 1
05C102	Chirurgie majeure de revascularisation, niveau 2
05C103	Chirurgie majeure de revascularisation, niveau 3
05C104	Chirurgie majeure de revascularisation, niveau 4
05C111	Autres interventions de chirurgie vasculaire, niveau 1
05C112	Autres interventions de chirurgie vasculaire, niveau 2
05C113	Autres interventions de chirurgie vasculaire, niveau 3
05C114	Autres interventions de chirurgie vasculaire, niveau 4
05K061	Endoprothèses vasculaires sans infarctus du myocarde, niveau 1
05K062	Endoprothèses vasculaires sans infarctus du myocarde, niveau 2
05K063	Endoprothèses vasculaires sans infarctus du myocarde, niveau 3
05K064	Endoprothèses vasculaires sans infarctus du myocarde, niveau 4
05K06T	Endoprothèses vasculaires sans infarctus du myocarde, très courte durée
	OPEN
05C101	Chirurgie majeure de revascularisation, niveau 1
05C102	Chirurgie majeure de revascularisation, niveau 2
05C103	Chirurgie majeure de revascularisation, niveau 3
05C104	Chirurgie majeure de revascularisation, niveau 4

Pour chacune des deux procédures, nous avons extrait l'ensemble des GHM d'intérêt entre le 1er mars 2009 et le 28 février 2010 (année de facturation), puis nous avons exclu ceux ne contenant

[27] Exemple : pour l'acte OPEN DGPA005, 1213 hospitalisations ont été enregistrées dans la base nationale du PMSI en 2008 sur plus de 20 GHM différents. 7 GHM relevaient de la CMD 05 mais on trouvait parmi eux 2 GHM de pontage aortocoronarien. Finalement, les 5 GHM restants représentaient 1167 hospitalisations avec l'acte DGPA005.

pas un acte CCAM caractéristique de la procédure. L'agrégation des nombres d'hospitalisations par GHM a abouti à un case-mix pour OPEN et pour EVAR. La répartition du nombre de séjours sélectionnés entre ces GHM définissait leur pondération.

II.B L'étude nationale des coûts a méthodologie commune (ENCC)

II.B.1 Description de l'ENCC et du référentiel national de coût

Le référentiel national de coûts produit des estimations de coûts moyens par GHM à partir d'un échantillon d'établissements participant à l'ENCC. Ce dernier se base sur des outils de comptabilité analytique qui permettent de ventiler les coûts sur différents postes de dépenses. Les coûts utilisés excluent les charges de structure (immobilier et financier). Les données 2008 ont été publiées pour des GHM au format v11b (i.e. en intégrant les niveaux de sévérité), permettant le traitement individuel des GHM.[28]

II.B.2 Guide technique des modalités de calcul du référentiel national de coûts

L'Agence Technique de l'Information sur l'Hospitalisation (ATIH) édite annuellement un guide technique des modalités de calcul du référentiel national de coûts. On y trouve notamment le « calcul de précision » dont le but est de quantifier l'erreur d'échantillonnage. Les éléments calculés sont les suivants :

- Le nombre de séjour total du GHM : N
- Le nombre de séjours échantillonnés dans l'ENCC : N_{ENCC}
- Le coût moyen du GHM estimé sur l'échantillon : $\bar{c}$
- La variance estimée du coût moyen[29] (erreur standard au carré) : $\sigma^2(\bar{c})$
- L'écart type estimé du coût moyen (erreur standard) : $\sigma(\bar{c})$

[28] Exemple : l'acte intitulé « Mise à plat d'un anévrisme aortique infrarénal non rompu avec remplacement prothétique aortoaortique infrarénal, par laparotomie avec clampage infrarénal » est codé DGPA005 en CCAM. Associé à un diagnostic principal d'affection de l'appareil circulatoire (par exemple : I71.4 « Anévrisme aortique abdominal, sans mention de rupture »), il orientera vers la racine de GHM 05C10 « Chirurgie majeure de revascularisation » (les 2 premiers chiffres correspondent à la Catégorie Majeure de Diagnostic (CMD), ie peu ou prou l'organe ou le groupe d'organes concernés, la lettre C signe le caractère chirurgical, les 2 chiffres suivants sont un compteur). L'algorithme permet ensuite le classement selon 4 niveaux de sévérités (05C101, 05C102, 05C103, 05C104).

[29] Cette quantité n'est pas disponible immédiatement dans l'ENCC mais son calcul est détaillé dans le référentiel.

- L'erreur relative (ERE) de l'estimateur du coût moyen : $ERE = \sigma(\bar{c}) / \bar{c}$
- L'intervalle de confiance à 95% du coût moyen : $IC_{95\%} = \bar{c} \pm 2 * \bar{c} * ERE = \bar{c} \pm 2 * \sigma(\bar{c})$

II.B.3 Calcul du coût moyen des procédures

Le coût moyen de la procédure OPEN correspond à la moyenne pondérée de chacun des GHM (k) :

$$\bar{C} = \frac{1}{\sum_k n_k} \left(\sum_k n_k * \bar{c_k} \right)$$

Le cas de la procédure EVAR est plus complexe. En effet, les tarifs des GHM ont vocation à couvrir toutes les dépenses engagées au cours d'un séjour moyen (dont celles liées aux dispositifs médicaux). Cependant, les DMI les plus onéreux bénéficient d'un mode de tarification spécifique qui s'ajoute aux tarifs des GHM. Les DMI concernés, dont font partie les endoprothèses aortiques, sont inscrits sur une liste dite « liste en sus ». On considère ici que le tarif en sus couvre le coût du DMI concerné. Le coût moyen issu de l'ENCC inclut un poste de charges directes pour les DMI facturables en sus. Les GHM sélectionnés n'étant pas spécifiques d'EVAR, on a choisi de retrancher ce poste au coût moyen et de lui substituer une estimation du coût d'un DMI EVAR.

II.B.4 Caractérisation de l'incertitude entourant les procédures recouvrant plusieurs GHM

La caractérisation de l'incertitude entourant le coût moyen de chacun des GHM n'est pas triviale et nécessite de comprendre les modalités du calcul de $\sigma(\bar{c})$. Ces modalités, détaillées dans le guide technique des modalités de calcul de l'ENCC de l'ATIH, sont résumées dans l'Encadré 3.

Caractérisation de l'incertitude de l'estimation du coût des GHM dans le référentiel national des coûts

Le coût moyen d'un GHM est établi par régression multiple sur les variables dites de calage (âge, nombre d'actes CCAM, mode d'entrée, mode de sortie, etc.). La variance est calculée à partir des résidus de la régression multiple :

- Pour chaque établissement (noté i), j séjours sont échantillonnés et la somme des résidus de cet établissement est calculée :

$$U_i = \sum_j U_{ij}$$

- La moyenne de la somme des résidus par établissement est calculée pour les m établissements échantillonnés dans l'ENCC :

$$\overline{U} = \frac{1}{m}\sum_{i=1}^{m} U_i$$

- L'erreur de prédiction est la différence entre U_i et $\overline{U}$ et on peut calculer une variance des coûts par établissement dans l'échantillon :

$$Var_{ech}(c) = \frac{1}{m-1}\sum_{i=1}^{m}(U_i - \overline{U})^2$$

- De même, on peut estimer la variance du coût moyen (équivalent de l'écart standard de la moyenne) en multipliant par un facteur $\frac{1}{m}$:

$$Var_{ech}(\bar{c}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}(U_i - \overline{U})^2$$

Ces valeurs ont néanmoins été obtenues dans le cadre de l'échantillonnage des m établissements participant à l'ENCC (sur un total de M établissements en France). L'erreur de prédiction concerne en réalité les établissements non échantillonnés.

- On doit tout d'abord repasser à la variance des coûts sur l'ensemble des établissements (en multipliant $Var_{ech}(\bar{c})$ par M. Puis on limite l'application de l'erreur de prédiction aux $\left(1 - \frac{m}{M}\right) * M$ établissements n'ayant pas été échantillonnés. Le calcul est le suivant :

$$\left(1 - \frac{m}{M}\right) * M * M * \left(Var_{ech}(\bar{c})\right) = \left(1 - \frac{m}{M}\right) * \frac{M^2}{m(m-1)}\sum_{i=1}^{m}(U_i - \overline{U})^2$$

- On s'intéresse à la variance de l'estimation du coût moyen des N séjours du GHM sur l'ensemble de la France qui est donc :

$$Var(\bar{c}) = \frac{1}{N^2}\left(\left(1 - \frac{m}{M}\right) * \frac{M^2}{m(m-1)}\sum_{i=1}^{m}(U_i - \overline{U})^2\right)$$

Le calcul est en réalité plus complexe car il prend en compte l'hétérogénéité des coûts liée à la différence de type d'établissement. Pour le calcul de la variance estimée du coût moyen, on considère que les coûts sont censés être homogènes au sein d'un même type d'établissement, mais qu'ils peuvent être hétérogènes pour deux établissements de types différents (ex : CHU et CH). On calcule donc les résidus moyens par type d'établissement.

- Au sein d'un même type d'établissement (noté h), la moyenne de la somme des résidus par établissement est calculée pour les m_h établissements de type h :

$$\overline{U_h} = \frac{1}{m_h}\sum_{i=1}^{m_h} U_i$$

- La formule de la variance de l'estimation du coût moyen est ensuite appliquée pour chacun des types d'établissements. Pour un type donné, m_h établissements ont été échantillonnés parmi les M_h établissements de même type :

$$Var(\bar{c}) = \frac{1}{N^2}\left(\sum_h \left(\left(1 - \frac{m_h}{M_h}\right) * \frac{M_h^2}{m_h(m_h-1)} \sum_{i=1}^{m_h} (U_i - \overline{U_h})^2\right)\right)$$

Encadré 3 : Calculs de précision dans le guide technique des modalités de calcul de l'ENCC (Agence Technique de l'Information sur l'Hospitalisation).

Nous disposons donc de la variance des coûts de chaque GHM $Var(c) = N * Var(\bar{c})$. Nous considérons la procédure comme un mélange de k séries statistiques, chacune correspondant à un GHM (pour la procédure OPEN, $k = \{1, \dots, 4\}$). Le calcul dans l'ENCC peut être redressé en utilisant le nombre de séjours OPEN dans chacun des GHM.

- La variance de ce mélange pourrait être calculée ainsi :

$$\sigma^2(c_{OPEN}) = \frac{1}{\sum_k n_k} \sum_k n_k (Var(c_k) + (\bar{C} - \overline{c_k})^2)$$

- Et l'erreur standard serait la suivante :

$$\boldsymbol{\sigma(\overline{c_{OPEN}}) = \sqrt{\frac{1}{(\sum_k n_k)^2} \sum_k n_k (Var(c_k) + (\overline{C} - \overline{c_k})^2)}}$$

Concernant la procédure EVAR, le même calcul est réalisé sur le coût des GHM amputés du coût du DMI en sus. Le prix moyen du DMI (endoprothèse aortique) étant quant à lui connu, il ne fait pas l'objet d'un calcul d'incertitude. Il est ajouté a postériori. On procède de même concernant les GHM de réinterventions.

II.C Application aux procédure EVAR et OPEN

II.C.1 Abord endovasculaire (procédure EVAR)

La procédure EVAR est répartie sur trois racines de GHM (05C10, 05C11 et 05K06) composées de 13 GHM. Les données fournies par le référentiel national de coûts sont résumées dans le Tableau 12 ci-dessus. Le nombre de procédures EVAR correspond au nombre de GHM contenant au moins l'un des actes sélectionnés (cf. Tableau 9).

Tableau 22 : Référentiel ENCC v11b (2008) pour les GHM 05C10 (Chirurgie majeure de revascularisation, niveaux 1 à 4), 05C11 (Autres interventions de chirurgie vasculaire, niveaux 1 à 4) et 05K06 (Endoprothèses vasculaires sans infarctus du myocarde, niveaux 1 à 4 et très courte durée). Le nombre de chirurgies EVAR (n) n'est pas présent

dans le référentiel mais est calculé à partir du nombre d'actes de pose d'endoprothèses aortiques dans une extraction de la base nationale du PMSI. Les valeurs sont arrondies pour clarifier la présentation.

k	GHM	Nb France (N)	Nb ENCC (N_{ENCC})	Nb EVAR (n)	Coût ($\bar{c}$)	$\sigma(\bar{c})$	ERE (%)	$IC_{95\%}$
1	05C101	6065	1500	207	8322	605	7.27	[7112 ; 9532]
2	05C102	4336	1249	427	11294	920	8.15	[9453 ; 13133]
3	05C103	2478	677	107	17132	1402	8.18	[14327 ; 19936]
4	05C104	1182	310	37	28191	1319	4.68	[25552 ; 30829]
5	05C111	3634	882	222	5873	414	7.05	[5045 ; 6701]
6	05C112	1787	481	200	9331	626	6.71	[8078 ; 10584]
7	05C113	979	223	54	15479	744	4.81	[13991 ; 16967]
8	05C114	354	82	17	25431	1181	4.65	[23068 ; 27793]
9	05K061	36011	8710	745	4078	267	6.54	[3544 ; 4611]
10	05K062	9599	2513	853	6651	487	7.33	[5675 ; 7625]
11	05K063	1530	395	360	11457	975	8.51	[9505 ; 13407]
12	05K064	261	76	79	18352	1552	8.45	[15248 ; 21455]
13	05K06T	6539	1991	4	3028	413	13.62	[2203 ; 3853]

Nous avons établi le coût moyen d'une endoprothèse à 5447€ sur la base des trois actes de pose et de leur proportion respective dans le case-mix EVAR. Le détail du calcul de coût est présenté dans le Tableau 13. Pour le calcul du coût moyen d'EVAR, nous avons donc déduit les charges directes pour les DMI facturables en sus puis ajouté le tarif moyen d'un DMI EVAR.

Tableau 23 : Coût d'une endoprothèse aortique. Le coût du corps, d'une extension/jambage est issu de la LPP 2008.

Acte CCAM	Libellé	Corps	Ext./Jamb.	Total	Proportion
DGLF001	Pose d'endoprothèse couverte bifurquée aortobisiliaque, par voie artérielle transcutanée	3577€	2254€	5831€	72%
DGLF002	Pose d'endoprothèse couverte aorto-uniiliaque, par voie artérielle transcutanée	3577€	1127€	4704€	22%

DGLF005	Pose d'endoprothèse couverte rectiligne dans l'aorte abdominale infrarénale, par voie artérielle transcutanée	3577€	0€	3577€	6%

II.C.2 Chirurgie ouverte (procédure OPEN)

La procédure OPEN est répartie sur une racine de GHM (05C10) composée de 4 GHM (Chirurgie majeure de revascularisation, niveaux 1 à 4). Les données fournies par le référentiel national de coûts sont résumées dans le Tableau 12 ci-dessus. Le nombre de chirurgies OPEN correspond au nombre de GHM contenant au moins l'un des actes sélectionnés (cf. Tableau 9).

Tableau 24 : Référentiel ENCC v11b (2008) pour les GHM 05C10 (Chirurgie majeure de revascularisation, niveaux 1 à 4). Le nombre de chirurgies OPEN (n) n'est pas présent dans le référentiel mais est calculé à partir du nombre d'actes de chirurgie ouverte d'anévrisme de l'aorte abdominale sous-rénale dans une extraction de la base nationale du PMSI. Les valeurs sont arrondies pour clarifier la présentation.

k	GHM	Nb France (N)	Nb ENCC (N_{ENCC})	Nb OPEN (n)	Coût ($\bar{c}$)	$\sigma(\bar{c})$	ERE (%)	$IC_{95\%}$
1	05C101	6065	1500	985	8322	605	7.27	[7112 ; 9532]
2	05C102	4336	1249	1030	11293	920	8.15	[9453 ; 13133]
3	05C103	2478	677	473	17131	1402	8.18	[14327 ; 19936]
4	05C104	1182	310	238	28191	1319	4.68	[25552 ; 30829]

II.C.3 Résultats

L'exploitation des données du référentiel national des coûts de l'ENCC aboutit aux résultats présentés dans le Tableau 14. Pour les conversions sur table (endovasculaire vers chirurgie ouverte), le PMSI ne permet pas de les détecter puisqu'un seul acte est pris en compte pour le groupage et qu'il n'existe ni acte ni diagnostic spécifique à ce cas de figure. Nous avons donc utilisé la somme des coûts d'OPEN et d'EVAR comme approximation. La variance de cette somme est estimée comme la somme des variances.

Nous avons évalué le coût moyen d'une réintervention à partir de l'extraction des séjours des patients pour les 4 années suivant l'intervention initiale. Seuls les GHM en rapport avec une

intervention de l'aorte étaient considérés. Le coût moyen et l'erreur standard étaient calculés selon la même méthode que précédemment.

Enfin le coût moyen d'une visite de suivi était la somme des coûts (en 2008) d'un scanner abdomino-pelvien (96.27€) et d'un acte de consultation de spécialiste (25€), à savoir **121.27€**.

Tableau 25 : Coûts moyens et Erreur Standard des procédure EVAR et OPEN issus du référentiel national de coûts 2008. Les valeurs sont arrondies pour clarifier la présentation.

Procédure	Cout moyen	Erreur Standard
EVAR	12 748€	824€
OPEN	12 708€	1 091€
Conversion	25 456€	1 915€
Réintervention	8 228€	2 114€

II.D Analyse reproductible

```
library(tidyverse)
library(readxl)
```

II.D.1 Extractions PMSI des séjours index d'hospitalisation (CCAM + GHM)

II.D.1.a Chargement des tableaux

```
index_open <- read.csv2(file = "donnees/I_OPEN_ANO.csv", as.is = T) %>% as_tibble()

index_evar <- read.csv2(file = "donnees/I_EVAR_ANO.csv", as.is = T) %>% as_tibble() %
>%  select(-id_conversion)
```

II.D.1.b Premières lignes des tableaux

II.D.1.c EVAR

```
knitr::kable(head(select(index_evar, -ghm_lib, -acte_lib)))
```

id_patient	age	sexe	ghm	acte
1	53	1	05C101	DGLF0010
2	79	1	05K061	DGLF0010
3	73	1	05K061	DGLF0020
4	90	1	05K063	DGLF0020
5	69	1	05C101	DGLF0020
6	65	1	05C111	DGLF0050

II.D.1.d OPEN

```
knitr::kable(head(select(index_open, -ghm_lib, -acte_lib)))
```

id_patient	age	sexe	ghm	acte
1	70	1	05C104	DGPA0050
2	80	1	05C102	DGPA0050
3	62	1	05C101	DGPA0050
4	68	1	05C103	DGPA0050
5	75	1	05C101	DGPA0120
6	60	1	05C101	DGPA0120

II.D.1.e Actes CCAM retenus

II.D.1.f EVAR

```
index_evar %>%
  group_by(acte, acte_lib) %>%
  summarise(n = n()) %>%
  select(-n) %>%
  knitr::kable()
```

acte	acte_lib
DGLF0010	Pose d'endoprothèse couverte bifurquée aortobisiliaque, par voie artérielle transcutanée
DGLF0020	Pose d'endoprothèse couverte aorto-uniiliaque, par voie artérielle transcutanée
DGLF0050	Pose d'endoprothèse couverte rectiligne dans l'aorte abdominale infrarénale, par voie artérielle transcutanée

II.D.1.g OPEN

```
index_open %>%
  group_by(acte, acte_lib) %>%
  summarise(n = n()) %>%
  select(-n) %>%
  knitr::kable()
```

acte	acte_lib
DGPA0050	Mise à plat d'un anévrisme aortique infrarénal non rompu avec remplacement prothétique aortoaortique infrarénal, par laparotomie avec clampage infrarénal
DGPA0100	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque non rompu avec remplacement prothétique aortobifémoral, par laparotomie avec clampage infrarénal

DGPA0120	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque non rompu avec remplacement prothétique aortobisiliaque, par laparotomie avec clampage infrarénal
DGPA0160	Mise à plat d'un anévrisme aorto-ilio-fémoral avec remplacement prothétique bifurqué aorto-ilio-fémoral, par laparotomie avec clampage infrarénal
DGPA0180	Mise à plat d'un anévrisme aortique infrarénal ou aortobisiliaque rompu avec remplacement prothétique, par laparotomie
EDPA0050	Mise à plat d'un anévrisme iliaque avec remplacement prothétique aorto-iliaque ou aortofémoral unilatéral, par laparotomie

II.D.1.h ENCC 2008

```
ghm <- read_excel(path = "donnees/ENCC V11b.xlsx") %>%
  select(ghm = `GHM V11`,
         nb_france = `Nombre de séjours National 2008`,
         nb_encc = `Nombre de séjours ENCC 2008`,
         cout = `Coût moyen du GHM`,
         cout_dmi_sus = cout_dmi_sus,
         se = `Ecart type`,
         ERE = `ERE (en %)`,
         IC_bas = `Borne basse de l'intervalle de confiance`,
         IC_haut = `Borne haute de l'intervalle de confiance`) %>%
  mutate(se = as.numeric(se),
         ERE = as.numeric(ERE),
         IC_bas = as.numeric(IC_bas),
         IC_haut = as.numeric(IC_haut))

knitr::kable(head(ghm))
```

II.D.2 Calcul des coûts

II.D.2.a EVAR

II.D.2.a.i Case-Mix

```
index_evar %>%
  left_join(ghm,by = "ghm") %>%
  group_by(ghm, ghm_lib) %>% summarize(n = n()) %>%
  knitr::kable()
```

ghm	ghm_lib	n
05C101	Chirurgie majeure de revascularisation, niveau 1	207
05C102	Chirurgie majeure de revascularisation, niveau 2	427
05C103	Chirurgie majeure de revascularisation, niveau 3	107
05C104	Chirurgie majeure de revascularisation, niveau 4	37
05C111	Autres interventions de chirurgie vasculaire, niveau 1	222
05C112	Autres interventions de chirurgie vasculaire, niveau 2	200

05C113	Autres interventions de chirurgie vasculaire, niveau 3	54
05C114	Autres interventions de chirurgie vasculaire, niveau 4	17
05K061	Endoprothèses vasculaires sans infarctus du myocarde, niveau 1	745
05K062	Endoprothèses vasculaires sans infarctus du myocarde, niveau 2	853
05K063	Endoprothèses vasculaires sans infarctus du myocarde, niveau 3	360
05K064	Endoprothèses vasculaires sans infarctus du myocarde, niveau 4	79
05K06T	Endoprothèses vasculaires sans infarctus du myocarde, très courte durée	4

II.D.2.a.ii ENCC

```
index_evar %>%
  group_by(ghm) %>% summarize(n = n()) %>%
  left_join(ghm,by = "ghm") %>%
  knitr::kable()
```

ghm	n	nb_france	nb_encc	cout	cout_dmi_sus	se	ERE	IC_bas	IC_haut
05C101	207	6065	1500	8322.470	747.62	605.08	7.2704	7112.28	9532.59
05C102	427	4336	1249	11293.697	1012.10	920.08	8.1469	9453.53	13133.86
05C103	107	2478	677	17131.816	928.86	1402.18	8.1847	14327.46	19936.19
05C104	37	1182	310	28191.140	913.23	1319.14	4.6793	25552.83	30829.39
05C111	222	3634	882	5873.361	457.27	413.89	7.0469	5045.60	6701.16
05C112	200	1787	481	9331.269	623.81	626.47	6.7137	8078.34	10584.22
05C113	54	979	223	15479.446	587.21	744.07	4.8068	13991.33	16967.60
05C114	17	354	82	25430.598	450.43	1181.27	4.6451	23068.07	27793.15
05K061	745	36011	8710	4077.655	1350.61	266.78	6.5424	3544.12	4611.23
05K062	853	9599	2513	6650.608	1509.77	487.45	7.3294	5675.67	7625.45
05K063	360	1530	395	11456.583	1607.77	975.48	8.5146	9505.62	13407.55
05K064	79	261	76	18352.159	1440.29	1551.81	8.4558	15248.53	21455.79
05K06T	4	6539	1991	3028.126	1353.11	412.51	13.6225	2203.10	3853.12

II.D.2.a.iii Coût moyen et Ecart standard

```
cout_dmi_evar <- 5447.148874

cout_moyen_evar <- index_evar %>%
  left_join(ghm,by = "ghm") %>%
  transmute(cout = cout - cout_dmi_sus) %>%
  summarize(cout = mean(cout)) %>%
  .$cout

couts_evar <- index_evar %>%
  group_by(ghm) %>%
  summarize(n = n()) %>%
  left_join(ghm,by = "ghm") %>%
```

```
  mutate(cout_ss_dmi = cout-cout_dmi_sus) %>%
  mutate(sd = se*sqrt(nb_france)) %>%
  mutate(somme = n*(sd^2+(cout_moyen_evar-cout_ss_dmi)^2))

se_evar <- sqrt(1/(sum(couts_evar$n)^2)*(sum(couts_evar$somme)))
```

II.D.2.b OPEN

II.D.2.b.i Case-Mix

```
index_open %>%
  left_join(ghm,by = "ghm") %>%
  group_by(ghm, ghm_lib) %>% summarize(n = n()) %>%
  knitr::kable()
```

ghm	ghm_lib	n
05C101	Chirurgie majeure de revascularisation, niveau 1	985
05C102	Chirurgie majeure de revascularisation, niveau 2	1030
05C103	Chirurgie majeure de revascularisation, niveau 3	473
05C104	Chirurgie majeure de revascularisation, niveau 4	238

II.D.2.b.ii ENCC

```
index_open %>%
  group_by(ghm) %>% summarize(n = n()) %>%
  left_join(ghm,by = "ghm") %>%
  knitr::kable()
```

ghm	n	nb_france	nb_encc	cout	cout_dmi_sus	se	ERE	IC_bas	IC_haut
05C101	985	6065	1500	8322.47	747.62	605.08	7.2704	7112.28	9532.59
05C102	1030	4336	1249	11293.70	1012.10	920.08	8.1469	9453.53	13133.86
05C103	473	2478	677	17131.82	928.86	1402.18	8.1847	14327.46	19936.19
05C104	238	1182	310	28191.14	913.23	1319.14	4.6793	25552.83	30829.39

II.D.2.b.iii Coût moyen et Ecart standard

```
cout_moyen_open <- index_open %>%
  left_join(ghm,by = "ghm") %>%
  transmute(cout = cout) %>%
  summarize(cout = mean(cout)) %>%
  .$cout

couts_open <- index_open %>%
  group_by(ghm) %>%
  summarize(n = n()) %>%
  left_join(ghm,by = "ghm") %>%
  mutate(sd = se*sqrt(nb_france)) %>%
  mutate(somme = n*(sd^2+(cout_moyen_open-cout)^2))

se_open <- sqrt(1/(sum(couts_open$n)^2)*(sum(couts_open$somme)))
```

II.D.2.c Réintervention

II.D.2.c.i ENCC

```
reint <- bind_rows(
  read.csv2(file = "donnees/R_EVAR_ANO.csv",as.is = T),
  read.csv2(file = "donnees/R_OPEN_ANO.csv",as.is = T)) %>%
  as_tibble() %>%
    filter(str_detect(ghm_lib,paste(collapse = '|', c(
    "Endoprothèses vasculaires sans infarctus du myocarde",
    "Chirurgie majeure de revascularisation",
    "Actes thérapeutiques par voie vasculaire sauf endoprothèses, âge supérieur à 17
ans",
    "Autres interventions de chirurgie vasculaire",
    "Actes thérapeutiques sur les artères par voie vasculaire, âge supérieur à 17 ans
",
    "Autres interventions sur le système circulatoire",
    "Autres interventions pour blessures ou complications d'acte")))) %>%
  filter(delai < 365*4) %>%
  mutate(delai = ceiling(delai/365)) %>%
  left_join(ghm,by = "ghm") %>%
  group_by(id_patient,id_sejour) %>%
  distinct(id_sejour,.keep_all = TRUE)
```

II.D.2.c.ii Coût moyen et Ecart standard

```
cout_moyen_reint <- mean(reint$cout, na.rm = TRUE)

couts_reint <- reint %>%
  group_by(ghm) %>%
  summarize(n = n()) %>%
  left_join(ghm,by = "ghm") %>%
  mutate(sd = se*sqrt(nb_france)) %>%
  na.omit() %>%
  mutate(somme = n*(sd^2+(cout_moyen_reint-cout)^2))

se_reint <- sqrt(1/(sum(couts_reint$n)^2)*(sum(couts_reint$somme)))
```

II.D.3 Synthèse

```
tibble(
  `Procédure` = c("EVAR","OPEN", "Conversion", "Réintervention"),
  `Cout moyen` = c(cout_moyen_evar+ cout_dmi_evar, cout_moyen_open, (cout_moyen_evar
+ cout_dmi_evar + cout_moyen_open),cout_moyen_reint),
  `Erreur Standard` = c(se_evar, se_open,(se_evar + se_open),se_reint)) %>%
  knitr::kable()
```

Procédure	Cout moyen	Erreur Standard
EVAR	12747.582	823.5443
OPEN	12708.357	1091.2594
Conversion	25455.939	1914.8038
Réintervention	8227.678	2114.2143