POLITECNICO
DI TORINO

Dipartimento
di Scienze Matematiche

DOCTORAL THESIS

Quantitative decision-making in drug development

Author:
Gaelle SAINT-HILARY

Supervisor:
Prof. Mauro GASPARINI
Scientific advisor:
Veronique ROBERT

POLITECNICO DI TORINO
Dipartimento di Scienze Matematiche "Giuseppe Luigi Lagrange"

16 July 2018

POLITECNICO DI TORINO
Dipartimento di Scienze Matematiche "Giuseppe Luigi Lagrange"

Abstract

Quantitative decision-making in drug development

by Gaelle SAINT-HILARY

The drug life cycle management is a long and continuous process, involving complex and critical decisions. Examples of strategic decisions would be continuing, terminating or expanding the development, dose(s) selection, choice of the targeted populations etc. New data and information coming from diverse sources collected throughout the development must be integrated with relevant previous information to inform and support decision-making.

Drug development teams need tools for structuring and analyzing decisions, with transparent processes that synthetize the whole available information to evaluate the probability of success and the risks associated to different options, in order to permit informed trade-offs between them.

A wide range of statistical methods could be used to develop optimal tools for decision analyses in drug development. This report focuses on two research axes: the drug benefit-risk assessment and the predictive probability of success.

Four methodologies are presented in this report: first, an extension of existing methodologies for drug benefit-risk assessment; second, a new tool for drug benefit-risk assessment that addresses the issues of the existing methods; third, a method to compute the predictive probability of a composite success including benefit-risk considerations; and fourth, how to predict the success of a future trial from data on surrogate endpoints.

These methodologies are valuable quantitative tools to support decision-making in the pharmaceutical development. They have strong theoretical foundations and were shown to have soundness in the context of healthcare decisions. They are simple, could be used in a wide range of applications throughout the drug life-cycle, from early development to post-marketing surveillance, and their utility has already been demonstrated in very concrete situations.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Mauro Gasparini for his guidance during my PhD study and related research, for his scientific rigor that considerably improved mine, for being accommodating but not lenient, for introducing me to many people and for all the collaborations that resulted from it. I also would like to thank him and his family for their warm welcome in Italy.

My sincere thanks also go to Veronique Robert, my scientific advisor at the Institut de Recherches Internationales Servier (IRIS), for her continuous support, her encouragements and all her contributions. The enthusiasm she had for my research was contagious and motivational for me. I could not have imagined having a better mentor.

I am especially grateful to Maylis Coste, head of the Biostatistics group at IRIS, who permitted this project to be possible. She showed a continued interest in my research, proposed stimulating challenges and greatly contributed to the visibility of my work.

My sincere thanks also go to my other co-authors, Pavel Mozgunov, Stephanie Cadour, Valentine Barboux, Matthieu Pannaux, Gianluca Mastrantonio and Thomas Jaki. Without their precious support it would not have been possible to conduct this research.

Lastly, I would like to thank my students, the clinicians I collaborated with, my friends and my family who contributed, even indirectly, to my professional or personal achievements during these three years.

Scientific production

Publications

Statistical publications

- **Saint-Hilary G**, Cadour S, Robert V, and Gasparini M. A simple way to unify multicriteria decision analysis (MCDA) and stochastic multicriteria acceptability analysis (SMAA) using a Dirichlet distribution in benefit-risk assessment. *Biometrical Journal*, 59(3):567–578, 2017. DOI: 10.1002/bimj.201600113. [112]

- **Saint-Hilary G**, Robert V, Gasparini M. Decision making in drug development using a composite definition of success. *Pharmaceutical Statistics*, online 2018:1-15. DOI: 10.1002/pst.1870. [110]

- **Saint-Hilary G**, Robert V, Gasparini M, Jaki T and Mozgunov P. A novel measure of drug benefit-risk assessment based on Scale Loss Score (SLoS). *Statistical Methods in Medical Research*, accepted 11 June 2018. [111]

- **Saint-Hilary G**, Barboux V, Pannaux M, Gasparini M, Robert V, Mastrantonio G. Predictive probability of success using surrogate endpoints. Submitted to *Statistics in Medicine*, major revisions.

Clinical publications

- D'Ascenzo F, Iannaccone M, **Saint-Hilary G**, Bertaina M, Schulz-Schupke S, (...) Gasparini M, Gaita F. Impact of design of coronary stents and length of dual antiplatelet therapies on ischaemic and bleeding events: a network meta-analysis of 64 randomized controlled trials and 102 735 patients. *European Heart Journal*, 38(42):3160-3172, 2017. DOI: 10.1093/eurheartj/ehx437. [27]

- Burrello J, Erhardt EM, **Saint-Hilary G**, Veglio F, Rabbia F, Mulatero P, Monticone S, D'Ascenzo F. Pharmacological treatment of arterial hypertension in children and adolescents: a network meta-analysis. *Hypertension*, 72(2):306-313, 2018. DOI: 10.1161/HYPERTENSIONAHA.118.10862. [11]

- Williams TA, Burrello J, (...), **Saint-Hilary G**, (...), Mulatero P, Reincke M. Computed tomography and adrenal venous sampling in the diagnosis and outcomes after adrenalectomy for unilateral primary aldosteronism: a retrospective international cohort study. *Hypertension*, online 2018. DOI: 10.1161/HYPERTENSIONAHA.118.11382. [143]

- Iannaccone M, **Saint-Hilary G**, Menardi D, (...), Gasparini M, Gaita F, D'Ascenzo F. Network meta-analysis of studies comparing closure devices for femoral access after percutaneous coronary intervention. *J Cardiovasc Med*. 19(10):586-596, 2018. DOI: 10.2459/JCM.0000000000000697. [63]

- Bertaina, M, Ferraro, I, Omedè, P, Conrotto, F, **Saint-Hilary, G**, (...), D'Ascenzo, F. Meta-analysis Comparing Complete or Culprit Only Revascularization in Patients With Multivessel Disease Presenting With Cardiogenic Shock. *The American Journal of Cardiology*, online 2018. DOI: 10.1016/j.amjcard.2018.08.003. [7]

- + 1 clinical manuscript submitted, under review.

Conference abstracts (see list of conferences below)

1. **Saint-Hilary G**, Robert V, and Gasparini M. Decision-making in drug development using a composite definition of success. 2017.

2. Cadour S, **Saint-Hilary G**, Robert V, and Gasparini M. An improved model for drug benefit-risk assessment: Dirichlet SMAA. 2017.

3. **Saint-Hilary G**, Barboux V, Pannaux M, Robert V, Gasparini M and Mastrantonio G. Predictive probability of success using surrogate endpoints. 2018.

4. **Saint-Hilary G**, Robert V, Gasparini M, Jaki T and Mozgunov P. A novel measure of drug benefit-risk assessment based on Scale Loss Score (SLoS). 2018.

Participation to conferences

- **PSI conference 2017** (Statisticians of the Pharmaceutical Industry). London, UK, May 2017. Oral presentation 1 and Poster 2.

- **ISCB conference 2017** (International Society for Clinical Biostatistics). Vigo, Spain, July 2017. Oral presentation 1 and Poster 2.

- **ISBS-CEN conference 2017** (jointly International Society for Biopharmaceutical Statistics and Central European Network). Vienna, Austria, August 2017. Oral presentation 1 and Poster 2.

- **SMB conference 2017** (Statistical Methods in Biopharmacy). Paris, France, September 2017. Oral presentation 1 and Poster 2.

- **PSI conference 2018** (Statisticians of the Pharmaceutical Industry). Amsterdam, Netherlands, June 2018. Oral presentation 3 and Poster 4.

- **IBC conference 2018** (International Biometric Conference). Barcelona, Spain, July 2018. Poster 4.

- **ISPOR Europe conference 2018** (International Society for Pharmacoeconomics and Outcomes Research). Barcelona, Spain, November 2018. Poster 4.

Seminars, webinars, workshops and training events

- Training *"Multiplicity issues in clinical trials"* at **Institut de Recherches Internationales Servier**, Suresnes, France, January 2016 and July 2017.

- Seminar *"Balancing benefits and risks of medicines"* at **Politecnico di Torino**, Turin, Italy, February 2016.

- Oral presentation *"Quantitative Decision-Making in Drug Development"* at the **Societe Francaise de Statistiques**, Journée Nationale Biopharmacie & Sante, Paris, France, November 2016.
- Seminar *"Quantitative benefit-risk assessment for decision-making in drug development"* at **AstraZeneca** and **Chalmers University**, Goteburg, Sweden, April 2017.
- Training *"Predictive probability of success"* at **Institut de Recherches Internationales Servier**, Suresnes, France, September 2017.
- Seminar *"The composite success"* at **Lancaster University**, Lancaster, UK, November 2017.
- Webinar *"Quantitative benefit-risk assessment using MultiCriteria Decision Analysis (MCDA) and its extensions: practical application"* for the **European Federation of Statisticians from the Pharmaceutical Industry (EFSPI)**, March 2018.
- Workshop *"Learn How to Swing: Hands-on workshop on patient preference elicitation in the age of personalised medicine"*, **PSI conference**, Amsterdam, Netherlands, June 2018.
- Training *"Futility analyses in clinical trials"* at **Institut de Recherches Internationales Servier**, Suresnes, France, June 2018.
- Oral presentation *"Experience sharing on the EFSPI Special Interest Group on Quantitative Decision-Making"* at the **EFSPI Statistical Leaders Meeting** (top managers of statistical departments from pharmaceutical industries in Europe), Louvain-la-Neuve, Belgium, July 2018.
- Seminar *"Predictive probability of success using surrogate endpoints"* at **Roche**, Welwyn, UK, September 2018.

Other activities

Co-chair of a **Special Interest Group (SIG)** of the **European Federation of Statisticians from the Pharmaceutical Industry (EFSPI)** on Quantitative Decision-Making, since October 2017.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Decision-making in drug development

The drug development is a long and continuous process, involving complex and critical decisions. The Figure 1.1 presents a simplified example of decision-making in a clinical drug development. The decisions, represented by the circles, are mainly based on the accumulated data collected on the considered drug, the portfolio and the resources of the company, and the competitors. Surprisingly, the pharmaceutical industry is far behind other major industries when it comes to use quantitative methods to support decision making, and there is a need for more focus on quantitative decision-making based on measurable parameters for that purpose. New data and information coming from diverse sources throughout the drug life cycle must be integrated with relevant previous information to inform and support decision-making.

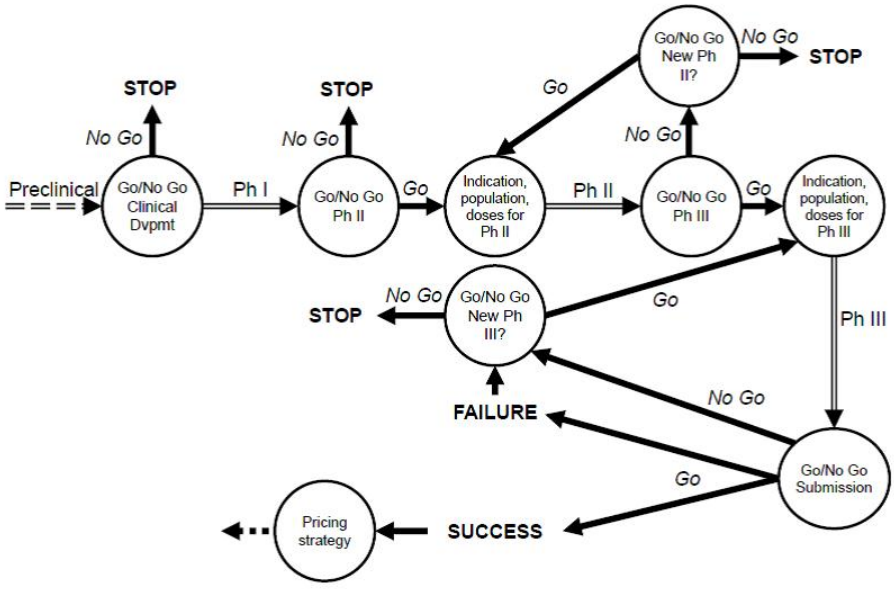

FIGURE 1.1: Example of decision-making in clinical drug development

For example, a quantitative decision tool for a Go/no Go Phase III would typically be the predictive probability of success of phase III studies based on results from phases II. The definition of success is critical here and generally includes efficacy considerations, but other criteria can be introduced such as the safety of the drug, its

overall benefit-risk balance, the time expected to complete the clinical development, the portfolio management or the budget.

The success of a submission being the market authorization granted by the health authorities, which is not quantifiable, the decision tools used for the Go/no Go Submission will mainly be based on the overall benefit-risk assessment of the drug and the clinical relevance of the available evidence. Quantitative methods will help in understanding how this evidence is expected to be recognized by the health authorities, and later by the scientific community and the prescribing physicians.

1.2 Who are the decision-makers in drug development?

The Figure 1.2 presents the hierarchy of decision-makers in drug development.

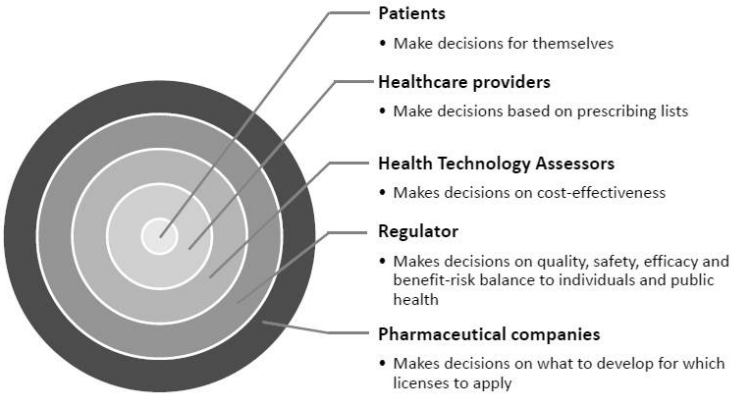

FIGURE 1.2: Decision makers – who are they?
Source: http://protectbenefitrisk.eu/

Patients are the ultimate decision-makers in health care: they make the choice to take or not the drug. They are diagnosed and advised by their physicians, who play a critical role in evaluating the benefits and risks of a therapy at an individual level.

Regulators and Health Technology Assessors (payers) assess the value of a drug at the population level, determine public policies and make recommendations to patients and physicians.

Finally, pharmaceutical companies make decisions on their pipeline and development strategies.They focus their research on diseases or therapies according to the medical need, their scientific experience, their ability to reach the patient population, and thus the economic value of their compounds.

Patients involvement in healthcare decisions is increasingly encouraged [114]. Several initiatives emerged to 'find out when and where patients want, can, and should be involved in drug development' [68]. Among them, we can cite the workshop on 'Advancing Use of Patient Preference Information as Scientific Evidence in Medical Product Evaluation' hosted in December 2017 by the Centers of Excellence in Regulatory Science and Innovation (CERSIs) and the Food and Drug Administration (FDA) [134], or the IMI-PREFER intiative [68, 53]. They both underline the need

to use and develop reliable methods to incorporate patients' preferences into drug benefit-risk assessments and decision-making processes.

1.3 Clinical development success rates

The low success rate and the high cost of developing a drug are major issues of the pharmaceutical industry. Despite large rewards when a new molecular entity gets a marketing authorization, an important part of the cost of developing a drug is due to the low success rate [29].

The reasons for this low success rate have been broadly explored in the literature [29, 32, 31, 58, 30, 144]. Apart from the benefit-risk balance of the drug itself (efficacy, safety, convenience of use etc.), some factors related to higher probabilities of drug development successes have been reported, among them: a short duration of development, a large number of patients enrolled in early phase trials, or a large-size company developing the compound. This illustrate the importance of reducing the uncertainty in early decisions, stopping unprofitable drug developments as early as possible, if the companies have the ability to focus and invest on other promising drugs in their pipeline.

The success rate varies across therapeutic areas [32, 31] and over time. Although a decrease has been observed before 2013, a positive trend is seen in recent years [144], which could be explained by a better identification of failures and earlier decisions to stop unpromising drugs. The increased use of biomarkers, and thus the development of targeted therapies, appear to be one major reason for this improvement, especially in the area of oncology [30, 22, 144].

Studying success rates, costs and durations of drug developments is important for improving pharmaceutical investments, and motivates researchers and industries to develop quantitative tools for better decision-making. Probabilities of success, predictive algorithms and portfolio simulations are critical when making scientific and economic decisions [22].

1.4 Three levels of decision

The decision-making process can be broken down into three levels: the "study level", the "development program level" and the "portfolio level". At the study level, decisions need to be made in particular regarding the population and the design of a trial (including the sample size, the number of arms, the choice of control treatments), given the information available from earlier studies and external data. For example, the doses to be tested in a phase II study can be selected based on their predicted efficacy and safety. The anticipated cost of a trial could be balanced against its predictive power. Statistical methods can also be used to assess whether the amount of information brought by the single trial, not only on the primary efficacy endpoint but on the overall benefit-risk of the drug, is sufficient to inform later decisions regarding the continuation of the development.

Then, at the development program level, there is a need to develop statistical methods to quantitatively assess the chances of success of an option, which could be the decision to continue or not the development, but also to develop the drug in one

or several indications, to target a population likely to respond to treatment etc. For example, these methods could be used to measure the joint probability of success of several subsequent clinical trials given the results of trials already conducted (e.g. success of phase III studies given the results of phase II and earlier studies).

Finally, if a drug development has a satisfactory chance of success and an acceptable risk, it would be worthwhile to extend the methods to decision-making across several drug developments, to integrate the preferences of the decision-maker in terms of portfolio management using a risk-based monitoring approach. New tools should also be developed.

1.5 Subjectivity in quantitative decision-making

Guidelines and good clinical practices recommend measures to minimize subjectivity in drug development [66, 67, 37], such as the choice of objective measurements as primary endpoints, the pre-specification of the statistical analysis and the collection of accurate, quality-checked data. However, the interpretation of these data and the decision resulting from their analysis necessarily require value judgments and a qualitative assessment: subjectivity is inherent to decision-making.

It is therefore important to make it transparent and consistent. Quantitative tools and processes are crucial in formalizing decision-making, providing strong, reproducible and explicit arguments to justify and communicate the decisions [22].

Being 'too subjective' is, however, one of the main criticisms of quantitative methods for decision-making [102, 113]. Indeed, introducing subjectivity is sometimes perceived as a failure in the reliability of a scientific process. Lawrence D. Phillips in [113, Chapter 5] provides a full discussion on the challenges faced by the use of explicit quantitative methods in benefit-risk assessment, and the advantages of overcoming these issues to enhance the decision-making process. He points out in particular that this objection comes from the difficulty to distinguish between objective data and their subjective interpretation, and that a benefit-risk assessment of drug cannot be performed by wholly objective means.

It may also be objected that clinical expertise cannot be reduced to numbers, and there is some skepticism in the ability of statistical methods to quantify expert opinion and value judgments [102]: human preferences ultimately incorporate some variability and arbitrariness. However, it is important to underline that quantitative tools are not intended to *replace* expert's decisions, but to support them. Clearly, both experience and judgment are required, but numbers may permit to express them in words, facilitating the communication. In particular, they may help in resolving disagreements by identifying underlying differences in the preferences of the decision-makers.

In summary, quantitative methodologies should be as transparent, robust and comprehensive as possible, but subjectivity cannot be avoided. It is important to ensure that all stakeholders understand their supporting purpose, their limitations and how they contribute to the decision-making process.

1.6 Data sources and data aggregation

Appropriately summarizing complex information from different sources is the first pillar of quantitative decision-making methodologies [3, 113]. The typical sources of evidence are clinical trials, epidemiological and health economics studies, public registries, and elicited stakeholders' preferences. Ideally, decisions should be based on an all relevant evidence, with an exhaustive collection of data and considering the validity of each data source, making explicit the associated uncertainties and assumptions that are part of decision-making.

The Cochrane Collaboration (`http://www.cochrane.org`) developed formalized processes to perform systematice reviews, in order to find pertinent data sources related to a particular question. Then, the quantitative analysis consists in aggregating the data using statistically rigorous methods, such as meta-analyses, in order to summarize the results [3, 142].

Some challenges arise from potential biases in the different sources of evidence and in the data collection [142, 141]. DiSantostefano *et al.* (2016) [33] recommends a set of principles on handling data from different sources to conduct benefit-risk assessment, including a thorough review of their limitations, and some recommendations on when it is appropriate or not to combine them to support decision-making . The robustness of the decisions to the selected data sources should also be carefully assessed, using for example sensitivity analyses.

1.7 Content of this report

A wide range of statistical methods could be used to develop optimal tools for decision analyses in drug development. This report focuses on two research axes: the drug benefit-risk assessment (Chapters 2 and 3) and the predictive probability of success (Chapters 4 and 5).

Each chapter is stand-alone and corresponds to an article published in or submitted to a statistical journal. Some background about the publication context is presented in the respective preambles.

Chapter 2 presents a simple unification and an extension of two existing methodologies for drug benefit-risk assessment. It has demonstrated its usefulness in concrete clinical applications, and the publication has already been cited by other authors working on benefit-risk assessment.

In Chapter 3, we present a novel, simple and valuable tool for drug benefit-risk assessment. It is based on strong theoretical principles, avoids the pitfalls of the existing methods and can lead to more meaningful conclusions. Although still quite new, this method generates a lot of interest in the scientific community and was awarded with the best poster prize at the PSI conference 2018 (Statisticians from the Pharmaceutical Industry).

The relationship between the two research axes is made in Chapter 4, where we propose a method to compute the predictive probability of a composite success which

includes benefit-risk considerations.

Finally, Chapter 5 presents a general, reliable and reproducible methodology to predict the success of a future trial from data on surrogate endpoints, in a way that makes the best use of all the available evidence.

Concluding remarks are provided in Chapter 6.

Chapter 2

Dirichlet Stochastic Multi-criteria Acceptability Analysis (Dirichlet SMAA)

Background

This chapter was published as:

Saint-Hilary G, Cadour S, Robert V and Gasparini M. A simple way to unify multi-criteria decision analysis (MCDA) and stochastic multicriteria acceptability analysis (SMAA) using a Dirichlet distribution in benefit-risk assessment. *Biometrical Journal*. 2017; **59(3)**:567-578. doi: 10.1002/bimj.201600113.

2.1 Introduction

Assessing the benefit-risk ratio of a treatment consists of balancing its favorable therapeutic effects versus the adverse reactions it may induce [21]. Obviously, only therapies with a favorable benefit-risk ratio should be considered, when the amount of benefits outweigh the amount of risks. The benefit-risk balance is therefore a strong predictor of the long-term viability of a therapy, and a key element for decision-making during the drug development, the regulatory approval process, and the post-marketing follow-up [41, 45, 136].
The assessment of the benefit risk is a complex process that requires the evaluation of a large amount of data coming from different sources. Still nowadays, it often consists of a qualitative description of the available evidence and of a discussion to establish whether the profile of the drug is favorable or not [35, 62]. However, structured qualitative frameworks and quantitative methodologies have been proposed to improve the assessment process [59, 92, 82, 133]. Even though quantitative analyses are not expected to replace qualitative judgements from experts, they can be useful for complex benefit-risk decision problems, and they can be used as tools to communicate them [90, 91].
In particular, Multi-Criteria Decision Analysis (MCDA) has been proposed to compare the benefit-risk of several drugs [93, 94, 98]. The European Medicine Agency Benefit-Risk Methodology Project suggested that MCDA is one of the most comprehensive among the quantitative methodologies [35, 36, 37, 38], and it is also recommended by the IMI PROTECT Work package 5 [69]. Its principle is to compare several treatments using utility scores calculated from multiple criteria of benefits and risks, taking into account their relative importance according to the preferences of the decision-makers. Although MCDA makes more explicit the drug benefit-risk

assessment, the scoring process of the treatments is deterministic and ignores the parameter uncertainty induced by the data sampling variation. Details are provided in Section 2.2.1, where the acronym dMCDA is used.
Therefore, some probabilistic models, often called Probabilistic MCDA (or Stochastic MCDA), were developed to take into account the uncertainty in the parameters. In particular, Waddingham et al. [138] propose a Bayesian MCDA model to estimate the score distributions based on the distributions of the criterion parameters, which are themselves estimated from the treatment effects observed in previous studies. This model is called pMCDA in Section 2.2.2.
The Stochastic Multicriteria Acceptability Analysis (SMAA in Section 2.2.3) is an extension of Probabilistic MCDA which considers the preferences of the decision-makers as another source of uncertainty. Instead of requiring the elicitation of exact weights for the criteria of benefits and risks, SMAA can be used with unknown or partially known preferences, and the distributions of the utility scores are estimated for all the possible combinations of weights [77, 130]. It allows for a more extended assessment of the drug benefit-risk ratios than Deterministic MCDA and Probabilistic MCDA, but it is often disregarded due to its increased complexity [90] or to the high degree of uncertainty in its results [138].
The aim of this paper is to build a simple unique model as a generalization of MCDA and SMAA, by applying a Dirichlet distribution to the weights of the criteria and by making its parameters vary. The Dirichlet distribution has a natural interpretation as the best-guess set of weights together with a precision parameter, which could be interpreted as the strength of confidence of the decision-makers in their elicitation of preferences. This model permits to explore all the possible ranges of confidence in the weight elicitation, from the deterministic weights in MCDA to the complete absence of information handled by SMAA.
In Section 2.2, we will introduce the models for Deterministic MCDA, Probabilistic MCDA and SMAA. The new model, called Dirichlet SMAA, is presented as a generalization of MCDA and SMAA in Section 2.3. It is applied in Section 2.4 on an example proposed by Tervonen et al. [131] of a published placebo-controlled trial in depression [96]. In this section, we will assess the impact of the precision of the Dirichlet distribution on the results, and we will also illustrate how the results from SMAA and Probabilistic MCDA can be retrieved from this unified model. A discussion and concluding remarks are given in Section 2.5.

2.2 Current models

2.2.1 Deterministic Multi-Criteria Decision Analysis (dMCDA)

Suppose m treatments $(i = 1, ..., m)$ are assessed on n criteria $(j = 1, ..., n)$. The model includes the following quantities and functions [93]:

(i) The performance of treatment i on criterion j is denoted by ξ_{ij}. Here, the ξ_{ij} are deterministic: they are usually taken as the point estimates observed in the clinical trials or from meta-analyses used to synthesize their results [3, 142, 141]. The vector of criterion measurements for treatment i is denoted by $\xi_i = (\xi_{i1}, ..., \xi_{in})$.

(ii) The partial value functions $u_j()$ are used to normalize the criterion measurements by mapping them into a 0 to 1 scale. These functions are not necessarily linear, but should be monotonically increasing with the preference on the considered criterion. Thus, $u_j(\xi_{ij}) > u_j(\xi_{i'j})$ indicates that the performance of the

treatment i is preferred to the performance of the treatment i' on criterion j, the worst value being $u_j(.) = 0$ and the best value being $u_j(.) = 1$. These partial value functions should be provided by the decision-makers to indicate the importance of a change on each criterion. For example, if ξ'_j and ξ''_j are considered respectively as the most preferable value and the least preferable value for ξ_{ij}, then the partial value functions could be defined as linear functions with (a) $u_j(\xi_{ij}) = \frac{\xi_{ij}-\xi'_j}{\xi''_j-\xi'_j}$ if an increase in ξ_{ij} represents an improvement on the criterion j and (b) $u_j(\xi_{ij}) = \frac{\xi'_j-\xi_{ij}}{\xi'_j-\xi''_j}$ if an increase in ξ_{ij} represents a worsening on the criterion j.

(iii) The weights indicating the relative importance of the criteria are denoted by w_j, with the constraint that $\sum_{j=1}^{n} w_j = 1$. Here, the w_j are deterministic: they should be provided by the decision-makers. The vector of weights used for the analysis is denoted by $w = (w_1, ..., w_n)$.

It is generally assumed that the criteria and their partial values functions are independent, which allows us to use an additive formula to calculate the global utility score:

$$u(\xi_i, w) = w_1 u_1(\xi_{i1}) + ... + w_n u_n(\xi_{in}) = \sum_{j=1}^{n} w_j u_j(\xi_{ij})$$

The utility score is the measure of benefit-risk, it permits to discriminate treatments according to their performances on the criteria of benefit and risk, and according to the weights attributed to these criteria. The treatment with the highest utility score is considered to be the treatment with the most preferable benefit-risk ratio.
The weights w_j and the performances ξ_{ij} being deterministic (implying that $u_j(\xi_{ij})$ are deterministic too), the results of the utility scores $u(\xi_i, w)$ for the different treatments are themselves deterministic, simply real numbers. The uncertainties in the treatment performances on the criteria and in the decision-makers' preferences are ignored.

2.2.2 Probabilistic Multi-Criteria Decision Analysis (pMCDA)

The probabilistic model for MCDA takes into account the data uncertainty. The notations and the formula for the utility score are the same as for dMCDA, but here the ξ_{ij} are random variables. Following the approach proposed by Waddingham et al. [138], we consider a Bayesian model and assign a probability distribution to the ξ_{ij} based on the data distributions observed in clinical trials or resulting from evidence synthesis of these trials.
Depending on the nature of the criteria, their underlying distribution will be different. For example, the proportion of patients achieving a clinical response will usually be distributed according to a Beta distribution $\xi_{ij} \sim Beta\left(a_{ij}, b_{ij}\right)$, the change from baseline on a measurement scale can be distributed according to a Normal distribution $\xi_{ij} \sim N\left(\mu_{ij}, \sigma^2_{ij}\right)$ etc. The parameters of these distributions are estimated from the data.
The utility scores $u(\xi_i, w)$ are now random variables, and their distributions can be obtained by simulating values from the distributions of the ξ_{ij}.
The benefit-risk assessment of the treatments can be assessed by comparing the distributions of their utility scores, using for example descriptive statistics and graphs, and by computing the probabilities to be the best treatment, the second best etc.

pMCDA is a major improvement on dMCDA since its permits to avoid declaring that a treatment has a better benefit-risk profile than another while the observed difference is only due to sampling error. On the other hand, the preferences of the decision-makers are still explicitly required to determine the partial value functions and the weights of the criteria.

2.2.3 Stochastic Multicriteria Acceptability Analysis (SMAA)

The SMAA method is an extension of the probabilistic MCDA model when no information, or partial information, is provided on the weights of the criteria by the decision-makers. The notations and the formula for the utility score are the same as for the previous models, the ξ_{ij} are random variables as for pMCDA, but here the weights w_j are also random variables. Typically, when no preference between the criteria could be elicitated, the space of weights W is defined as an $n-1$-dimensional simplex in n-dimensional space [130]:

$$W = \left\{ (w_1, ..., w_n) : w_i \geq 0 \; \forall i \text{ and } \sum_{i=1}^{n-1} w_i \leq 1 \text{ and } w_n = 1 - \sum_{i=1}^{n-1} w_i \right\} \tag{2.1}$$

Constraints could be applied to this space of weights to handle partial preferences of the decision-makers, for example:

(i) Applying lower and upper bounds to the weights:
$W' = \left\{ w \in W \quad | \quad w_j^{min} \leq w_j \leq w_j^{max}, \quad j = 1, ..., n \right\}$

(ii) Ranking the criteria according to their importance:
$W' = \left\{ w \in W \quad | \quad w_{j_1} \geq w_{j_2} \geq ... \geq w_{j_n} \right\}$

(iii) Enforce the overall benefits and the overall risks to have the same weight:
$W' = \left\{ w \in W \quad | \quad \sum_{j \in Benefits} w_j = \sum_{j \in Risks} w_j = 0.5 \right\}$

Again, the utility scores $u(\xi_i, w)$ are random variables, and now their distributions can be obtained by simulating values from the distributions of the ξ_{ij} and uniformly from the space of weights W (or a restricted space of weights W'). The distributions of the utility scores account not only for the sampling variation in the criterion measurements but also for the uncertainty in the decision-makers' preferences regarding the relative importance of the criteria.
It should be noted that the benefit-risk ratio of two treatments i and i' should only be compared when the same preferences are used for the weights. That is, we should not compare the distributions of two utility scores $u(\xi_i, w)$ and $u(\xi_{i'}, w)$ on the whole weight space, but only when using the same weight vector. Therefore, the distribution of the difference between two utility scores using the same weight vector can be computed by simulating values for w^* in the considered weight space:

$$\Delta u(\xi_i, \xi_{i'}, w^*) = u(\xi_i, w^*) - u(\xi_{i'}, w^*) \tag{2.2}$$

The benefit-risk ratios of the treatments can then be assessed by comparing these distributions using descriptive statistics and graphs. The probabilities to be the best treatment, the second best etc. can be computed. These probabilities are called *rank acceptability index* in the literature. The decision-makers can also make use of the *central weight vectors* (expected center of gravity of all possible weight vectors that rank the treatment at the first place) and their *confidence factors* (probability for a treatment

to obtain the first rank when the central weight vector is chosen) described for example in Tervonen *et al.*, 2011.
SMAA is a useful method to account for both the sampling variation and the lack of preferences to weight the criteria. Obviously, the results will have a lower degree of precision than with pMCDA, since the amount of uncertainty increased. The preferences of the decision-makers are still explicitly required to determine the partial value functions, but one could easily take into account their uncertainty by applying probability distributions to the supposed minimum and maximum values ξ_j' and ξ_j'', or in a more complex way to use classes of functions for the $u_j(\xi_{ij})$.

2.3 General model: Dirichlet SMAA

2.3.1 The Dirichlet distribution

The Dirichlet distribution is the reference distribution to model vectors of weights summing to unity. It represents the maximum degree of independence components can get subject to the constraint of unit sum. The Dirichlet density with parameters $\alpha_1, ..., \alpha_n$ of a random vector $(w_1, ..., w_n)$ is

$$f(w_1, ..., w_n) = \frac{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n w_i^{\alpha_i - 1} \left((w_1, ..., w_n) \in W\right)$$

where W is the $n-1$-dimensional simplex in n-dimensional space as defined in Equation 2.1, and $\Gamma(s)$ denotes the gamma function. We write in short $(w_1, ..., w_n) \sim Dirichlet(\alpha_1, ..., \alpha_n)$.

Two special cases of Dirichlet distributions are:

(i) The Beta distribution $Beta(\alpha_1, \alpha_2)$, which is a Dirichlet distribution with $n = 2$ and parameters α_1 and α_2.

(ii) The standard Uniform distribution $Unif(0, 1)$, which is a Dirichlet distribution with $n = 2$ and $\alpha_1 = \alpha_2 = 1$. More generally, generating values uniformly over the space of weights W corresponds to generating values according to a Dirichlet distribution with $\alpha_1 = \alpha_2 = ... = \alpha_n = 1$.

An interesting property of the Dirichlet distribution is that the means of all the w_i stay the same if all α_i are scaled with the same multiplicative constant. The variances will, however, get smaller as the parameters α_i grow. In the following sections, we will write the parameters of the Dirichlet distribution as $c \times \left(w_1^0, ..., w_n^0\right)$ with

(i) $0 \leq w_1^0, ..., w_n^0 \leq 1$ with $\sum_{i=1}^n w_i^0 = 1$,

(ii) c a scaling constant that can vary from 0 to infinity. The variances of the w_i are inversely proportional to c. They equal to infinity when $c = 0$ and to zero (deterministic approach) when $c = +\infty$.

2.3.2 The model

The same notations and formula as for the previous models are used, and we specify that

$$(w_1, ..., w_n) \sim Dirichlet\left(c \times \left(w_1^0, ..., w_n^0\right)\right) \left((w_1, ..., w_n) \in W\right)$$

We can easily see that this model is a generalization of the models presented before:

(i) It corresponds to the pMCDA model defined in Section 2.2.2 when $c = +\infty$, since in this case the weights are deterministic with $w_j = w_j^0$ for $j = 1, ..., n$. pMCDA is itself a generalization of dMCDA that takes into account the data uncertainty.

(ii) It corresponds to the SMAA model defined in Section 2.2.3 without preference information when the weight values are uniformly distributed with all Dirichlet parameters equal to 1, i.e. with our notations when $w_1^0 = ... = w_n^0 = 1/n$ and $c = n$.

(iii) It corresponds to the SMAA model with partial preferences of the decision-makers when the weight values are uniformly distributed in a restricted space of weights W' which reflects the constraints applied to W, as presented in Section 2.2.3.

The benefit-risk assessment of the treatments is done in the same way as for SMAA, by comparing the utility scores when using the same vector of parameters $w_1^0, ..., w_n^0$ for all the treatments.
The constant c can be interpreted as a sample size reflecting the strength of belief, by considering the Dirichlet prior distribution as a posterior distribution under an implicit sample of multinomial data. It corresponds to the confidence level of the decision-makers in the elicitation of their preferences.

2.4 Application to a placebo-controlled trial on two antidepressants

2.4.1 Data and model

The treatments to be compared, the criteria and their distributions are the same as those presented by Tervonen et al. [131] from a published study [96]. The data are summarized in Table 2.1.
All the criteria are binary events, therefore a Bayesian model is built with binomial likelihoods $r_{ij} \sim Bin\left(n_{ij}, \pi_{ij}\right)$ and flat conjugate priors (to be consistent with the approach used in Tervonen et al. [131]) $\pi_{ij} \sim Beta\left(1, 1\right)$, r_{ij}, n_{ij} and π_{ij} being respectively the number of events, the number of patients and the probability of event for the treatment i and the criterion j. We therefore obtain the posterior distribution $\pi_{ij} \sim Beta\left(r_{ij} + 1, n_{ij} - r_{ij} + 1\right)$ which takes into account the data sampling variation. The probabilities range between 0 and 1, so the partial value functions are naturally defined as linear functions $u(\pi_{ij}) = \pi_{ij}$ for benefits and $u(\pi_{ij}) = 1 - \pi_{ij}$ for risks. The median and 95% credibility intervals of the posterior distributions and the partial value functions are summarized in Table 2.2.
In order to compare Dirichlet SMAA with pMCDA and SMAA, we performed several analyses with different weight vectors:

(i) Three analyses with deterministic weight vectors, corresponding to pMCDA. Since no precise weight preferences were elicited in Tervonen et al. [131], we supposed that three different decision-makers provided different preferences, respectively (0.25, 0.25, 0.25, 0.25), (0.58, 0.11, 0.15, 0.15) and (0.18, 0.28, 0.25, 0.29) for the criteria (treatment response, nausea, insomnia, anxiety). Therefore, the first decision-maker does not have preferences between the criteria, the second decision-maker favors the efficacy criterion, and the third decision-maker gives more weight to safety in general, but does not have a marked preference for one specific criterion.

TABLE 2.1: Number of patients and number of events for the criteria

	Placebo	Fluoxetine	Venlafaxine
Benefit criterion			
Treatment response	37/101	45/100	51/96
Risks criteria			
Nausea	8/102	22/102	40/100
Insomnia	14/102	15/102	22/100
Anxiety	1/102	7/102	10/100

TABLE 2.2: Median and 95% credibility interval (CI) of the Beta posterior distributions of the parameters π_{ij} and their partial value functions

	Posterior distributions Median (95% CI)			Partial value functions
	Placebo	Fluoxetine	Venlafaxine	
Benefits criteria				
Treatment response	0.37 (0.28;0.46)	0.45 (0.36;0.55)	0.53 (0.43;0.63)	$u_1(\pi_{i1}) = \pi_{i1}$
Risks criteria				
Nausea	0.08 (0.04;0.15)	0.22 (0.15;0.31)	0.40 (0.31;0.50)	$u_2(\pi_{i2}) = 1 - \pi_{i2}$
Insomnia	0.14 (0.08;0.22)	0.15 (0.09;0.23)	0.22 (0.15;0.31)	$u_3(\pi_{i3}) = 1 - \pi_{i3}$
Anxiety	0.02 (0.00;0.05)	0.07 (0.03;0.04)	0.11 (0.06;0.17)	$u_4(\pi_{i4}) = 1 - \pi_{i4}$

(ii) Three analyses corresponding to different versions of the Dirichlet SMAA model, with

$$(w_1, w_2, w_3, w_4) \sim Dirichlet\,(c \times (0.25, 0.25, 0.25, 0.25))$$

$$(w_1, w_2, w_3, w_4) \sim Dirichlet\,(c \times (0.58, 0.11, 0.15, 0.15))$$

$$(w_1, w_2, w_3, w_4) \sim Dirichlet\,(c \times (0.18, 0.28, 0.25, 0.29))$$

c ranging from 1 to 10^5: the same preferences as for pMCDA are used but we make the strength of confidence of the decision-makers vary. When $c = 4$, the first vector of Dirichlet parameters is equal to $(1, 1, 1, 1)$ which corresponds to the SMAA model without preference information.

In each case and for each treatment, the benefit-risk utility scores are calculated using Monte Carlo simulations. We compare the models by describing the distributions of the pairwise differences in utility scores defined in Equation 2.2 and the probability for each treatment i' to be the best treatment, estimated as the proportion of simulations in which its utility score is greater than those from the other treatments:

$$Prob\left(i' = best\right) \approx \frac{1}{K} \sum_{k=1}^{K} \sum_{w^*} \mathbb{1}\left[u_k(\pi_{i'}, w^*) > u_k(\pi_i, w^*) \;\; \forall i \neq i'\right]$$

where K is the total number of iterations (a large number), $\mathbb{1}[true] = 1$ and $\mathbb{1}[false] = 0$, and w^* are simulated for each iteration using the distributions defined above.

2.4.2 Results

The analyses were conducted using R, and 20,000 Monte Carlo simulations were run to estimate the parameters and the utility scores.

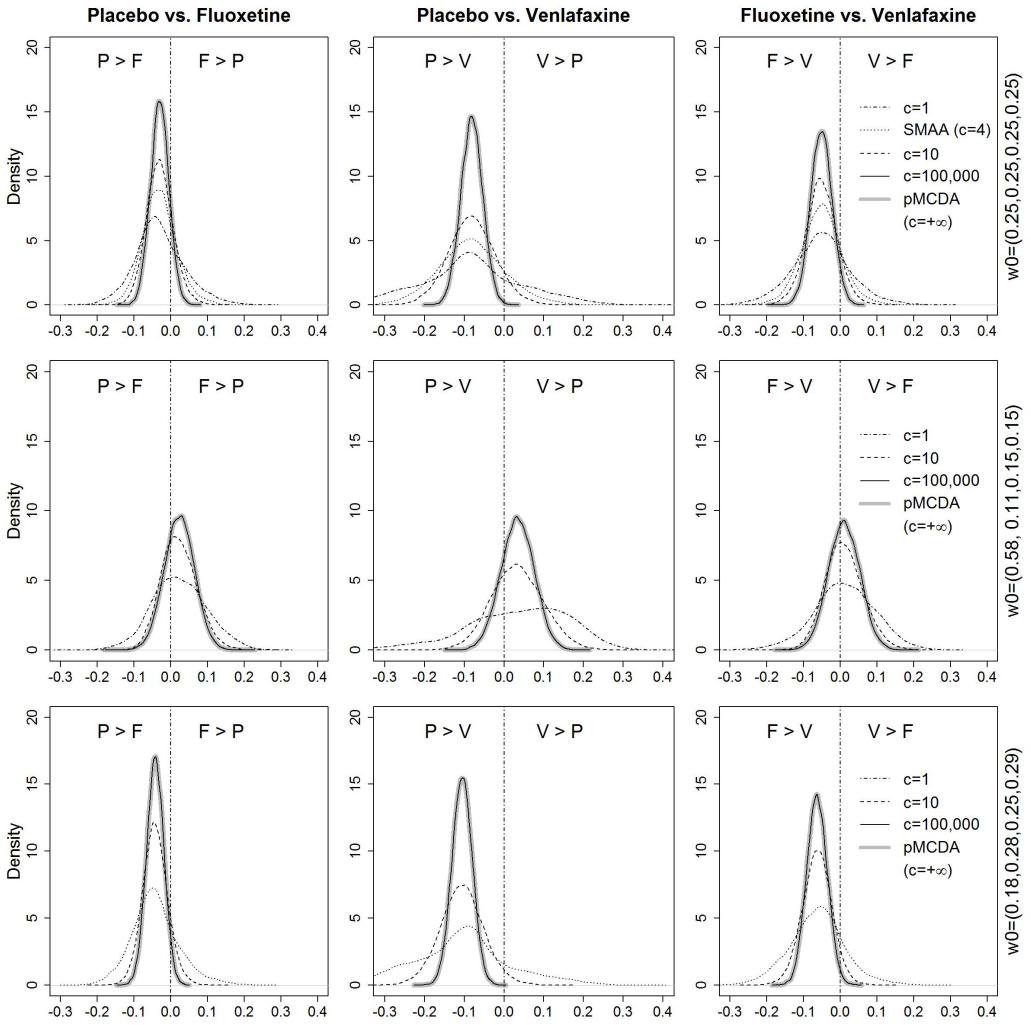

FIGURE 2.1: Distributions of the pairwise differences in utility scores when using the same vector of parameters, for Dirichlet SMAA with $c = 1$, $c = 10$ and $c = 10^5$, SMAA without preference information and pMCDA. Top: results with the preferences of the first decision-maker with $w^0 = (0.25, 0.25, 0.25, 0.25)$. Middle: results with the preferences of the second decision-maker with $w^0 = (0.58, 0.11, 0.15, 0.15)$. Bottom: results with the preferences of the third decision-maker with $w^0 = (0.18, 0.28, 0.25, 0.29)$. P = Placebo, F = Fluoxetine, V = Venlafaxine.

The distributions of the pairwise differences in utility scores when using the same vector of weights are presented in Figure 2.1. As expected, the precision of the differences increases with the strength of the decision-makers' confidence c. In all cases, the distributions for $c = 10^5$ and for pMCDA ($c = +\infty$) are superimposed, indicating that the results appropriately converge. As outlined earlier, SMAA without elicitation of preferences actually corresponds to a Dirichlet SMAA model with the vector of preferences of the first decision-maker, $w^0 = (0.25, 0.25, 0.25, 0.25)$, and a confidence factor $c = 4$. Its precision lies between the precision of Dirichlet SMAA for $c = 1$ and the precision of pMCDA. When the preferences of the second decision-maker are used, the conclusions are not much impacted by its strength of confidence. For example, the probability that Venlafaxine is better than Fluoxetine is 55% for

$c = 1$ and 60% for $c = 10^5$ and pMCDA. In this case, all the treatments have a similar benefit-risk ratio, the mean differences are close to zero and are not affected by the confidence the decision-maker has in his elicitation. On the other hand, the preferences of the first and the third decision-makers allow to discriminate between the treatments, and decisions can be taken with more assurance when they are more confident in their preferences. For example for the third decision-maker, the probability for the placebo to be better than Fluoxetine equals 77% when $c = 1$, with still a non-negligible chance that Fluoxetine is actually better, while it equals 96% when $c = 10^5$.

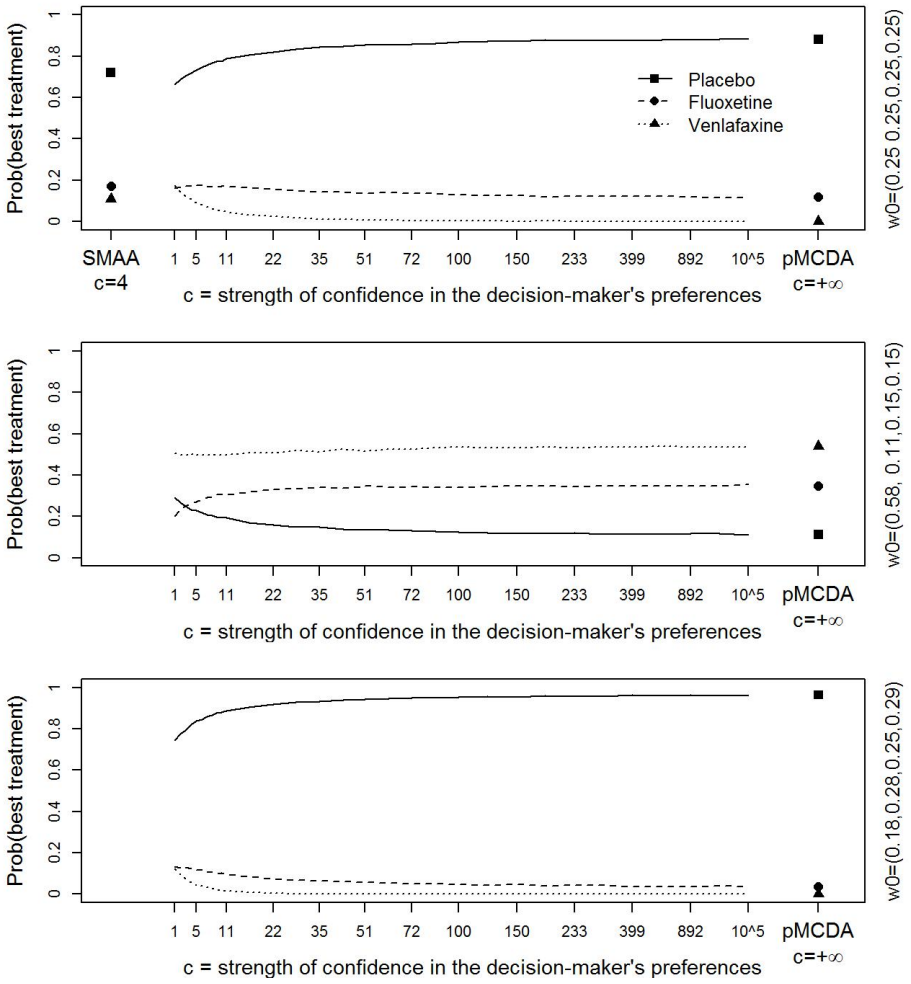

FIGURE 2.2: Probabilities to be the best treatment for the models SMAA without preference information (symbols, left), pMCDA (symbols, right) and Dirichlet SMAA (curves) with c ranging from 1 to 10^5. Top: results with the preferences of the first decision-maker with $w^0 = (0.25, 0.25, 0.25, 0.25)$. Middle: results with the preferences of the second decision-maker with $w^0 = (0.58, 0.11, 0.15, 0.15)$. Bottom: results with the preferences of the third decision-maker with $w^0 = (0.18, 0.28, 0.25, 0.29)$.

The probabilities for each treatment to be the best treatment are presented in Figure 2.2. When using the vector of preferences of the first decision-maker, we notice that the probabilities to be the best treatment from Dirichlet SMAA with $c = 1$ and from SMAA ($c = 4$) slightly differ: 66%, 16% and 18% for Dirichlet SMAA with $c = 1$

and 72%, 17% and 11% for SMAA, respectively for the placebo, Fluoxetine and Venlafaxine. This indicates that an amount of precision $c = 4$ is non-negligible and actually reflects some confidence of the decision-maker that the criteria are equally important. These results are much different from those obtained with the preferences of the second decision-maker with the strength of confidence $c = 1$ (29%, 20% and 50% respectively) but are closer to those obtained with the preferences of the third decision-maker (74%, 13% and 12% respectively). According to the second decision-maker, the criteria clearly have different importances, favoring the efficacy criterion, while the third decision-maker globally favors safety but without a marked difference between the criteria. Therefore, the results are affected as soon as some clear preferences are made between the criteria, even with a high degree of uncertainty.
The probabilities to be the best treatment with Dirichlet SMAA change as the strength of confidence increases, converging quickly toward pMCDA where deterministic weights are used. For example, they are respectively equal to 95%, 5% and 0% for the placebo, Fluoxetine and Venlafaxine with a Dirichlet SMAA using the preferences of the third decision-maker and a strength of confidence $c = 50$, very close to those obtained with pMCDA (94%, 6% and 0% respectively).

2.5 Discussion

SMAA is a very complete decision tool to assess and compare benefit-risk ratios of treatments, that takes into account both the sampling variation inherent in criterion measurements and the lack of preferences of the decision-makers regarding the relative importance of the criteria. However, it is still rarely used for decision-making in drug development, mainly due to the high degree of uncertainty in its results and its alleged complexity resulting in some difficulty faced by the decision-makers to interpret these results. Probabilistic MCDA, which can be seen as a simplified version of SMAA, is more popular, but it requires the exact elicitation of the criteria weights by the decision-makers, which could be difficult to achieve in practice, mainly due to the lack of confidence the decision-makers have in their tentative elicitations.
The Dirichlet SMAA model is a simple generalization and unification of pMCDA (which is a generalization of dMCDA) and SMAA, enabling to fit both of them but also allowing for more flexibility by using various levels of the scaling constant c. It permits to assess the sensitivity of the results of pMCDA when the decision-makers are able to provide some preferences but with some uncertainty. The results can be explained as a function of the decision-makers' preferences and their degree of confidence.
It is still not clear how the scaling constant c can actually reflect the decision-makers' confidence in their elicitations. Ideally, c should be provided by the decision-makers themselves, but it is unlikely they are able to quantify their degree confidence with such precision and without guidance. In our example in depression, we have seen that the results converge quickly, remaining stable for c greater than 50, which results are very close to those obtained with pMCDA where $c = +\infty$. In this particular case, the impact of the strength of confidence on the results can therefore be explored and discussed with the decision-makers for $c < 50$, rather than presenting them the results for one specific, and hypothetical, value of c.
Compared to pMCDA, both SMAA and Dirichlet SMAA permit to introduce some uncertainty in the subjectivity inherent to decision-making analyses. Nevertheless, two other sources of subjectivity remain present. First, the choice of the criteria used

to assess the benefits and the risks can strongly affect the results. A considerable effort has been made in the past years [69, 92] to increase the transparency of the whole benefit-risk assessment process, including the use of framework approaches such as BRAT (PhRMA Benefit-Risk Action Team) [103, 104] and PrOACT-URL (Problem formulation, Objectives, Alternatives, Consequences, Trade-Offs, Uncertainties, Risk Attitude and Linked Decisions) [56]. These approaches guide and structure the discussions between the decision-makers, ensuring they have a common understanding of the objectives of the assessment, and therefore helping them in identifying the key benefits and the key risks to address these objectives [99]. The second source of subjectivity is the definition of the partial value functions to map the criterion measurements into a 0 to 1 scale, which should reflect the importance of a change on each criterion. The definition of the partial value functions could be very simple in some cases, as in our example where the amounts of benefits and risks are assumed to be linearly related to the value of their respective criterion, and where the natural range of their spread $[0, 1]$ is taken as the reference range. Non-linear functions are more sensible when only some values, or ranges of values, actually represent an increased benefit or risk. When the criterion is an average over patients, the linear assumption can be interpreted at the population level: for a binary event, it means that two patients experiencing an event is twice as good (for benefits) or twice as bad (for risks) as one patient experiencing the event. A non-linear function is more likely to be needed at the patient level for a quantitative outcome, and one could consider to transform the outcome before averaging across patients. In general, non-linear partial value functions are certainly more difficult to define in practice and, in this case, the uncertainty of the decision-makers could be accounted by applying probability distributions to the parameters of the partial value functions, or by using classes of functions.

In conclusion, we believe that the Dirichlet SMAA is a simple model to unify MCDA and SMAA, and allows for a more extended exploration of the benefit-risk assessment of treatments. All the results can be presented according to the parameters which have a natural interpretation: treatment effects, decision-makers' preferences and strength of confidence.

Chapter 3

Scale Loss Score (SLoS): a novel measure of drug benefit-risk assessment

Background

This chapter was published as:

Saint-Hilary G, Robert V, Gasparini M, Jaki T, Mozgunov P. A novel measure of drug benefit–risk assessment based on Scale Loss Score. *Statistical Methods in Medical Research*. 2018; 1-16. doi: 10.1177/0962280218786526. [111]

3.1 Introduction

A drug benefit-risk assessment consists of balancing its favourable therapeutic effects versus adverse reactions it may induce [21]. The benefit-risk balance is a strong predictor of the therapy's long-term viability and a key element for decision-making during the drug's development, the regulatory approval process, and the post marketing follow-up [41, 45, 136]. For many years, a qualitative description of evidences had been the main approach to establish a drug's profile [62, 35]. This approach, however, tends to lack transparency since the decision of taking (dropping) a drug is based on a large amount of data coming from different sources and on criteria which can vary for different experts. Structured frameworks and quantitative methodologies have been recently proposed to make a benefit-risk assessment more comprehensive and consistent [90, 91, 79, 92, 133, 82].

According to the European Medicine Agency Benefit-Risk Methodology Project [35, 36, 37, 38], one of the most comprehensive quantitative approaches is MultiCriteria Decision Analysis (MCDA) [93, 94, 98, 81]. It has also been recommended by several highly profiled expert groups, e.g. see IMI PROTECT Work package 5 [69]. The main idea of MCDA is to calculate a single *utility score* using multiple criteria and taking into account the importance of each criterion. While non-linear forms of the utility score are recognized in various application areas of MCDA [78, 84], a linear aggregation of treatment's effects on benefits and risks remains the most common choice for the drug development [94, 131, 99, 80, 112]. The major advantage of the linear model is its intuitive interpretation: a poor efficacy can be compensated by a good safety, and vice-versa. However, the linear utility score can result in the recommendation of highly unsafe or poorly effective drugs [87, 83] and, consequently, in a counter-intuitive conclusion. Moreover, the linearity implies that the relative tolerance in the

toxicity increase is constant for all levels of benefit. This leads to implicit assumptions on decision-makers' preferences which might not hold for all drugs. Avoiding these pitfalls is possible with, first, adopting good practices to ensure that the modeling approach makes sense [87], and second by using non-additive and non-linear models [82, 74]. The main objectives of this work are to explicitly illustrate the issues of the additive linear MCDA model through a comprehensive simulation study, and to provide an alternative approach, namely, Scale Loss Score, for aggregation of treatment's effect overcoming these issues. The proposed approach is based on recent developments in the theory of estimation in restricted parameter spaces [71, 89] and is shown to have soundness in the context of drug evaluation.

The case study of telithromycin (Ketek®) raises questions regarding the suitability of a linear MCDA utility score for the drug benefit-risk assessment. Telithromycin was approved for the treatment of infections in several indications in 2001 by the EMA [39] and in 2005 by the FDA [135]. It was (qualitatively) re-assessed in 2006-2007 by both agencies based on updated safety data. In particular, some serious visual adverse reactions, syncopes and acute liver failures have been reported. The terms of the marketing authorizations were varied in order to better describe the drug safety profile, and two indications were removed from the labeling by the FDA, among them Acute Bacterial Sinusitis (ABS). More recently, the IMI PROTECT Benefit-Risk Group [70] applied MCDA to this clinical example. Even if this assessment was performed solely for the purpose of testing the methodology, the main results indicated a fairly strong superiority of telithromycin versus the comparators in ABS, which is not consistent with the concerns expressed by the health authorities. Consequently, alternative methods more accurately reflecting decision-makers' preferences are of great interest.

In this work, we extend the assumption of non-linearity of preferences, which is well established in other fields such as microeconomics or ecology [78, 6, 137], to the drug development context. We advocate two properties that a desirable measure of drug benefit-risk assessment should have:

1. Decreasing level of risk tolerance relative to benefits: an increase in risk could be more tolerated when benefit improves from 'very low' to 'moderate', compared to from 'moderate' to 'very high'.

2. Non-effective or/and extremely unsafe treatments should never be recommended.

Motivated by recent developments in the theory of the weighted information measures [75, 88] and in the theory of estimation in restricted parameter spaces [89], we propose Scale Loss Score (SLoS) as a novel measure for the benefit-risk assessment which shares both of these properties. The first property is achieved through convex preferences between efficacy and safety and the second one by a strong penalization of extremely low benefit and high risk values.

We perform a comprehensive simulation study investigating the performances of SLoS and MCDA in many different scenarios. Note that this is, to our knowledge, the first time the properties of MCDA are systematically explored by simulations in the medical context. We also apply the new measure to the motivating clinical context of telithromycin. The elicitation of criterion weights for linear MCDA utility scores is widely discussed in the literature [133, 82, 54, 107, 113, 8, 61, 132, 9]. Therefore, we provide an algorithm of mapping MCDA weights to SLoS weights so that

the same elicitation process could be followed while preserving the weight interpretation.

The rest of the paper is organized as follows. The MCDA utility score and the novel measure are detailed in Section 3.2. Section 3.3 describes the application of both measures in the real case study (telithromycin). We present a simulation study in Section 3.4 and conclude with discussion in Section 3.5. Additional information may be found in Appendix A.

3.2 Methods

3.2.1 MCDA utility score

The original proposal of MCDA [93, 94] ignores the uncertainty of parameter estimates. As this uncertainty can bare crucial information, an extension of MCDA taking into account the variability of estimates was proposed by Waddingham *et al.* [138]. This approach is often called Probabilistic MCDA (or Stochastic MCDA) and is described below.

Utility score

Consider m treatments (indexed by i) which are assessed on n criteria (indexed by j). We adopt the following notation:

(i) ξ_{ij} is the performance of treatment i on criterion j, so that treatment i is characterized by the vector $\boldsymbol{\xi}_i = (\xi_{i1}, ..., \xi_{in})$.

(ii) The monotonically increasing partial value functions $0 \leq u_j(\cdot) \leq 1$ are used to normalize the criterion performances. Let ξ'_j and ξ''_j be the most and the least preferable values, then $u_j(\xi''_j) = 0$ and $u_j(\xi'_j) = 1$. The inequality $u_j(\xi_{ij}) > u_j(\xi_{hj})$ indicates that the performance of the treatment i is preferred to the performance of the treatment h on criterion j. In this work, we focus on linear partial value functions, one of the most common choice in drug benefit-risk assessment [133, 93, 131, 80, 138]. They can be written as

$$u_j(\xi_{ij}) = \frac{\xi_{ij} - \xi''_j}{\xi'_j - \xi''_j}. \tag{3.1}$$

(iii) The weights indicating the relative importance of the criteria are known constants denoted by w_j such that $\sum_{j=1}^{n} w_j = 1$. The vector of weights used for the analysis is denoted by $\boldsymbol{w} = (w_1, ..., w_n)$.

The MCDA utility score is obtained as

$$u(\boldsymbol{\xi}_i, \boldsymbol{w}) := \sum_{j=1}^{n} w_j u_j(\xi_{ij}). \tag{3.2}$$

The higher the utility score, the more preferable the benefit-risk ratio. Then, the comparison of treatments i and h is based on

$$\Delta u(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w}) := u(\boldsymbol{\xi}_i, \boldsymbol{w}) - u(\boldsymbol{\xi}_h, \boldsymbol{w}).$$

While maximizing utility is common in economics [137], the concept of a loss function is usually preferred in statistical decision theory and Bayesian analysis for parameter estimation [6]. The complement of the MCDA utility score, $\bar{u}(\boldsymbol{\xi}_i, \boldsymbol{w}) = 1 - u(\boldsymbol{\xi}_i, \boldsymbol{w})$, could be considered as a *linear loss score* to be minimized, and it can be used equivalently as a measure of discrimination.

Although the term 'MCDA' outside of the health domain refers to the general methodology to summarize several characteristics in a single aggregated score, in this work we adopt the notation 'MCDA' for the additive utility score with linear partial values functions corresponding to the conventional model adopted so far in the drug benefit-risk assessment [82].

Estimation

Within a Bayesian approach, the utility score $u(\boldsymbol{\xi}_i, \boldsymbol{w})$ is a random variable having a prior distribution. Given observed outcomes $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{in})$ and $\boldsymbol{x_h} = (x_{h1}, \ldots, x_{hn})$ (corresponding to treatment performances $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_h$, respectively) for i and h, one can obtain the posterior distribution of $\Delta u(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w})$. The inference is based on the complete posterior distribution and the conclusion on the benefit-risk balance is supported by the probability of treatment i to have a greater utility score than treatment h:

$$\mathcal{P}_u^{ih} = \mathbb{P}(\Delta u(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w}) > 0 \mid \boldsymbol{x_i}, \boldsymbol{x_h}). \tag{3.3}$$

The probability (3.3) is used to guide a decision on taking/dropping a drug. A possible way to formalize the decision based on this probability is to compare it to a threshold confidence level $0.5 \leq \psi \leq 1$. Then, $\mathcal{P}_u^{ih} > \psi$ would mean that one has enough evidence to say that treatment i has a better benefit-risk balance than h with a level of confidence ψ. Note that $\mathcal{P}_u^{ih} = 0.5$ corresponds to the case where the benefit-risk profiles of i and h are equal according to MCDA.

Weight elicitation

Weighting is a structured way to capture the stakeholders' preferences between the criteria. It is recognized as a complex problem since it involves both clinical and societal value judgments [113]. Methods for quantifying subjective preferences have been widely studied in the literature [133, 82, 54, 107, 8, 61, 9], among which Discrete Choice Experiment and Swing-Weighting appeared to be appropriate in terms of theoretical foundations, cognitive burden, feasibility and robustness [93, 132, 73, 57]. In the MCDA framework, the weight assigned to one criterion is interpreted as a scaling factor which relates one increment on this criterion to increments on all other criteria.

MCDA illustration: two criteria

Let us consider an example with two criteria (one benefit indexed by 1, one risk indexed by 2) to illustrate an insight on the linear utility score in Equation (3.2). The utility score for treatment i at fixed parameter values θ_{i1}, θ_{i2} takes the form

$$u(\theta_{i1}, \theta_{i2}, w) := w u_1(\theta_{i1}) + (1 - w) u_2(\theta_{i2}). \tag{3.4}$$

As values $u_1(\theta_{i1}), u_2(\theta_{i2}) \in (0,1)$, one can interpret $u_1(\theta_{i1})$ as a probability of benefit and $1 - u_2(\theta_{i2})$ as a probability of risk. The contours of equal linear loss score $\bar{u}(\theta_{i1}, \theta_{i2}, w) = 1 - u(\theta_{i1}, \theta_{i2}, w)$ for all values of $u_1(\theta_{i1})$ and $(1 - u_2(\theta_{i2}))$ using $w = 0.5$ (left panel) and $w = 0.25$ (right panel) are given in Figure 1.

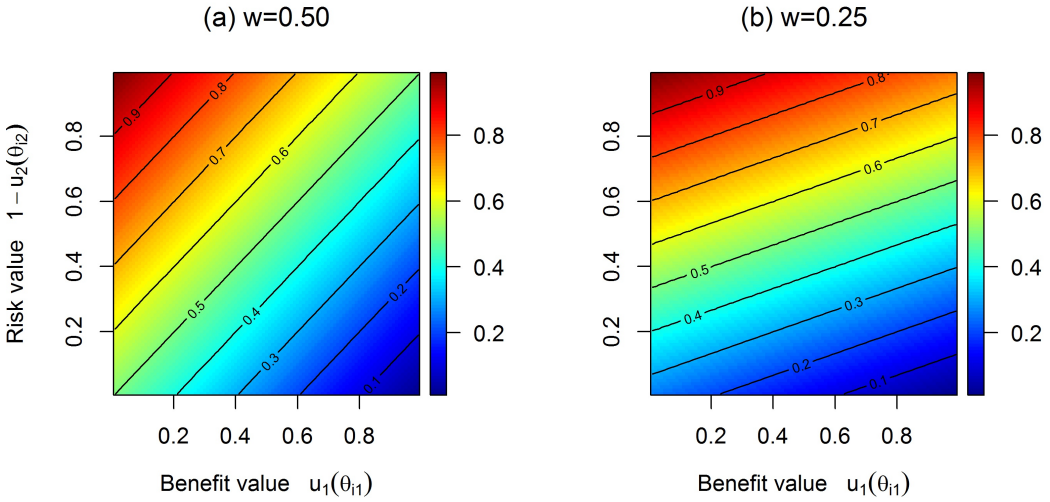

FIGURE 3.1: Left panel: contours of equal linear loss score $\bar{u}(\theta_{i1}, \theta_{i2}, w = 0.5)$. Right panel: contours of equal linear loss score $\bar{u}(\theta_{i1}, \theta_{i2}, w = 0.25)$

.

Lower values of $\bar{u}(\theta_{i1}, \theta_{i2}, w)$ correspond to better drug benefit-risk profiles. It is minimized (right bottom corner) when the maximum possible benefit is reached ($u_1(\theta_{i1}) = 1$) with no risk ($1 - u_2(\theta_{i2}) = 0$). The contours of (3.4) are linear, with a constant slope $w/(1-w)$. It implies that if one treatment has an increased probability of risk of x% compared to another, its benefit probability should be increased by $(1-w)/w\times$ x% to have the same utility score. This holds for all values of benefit and risk. While the linear form of the utility score makes the interpretation simple, it might lead to some counter-intuitive conclusions. Below, we illustrate possible paradoxes for $w = 0.25$, i.e. when the importance of the risk is three times higher than the importance of the benefit.

1. The benefit-risk trade-off is the same for all values of the risk/benefit.

   Consider two cases where a drug increases the benefit probability from (a) 0.15 to 0.30 and (b) from 0.80 to 0.95 compared to another therapy. In case (a) the increase doubles the benefit probability and a higher increase in toxicity can usually be tolerated. At the same time, the same increase in case (b) is not as relatively large, therefore it can be argued that only a smaller increase in the risk probability may be tolerated. However, the linear utility score implies that the same increase in risk to match the benefit increase can be sacrificed.

2. Drugs with 0% benefit or 100% risk can be recommended.

   Consider the first example in Table 3.1: drug 1 that cannot treat patients and causes adverse events only would be preferred. At the same time, drug 2 that adds 11% toxicity, but 30% efficacy would not be chosen. Similarly, in the second example in Table 3.1, drug 1 that leads to an adverse event for 100% patients would be preferred.

TABLE 3.1: Examples of MCDA linear utility scores with two criteria and $w = 0.25$.

	Example 1		Example 2	
	Drug 1	Drug 2	Drug 1	Drug 2
Benefit: $u_1(\theta_{i1})$	0.00	0.30	0.96	0.50
Risk: $1 - u_2(\theta_{i2})$	0.09	0.20	1.00	0.85
Utility score $u(\theta_{i1}, \theta_{i2}, w = 0.25)$	0.6825	0.6750	0.2400	0.2375

Even if none of those drugs are likely to be taken to the market, the goal of MCDA is to rank treatments and these examples reveal some counter-intuitive conclusions to which MCDA can lead. Note that decreasing values of w would help to solve the paradox in Example 1, but would worsen it in Example 2.

We advocate two properties of a benefit-risk analysis measure: (i) for a given increase in benefit, one can tolerate a larger increase in risk if the amount of benefit is small than if it is high, and (ii) one is not interested in the level of risk (benefit) if the drug does not treat (harm all) patients. Formally, these properties correspond to (i) the concavity of equal loss score contours (or, equivalently, the convexity of equal utility score contours) and to (ii) a strong penalization of extreme low benefit values and extreme high risk values. We would like to stress that the convexity of utility (concavity of loss) is widely advocated in microeconomics and is believed to reflect preferences in a more adequate way than linear ones in many applications [78, 137].

One can check that none of these properties are satisfied for MCDA due to its linearity. There are two forms of linearity in MCDA: in the partial value functions (3.1) and in the utility score (3.2). Note that property (i) of decreasing level of risk tolerance relative to benefits can be achieved by varying the shape of the partial value functions (for instance, using concave functions for benefit and linear functions for risk). However, the explicit elicitation of non-linear forms for the partial value functions may be challenging. As the linear partial function remains a common choice in drug benefit-risk assessment, we propose a novel measure of aggregation which allows for both properties to be achieved even under linear partial value functions.

3.2.2 Scale Loss Score

Derivation

As an alternative to the linear MCDA utility score (3.2), we define Scale Loss Score (SLoS) for aggregation of treatment's performances as

$$l(\boldsymbol{\xi}_i, \tilde{\boldsymbol{w}}) := \sum_{j=1}^{n} \left(\frac{1}{u_j(\xi_{ij})} \right)^{\tilde{w}_j} \tag{3.5}$$

where $\tilde{w}_j$ is the weight indicating the average relative importance of criterion j compared to the others and $u_j(\cdot)$ is a linear partial value function (3.1). The form of SLoS is motivated by the scale symmetric loss function [89, 85] and the precautionary loss function [71]. These functions allow to stay away from 'boundary' values $u_j(\cdot) = 0$. In the context of the benefit-risk assessment, they correspond to an extremely undesirable performance of a drug: low benefit or high risk. SLoS can be interpreted as a

divergence between drug i characteristics $\boldsymbol{\xi}_i$ and the 'perfect' benefit-risk characteristics $(1,\ldots,1)_{n\times 1}$. As a *loss score* is used rather than a utility score, lower values of $l(\boldsymbol{\xi}_i, \tilde{\boldsymbol{w}})$ correspond to more desirable performances of the drug.

Clearly, $l(\boldsymbol{\xi'}, \tilde{\boldsymbol{w}})$ is minimized for $\boldsymbol{\xi'}$ such that $u_j(\xi'_j) = 1$ for all $j = 1, ..., n$, i.e. at the point of the ideal benefit-risk profile. Additionally, $l(\boldsymbol{\xi''}_{(k)}, \tilde{\boldsymbol{w}}) = +\infty$ for $\boldsymbol{\xi''}_{(k)}$ a vector of parameters containing ξ''_k such that $u_k(\xi''_k) = 0$, for at least one $k \in \{1, ..., n\}$, so the loss score for a treatment with at least one extreme negative performance is equal to infinity. The lower bounds are determined by the least preferred values ξ''_k used in the partial value functions, and correspond to *unacceptable* levels of benefit or risk. It should be noted that SLoS is intentionally sensitive to these unacceptable values, therefore their choice could have a non-negligible impact on the results. While unacceptable values of 0 (for benefit) or 1 (for risk) may be obvious choices for probabilities of a binary outcome, the unacceptable value for a continuous outcome may be more subjective and requires a careful investigation.

SLoS is a measure of the benefit-risk balance permitting to discriminate treatments according to their performances and according to the weights attributed to the criteria. The lower the SLoS, the more preferable the benefit-risk ratio, and the comparison of treatments i and h is based on

$$\Delta l(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \tilde{\boldsymbol{w}}) := l(\boldsymbol{\xi}_i, \tilde{\boldsymbol{w}}) - l(\boldsymbol{\xi}_h, \tilde{\boldsymbol{w}}).$$

Estimation

Similarly to MCDA, we consider a Bayesian model and assign a prior probability distribution to ξ_{ij}. Given the observed outcomes $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{in})$ and $\boldsymbol{x_h} = (x_{h1}, \ldots, x_{hn})$ for the treatments i and h, one can obtain a posterior distribution of $\Delta l(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \tilde{\boldsymbol{w}})$. Again, the inference is based on the complete posterior distribution and the conclusion on the benefit-risk balance is supported by the probability of treatment i to have a smaller SLoS than treatment h:

$$\mathcal{P}_l^{ih} = \mathbb{P}(\Delta l(\boldsymbol{\xi}_i, \boldsymbol{\xi}_h, \boldsymbol{w}) < 0 \mid \boldsymbol{x_i}, \boldsymbol{x_h}).$$

This probability can be compared to a fixed confidence threshold ψ as in the MCDA approach.

SLoS illustration: two criteria

To illustrate the properties of SLoS, consider the example presented in Section 3.2.1 with one benefit and one risk. The SLoS for treatment i in the point of fixed parameter values takes the form

$$l(\theta_{i1}, \theta_{i2}, w) := \left(\frac{1}{u_1(\theta_{i1})}\right)^{\tilde{w}} + \left(\frac{1}{u_2(\theta_{i2})}\right)^{1-\tilde{w}}. \tag{3.6}$$

The Figure 2 presents the contours of SLoS (3.6) for all pairs of $u_1(\theta_{i1})$ and $1 - u_2(\theta_{i2})$ using $\tilde{w} = 0.5$ (left panel) and $\tilde{w} = 0.25$ (right panel). The tangents of the contours at the point (0.5,0.5) are presented on the graph for the purpose of the weight mapping

detailed in the next section.

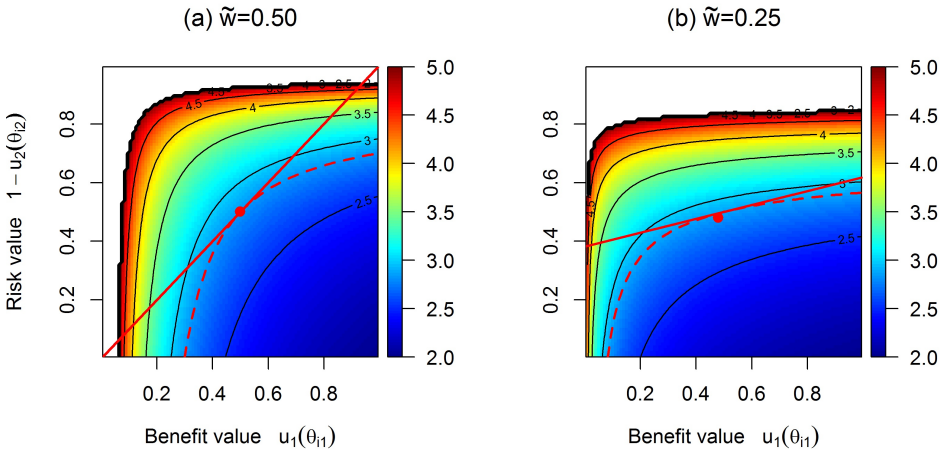

FIGURE 3.2: Left panel: contours of $l(\theta_{i1}, \theta_{i2}, \tilde{w} = 0.5)$. Right panel: contours of $l(\theta_{i1}, \theta_{i2}, \tilde{w} = 0.25)$. Red lines correspond to tangents at the point $(0.5, 0.5)$.

SLoS is minimized when the benefit-risk balance of the drug is maximized, at the point (1,0) (right bottom corner), where the maximum possible benefit is reached with no risk. The loss score is infinite for extreme low benefit values and extreme high risk values, thus non-effective or extremely unsafe treatments could never be recommended. Considering the cases presented in Table 3.1, the drug 2 had a SLoS equal to 2.53 for the first example and of 5.34 for the second example, and it is preferred to drug 1 which SLoS is infinite in both cases.

The contour lines of equal loss are concave, which is equivalent to having convex preferences between additional benefit and avoided risk, and have the form:

$$1 - u_2(\theta_{i2}) \;=\; 1 - (z - u_1(\theta_{i1})^{-\tilde{w}})^{-\frac{1}{1-\tilde{w}}}, \text{ for } u_1(\theta_{i1}) > z^{-1/\tilde{w}}$$

for a fixed value $l(\theta_{i1}, \theta_{i2}, \tilde{\boldsymbol{w}}) = z$. The slope of the tangent of the contour at a given $u_1(\theta_{i1})$ for the loss score value z is

$$\frac{\tilde{w}}{1-\tilde{w}} \left(z - u_1(\theta_{i1})^{-\tilde{w}}\right)^{\frac{\tilde{w}-2}{1-\tilde{w}}} \cdot u_1(\theta_{i1})^{-(\tilde{w}+1)}. \tag{3.7}$$

The slope decreases as benefit increases. It follows that the relative importance of the benefit criterion over the risk criterion decreases with the amount of benefit itself. In other words, an increase in toxicity is more tolerated if, in parallel, efficacy improves from 'very low' to 'moderate', compared to from 'moderate' to 'very high'.

Weight elicitation

Since comprehensive work has been published and is currently being continued on the weight elicitation for MCDA, we present a way to map MCDA weights w_j to SLoS weights $\tilde{w}_j$. Note that the slope of MCDA contour tangents is *constant* for all values of parameters and defined by the weights w_j only, while the slope of SLoS

contour tangents is *non-constant* and defined by both $\tilde{w}_j$ and values of the criteria. To map weights, we would interpret $\tilde{w}_j$ as an *average* relative importance of each criterion over the others. With two criteria, the weight $\tilde{w}_j$ corresponding to the MCDA weight w_j can be found from the equality of the slopes of the tangents of MCDA and SLoS contours in the middle point $u_1(\theta_{i1}) = u_2(\theta_{i2}) = 0.5$ of treatment i performances,

$$\frac{\tilde{w}_j}{1-\tilde{w}_j} \cdot 2^{2\tilde{w}_j - 1} = \frac{w_j}{1-w_j}, \tag{3.8}$$

where the slopes of SLoS and MCDA contour tangents in this point are given on the left and right hand sides, respectively.

The weight mapping (3.8) does not have an analytical solution, but the approximate value of $\tilde{w}_j$ can be obtained by line search. The mapping of the weights is illustrated in Figure 3. The weights are the same when both criteria are considered equally important ($w = \tilde{w} = 0.5$), while $w < 0.5$ corresponds to slightly greater values of $\tilde{w}$. For instance, $\tilde{w} = 0.30$ corresponds to $w = 0.25$.

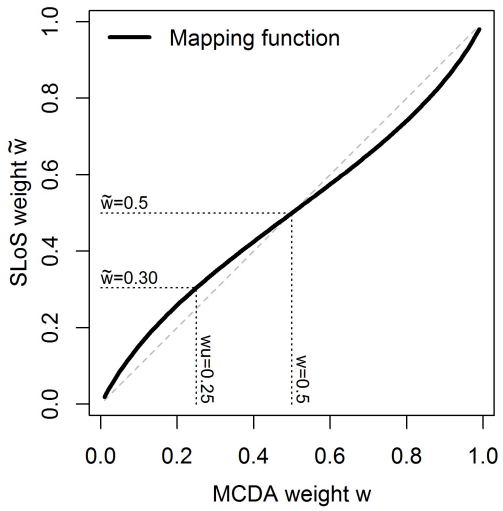

FIGURE 3.3: Weight mapping.

Considering an arbitrary number of criteria, the mapping (3.8) can be applied to each value of the MCDA weights. For instance, using four criteria with a weight vector $\boldsymbol{w} = (0.30,\ 0.15,\ 0.15,\ 0.15,\ 0.25)$, the vector of SLoS weights is equal to $\tilde{\boldsymbol{w}} = (0.35,\ 0.21,\ 0.21,\ 0.21,\ 0.30)$. It should be noted that, in this case, the weights $\tilde{w}_j$ do not necessarily sum to 1, but this does not prevent from calculating SLoS, for which the formula (3.5) still applies.

Mapping weights to the middle point of the benefit and risk treatment performance range relies on the assumption that MCDA weights were elicited across the entire range, or that the trade-off between criteria was anchored on average at the middle point. However, in practice, MCDA weights could have been elicited at any other point and extrapolated. In this case, the mapping procedure above could be performed accordingly by finding the SLoS weight satisfying the equality of the slopes of MCDA and SLoS contour tangents in any other point of interest.

3.3 Case study: telithromycin

We illustrate the use of SLoS and MCDA in a real clinical context on the case-study thelithromycin (Ketek®) reported by the IMI PROTECT Benefit-Risk Group [70].

Telithromycin was approved in 2001 for several indications as an alternative when beta-lactam antibiotics are not appropriate, and we will focus on the indications Community Acquired Pneumonia (CAP) and Acute Bacterial Sinusitis (ABS) as they well illustrate similarities and differences between the two methods. A Probabilistic MCDA model was considered in the IMI PROTECT report [70] (called Stochastic Multicriteria Acceptability Analysis with fixed weights), and MCDA utility scores presented here are derived from the original report.

Telithromycin is compared to a single alternative called 'comparator', which comprises amoxicillin-clavulanic acid, cefuroxime and clarithromycin, used as comparators in clinical studies and pooled together. The probabilities of five binary criteria, one benefit and 4 adverse events (AE), were transformed using linear partial value functions (Equation (3.1)) with the following most and least preferred probabilities of occurrence ξ'_j and ξ''_j [70]:

- Benefit: cure rate (CAP: $\xi'_1 = 1$, $\xi''_1 = 0.4$; ABS: $\xi'_1 = 0.86$, $\xi''_1 = 0.71$),
- Risks:
  - Hepatic AE (CAP: $\xi'_2 = 0$, $\xi''_2 = 0.1$; ABS: $\xi'_2 = 0$, $\xi''_2 = 0.02$),
  - Cardiac AE (CAP: $\xi'_3 = 0$, $\xi''_3 = 0.1$; ABS: $\xi'_3 = 0$, $\xi''_3 = 0.01$),
  - Visual AE (CAP: $\xi'_4 = 0$, $\xi''_4 = 0.1$; ABS: $\xi'_4 = 0$, $\xi''_4 = 0.02$),
  - Syncope (CAP: $\xi'_5 = 0$, $\xi''_5 = 0.1$; ABS: $\xi'_5 = 0$, $\xi''_5 = 0.01$).

Using uniform priors and given the number of cures and AE (see Appendix A, Table A.1), Beta posterior distributions for the event probabilities are approximated using 100,000 simulations in `R` [106], and are used to compute the corresponding distributions of the partial value functions. Means and 95% Credibility Interval (CrI) of the probabilities and of the partial value functions, and the MCDA weights, are summarized in Table 3.2.

This information was used to approximate the posterior distributions of MCDA linear utility score and SLoS. The mapped SLoS weights corresponding to the MCDA weights are $\tilde{\boldsymbol{w}} = (0.35,\ 0.21,\ 0.21,\ 0.21,\ 0.30)$.

For the CAP indication, MCDA and SLoS provide similar results, with probabilities that telithromycin is better than the comparator equal to 59% and 51%, respectively. These results indicate that telithromycin has a slightly better benefit-risk profile than the comparator, but with large uncertainty.

For the ABS indication, the probability that the benefit-risk balance of telithromycin is better than the comparator is equal to 71% using MCDA and 55% using SLoS. While they both indicate results in favour of telithromycin, this advantage appears to be much more uncertain with SLoS than with MCDA. The difference between the methods can be mainly explained by a higher rate of Visual AE with telithromycin (1.3% versus 0.5%), which is close to the least preferred value for this criterion in this indication ($\xi''_4 = 2\%$). This leads to low values of the corresponding partial

TABLE 3.2: Mean and 95% CrI of the Beta posterior distributions of benefit and risk parameters and of corresponding partial value functions, with their MCDA weight, for Telithromycin (Teli.) and Comparator (Comp.)

		CAP		ABS		MCDA
		Teli.	Comp.	Teli.	Comp.	weight
Cure rate						
ξ_{i1}	Mean	0.908	0.877	0.828	0.772	30%
	95% CrI	[0.896; 0.919]	[0.855; 0.897]	[0.800; 0.855]	[0.715; 0.824]	
$u_1(\xi_{i1})$	Mean	0.846	0.795	0.787	0.414	
	95% CrI	[0.827; 0.864]	[0.759; 0.829]	[0.601; 0.964]	[0.036; 0.760]	
Hepatic AE						
ξ_{i2}	Mean	0.044	0.042	0.011	0.004	15%
	95% CrI	[0.034; 0.056]	[0.031; 0.054]	[0.006; 0.017]	[0.001; 0.009]	
$u_2(\xi_{i2})$	Mean	0.561	0.582	0.468	0.789	
	95% CrI	[0.444; 0.664]	[0.457; 0.691]	**[0.158; 0.707]**	**[0.542; 0.942]**	
Cardiac AE						
ξ_{i3}	Mean	0.005	0.004	0.002	0.002	15%
	95% CrI	[0.002; 0.01]	[0.001; 0.009]	[0.000; 0.004]	[0.000; 0.006]	
$u_3(\xi_{i3})$	Mean	0.947	0.956	0.849	0.790	
	95% CrI	[0.902; 0.979]	[0.909; 0.985]	[0.579; 0.982]	[0.414; 0.974]	
Visual AE						
ξ_{i4}	Mean	0.011	0.004	0.013	0.005	15%
	95% CrI	[0.006; 0.018]	[0.001; 0.009]	[0.008; 0.020]	[0.002; 0.011]	
$u_4(\xi_{i4})$	Mean	0.887	0.956	0.357	0.736	
	95% CrI	[0.823; 0.937]	[0.909; 0.986]	**[0.016; 0.625]**	**[0.461; 0.914]**	
Syncope						
ξ_{i5}	Mean	0.002	0.004	0.001	0.002	25%
	95% CrI	[0.000; 0.005]	[0.001; 0.008]	[0.000; 0.003]	[0.000; 0.006]	
$u_5(\xi_{i5})$	Mean	0.977	0.964	0.924	0.789	
	95% CrI	[0.945; 0.995]	[0.922; 0.990]	[0.719; 0.998]	[0.414; 0.974]	

value function (mean (95% CrI) $u_4(\xi_{14})$: 0.36 [0.02; 0.63]), and values at the lower end of the distribution are strongly penalized by SLoS. At the same time, the mass of the corresponding partial value function distribution of the comparator (mean (95% CrI) $u_4(\xi_{24})$: 0.74 [0.46; 0.91]) is shifted further from the bound, which results in lower value of SLoS. A similar argument could be applied to Hepatic AE, and the combination of these safety issues is more penalized by SLoS than by MCDA, despite the worse cure rate of the comparator. Even if the benefit-risk assessment by IMI PROTECT [70] was performed in order to test the methodologies and may have been conducted differently in the actual regulatory context, it is worth noting that the conclusion obtained using SLoS for the ABS indication is more in line with the concerns expressed by the Committee for Medicinal Products for Human Use (CHMP) regarding the atypical safety profile of the drug [40] and the removal of this indication from the labeling by the FDA [135]. This could be an example of SLoS reflecting the decision-makers' preferences more accurately than MCDA.

A sensitivity analysis was conducted using MCDA weights to compute SLoS (omitting the weight mapping) and the conclusions are globally robust, with the probability of telithromycin being better than the comparator equal to 57% for CAP and 62% for ABS.

In the next section, we present a simulation study illustrating the properties of SLoS and MCDA in many different scenarios.

3.4 Simulation study

3.4.1 Setting

To investigate the performances of SLoS and MCDA, we simulated randomised controlled clinical trials with two treatments $i = 1, 2$, named T_1 and T_2, $N = 100$ patients per group, and two uncorrelated binary criteria ($j = 1$ for benefit and $j = 2$ for risk). We assume that benefit events are desirable (e.g. treatment response), while risk events should be avoided (e.g. adverse event), with the performance parameters ξ_{ij} being their probability of occurrence. The partial value functions are defined as $u_1(\xi_{i1}) = \xi_{i1}$ and $u_2(\xi_{i2}) = 1 - \xi_{i2}$. Equally important criteria with weights $w_j = \tilde{w}_j = 0.5$, $j = 1, 2$, are considered.

The investigated scenarios are summarized in Table 3.3, where the expected probabilities of event θ_{ij} are presented for T_1 (●) and T_2 (◇). Nine sets of T_1 characteristics are fixed. For each set, all possible combinations of T_2 characteristics with $\theta_{21}, \theta_{22} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ are considered. This results in 81 profiles for T_2 and in 729 cases in total: we explored the grid of treatment performances in order to identify under which conditions MCDA and SLoS lead to different conclusions. For example, the first scenario corresponds to fixed expected probabilities of benefit and risk for treatment T_1 $\theta_{11} = \theta_{12} = 0.5$ compared to all considered combinations of probabilities of event for T_2. In this scenario, we expect SLoS to recommend T_1 more than MCDA when T_2 is associated with an extreme risk or no benefit. In the other scenario where $\theta_{11} = \theta_{12} = 0.1$, T_1 has almost no benefit, so it should not be recommended despite its good safety profile. Indeed, even if T_1 does not harm the patients (it is similar to a placebo), administrating it to the patients implies we make the assumption that it has a positive effect, while it has not in reality. This interpretation is close to the usual type I error, and is not acceptable from a regulatory and health economics perspective. Similarly, when $\theta_{11} = \theta_{12} = 0.9$, T_1 should not be recommended despite its outstanding efficacy as it is associated with an extreme risk. All intermediate cases are considered.

TABLE 3.3: Simulation scenarios with two criteria

		Probability of Benefit θ_{i1}								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Probability of Risk θ_{i2}	0.9	◇●	◇	◇	◇	◇	◇	◇	◇	◇●
	0.8	◇	◇	◇	◇	◇	◇	◇	◇	◇
	0.7	◇	◇	◇●	◇	◇	◇	◇●	◇	◇
	0.6	◇	◇	◇	◇	◇	◇	◇	◇	◇
	0.5	◇	◇	◇	◇	◇●	◇	◇	◇	◇
	0.4	◇	◇	◇	◇	◇	◇	◇	◇	◇
	0.3	◇	◇	◇●	◇	◇	◇	◇●	◇	◇
	0.2	◇	◇	◇	◇	◇	◇	◇	◇	◇
	0.1	◇●	◇	◇	◇	◇	◇	◇	◇	◇●

● = treatment T_1 ; ◇ = treatment T_2

Let S be the total number of simulated trials and K the number of samples generated to approximate the distributions of interest. In each trial $s = 1, ..., S$, the

number of events for each criterion was simulated using a Binomial likelihood $x_{ij}^s \sim Bin(N, \theta_{ij})$. Then, values ξ_{ij}^{sk} are sampled from the posterior distribution of the parameters $\mathcal{B}(x_{ij}^s, N - x_{ij}^s)$ for $k = 1, ..., K$, assuming implicitly an improper conjugate prior $\mathcal{B}(0, 0)$. The posterior distributions of the utility score and the loss score are approximated by the samples $u(\boldsymbol{\xi}_i^{sk}, \boldsymbol{w})$ and $l(\boldsymbol{\xi}_i^{sk}, \boldsymbol{w})$.

Assuming the threshold confidence level $\psi = 0.8$, MCDA and SLoS are compared using $\mathbb{P}[\mathcal{P}_u^{1,2} > 0.8]$, $\mathbb{P}[\mathcal{P}_l^{1,2} > 0.8]$ and $\phi = \mathbb{P}[\mathcal{P}_l^{1,2} > 0.8] - \mathbb{P}[\mathcal{P}_u^{1,2} > 0.8]$. As a difference between probabilities, ϕ ranges in $(-1, 1)$. A value $-1 \leq \phi < 0$ indicates that SLoS recommends treatment T_1 more often than MCDA and $0 < \phi \leq 1$ that SLoS recommends T_1 less often than MCDA. The two approaches are in agreement when $\phi = 0$. A similar analysis for $\mathbb{P}[\mathcal{P}_l^{2,1} > 0.8]$ and $\mathbb{P}[\mathcal{P}_u^{2,1} > 0.8]$ is presented in Supplemental Material (Figure S1). Simulations with other choices of ψ led to similar conclusions on the comparison between the two methods and are not presented here.

The analyses were conducted using R, with $S = 2,500$ simulated clinical trials and $K = 2,000$ simulations to estimate the parameter distributions.

3.4.2 Results

The results are presented in Figure 4. All 9 scenarios for treatment T_1 are presented in rows and numbered (1)-(9). Each graph corresponds to fixed expected probabilities of event for treatment T_1 (•), and each cell corresponds to a combination of expected probabilities of benefit and risk for T_2. The probabilities $\mathbb{P}[\mathcal{P}_l^{1,2} > 0.8]$ are presented on the left panel, $\mathbb{P}[\mathcal{P}_u^{1,2} > 0.8]$ on the middle panel and ϕ on the right panel, for which positive values are displayed in blue, negative values are in red and null values are in white.

In scenario 1, the two measures are in agreement to recommend T_1, which has moderate benefit and risk ($\theta_{11} = \theta_{12} = 0.5$), when T_2 has less benefit and more risk. On the diagonal, SLoS favours T_1 to more effective treatments but with very high risk (respectively, to safer treatments but with very low benefit). In contrast, MCDA recommends more effective but highly unsafe treatments, or safer but no effective treatments, compared to T_1. For example, when $\theta_{21} = 0.8$ and $\theta_{22} = 0.9$ (large benefit but high risk), SLoS favours T_1 in 100% of the trials while MCDA recommends it in 62% only, resulting in $\phi = 0.38$. Also, when $\theta_{21} = 0.6$ and $\theta_{22} = 0.7$ (increased benefit by 0.1 and risk by 0.2 compared to T_1), SLoS favours T_1 in 80% and MCDA in 57% of the cases ($\phi = 0.23$). This reflects the property of SLoS that increases in risk are less tolerated when the amount of benefit is large enough. Similar patterns are observed in scenarios 2 and 3 where treatment T_1 has either a low benefit and a large risk, or a large benefit and a low risk, but not extreme probabilities of event.

In scenario 4, T_1 has almost no benefit nor risk, with $\theta_{11} = \theta_{12} = 0.1$. As expected, it is almost never recommended by SLoS, but it could be recommended by MCDA in scenarios where the alternative T_2 has some benefit but a higher increase in risk. For example, when $\theta_{21} = 0.2$ and $\theta_{22} = 0.3$ (increased benefit by 0.1 and risk by 0.2 compared to T_1), MCDA recommends T_1 in 70% of the cases while it is never recommended by SLoS ($\phi = -0.70$). This is consistent with the stated desirable property that we are not interested in the level of risk if the drug does not treat the patients.

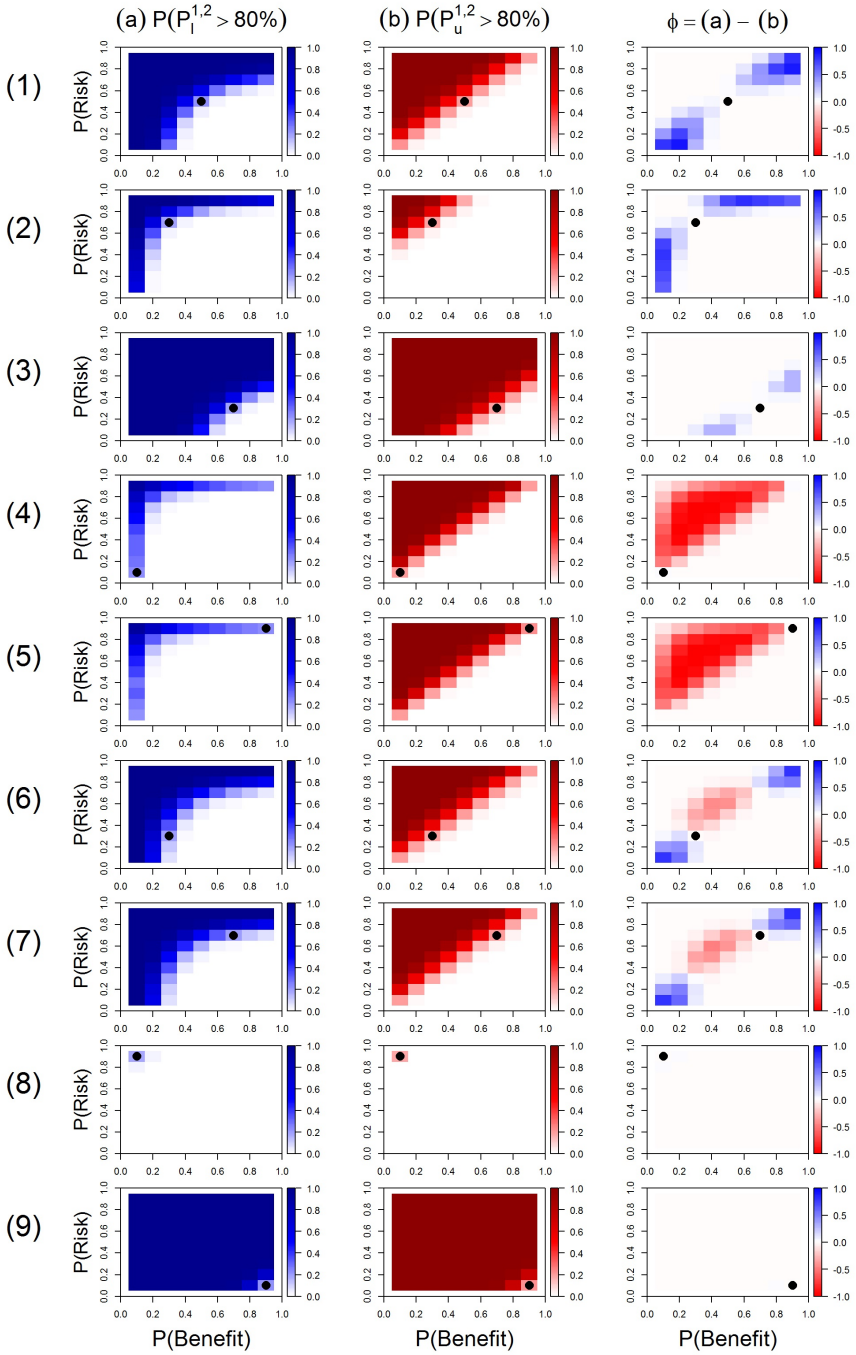

FIGURE 3.4: Results of the simulation scenarios for two equally important criteria ($w_j = \tilde{w}_j = 0.5$ for $j = 1, 2$). • = T_1. Left panel: $P[\mathcal{P}_l^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_l^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

On the other hand, when $\theta_{21} = 0.3$ and $\theta_{22} = 0.2$ (increased benefit by 0.2 and risk by 0.1 compared to T_1), SLoS discriminates better the treatments and recommends T_2 in 100% of the cases while MCDA recommends it in only 68% (Appendix A, Figure A.1). Similar conclusions are obtained in scenario 5, where T_1 has both extreme

efficacy and risk ($\theta_{11} = \theta_{12} = 0.9$): SLoS never recommends the unsafe treatment T_1 if alternative treatments T_2 have lower risk and at least some small benefit, while MCDA recommends T_1 as compared to treatments with a larger decrease in benefit than in risk. This is the case for instance when $\theta_{21} = 0.6$ and $\theta_{22} = 0.7$ (decreased benefit by 0.3 and risk by 0.2 compared to T_1), where T_1 is recommended in 65% of the cases by MCDA and never recommended by SLoS ($\phi = -0.65$). In contrast, when $\theta_{21} = 0.7$ and $\theta_{22} = 0.6$ (decreased benefit by 0.2 and risk by 0.3 compared to T_1), SLoS favours T_2 in 100% and MCDA in 67% of the cases (Supplemental Material, Figure S1).

Scenarios 6 and 7 correspond to treatment T_1 with either both low benefit and risk ($\theta_{11} = \theta_{12} = 0.3$) or large benefit and risk ($\theta_{11} = \theta_{12} = 0.7$) but where the probabilities of event are not extreme. The measures are in agreement to recommend T_1 when T_2 is indisputably worse. On the diagonal, T_1 is more often recommended by SLoS when T_2 has no benefit nor risk ($\theta_{21} = \theta_{22} = 0.1$) or very large benefit and risk ($\theta_{21} = \theta_{22} = 0.9$). On the other hand, SLoS favours more treatments with benefit and risk probabilities closer to 50%. For example, in scenario 6, when $\theta_{21} = 0.4$ and $\theta_{22} = 0.5$ (increased benefit by 0.1 and risk by 0.2 compared to T_1), SLoS recommends T_1 in only 17% of the cases, but MCDA in 59% ($\phi = -0.42$). Similarly, when $\theta_{21} = 0.5$ and $\theta_{22} = 0.4$ (increased benefit by 0.2 and risk by 0.1 compared to T_1), T_1 is not favoured by any of the methods, but SLoS recommends the alternative T_2 in 88% of the cases and MCDA in only 59% (Supplemental Material, Figure S1). Similar results are observed in scenario 7 for the same examples.

In all scenarios, both methods are in agreement to recommend T_1 when it is indisputably better than T_2, i.e. more effective and safer (or to recommend T_2 when T_1 is indisputably worse, i.e. less effective and more toxic, see Supplemental Material Figure S1). This is well illustrated in scenarios 8 ($\theta_{11} = 0.1$ and $\theta_{12} = 0.9$) and 9 ($\theta_{11} = 0.9$ and $\theta_{12} = 0.1$). In scenario 8, MCDA discriminates slightly better treatments with no efficacy or high risk between themselves, while SLoS penalizes them equally, as they should not be recommended anyway.

Overall, both MCDA and SLoS have good performances to discriminate the benefit-risk balance of the treatments. They provide similar conclusions in many situations, and the cases where they differ highlight the two desirable properties of SLoS. Over all possible scenarios, SLoS recommends safer treatments than MCDA in half of the scenarios, and less safe treatments in the other half.

3.4.3 Sensitivity analyses

While the case of equally important and uncorrelated criteria is considered above, we investigated the robustness of the results in cases of:

- Equally important criteria $w_j = \tilde{w}_j = 0.5$ for $j = 1, 2$ and strongly correlated criteria: $\rho = 0.8$ (positive correlation) and $\rho = -0.8$ (negative correlation).
- More weight on the risk criterion, using MCDA weights (w_1, w_2)=(0.25, 0.75) and mapped SLoS weights $(\tilde{w}_1, \tilde{w}_2)$=(0.30, 0.70), no correlation between the criteria.
- More weight on the risk criterion, with $(w_1, w_2) = (\tilde{w}_1, \tilde{w}_2)$=(0.25, 0.75) (no mapping), no correlation between the criteria. This scenario aims at evaluating

the impact of the weight mapping on the results, by comparing its results to those of the previous case.

The results of the sensitivity analyses are given in Supplemental Material.

Both measures are robust to positive and negative correlations between the outcomes, with very similar results (Supplemental Material, Figures S2-S5). When an MCDA weight of 25% is given to the benefit, both measures penalize more the risk, but analogous differences and similarities as before could be observed between them (Supplemental Material, Figures S6-S7). Since the mapping is not far from an identity transformation, omitting it does not have a major impact on the results (Supplemental Material, Figures S8-S9).

A simulation study was also conducted with four criteria (two benefits $j = 1, 2$ and two risks $j = 3, 4$), for which the investigated scenarios are summarized in Supplemental Material, Table S1, under the following assumptions:

- Equally important criteria with weights $w_j = \tilde{w}_j = 0.25$ for $j = 1, ..., 4$, no correlation between the criteria.
- Equally important criteria with weights $w_j = \tilde{w}_j = 0.25$ for $j = 1, ..., 4$, correlated criteria (see correlation matrices in Supplemental Material).
- More weight on the risk criteria, with MCDA weights (w_1, w_2, w_3, w_4)=(0.10, 0.10, 0.40, 0.40) and mapped SLoS weights $(\tilde{w}_1, \tilde{w}_2, \tilde{w}_3, \tilde{w}_4)$=(0.15, 0.15, 0.43, 0.43), no correlation between the criteria.
- More weight on the risk criteria, with (w_1, w_2, w_3, w_4)=$(\tilde{w}_1, \tilde{w}_2, \tilde{w}_3, \tilde{w}_4)$=(0.10, 0.10, 0.40, 0.40) (no mapping), no correlation between the criteria.

Similar conclusions could be drawn when comparing MCDA and SLoS using four criteria, even if the interpretation of the simulation scenarios is somewhat less straightforward as the amount of possible situations (low/moderate/high benefits and risks) increases (Supplemental Material, Figures S10-S19).

Overall, the conclusions are robust to correlations, number of criteria, weighting and weight mapping for both measures.

3.5 Discussion

In this paper, we propose SLoS as a new tool for drug benefit-risk assessment. It offers the same advantages as MCDA to summarize the benefit-risk balance of the treatments in a single measure, but it has additional desirable properties permitting to avoid recommendations of non-effective or extremely unsafe treatments, and to tolerate larger increases in risk for a given increase in benefit when the amount of benefit is small than when it is high. In contrast, we have shown that the linear form of the MCDA utility score involves implicit assumptions of the decision-makers, such as a constant benefit-risk trade-off for all values of benefit or risk, and might lead to counter-intuitive conclusions. It is worth noting that these additive and linear properties were shown to be inadequate in other application areas of MCDA [78, 84], and its limitations in the health domain have been highlighted as well [87, 83].

The independence of the benefit and risk criteria is usually assumed for the sake of simplicity. Correlations could be taken into account in the analyses, however our simulation study shows that both measures are robust to correlations between outcomes.

Importantly, SLoS penalizes drugs with no efficacy, which is sensible for comparisons between active treatments. Indeed, a 'no treatment/placebo' option, in the absence of placebo effect, will most likely be strongly penalized by SLoS due to its lack of efficacy, although it may be preferable to any active treatment with a small amount of efficacy but that causes more harm overall. Therefore, MCDA's recommendations may be more reliable in such cases and this should be carefully considered before choosing the method and when interpreting the results. However, the area of application of SLoS remains large, as many drug comparisons involve a standard of care, or a placebo with expected effects that are non-negligible [55].

The MCDA weights of the criteria should be elicited according to the preferences of the decision-makers (regulators, experts, patients...) and methods have been proposed in the literature for this purpose [133, 82, 93, 132, 73, 57, 8, 9]. We propose a simple mapping to obtain SLoS weights from MCDA weights, so that the same elicitation process could be followed while preserving the weight interpretation. It should be noted that the mapping is not far from an identity transformation, and omitting it does not strongly affect the results. We considered in this paper fixed weights, but extended models have been proposed where the weights are treated as random variables to account for some uncertainty in their assignment [131, 112].

As an aggregation method involving multiple criteria, SLoS could be included within the family of non-linear MCDA models. It was shown that SLoS has the desirable properties even under the linear partial value functions on which this work has focused only. An alternative approach between linear MCDA and SLoS could be to handle the decreasing level of risk tolerance relative to benefits by varying the shape of the partial value functions. For instance, one can derive linearly-weighted partial value functions used in the linear utility score that exhibit the same degree of decreasing risk tolerance as SLoS. This, however, seems to be non-trivial and requires extensive attention. Furthermore, as stated above, the explicit elicitation of non-linear forms for partial value functions may be difficult for project teams. The weight elicitation and their interpretation appear also more challenging, in particular if the shapes of the partial value functions are different from one criterion to another. Meanwhile, an exploration of the use of non-linear partial value functions both in the framework of the additive utility score and SLoS is of great practical interest and is to be investigated.

In many cases, SLoS and MCDA provide similar conclusions, but SLoS shows clear advantages when treatments have no benefit or extreme risk. In general, this situation may occur in early stage drug developments, or at least before the time of marketing authorization application, since treatments with no evidence of efficacy or high toxicities usually do not reach this point and are stopped before. Until now, benefit-risk assessments were mainly conducted in late stage by the sponsor and/or regulatory agencies, but it is recommended to initiate the benefit-risk assessment earlier in order to better support internal decisions and discussions with health authorities about the development strategy [113]. Therefore, SLoS could be used in

early development, and then updated during the following phases and the regulatory process until post-marketing surveillance, in order to ensure a transparent and consistent benefit-risk assessment throughout the drug life-cycle.

Chapter 4

Decision-making using a composite definition of success

Background

This chapter was published as:

Saint-Hilary G, Robert V, Gasparini M. Decision-making in drug development using a composite definition of success. *Pharmaceutical Statistics*; 2018;1–15. doi: 10.1002/pst.1870. [110]

4.1 Introduction

Decision-making in pharmaceutical development aims at making an optimal choice between several alternatives, at multiple time points during a drug life-cycle, based on the current knowledge of the investigational product. For example, go/no-go decisions are made at the end of phase I and of phase II clinical trials, according to the evidence from the accumulated data and the market potential of the experimental drug compared to other compounds for the same disease. However, decisions are not limited to the continuation or the termination of the development, but are also needed to choose the targeted indication, the patient population, the doses or the study designs.

The success of a drug development is driven by the conjunction between a valuable product and a successful development strategy. A marketing authorization is usually conditioned by the success of the pivotal clinical trials, which must reach statistical significance on their primary endpoint while showing a clinically meaningful effect of the drug (see for example [43, 44, 42]). On the other hand, the benefit-risk balance is a strong predictor of the long-term viability of a medicine, and a key element for the regulatory approval process [41, 45, 136]. Indeed, only medicines with a favorable benefit-risk ratio should be considered, i.e. when the benefits outweigh the risks.

Moreover, even though the final decisions always involve a qualitative judgement from the decision-makers, the project teams need tools to summarize the available information and to assess the chances of success of the drug development. Evidence-based quantitative methodologies have been proposed to inform decision-making, either to develop metrics and standard processes to make go/no-go decisions [23, 49], to assess the benefit-risk balance of the treatments [59, 133, 82, 92] or to predict the statistical significance or the futility of clinical trials [52, 72, 101, 127, 129, 128, 139, 145]. So far, predictions of success and benefit-risk assessments were both used for decision-making, but were considered separately.

The aim of this paper is to propose a comprehensive approach to predict the success of a drug development strategy. We define success as a composite event based on the statistical significance of the treatment effect on the primary endpoint, its clinical relevance and a favorable benefit-risk balance versus the comparator(s) in the next pivotal studies. Using a Bayesian framework, we account for the dependence between the different components, and we also present their marginal predictive probability of occurrence separately for a transparent assessment of the strategies.
The statistical methods to predict the composite success of a drug development strategy and of its components are detailed in section 4.2. In Section 4.3, we present a case-study to compare the chances of success of different development strategies in Major Depressive Disorder. This example is fictive but inspired by a real case where the same statistical methods were used. A discussion and concluding remarks are given in Section 4.4. Additional information may be found in Appendix B.

4.2 Methods

In this section, we suppose that some evidence on the efficacy and safety endpoints is available from one or several clinical non-pivotal trials, and that the future clinical development strategy has been defined with one or several future pivotal trials (Figure 4.1). The future trials are already designed and powered to show superiority of an experimental treatment against a control on a primary endpoint. It is assumed that this primary endpoint was one of the efficacy criteria assessed in the previous trials. First, we will present in Sections 4.2.1 to 4.2.3 how to predict the success of one future trial using our composite definition of success. The extension to drug developments including several future trials is presented in Section 4.2.4.

FIGURE 4.1: Decision-making timepoint

The methods presented here can be simply extended to earlier decision-making timepoints, when some non-pivotal clinical trials are still to be conducted, or later, when results for some pivotal trials have been observed. In the first case, one should expect more uncertainty, while in the second case, the variability is reduced since the outcome of some pivotal trials is observed.
We declare the success of a drug development strategy if, in each pivotal study, the observed treatment effect on the primary endpoint is statistically significant, if it is also clinically relevant, and if the observed benefit-risk balance is better than the comparator(s). If several pivotal trials are planned, we assume that the criteria should be fulfilled in all of them and not only at the development level (using for example meta-analyses or a full Bayesian approach), because one pivotal trial failing to satisfy these criteria is likely to cast some doubts on the replicability of the

results [46]. It should be noted however that, when the safety of a new drug is evaluated for marketing authorization, the individual study safety results are important but pooled analyses should also be provided in order to incorporate long-term, less common and rare outcomes in the overall safety profile. These data are usually not available at the time of the decision-making timepoint considered in this paper and are not incorporated in our composite definition of success.
The predictive probabilities are called respectively $PPoS_1$, $PPoS_2$, $PPoS_3$ and $PPoS$ for the statistical significance on the primary endpoint, its clinical relevance, the positive benefit-risk balance and the overall composite success.

4.2.1 Success criteria based on the primary endpoint

Suppose that the planned analysis on the primary endpoint in the next study follows a conventional frequentist approach testing the null hypothesis $H_0 : \delta \leq 0$ against the alternative, $H_1 : \delta > 0$, where δ is a measure of difference between the experimental treatment and the control. Suppose we have the prior distribution density $f(\delta)$, then its posterior distribution obtained from the data $\mathbf{Y} = \mathbf{y}$ observed in one previous clinical trial or resulting from evidence synthesis of several trials [3, 141, 142] can be calculated according to Bayes theorem as:

$$f(\delta \mid \mathbf{Y} = \mathbf{y}) = \frac{f_{\mathbf{Y}}(\mathbf{y} \mid \delta) f(\delta)}{f(\mathbf{y})}, \tag{4.1}$$

where $f_{\mathbf{Y}}$ is the density of $\mathbf{Y}$ conditional on δ and $f(\mathbf{y}) = \int f_{\mathbf{Y}}(\mathbf{y} \mid \delta) f(\delta) d\delta$.
Let d^* be the difference between treatments that will be observed on the primary endpoint in the next trial, and f_{d^*} its density conditional on δ. The probability to have d^* greater than a pre-defined threshold D in the next trial conditional on δ is:

$$P(d^* > D \mid \delta) = \int_{z>D} f_{d^*}(z \mid \delta) dz.$$

Its predictive probability after observing the data from the previous trials can therefore be calculated using the posterior distribution $f(\delta \mid \mathbf{Y} = \mathbf{y})$, under the usual assumption of conditional independence of the next trial from the previous ones given δ:

$$P(d^* > D \mid \mathbf{Y} = \mathbf{y}) = \int \int_{z>D} f_{d^*}(z \mid \delta) f(\delta \mid \mathbf{Y} = \mathbf{y}) dz d\delta.$$

Using Equation (4.1), it can be re-written as:

$$P(d^* > D \mid \mathbf{Y} = \mathbf{y}) = \frac{\int \int_{z>D} f_{d^*}(z \mid \delta) f_{\mathbf{Y}}(\mathbf{y} \mid \delta) f(\delta) dz d\delta}{\int f_{\mathbf{Y}}(\mathbf{y} \mid \delta) f(\delta) d\delta}. \tag{4.2}$$

For example, assume that the current posterior distribution of δ based on the available evidence (i.e., having seen $\mathbf{Y} = \mathbf{y}$) is normal $N(d, s^2)$, and the distribution of d^* conditional on δ is normal with $d^* \mid \delta \sim N(\delta, s^{*2})$, where s^{*2} is its variance in the next trial. From the posterior distribution of δ and the distribution of $d^* \mid \delta$, we obtain the predictive distribution:

$$d^* \mid \mathbf{Y} = \mathbf{y} \sim \int f_{d^*}(z \mid \delta) f(\delta \mid \mathbf{Y} = \mathbf{y}) d\delta = N(d, s^2 + s^{*2}).$$

Therefore the predictive probability of $d^* > D$ is:

$$P\left(d^* > D \mid d, s^2\right) = 1 - \Phi\left(\frac{D - d}{\sqrt{s^2 + s^{*2}}}\right),$$

where Φ denotes the cumulative distribution function of the standard normal distribution.

We define the predictive probabilities of two success criteria based on the primary endpoint:

- **Statistical significance.** When $D = c$, with $c > 0$ the critical value at which the null hypothesis H_0 is rejected at a pre-specified significance level α, the probability in Equation (4.2) is the predictive probability of statistical significance on the primary endpoint in the next trial, and we note it $PPoS_1$. Its closed formula has been derived in earlier work, where it is also called *assurance* [101] or *Bayesian predictive power* [126, 109]. In the example with a normal distribution presented above, we have $c = z_\alpha s^*$, where z_α is the $(1-\alpha)100^{th}$ percentile of the standard normal distribution.

- **Clinical relevance.** While the statistical significance is a gatekeeper to declare the success of a trial, the clinical relevance of the observed difference between treatments on the primary endpoint is also required for success (see for example [43, 44, 42]). We define the probability of clinical relevance on the primary endpoint, $PPoS_2$, as the probability in Equation (4.2) for $D = d_T$ a pre-defined minimal clinically relevant threshold.

According to the regulatory recommendations, the study should be powered such that the anticipated treatment effect is equal to or larger than d_T [67]. Statistical significance is easier to reach than clinical relevance ($PPoS_1 > PPoS_2$) if $c < d_T$, when for example the study is powered with an anticipated treatment effect that is the minimal clinically relevant difference. Clinical relevance is easier to reach than statistical significance ($PPoS_1 < PPoS_2$) if $c > d_T$, when for example a treatment effect greater than d_T is anticipated and c is large due to the management of multiplicity issues. $PPoS_1$ and $PPoS_2$ are equal if $c = d_T$, i.e. if c is just clinically meaningful.

4.2.2 Success criterion based on the benefit-risk balance

While the success of a pivotal clinical trial is often focused on the primary efficacy endpoint, the decisions regarding the drug development and its licensing are taken considering several efficacy and safety endpoints, i.e. by assessing the benefit-risk balance of the new drug versus comparator(s). Several quantitative methodologies have been proposed [59, 133, 82, 92, 90, 91] and provide an explicit quantitative information on benefits and risks in order to assist the decision-making process. In this paper, we choose a Multi-Criteria Decision Analysis (MCDA) [93, 94, 98], since the European Medicine Agency Benefit-Risk Methodology Project suggested that it is one of the most comprehensive among the quantitative methodologies they considered [35, 36, 37, 38], and it is also recommended by the IMI PROTECT Work package 5 [69]. Other methodologies can be chosen and the methods described in this paper can be adapted accordingly. In this section, we first briefly present the MCDA model, and then show how it can be used to calculate another component of the

predictive probability of success in a next trial.

The principle of MCDA is to compare several treatments using utility scores calculated from multiple criteria of benefit and risk, and taking into account their relative importance according to the preferences of the decision-makers. In the initial version of MCDA [93, 94], the scoring process of the treatments is deterministic and ignores the parameter uncertainty induced by the data sampling variation. Instead, we use a probabilistic model, often called Probabilistic MCDA (or Stochastic MCDA), developed by Waddingham *et al.* [138] which estimates the score distributions based on the distributions of the criterion parameters, which are themselves estimated from the treatment effects observed in previous studies.

Consider the experimental treatment and the control denoted by $i = 1, 2$ respectively, assessed on n criteria $(j = 1, ..., n)$, and the following quantities and functions [93]:

(i) The performance of treatment i on criterion j is denoted by ξ_{ij}. The vector of criterion performances for the treatment i is denoted by $\boldsymbol{\xi}_i = (\xi_{i1}, ..., \xi_{in})$.

(ii) The monotonically increasing partial value functions $0 \leq u_j(\cdot) \leq 1$ are used to normalize the criterion performances. Let ξ_j' and ξ_j'' be the most and the least preferable values, then $u_j(\xi_j'') = 0$ and $u_j(\xi_j') = 1$. The inequality $u_j(\xi_{ij}) > u_j(\xi_{hj})$ indicates that the performance of the treatment i is preferred to the performance of the treatment h on criterion j. A common choice for the function [133, 93, 138, 80] is

$$u_j(\xi_{ij}) = \frac{\xi_{ij} - \xi_j''}{\xi_j' - \xi_j''}.$$

(iii) The weights indicating the relative importance of the criteria are known constants denoted by w_j, with the constraint that $\sum_{j=1}^{n} w_j = 1$. The w_j should be provided by the decision-makers. The vector of weights used for the analysis is denoted by $\mathbf{w} = (w_1, ..., w_n)$.

It is generally assumed that the criteria are independent, which allows us to use an additive formula to calculate the global utility score:

$$u_i = u(\boldsymbol{\xi}_i, \mathbf{w}) = w_1 u_1(\xi_{i1}) + ... + w_n u_n(\xi_{in}) = \sum_{j=1}^{n} w_j u_j(\xi_{ij}).$$

The utility score is a measure of benefit-risk, which permits to discriminate the treatments according to their performances, and according to the weights attributed to the criteria. The highest the utility score, the most preferable the benefit-risk ratio, therefore a treatment has a positive benefit-risk balance compared to the control if the difference between the two utility scores is positive:

$$\Delta u_{12} = \Delta u(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \mathbf{w}) = u(\boldsymbol{\xi}_1, \mathbf{w}) - u(\boldsymbol{\xi}_2, \mathbf{w}) > 0.$$

Following the approach proposed by Waddingham *et al.* [138], we consider a Bayesian model and assign a probability distribution to the ξ_{ij}, which are considered as unknown parameters. Suppose the information we have about ξ_{ij} prior to the clinical development is expressed through the prior distribution density $f(\xi_{ij})$. Its posterior distribution can be obtained from the data $X_{ij} = x_{ij}$ summarizing the available

evidence, according to Bayes theorem:

$$f(\xi_{ij} \mid X_{ij} = x_{ij}) = \frac{f_{X_{ij}}\left(x_{ij} \mid \xi_{ij}\right) f\left(\xi_{ij}\right)}{f(x_{ij})}, \tag{4.3}$$

where $f_{X_{ij}}$ is the density of X_{ij} conditional on ξ_{ij} and $f(x_{ij}) = \int f_{X_{ij}}\left(x_{ij} \mid \xi_{ij}\right) f\left(\xi_{ij}\right) d\xi_{ij}$. It follows that the utility scores u_i and their difference between two treatments Δu_{12} are unobservable random variables.

At the sampling level, on the other hand, there will usually exist observable random variables x_{ij}^* which are estimates of the ξ_{ij} in the next trial, much like in the discussion about efficacy there exists an observable random variable d^* which is an estimate of δ. Let $\mathbf{x_i^*}$ be the vectorized notation of x_{ij}^* accross the criteria.

To fulfill our stated goal of requiring a positive benefit-risk balance at the trial level on each pivotal study, consider then $\Delta^* u_{12} = \Delta u(\mathbf{x_1^*}, \mathbf{x_2^*}, \mathbf{w})$ the observed difference between the utility scores of the experimental treatment and the control in the next trial. Let $f_{\Delta^* u_{12}}$ be its density conditional on unknown true values of the parameters $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$: $f_{\Delta^* u_{12}}$ takes into account the data sampling variation in the next study. The probability of observing a positive benefit-risk balance of the experimental treatment versus the control in the next trial conditional on $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ is calculated as

$$P(\Delta^* u_{12} > 0 \mid \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \int_{v>0} f_{\Delta^* u_{12}}(v \mid \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) dv.$$

Its predictive probability after observing the data from the previous trials can therefore be calculated using the posterior distributions $f(\boldsymbol{\xi}_1 \mid \mathbf{X_1} = \mathbf{x_1})$ and $f(\boldsymbol{\xi}_2 \mid \mathbf{X_2} = \mathbf{x_2})$, given that $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are assumed to be independent:

$$\begin{aligned} PPoS_3 &= P(\Delta^* u_{12} > 0 \mid \mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}) \\ &= \int\int\int_{v>0} f_{\Delta^* u_{12}}(v \mid \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) f(\boldsymbol{\xi}_1 \mid \mathbf{X_1} = \mathbf{x_1}) f(\boldsymbol{\xi}_2 \mid \mathbf{X_2} = \mathbf{x_2}) dv d\boldsymbol{\xi}_1 d\boldsymbol{\xi}_2. \end{aligned}$$

Using Equation (4.3) and the vectorized notations, it can be re-written as:

$$PPoS_3 = \frac{\int\int\int_{v>0} f_{\Delta^* u_{12}}(v \mid \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) f_{\mathbf{X_1}}\left(\mathbf{x_1} \mid \boldsymbol{\xi}_1\right) f_{\mathbf{X_2}}\left(\mathbf{x_2} \mid \boldsymbol{\xi}_2\right) f\left(\boldsymbol{\xi}_1\right) f\left(\boldsymbol{\xi}_2\right) dv d\boldsymbol{\xi}_1 d\boldsymbol{\xi}_2}{\int f_{\mathbf{X_1}}\left(\mathbf{x_1} \mid \boldsymbol{\xi}_1\right) f\left(\boldsymbol{\xi}_1\right) d\boldsymbol{\xi}_1 \int f_{\mathbf{X_2}}\left(\mathbf{x_2} \mid \boldsymbol{\xi}_2\right) f\left(\boldsymbol{\xi}_2\right) d\boldsymbol{\xi}_2}.$$

While these formula are likely to be difficult to resolve analytically, the results can be easily obtained by simulations according to the following steps:

(i) The posterior distributions $f(\boldsymbol{\xi}_1 \mid \mathbf{X_1} = \mathbf{x_1})$ and $f(\boldsymbol{\xi}_2 \mid \mathbf{X_2} = \mathbf{x_2})$ of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are obtained using classical Bayesian methods [125], either analytically or with Markov Chain Monte Carlo (MCMC) simulations.

(ii) Values $\boldsymbol{\xi}_1^{*(k)}$ and $\boldsymbol{\xi}_2^{*(k)}$ are sampled from $f(\boldsymbol{\xi}_1 \mid \mathbf{X_1} = \mathbf{x_1})$ and $f(\boldsymbol{\xi}_2 \mid \mathbf{X_2} = \mathbf{x_2})$, for $k = 1, ..., K$ where K is the total number of simulations (a large number). These simulations can come from the MCMC simulations, after the chain(s) converged.

(iii) Observed values $\mathbf{x_1^{*(k)}}$ and $\mathbf{x_2^{*(k)}}$ of the performances of the treatments in the next trial are simulated from $f_{\mathbf{X_1}}(\mathbf{x_1} \mid \boldsymbol{\xi}_1^{*(k)})$ and $f_{\mathbf{X_2}}(\mathbf{x_2} \mid \boldsymbol{\xi}_2^{*(k)})$, according to the study design and in particular the planned number of patients.

(iv) The difference between treatment utility scores is calculated for each simulated trial k as $\Delta^{*(k)} u_{12} = u(\mathbf{x_1^{*(k)}}, \mathbf{w}) - u(\mathbf{x_2^{*(k)}}, \mathbf{w})$.

(v) The predictive probability of positive benefit-risk balance of the experimental treatment versus the control in the next trial is approximated by

$$PPoS_3 \approx \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}\left[\Delta^{*(k)} u_{12} > 0\right],$$

where $\mathbb{1}[true] = 1$ and $\mathbb{1}[false] = 0$.

4.2.3 Composite success

We define the success of a drug development strategy as the simultaneous fulfillment of the following criteria in all the pivotal studies:

(i) The statistical significance on the primary endpoint.

(ii) A clinically meaningful effect on the primary endpoint.

(iii) A positive benefit-risk balance versus the comparator(s).

Therefore, the predictive probability of composite success of a drug development strategy, with one future pivotal study, can be written as:

$$PPoS \quad = \quad P\left[(d^* > \max(c, d_T)) \cap (\Delta^* u_{12} > 0) \mid \mathbf{Y} = \mathbf{y}, \mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}\right].$$

It is highly unlikely that δ is independent of ξ_1 and ξ_2, since the primary endpoint is almost always one of the criteria considered in the benefit-risk assessment. Therefore, we consider the joint distribution of (δ, ξ_1, ξ_2) to write explicitely the formula of the PPoS, following the same principle as in the previous sections:

$$PPoS = \frac{\int\int_Z f_{d^*,\Delta^* u_{12}}(z, v \mid \delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2) f_{\mathbf{Y},\mathbf{X_1},\mathbf{X_2}}(\mathbf{y}, \mathbf{x_1}, \mathbf{x_2} \mid \delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)\, f(\delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)\, d(z, v) d(\delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)}{\int f_{\mathbf{Y},\mathbf{X_1},\mathbf{X_2}}(\mathbf{y}, \mathbf{x_1}, \mathbf{x_2} \mid \delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)\, f(\delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)\, d(\delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2)},$$

where $Z = \{z > \max(c, d_T), v > 0\}$.

It should be noted that, for a fixed c (which usually depends on a fixed type I error α, the estimation from previous evidence of the variability of the primary endpoint and the number of patients or of events in the next study) and a pre-defined threshold d_T, one should know in advance the maximum between c and d_T, and only one of the two criteria is actually needed in the formula. On the other hand, these two criteria are useful to communicate with non-statisticians on the definition of success. Using both thresholds permits to calculate several $PPoS$ separately and to discuss them with the project team while the discussions regarding the sample size of the next study or the choice of the threshold d_T are still on-going, without changing the formula itself. For transparency, the $PPoS$ should be provided along with its components $PPoS_1$, $PPoS_2$ and $PPoS_3$, to present which ones are the most restrictive and have the greatest impact on the predictive probability of composite success. The predictive probabilities of achieving two components out of three can also be calculated and discussed.

4.2.4 Development strategies with more than one future studies

Suppose now that the future development strategy consists in S future pivotal trials. We assume that the development strategy will be successful if the criteria of

statistical significance, clinical relevance and positive benefit-risk balance are fulfilled in each of the pivotal trials. The estimates of the efficacy and safety criterion performances in the next trials are conditionally independent, between trials, given the posterior distribution of their parameters. The predictive probabilities can be obtained by marginalizing over the parameters, using the posterior distributions:

$$
\begin{aligned}
PPoS_1 &= \int \left(\prod_{m=1}^{S} P\left[d^{*m} > c^m \mid \delta \right] \right) f(\delta \mid \mathbf{Y} = \mathbf{y}) d\delta, \\
PPoS_2 &= \int \left(\prod_{m=1}^{S} P\left[d^{*m} > d_T^m \mid \delta \right] \right) f(\delta \mid \mathbf{Y} = \mathbf{y}) d\delta, \\
PPoS_3 &= \int \int \left(\prod_{m=1}^{S} P\left[\Delta^{*m} u_{12} > 0 \mid \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \right] \right) f(\boldsymbol{\xi}_1 \mid \mathbf{X_1} = \mathbf{x_1}) f(\boldsymbol{\xi}_2 \mid \mathbf{X_2} = \mathbf{x_2}) d\boldsymbol{\xi}_1 d\boldsymbol{\xi}_2, \\
PPoS &= \int \left(\prod_{m=1}^{S} P\left[(d^{*m} > \max(c^m, d_T^m)) \cap (\Delta^{*m} u_{12} > 0) \mid \delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \right] \right) \times \\
&\qquad f(\delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \mid \mathbf{Y} = \mathbf{y}, \mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}) d\left(\delta, \boldsymbol{\xi}_1, \boldsymbol{\xi}_2\right),
\end{aligned}
$$

where d^{*m}, c^m, d_T^m and $\Delta^{*m} u_{12}$ are respectively the observed difference between treatments on the primary endpoint, the critical value at which the null hypothesis will be rejected, the clinical threshold and the observed difference between treatment utility scores in study m. As before, these formulas are likely to be difficult to resolve analytically, but the results can be easily obtained by simulations.

4.3 Example in Major Depressive Disorder

In this section, we illustrate the use of the above methods to support decision-making between different future strategies of development in Major Depressive Disorder. This example is fictive but inspired by a real case discussed with a project team: the clinical context, the indication and the data have been changed for confidentiality reasons, but the essence of the problem and the statistical methods are the same.

4.3.1 Context and data

We assume that the results of one Phase II trial are available, which compared a Low dose and a High dose of an experimental treatment versus placebo. Suppose that only one pivotal two-arm Phase III study is planned to compare this experimental treatment versus placebo. The dose or regimen of the experimental treatment group needs to be chosen, according to its probability to reach statistical significance and clinical relevance on the primary endpoint and to have a positive benefit-risk balance versus placebo in the next trial.

The primary efficacy endpoint for both the Phase II and the Phase III trials is the total score on the Hamilton Depression Rating Scale 17 items (HAM-D_{17}) after 6 weeks of treatment. The HAM-D_{17} total score ranges from 0 to 52, with higher values indicating a higher severity of illness. The safety of the treatment is mainly assessed by the proportion of patients experiencing emergent adverse events during the study. Descriptive statistics of the results of the Phase II study on the HAM-D_{17} total score

and on the five more frequent adverse events are presented in Table 4.1. A dose-response relationship is observed, with the Higher dose showing a better efficacy but also more adverse events than the Low dose. In particular, Hypokalemia are observed in 71% of the patients at High dose: this adverse event may be a safety concern for this dose.

TABLE 4.1: Results of the Phase II study for the primary efficacy endpoint and the five more frequent adverse events (descriptive statistics)

	Low dose	High dose	Placebo
Efficacy (Intent-To-Treat population)			
N	50	48	51
HAM-D_{17} – Mean (SD)	14.0 (6.9)	12.6 (7.1)	16.9 (6.9)
Safety (Safety population)			
N	50	49	52
Hypokalemia — n (%)	1 (2%)	35 (71%)	0 (0%)
Nausea — n (%)	8 (16%)	14 (29%)	2 (4%)
Diarrhea — n (%)	4 (8%)	8 (16%)	1 (2%)
Dizziness — n (%)	5 (10%)	9 (18%)	0 (0%)
Headache — n (%)	7 (14%)	7 (14%)	3 (6%)

HAM-D_{17}: HAM-D_{17} total score at 6 weeks ; SD = Standard Deviation

The next Phase III study is designed as a two-arm trial comparing one regimen of the experimental treatment, to be chosen, versus placebo on the HAM-D_{17} total score. A sample size of 228 patients (114 per arm) is planned to reach a power of 90%, based on an assumed difference of 3 points on the HAM-D_{17} total score at 6 weeks, a standard-deviation of 7 and a one-sided α of 2.5%. Both statistical significance and clinical relevance on this endpoint should be achieved in this trial to apply for a marketing authorization. There is no consensus on the minimally relevant effect but the clinical relevance would be indisputable for a threshold $d_T = 3$ points. For the MCDA analysis, we consider the HAM-D_{17} total score as the only criterion of benefit, and the occurrence of the five more frequent adverse events as the risk criteria.

4.3.2 Bayesian model

The prior distributions, the sampling distributions (likelihoods) and the posterior distributions of all the parameters used in the model are summarized in Table 4.2.
The HAM-D_{17} total score is usually and reasonably assumed to be normally distributed [16]. The mean effects in each arm for the Low dose, the High dose and the placebo ($i = 1, 2, 3$ respectively) are denoted by the parameters ξ_{i1}. Their posterior distributions are obtained from a weakly informative conjugate prior $\xi_{i1} \sim N(0, 10^4)$ and the sample means $m_1 = 14.0$, $m_2 = 12.6$ and $m_3 = 16.9$ observed in the Phase II study, which are realizations of the normal distributions $N(\xi_{i1}, \sigma_i^2)$ with $\sigma_1 = 0.98$, $\sigma_2 = 1.02$ and $\sigma_3 = 0.97$.
The parameters of the treatment differences versus placebo for each dose $i = 1, 2$ are $\delta_i = \xi_{31} - \xi_{i1}$. Their posterior distributions are obtained from a weakly informative prior $\delta_i \sim N(0, 2 \times 10^4)$ (induced by the priors on ξ_{i1}) and from the observed differences between treatments $d_i = m_3 - m_i$ which are realizations of the normal distributions $N(\delta_i = \xi_{31} - \xi_{i1}, s_i^2)$ for $i = 1, 2$, where $s_1 = 1.37$ and $s_2 = 1.41$ are the standard errors of the differences.
The five more frequent adverse events are binary events. We note r_{ij}, n_{ij} and ξ_{ij} respectively the number of events, the number of patients and the probability of event

TABLE 4.2: Distributions of the parameters

Parameter	Estimate	Prior	Likelihood	Posterior
HAM-D_{17} mean total score in each arm ($i = 1, ..., 3$)				
ξ_{i1}	m_i	$N(0, \sigma_0^2)$	$N(\xi_{i1}, \sigma_i^2)$	$N\left(\frac{\sigma_0^2}{\sigma_0^2+\sigma_i^2} m_i, \frac{\sigma_0^2\sigma_i^2}{\sigma_0^2+\sigma_i^2}\right)$
HAM-D_{17} mean total score, difference versus placebo ($i = 1, 2$)				
$\delta_i = \xi_{31} - \xi_{i1}$	$d_i = m_3 - m_i$	$N(0, s_0^2)$	$N(\delta_i, s_i^2)$	$N\left(\frac{s_0^2}{s_0^2+s_i^2} d_i, \frac{s_0^2 s_i^2}{s_0^2+s_i^2}\right)$
Occurrence of adverse events in each arm ($i = 1, ..., 3$; $j = 2, ..., 6$)				
ξ_{ij}	r_{ij}/n_{ij}	$Beta(1,1)$	$Bin(n_{ij}, \xi_{ij})/n_{ij}$	$Beta(r_{ij}+1, n_{ij}-r_{ij}+1)$
$\sigma_0^2 = 10^4$; $s_0^2 = 2 \times \sigma_0^2 = 2 \times 10^4$				.

for treatment i ($i = 1, 2, 3$) and safety criterion j ($j = 2, ..., 6$). We obtain the posterior distributions of the parameters ξ_{ij} from the realizations r_{ij} of the binomial densities $Bin\left(n_{ij}, \xi_{ij}\right)$ and uniform conjugate priors $\xi_{ij} \sim Beta\left(1, 1\right)$.
The partial value functions of all criteria are defined as linear functions as presented in Section 4.2.2. The best and the worst values of the HAM-D_{17} mean total score at 6 weeks in the patient population are assumed to be 10 and 25 respectively. The range of the probabilities of adverse event is $[0, 1]$, so the best and the worst values for the risk criteria are naturally defined as 0 and 1 respectively.
Benefits and risks are assumed to have an equal importance, with a weight of 50% attributed to the HAM-D_{17} total score and 50% in total for the safety criteria, split as 20% for Hypokalemia and 7.5% for each of the other adverse events. The median and 95% credible intervals of the posterior distributions, the partial value functions and the weights are summarized in Table 4.3.

TABLE 4.3: Median and 95% credible interval (CrI) of the posterior distributions of the benefit and risk parameters, their partial value functions and their weight

	Posterior distribution Median (95% CrI)			Partial value function	Weight
	Low dose	High dose	Placebo		
Benefit criterion					
HAM-D_{17}	14.00 (12.13;15.86)	12.59 (10.54;14.66)	16.90 (15.07;18.73)	$u_1(\xi_{i1}) = \frac{25-\xi_{i1}}{25-10}$	50%
Risk criteria					
Hypokalemia	0.02 (0.00;0.10)	0.71 (0.58;0.82)	0.00 (0.00;0.07)	$u_2(\xi_{i2}) = 1 - \xi_{i2}$	20%
Nausea	0.16 (0.08;0.29)	0.29 (0.18;0.42)	0.04 (0.01;0.13)	$u_3(\xi_{i3}) = 1 - \xi_{i3}$	7.5%
Diarrhea	0.08 (0.03;0.19)	0.16 (0.09;0.29)	0.02 (0.00;0.10)	$u_4(\xi_{i4}) = 1 - \xi_{i4}$	7.5%
Dizziness	0.10 (0.04;0.21)	0.18 (0.10;0.31)	0.00 (0.00;0.07)	$u_5(\xi_{i5}) = 1 - \xi_{i5}$	7.5%
Headache	0.14 (0.07;0.26)	0.14 (0.07;0.27)	0.06 (0.02;0.16)	$u_6(\xi_{i6}) = 1 - \xi_{i6}$	7.5%

The results of the next Phase III study are simulated conditional on the parameters ξ_{ij} and δ_i, which have the posterior distributions defined in Table 4.2, and assuming that 114 patients per arm are included:

(i) Means HAM-D_{17} total score: $m_i^* \mid \xi_{i1} \sim N\left(\xi_{i1}, \sigma_i^{*2}\right)$ for $i = 1, 2, 3$, with the standard errors in the new trial σ_i^* fixed to $7/\sqrt{114} \approx 0.66$, i.e. with a standard deviation in all arms equal to 7 according to the literature and to the data

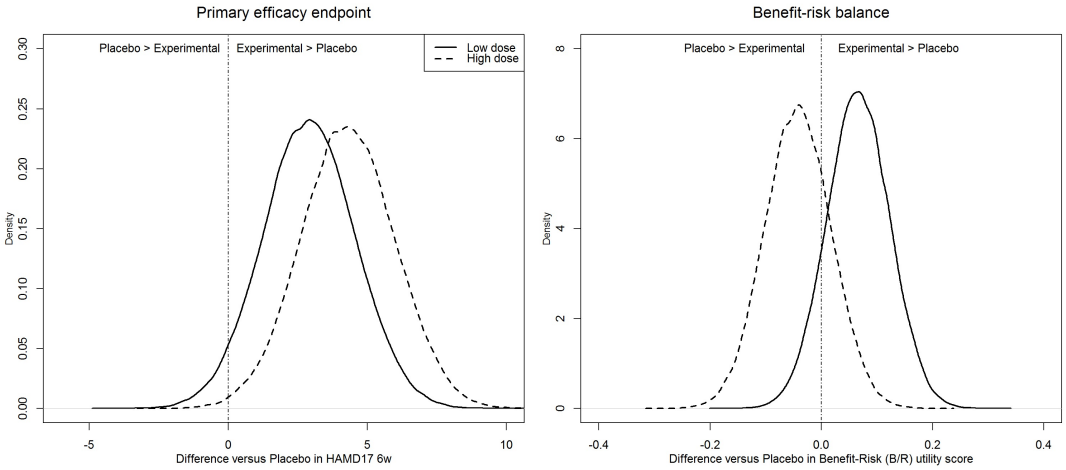

FIGURE 4.2: Left: predictive distributions of the differences in Hamilton Depression Rating Scale 17 items (HAM-D_{17}) mean total score of each dose versus placebo in the next phase 3 study. Right: predictive distributions of the differences in benefit-risk (B/R) utility scores of each dose versus placebo in the next phase 3 study

observed in the Phase II study.

(ii) Differences in HAM-D_{17} mean total score versus placebo: $d_i^* \mid \delta_i \sim N\left(\delta_i, s_i^{*2}\right)$ for $i = 1, 2$, with the standard errors in the new trial $s_i^* = \sqrt{2}\sigma_i^* \approx 0.93$.

(iii) Proportions of adverse events: $p_{ij}^* \mid \xi_{ij} = r_{ij}^*/114$ with $r_{ij}^* \sim Bin\left(114, \xi_{ij}\right)$ for $i = 1, 2, 3$ and $j = 2, ..., 6$.

(iv) Benefit-risk utility scores: $u(m_i^*, p_{i2}^*, ..., p_{i6}^*, \mathbf{w}) = w_1 u_1(m_i^*) + w_2 u_2(p_{i2}^*) + ... + w_6 u_6(p_{i6}^*)$ with $\mathbf{w} = (0.5,\ 0.2,\ 0.075,\ 0.075,\ 0.075,\ 0.075)$. As before, for simplicity, we note $u(m_i^*, p_{i2}^*, ..., p_{i6}^*, \mathbf{w}) = u_i^*$.

(v) Differences in benefit-risk utility score versus placebo: $\Delta^* u_{i3} = u_i^* - u_3^*$ for $i = 1, 2$.

The analyses were conducted using `R`, and 100,000 simulations were run to estimate the parameter distributions and the probabilities of success.

4.3.3 First results

The predictive distributions of the differences in HAM-D_{17} mean total score, d_1^* and d_2^*, and the predictive distributions of the differences in benefit-risk utility score, $\Delta^* u_{13}$ and $\Delta^* u_{23}$, of each dose versus placebo in the next Phase III study are presented in Figure 4.2. The predictive probability of composite success of the development strategies, $PPoS$, along with the predictive probabilities of its components, $PPoS_1$, $PPoS_2$ and $PPoS_3$ are presented in Table 4.4.
Regarding the primary efficacy endpoint, statistical significance is reached if the difference between treatments on the primary endpoint in the next study is greater than $1.96 s^* \approx 1.82$ and the clinical relevance is indisputably achieved if it is greater than $d_T = 3$, therefore the statistical significance is easier to achieve than the clinical relevance ($PPoS_1 > PPoS_2$). The predictive probabilities for the High dose to fulfill

TABLE 4.4: Predictive probabilities of success

Dose	$PPoS_1$ (statistical significance)	$PPoS_2$ (clinical relevance)	$PPoS_3$ (positive B/R balance)	$PPoS$ (overall)
Low dose	74%	48%	88%	48%
High dose	93%	78%	24%	24%

these criteria are high (93% and 78% respectively). The predictive probability for the Low dose to achieve the statistical significance is also encouraging (74%), but its capacity to reach the clinical relevance could be questionable (48%). Therefore, if the choice between the two doses was based only on the primary efficacy endpoint, the High dose would be preferred.
On the other hand, the Low dose has a high predictive probability of positive benefit-risk balance versus placebo (88%). In contrast, despite its encouraging efficacy results, the High dose has a safety profile which leads to a probability of only 24% to show a better benefit-risk balance than placebo in the next Phase III.
Overall, the predictive probabilities of composite success of the drug development strategies are only 48% and 24% respectively for the Low dose and the High dose. It should be noted, and emphasized during the discussions with the decision-makers, that the probability of success of the Low dose is bounded by its probability to achieve the clinical relevance on the primary endpoint with $d_T = 3$ points, while the success of the High dose is compromised by potential safety concerns.

4.3.4 Strategy refinement

Based on the previous results, the project team can consider either stopping the development, choosing the Low dose despite its low predictive probability of composite success if the chosen clinical threshold is considered to be an ambitious target, or changing of strategy. Indeed, the unfavorable benefit-risk balance of the High dose prevents from choosing it for further development. However, it is observed that the most frequent adverse event at this dose is Hypokalemia, which could be managed for example by a supplementation in potassium co-administered with the drug. The project team may also consider another strategy which consists in initiating all patients at the Low dose, and to increase at the High dose only those not responding to treatment at short term. This would permit to limit, although not completely preventing, the occurrence of Hypokalemia, while increasing the overall efficacy of the regimen compared to the Low dose only. Since no data were available for these two regimen, the clinical assumptions were incorporated in the model as follows:

(i) **High dose with potassium supplements.** The predictions are based on the posterior distribution of ξ_{32} obtained for the placebo for Hypokalemia, and on the posterior distributions obtained for the High dose (as in the previous section) for all the other criteria.

(ii) **Dose increase.** According to the clinicians, 30% to 40% of the patients would increase to the High dose in the Phase III study, therefore a new parameter with a uniform prior distribution $\zeta \sim U[0.3, 0.4]$ is used in the model as the proportion of patients receiving the High dose. We make the assumption that the expected efficacy and safety of the experimental treatment in the subpopulation of responder patients staying at Low dose are the same as those observed for all patients receiving Low dose in the Phase II study. Similarly, we

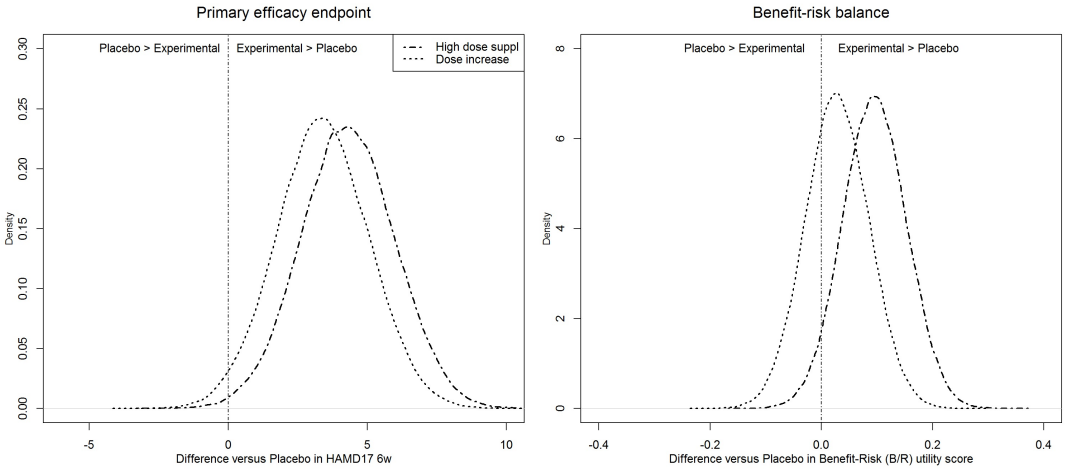

FIGURE 4.3: Left: predictive distributions of the differences in Hamilton Depression Rating Scale 17 items (HAM-D_{17}) mean total score of each dose versus placebo in the next phase 3 study. Right: predictive distributions of the differences in benefit-risk (B/R) utility scores of each regimen versus placebo in the next phase 3 study. High dose suppl = High dose with potassium supplements

suppose that the expected efficacy and safety in the subpopulation of nonresponder patients increasing to High dose are the same as those observed for all patients receiving High dose in the Phase II study. This could be debatable, however it is considered to be a reasonable assumption and no other objective hypothesis could be made. As a consequence, the parameters associated to the efficacy and safety criteria are assumed to be linear combinations of the initial parameters: $(1-\zeta) \times \xi_{1j} + \zeta \times \xi_{2j}$ for $j = 1, ..., 6$.

The predictive distributions of the differences in HAM-D_{17} mean total score and of the differences in benefit-risk utility score of each new regimen versus placebo in the next Phase III study are presented in Figure 4.3, and the predictive probabilities of success are summarized in Table 4.5.

TABLE 4.5: Predictive probabilities of success

Regimen	$PPoS_1$ (statistical significance)	$PPoS_2$ (clinical relevance)	$PPoS_3$ (positive B/R balance)	$PPoS$ (overall)
High dose suppl	93%	78%	95%	78%
Dose increase	83%	59%	69%	58%

High dose suppl = High dose with potassium supplements

The supplementation in potassium substantially improves the benefit-risk balance of the High dose, which is now predicted to be positive versus placebo with a probability of 95%, leading to a predictive probability of composite success of 78% for this regimen. The dose increase, as expected, improves the chances to observe a clinically relevant difference on the primary endpoint compared to the Low dose. However, its predictive probability of composite success is only 58%. Given these results, the best strategy seems to choose for further development the High dose with a co-administration of potassium supplements, if the external factors (feasibility, quality of life, price...) do not alter this conclusion.

4.3.5 Sensitivity analyses

We investigated the robustness of the results in cases of:

- Uncertainty in the weight elicitation, by applying a Dirichlet Stochastic Multicriteria Acceptability Analysis (Dirichlet SMAA) model [112], where the weights are treated as random variables, and their variance depends on the decision-makers' confidence in their elicitation.
- Correlated criteria, by considering correlation patterns where (i) all criteria are positively correlated, or (ii) the benefit criterion is negatively correlated with the risk criteria, and the risk criteria are positively correlated between themselves.
- Departure from the clinical assumptions for the strategy refinement, where the priors on the corresponding parameters (probability of Hypokalemia, proportion of patients receiving the High dose) are changed.

The results of the sensitivity analyses are given in Appendix B. Overall, the conclusions are robust to uncertainty in weight elicitation, correlations, and departure from clinical assumptions.

4.3.6 Alternative example

An alternative example is presented in Appendix B, with the two following changes:

- The threshold of minimal clinical relevance is fixed at $d_T = 2$ points. This could be relevant if, for example, the drug is an add-on therapy administered on top of a standard therapy, so the difference versus the control group may not need to be as large as for a monotherapy.
- Three experimental arms are to be included in the next Phase III trial, and they should be selected among the four possible regimen (Low dose, High dose, High dose with potassium supplements and Dose increase).

This example illustrates a case where clinical relevance is easier to reach than statistical significance ($PPoS_1 < PPoS_2$). Since $PPoS_3$ is unchanged, the results indicate that the High dose could be excluded from the selected regimen for Phase III, as in the initial example, due to its low probability to show a positive benefit-risk balance versus the control.

4.4 Discussion

The approach described in the paper provides some new quantitative methods for predicting the success of a drug development by comparing several development strategies using a composite definition of success, including the statistical significance of the future trial(s) on the primary efficacy endpoint, the clinical relevance of the treatment effect and a positive benefit-risk balance of the drug. The methods are based on the available evidence from previous trials, which could be combined with new additional hypotheses on the future development (such as a modification of the regimen of a drug) using priors. The resulting predictive probability of composite success and its components have demonstrated their utility in an actual go/no-go decision setting, which inspired us to present a fictive, but realistic, example. Other

applications could be considered, such as a decision-making tool for the selection between several doses at the interim analysis of an adapative design trial, or a measure of development risks to be incorporated in financial tools for portfolio management and valuation of investments [2].

Quantitative benefit-risk assessment requires many assumptions that may appear at first difficult to elicitate, and the important role of value judgments in this undertaking needs to be emphasized. This actually reflects the complexity of the context in which drugs are evaluated, and the cognitive load required for health care decisions (see [113, Chapter 5] for a full discussion on the challenges faced by the use of explicit quantitative methods in benefit-risk assessment, and the advantages of overcoming these issues to enhance the decision-making process). Guidance and practical recommendations on the implementation of MCDA in the medical context [82, 93] are valuable tools to help building such models, and some support could also be found from the general literature on MCDA [5, 47].

Sensitivity analyses should be conducted as part of the decision-making process:

- The influence of subjectivity on the conclusions from MCDA should be investigated. First, the choice of the criteria used to assess the benefits and risks can strongly affect the results, and a considerable effort has been made in the past years to propose framework approaches that help in identifying the key benefits and the key risks [92, 69, 99]. The second source of subjectivity is the definition of the partial value functions to map the criterion measurements into a 0-1 scale, which should reflect the importance of a change on each criterion. Partial value functions could be very simple in some cases, as in our example where they are assumed to be linear, but nonlinear functions are more sensible when only some values, or ranges of values, actually represent an increased benefit or risk. Third, MCDA requires the exact elicitation of weights to quantify the relative importance of the criteria according to the preferences of the decision makers. Extended models have been proposed where the weights are considered to be random variables [112, 131], and sensitivity analyses could be conducted by varying the variance of weights. Finally, the independence of the criteria for benefits and risks is usually assumed for the sake of simplicity, but the impact of possible correlations should be assessed [138].

- The sensitivity of the results to the choice of the priors used in the Bayesian analysis should be evaluated [109]. In particular, our example presents a situation where some of the strategies considered for future development differ from the past ones, and are not yet experimented. The success of these strategies is predicted using together previous evidence on other regimens and clinical assumptions, which are translated into priors on some parameters. The impact of these assumptions on the reliability of the conclusions was evaluated.

One may prefer to use a Frequentist framework instead of a Bayesian one where only vague priors are used, and to present the same success component criteria on different scales such as standardized differences or conditional powers [51]. These are common approaches when the success definition is based solely on the primary efficacy endpoint, but some difficulty arises when trying to derive a single Frequentist test statistic on multiple outcomes of benefit and risk, which often have different distributions.

Since the methods described here are evidence-based, they require that some clinical data on the efficacy and the safety of the experimental treatment are available. Therefore, these methods may not be appropriate in very early development, when the knowledge about the drug comes mainly from the pre-clinical development or pharmacokinetics trials. In this case, extrapolation models or beliefs from experts or literature could be used and incorporated in the model using priors to substitute or complement the clinical data. Priors could also be elicited by borrowing information from very similar compounds, if any. The advantage of the Bayesian framework of our approach is that the predictions of success can be updated with the accumulation of knowledge from trial to trial.

Moreover, predicting the efficacy and the safety in future trials from the posterior distribution of parameters assessed in previous trials supposes that the future and the previous trials use the same endpoints in the same clinical context (patient population, assessment timepoint(s)...). While this assumption is realistic for some diseases, like Major Depressive Disorders in our example, for other diseases early clinical trials may use a surrogate or a predictive endpoint as primary endpoint. In such situation, the predictive distribution of the clinical endpoint in a future trial may be estimated from the posterior distribution of a surrogate endpoint of a previous trial, taking into account the dependence between the two endpoints [140]. If some limited data have also been collected on the clinical endpoint in early trials, these data may be combined with those of the surrogate data to be integrated in the decision-making process [60, 19].

In the composite definition of success, we have seen that the two components of statistical significance and clinical relevance may be seen as redundant. However, both aspects are important to achieve success, and knowing which one is the most restrictive may not be obvious in advance, in particular for non-statisticians. Moreover, the clinical relevant threshold and the sample size of the next studies could be subject to discussions, and keeping both rules permits to perform several analyses using different thresholds or different sample sizes without changing the definition of composite success. In any case, presenting the marginal predictive probabilities of all the success components can help the decision-makers in choosing between different strategies, when some uncertainty remains on some but not all of the components.

We defined the success components using observable statistics (observed treatment differences in efficacy and in benefit-risk balance) in each pivotal study. One could consider defining criteria at the development level rather than at the trial level, using for example meta-analyses and/or hierarchical models, in a full Bayesian approach, after completion of all the trials. However, we believe that our method addresses a general demand for replication of the study results when medicinal products are evaluated for marketing authorization [46]. Once the development is completed, a synthesis of the results at the development level is usually worthwhile to complement the individual study results. In particular, the overall safety profile is estimated considering data from multiple sources (pivotal and non-pivotal clinical trials, pharmacovigilance...) to incorporate for example long-term, less common and rare outcomes.

Finally, without a large experience using this composite definition of success, no

clear threshold could be provided yet to indicate whether its predictive probability supports a go or a no-go decision. The results depend on the precision of the available evidence and on how promising (or non-promising) the strategy is: the predictive probabilities are expected to be close to 50% when the amount of evidence is very low, and decision-making is challenging in this case ; they are expected to increase for promising strategies (or, respectively, to decrease for non-promising strategies) with the time of development and the accumulation of knowledge ; and they are expected to remain close to 50% for average strategies, whatever the amount of available evidence. Depending on the therapeutic area and the phase of development, some thresholds could be defined using pre-specified targeted levels of evidence following for example the concepts developed by Neuenschwander *et al.* [97] or Frewer *et al.* [49]. In any case, one should be careful in making decisions based on a direct, intuitive, interpretation of the PPoS as a chance of success for the development [51]. The probability of composite success presented here rather corresponds to a 'probability of technical success [...] defined as the probability of a compound generating favorable data to support a filing to regulators' [22], and supports decision in favor of one development strategy when the whole set of results (on the three components and the composite, for the main analysis and the sensitivity analyses) supports the belief of a positive outcome.

In conclusion, the predictive probabilities of composite success and of its components are helpful tools to compare development strategies and to inform decision-making in the pharmaceutical development. Since it is an evidence-based approach to make predictions, the similarity between the previous and the future studies (e.g. in terms of endpoints, patient population, doses) is an important condition that may be bypassed by appropriate assumptions. Although the composite definition of success provides a useful summary of the potential of a strategy, it is recommended to present it along with its different components, to appropriately support the discussions of the decision-makers. In particular in therapeutic areas with unmet medical needs, the project team may be willing to take a certain amount of risk to continue the development, even when some uncertainty remains regarding the chances to reach some of the success criteria.

Chapter 5

Predictive probability of success using surrogate endpoints

Background

This chapter was submitted and is currently under review by *Statistics in Medicine* (major revisions):

Saint-Hilary Gaelle, Barboux V, Pannaux M, Gasparini M, Robert V, Mastrantonio G. Predictive probability of success using surrogate endpoints.

5.1 Introduction

The drug development is a long and continuous process involving complex and critical decisions in order to make an optimal choice between several alternatives, such as go/no-go decisions at the end of phase I and phase II clinical trials, based on the current knowledge on the experimental drug and its potential on the market compared to other compounds for the same disease. Decision-makers need quantitative methods to support informed decisions, with transparent processes that synthesize the whole available information in order to evaluate the probabilities of success associated to different options.

The Predictive Probability of Success (PPoS) of a future clinical trial is a key decision-making tool. It is commonly defined as the predictive probability of statistical significance on the primary endpoint in the next trial. It has first been introduced by Spiegelhalter *et al.* [126] in 1986 as the *Bayesian predictive power*, and it was detailed later in various contexts by O'Hagan *et al.*, 2005 [101] (where it is called *assurance*) and Gasparini *et al.*, 2013 [52]. It is usually obtained using both frequentist and Bayesian methods, since the distribution of the treatment effect parameter is computed from prior knowledge and available evidence (a Bayesian concept), and then the power (a frequentist concept) of the next study is averaged over this distribution. For this reason, PPoS has also been called *expected power* or *average power* [20, 22].

In the current practice, the available evidence typically comes from the accumulated data on the clinical endpoint of interest in previous clinical trials [145, 139, 129, 109]. However, a surrogate endpoint could be used in early development, while no or limited data are collected on the relevant clinical endpoint of interest, which we call the final endpoint. The surrogate endpoint can usually be assessed more frequently and/or earlier than the final endpoint, and is intended to detect the activity of a drug in early phases before pursuing (or not) the development in confirmatory

phases with longer, bigger and more expensive trials. For example in oncology, the response to treatment or the progression-free survival (PFS, defined as the time from randomization to disease progression or death due to any cause) are frequently used as primary endpoint in phase I/II studies, while the overall survival (OS) is the true endpoint of interest that will be used in phase III. In this paper, we call surrogate endpoint any marker used in early phase as a measure of the treatment effect, but which is not necessarily accepted for confirmatory phase from a regulatory perspective.

It is of interest to use the available evidence from data on the surrogate endpoint to compute the PPoS of future trials designed with the final endpoint as primary endpoint. This could be achieved by modeling the relationship between the two endpoints. Hong *et al.*, 2012 [60] present a methodology to calculate the PPoS of a phase III trial with OS as primary endpoint from phase II data on PFS, using a bivariate normal model in phase II to estimate the joint distribution of the log hazard ratios on the two endpoints. While this proposal is a major advance in the generalization of the use of the PPoS, it may be limiting since a relationship between endpoints estimated from a single trial is insufficient to support predictions across trials [48, 4, 13, 14, 1]. Indeed, it focuses on the patient level association, while we are interested in the relationship between the treatment effects on the endpoints at the trial level. To overcome this issue, meta-analytic approaches have been proposed [26, 15, 50] where the trial level association is assessed from several groups of patients (multi-center trials or meta-analyses). A nice example of a PPoS calculation using this approach is shortly presented in Wang *et al.*, 2013 [140], with PFS and OS, in metastatic breast cancer. Since the OS data from the phase II trial were sparse, the authors used only the PFS data to make the predictions. Again in oncology, Chen *et al.*, 2011 [19] proposed a joint test-statistic for PFS and OS in order to support accelerated approval by the regulatory agencies, focusing on hypothesis testing rather than on predictions.

While these examples are instructive, it is worth mentioning that, to our knowledge, the PPoS calculation based on surrogate endpoints has never been widely explored nor formalized. For this reason, the aim of this paper is to propose a general and reliable approach to compute the PPoS of one future trial on the final endpoint, when data on the surrogate endpoint, and possibly a limited amount of data on the final endpoint, are available from previous trials (Figure 5.1). Its principle is to estimate the joint distribution of the treatment effects on the endpoints using a meta-analytic approach on external data, and to use this joint distribution with the past observations on the surrogate endpoint to build an informative prior distribution for the final endpoint parameter. If data are also available on the final endpoint, they can be combined with the prior using a classical Bayesian approach, and in this case we propose two methods to address a potential discordance between the prior and the data on the final endpoint parameter. Finally, the predictive distribution of the estimate of the final endpoint parameter is used to calculate the PPoS of the next trial. We extend this approach to cases where information can be borrowed from multiple surrogate endpoints.

The rest of the paper is organized as follows. The methods are detailed in Section 5.2, where first some reminders are made on PPoS calculation based on one single endpoint. Then the calculation of the PPoS is presented based on either one surrogate endpoint only, one surrogate and the final endpoint, or multiple surrogate

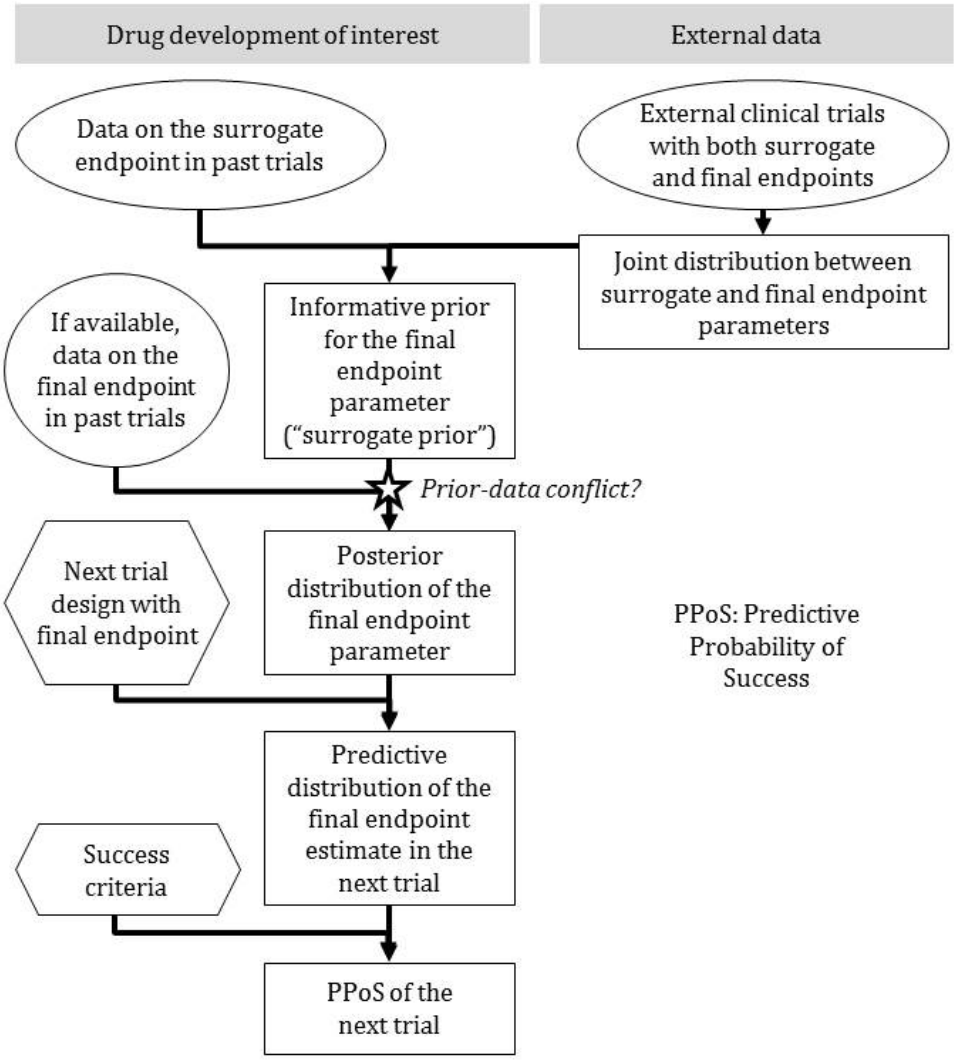

FIGURE 5.1: Proposed approach to compute the PPoS of one future trial with the final endpoint as primary endpoint, using data on the surrogate endpoint.

endpoints and the final endpoint. In Section 5.3, we provide the results of a comprehensive simulation study investigating the patterns of behavior of the PPoS according to the precision of the estimated association between the endpoints, the validity of the surrogate endpoint, the amount of information available in past trials on each endpoint and the extent of a potential prior-data conflict. Section 5.4 describes the application of this approach to a drug development in Multiple Sclerosis, where one or two endpoints are used as surrogate. A discussion and concluding remarks are given in Section 5.5.

5.2 Methods

5.2.1 PPoS based on the final endpoint only (reminders)

Consider a future clinical trial, with the primary objective to compare two treatments on the final endpoint, and let θ denote the corresponding true treatment difference.

Suppose that the trial success is defined as the rejection of the null hypothesis $H_0 : \theta \leq 0$ against the alternative, $H_1 : \theta > 0$, following a conventional frequentist approach. In most cases, there will exist in the next trial an estimate $\hat{\theta}_f$ (indexed by f for *future*) of the treatment difference reasonably assumed to be normally distributed with

$$\hat{\theta}_f \mid \theta \sim N(\theta, \sigma_f^2),$$

with density $f_{\hat{\theta}_f|\theta}(\cdot)$, and where σ_f^2 is its sampling variance, supposed to be known. For example, $\hat{\theta}_f$ could be the difference between means, the log odds ratio, the log hazard ratio or the log rate ratio if the final endpoint is a continuous variable, a binary or ordinal variable, a time-to-event variable or a count variable respectively. Therefore, the null hypothesis H_0 will be rejected if $\hat{\theta}_f > z_\alpha \sigma_f^2$, where z_α is the $(1-\alpha)100^{th}$ percentile of the standard normal distribution.

Suppose moreover that some information on the final endpoint has been collected already in the past, and it has been summarized with $\hat{\theta}$, the estimate of the treatment difference observed in one single past trial or resulting from evidence synthesis [3, 141, 142], and σ^2, its sampling variance which is assumed to be known. The sampling distribution (likelihood) could be written

$$\hat{\theta} \mid \theta \sim N(\theta, \sigma^2), \tag{5.1}$$

and we note its density $f_{\hat{\theta}|\theta}(\cdot)$. Using a classical Bayesian approach, assuming there is no prior knowledge on the treatment effect, we apply a normal vague prior on the parameter

$$\theta \sim N(\theta_0, \sigma_0^2) \tag{5.2}$$

which density is denoted by $\pi_\theta^V(\cdot)$ (with the superscript V for *vague*), and we obtain the posterior distribution

$$\theta \sim N(\theta_p, \sigma_p^2) \tag{5.3}$$

where $\theta_p = \hat{\theta}\sigma_0^2/(\sigma_0^2+\sigma^2) + \theta_0\sigma^2/(\sigma_0^2+\sigma^2)$ and $\sigma_p^2 = \sigma_0^2\sigma^2/(\sigma_0^2+\sigma^2)$ (indexed by p for *posterior*). We note its density $g_\theta^V(\cdot)$.

From this posterior distribution and the distribution of $\hat{\theta}_f \mid \theta$, under the usual assumption of conditional independence given θ, we obtain the predictive distribution of $\hat{\theta}_f$:

$$\hat{\theta}_f \sim N\left(\theta_p, \sigma_p^2 + \sigma_f^2\right)$$

and we note its density $h_{\hat{\theta}_f}^V(\cdot)$. The probability of rejecting H_0 based on the available evidence can therefore be calculated as

$$PPoS^V = P\left(\hat{\theta}_f > z_\alpha \sigma_f^2\right) = \int_{u > z_\alpha\sigma_f^2} h_{\hat{\theta}_f}^V(u)du = 1 - \Phi\left(\frac{z_\alpha\sigma_f^2 - \theta_p}{\sqrt{\sigma_p^2+\sigma_f^2}}\right),$$

where Φ denotes the cumulative distribution function of the standard normal distribution.

5.2.2 PPoS based on the surrogate endpoint only

Suppose now that some information has been collected in the past on one surrogate endpoint but *not* on the final endpoint. The information on the true treatment difference γ on the surrogate endpoint can be summarized with its estimate $\hat{\gamma}$, obtained from the past trial(s), and δ^2, its sampling variance which is assumed to be known. As before, using a Bayesian approach with a normal prior $\gamma \sim N(\gamma_0, \delta_0^2)$ and the sampling distribution $\hat{\gamma} \mid \gamma \sim N(\gamma, \delta^2)$, we obtain the posterior distribution

$$\gamma \sim N(\gamma_p, \delta_p^2) \tag{5.4}$$

where $\gamma_p = \hat{\gamma}\delta_0^2/(\delta_0^2+\delta^2) + \gamma_0\delta^2/(\delta_0^2+\delta^2)$ and $\delta_p^2 = \delta_0^2\delta^2/(\delta_0^2+\delta^2)$. We wish to use this information to calculate the PPoS of the future trial designed with the final endpoint as primary endpoint.

The first step is to quantify the relationship between the treatment effects on the two endpoints, using external data. We propose to apply the Bayesian meta-analytic approach for the evaluation of surrogate endpoints developed by Daniels and Hughes, 1997 [26], which uses summary data from multiple clinical trials (or centers). Alternatively, the two-stage model introduced by Buyse *et al.*, 2000 [15] can be used if individual patient data are available.

Consider $i = 1, ..., N$ external randomized clinical trials where the surrogate and the final endpoints were evaluated. The relationship between the endpoints in the selected trials should be thought to be comparable with what is expected in our drug development: they should be conducted in an analogous clinical context (e.g. the same patient population), and/or involve a similar class of treatments. For simplicity, we assume here that all trials compared two treatment arms, but the extension to multi-arm trials is presented in Appendix C, Section C.1. Following [26], let us denote θ_i and γ_i the true unknown treatment differences on the final endpoint and the surrogate endpoint, respectively, in the trial i. For each trial, we suppose that their estimates $\hat{\theta}_i$ and $\hat{\gamma}_i$ are available with their sampling variance σ_i^2 and δ_i^2 and their correlation ρ_i. We assume the following within-trial model:

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\gamma}_i \end{pmatrix} \Bigg| \begin{pmatrix} \theta_i \\ \gamma_i \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_i \\ \gamma_i \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_i\sigma_i\delta_i \\ \rho_i\sigma_i\delta_i & \delta_i^2 \end{pmatrix} \right). \tag{5.5}$$

Usually, the estimated variances of $\hat{\theta}_i$ and $\hat{\gamma}_i$ reported in the clinical trial publications are treated as if they were the true variances σ_i^2 and δ_i^2 [142]. On the other hand, the correlations ρ_i are rarely reported in the literature. They could be estimated from the individual patient data if they are available, otherwise it has been suggested to use a range of plausible values [26, 108, 105], with sensitivity analyses across the entire correlation range.

Considering the true treatment effect on the surrogate endpoint γ_i as a fixed effect, we assume that the conditional distribution of $\theta_i \mid \gamma_i$ is normal, in which θ_i linearly depends on γ_i:

$$\theta_i \mid \gamma_i, a, b, \tau \sim N(a + b\gamma_i, \tau^2). \tag{5.6}$$

The generalization of this model where γ_i is assumed to be a normally distributed random effect could be found in McIntosh, 1996 [86]. It follows from (5.5) and (5.6)

that

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\gamma}_i \end{pmatrix} \Bigg| \gamma_i, a, b, \tau \sim N\left(\begin{pmatrix} a + b\gamma_i \\ \gamma_i \end{pmatrix}, \begin{pmatrix} \sigma_i^2 + \tau^2 & \rho_i\sigma_i\delta_i \\ \rho_i\sigma_i\delta_i & \delta_i^2 \end{pmatrix}\right). \tag{5.7}$$

In this model, b measures the association between the treatment effects on the two endpoints, and is expected to be different from zero if the marker is actually a surrogate endpoint. On the other hand, the parameter a represents the part of the treatment effect on the final endpoint unexplained by its effect on the surrogate endpoint, and is expected to be close to zero. In a Bayesian framework, choosing a normal-inverse-gamma distribution as conjugate prior for the parameters (a, b, τ), their joint posterior distribution $f_{a,b,\tau}(\cdot)$ given the data from the meta-analysis can be derived analytically as a normal-inverse-gamma distribution too (Appendix C, Section C.2).

As a second step, we use these results for our drug development of interest. The distribution of the treatment effect on the final endpoint, conditional on the regression parameters, is obtained from Equations (5.4) and (5.6) as

$$\theta \mid a, b, \tau \sim N(a + b\gamma_p, \tau^2 + b^2\delta_p^2),$$

and we note its density $f_{\theta|a,b,\tau}(.)$. Integrating over the joint distribution of the regression parameters, we obtain the following density of θ:

$$\pi_\theta^S(\cdot) = \int f_{\theta|a,b,\tau}(.) f_{a,b,\tau}(x,y,z) d(x,y,z). \tag{5.8}$$

We call this distribution the surrogate prior (and use the superscript S for *surrogate*), as it could be used as a prior distribution for the parameter of interest θ. In this section, since no data are available on the final endpoint, the posterior distribution is the same as the prior distribution. The predictive distribution of the estimate $\hat{\theta}_f$ in the next trial has the density:

$$\begin{aligned} h_{\hat{\theta}_f}^S(\cdot) &= \int f_{\hat{\theta}_f|\theta=t}(\cdot)\pi_\theta^S(t)dt \\ &= \int\int f_{\hat{\theta}_f|\theta=t}(.) f_{\theta|a,b,\tau}(t) f_{a,b,\tau}(x,y,z) dt d(x,y,z). \end{aligned}$$

Marginalizing over θ, it could be simplified in:

$$h_{\hat{\theta}_f}^S(\cdot) = \int f_{\hat{\theta}_f|a,b,\tau}(.) f_{a,b,\tau}(x,y,z) d(x,y,z),$$

where $f_{\hat{\theta}_f|a,b,\tau}(.)$ is the density of the distribution $N\left(a + b\gamma_p, \tau^2 + b^2\delta_p^2 + \sigma_f^2\right)$, since $f_{\hat{\theta}_f|\theta=t}(\cdot)$ and $f_{\theta|a,b,\tau}(\cdot)$ are both normal densities. The predictive probability of success of the next trial can then be calculated as

$$\begin{aligned} PPoS^S = P\left(\hat{\theta}_f > z_\alpha\sigma_f^2\right) &= \int_{u > z_\alpha\sigma_f^2} h_{\hat{\theta}_f}^S(u) du \\ &= \int \left[1 - \Phi\left(\frac{z_\alpha\sigma_f^2 - (a + b\gamma_p)}{\sqrt{\tau^2 + b^2\delta_p^2 + \sigma_f^2}}\right)\right] f_{a,b,\tau}(x,y,z) d(x,y,z). \end{aligned} \tag{5.9}$$

It is worth noting that the prediction variance of $\hat{\theta}_f$ depends on:

a) The variance of the joint posterior distribution $f_{a,b,\tau}(\cdot)$ obtained from the linear

regression on the meta-analysis data (small if the amount of data in the meta-analysis is large);

b) The dependence between the treatment effects on the surrogate and the final endpoints (τ^2 is small if the surrogate is a 'good surrogate');

c) The precision of the available evidence on the surrogate endpoint (δ_p^2 is small if the amount of data on the surrogate is large in the past trials);

d) The precision of the information that will be collected in the future trial on the final endpoint (σ_f^2 is small if the planned number of patients or number of events is large in the future trial).

Note that (d) is a component of the prediction variance that was not used in similar findings already reported in the literature (in a frequentist approach) [13, 14, 12]. The reason why we add it is because we perform predictions across trials, while others focused on within-trial predictions.

5.2.3 PPoS based on the surrogate and the final endpoints

Suppose now that some information has been collected in the past on both the surrogate and final endpoints. If the amount of data collected on the final endpoint is large enough, it could be used alone to predict the success of the next trial, as presented in Section 5.2.1. However, past trials are often designed and powered with the surrogate endpoint as primary endpoint, while some sparse or incomplete data on the final endpoint are collected for exploratory purpose.

In order to make the most of the available data, we propose to use both sources of information to calculate the PPoS of the next trial (assuming the quality of the data on both endpoints is satisfactory). The general approach is presented first. It should be noted, however, that the evidences on the surrogate endpoint and on the final endpoint may be conflicting if the surrogacy relationship established from external data does not hold in our drug development. While this is an unexpected result, it should be anticipated in order to avoid wrong decisions. For example, we could be interested to continue the development if an overwhelming efficacy is observed on the final endpoint, even if no effect is observed on the surrogate. We propose two approaches to handle this issue, using the methods for addressing potential prior-data conflict presented in Mutsvari *et al.*, 2016 [95]. The principle of the first approach, referred to as the 'testing approach', is to assess the level of conflict between the two sources of evidence, and to discard the information coming from the surrogate endpoint if this level of conflict exceeds a certain predefined threshold. On the other hand, the second approach, referred to as the 'mixture prior approach', always consider both sources of evidence in the predictions but permits to down-weight the information coming from the surrogate endpoint in case of conflict. This latter approach is also detailed in a different context and called the 'robust prior' in Schmidli *et al.*, 2014 [115].

General model

In a Bayesian framework, the information coming from the surrogate endpoint could be combined with the data collected on the final endpoint, using (5.8) as surrogate

prior on θ, and the sampling distribution (5.1). The posterior distribution is then

$$g_\theta^S(\cdot) = \frac{f_{\hat{\theta}|\theta}(d)\pi_\theta^S(\cdot)}{\int f_{\hat{\theta}|\theta=t}(d)\pi_\theta^S(t)dt} \tag{5.10}$$

where d is the observed outcome of $\hat{\theta}$.

From this posterior distribution and the distribution of $\hat{\theta}_f \mid \theta$, we obtain the predictive distribution of $\hat{\theta}_f$

$$h_{\hat{\theta}_f}^S(\cdot) = \int f_{\hat{\theta}_f|\theta=t}(\cdot)g_\theta^S(t)dt,$$

and the predictive probability of success of the next trial is obtained as in Equation (5.9) by integrating $h_{\hat{\theta}_f}^S(\cdot)$ over the range of values greater than $z_\alpha\sigma_f^2$.

Prior-data conflict: testing approach

The principle of this approach is to find how large is the probability to get a more extreme result than the estimate $\hat{\theta} = d$ (obtained from the past trials), considering its prior predictive distribution, i.e. its predictive distribution based solely on the surrogate prior:

$$\begin{aligned} h_{\hat{\theta}}^S(\cdot) &= \int f_{\hat{\theta}|\theta=t}(\cdot)\pi_\theta^S(t)dt \\ &= \int\int f_{\hat{\theta}|\theta=t}(\cdot)f_{\theta|a,b,\tau}(t)f_{a,b,\tau}(x,y,z)d(x,y,z)dt \\ &= \int f_{\hat{\theta}|a,b,\tau}(\cdot)f_{a,b,\tau}(x,y,z)d(x,y,z) \end{aligned} \tag{5.11}$$

where $f_{\hat{\theta}|a,b,\tau}(\cdot)$ is the density of $N(a+b\gamma_p, \tau^2+b^2\delta_p^2+\sigma^2)$.

A prior-data conflict is declared if $\hat{\theta} = d$ lies outside the $\xi 100^{th}$ and $(1-\xi)100^{th}$ percentiles of this prior predictive distribution, i.e. if

$$\min\left[\int_{t<d} h_{\hat{\theta}}^S(t)dt, \int_{t>d} h_{\hat{\theta}}^S(t)dt\right] < \xi,$$

where ξ is a predefined testing level, usually 5% or 10% [95]. It could be chosen by computing the operational characteristics of simulated data scenarios, to ensure that clinically pertinent prior-data conflicts are detected with this approach.

If no prior-data conflict is declared, both sources of evidence are used to calculate the PPoS, and the general model applies. Otherwise, the established relationship between the endpoints is likely to be non-relevant for our drug development, and the PPoS is computed from the final endpoint only, as described in Section 5.2.1. Indeed, in this latter case, the data collected on the final endpoint are considered to be the only reliable information to predict the outcome of the future trial.

Prior-data conflict: mixture prior approach

The testing approach corresponds to an 'all-or-nothing' process, where the evidence on the surrogate endpoint is either included as a whole if it matches with the evidence on the final endpoint, or entirely discarded from the analysis in case of conflict.

On the other hand, the mixture approach permits to down-weight the information coming from the surrogate endpoint but not to ignore it completely.

We consider a mixture prior that is conjugate to the distribution of the data, defined as the sum of a vague proper normal prior (Equation (5.2)) and the surrogate prior (Equation (5.8)). The mixture prior distribution (with a superscript M for *mixture*) is written

$$\pi_\theta^M(\cdot) \;=\; w\,\pi_\theta^V(\cdot) \;+\; (1-w)\,\pi_\theta^S(\cdot)$$

where w is the prior probability that the assumption of surrogacy does not hold in our drug development. Since we place us in a context where the project teams expect to detect a treatment effect on the surrogate endpoint in early phases before pursuing the development with the final endpoint, this prior probability is usually low, say 0.05 or 0.1. Higher w will lead to higher discounting of the surrogate information in the posterior distribution.

It follows that the posterior distribution is also a mixture of distributions

$$g_\theta^M(\cdot) \;=\; \tilde{w}\,g_\theta^V(\cdot) \;+\; (1-\tilde{w})\,g_\theta^S(\cdot)$$

where $g_\theta^V(\cdot)$ and $g_\theta^S(\cdot)$ are the posterior densities obtained in Equations (5.3) and (5.10). The updated posterior probability $\tilde{w}$ is calculated as

$$\tilde{w} \;=\; \frac{w\,h_{\hat{\theta}}^V(d)}{w\,h_{\hat{\theta}}^V(d) \;+\; (1-w)\,h_{\hat{\theta}}^S(d)}$$

where $h_{\hat{\theta}}^V(\cdot)$ is the density of the prior-predictive distribution of $\hat{\theta}$ based on the vague prior, $N(\theta_0, \sigma_0^2+\sigma^2)$, and $h_{\hat{\theta}}^S(\cdot)$ is defined in Equation (5.11).

From the mixture posterior distribution and the distribution of $\hat{\theta}_f \mid \theta$, we obtain the predictive distribution of $\hat{\theta}_f$:

$$h_{\hat{\theta}_f}^M(\cdot) \;=\; \tilde{w}\,h_{\hat{\theta}_f}^V(\cdot) \;+\; (1-\tilde{w})\,h_{\hat{\theta}_f}^S(\cdot),$$

and the predictive probability of success of the next trial:

$$\begin{aligned} PPoS^M \;=\; P\left(\hat{\theta}_f > z_\alpha\sigma_f^2\right) \;&=\; \int_{u>z_\alpha\sigma_f^2} h_{\hat{\theta}_f}^M(u)du \\ &=\; \tilde{w}\,PPoS^V \;+\; (1-\tilde{w})\,PPoS^S. \end{aligned}$$

While the surrogate prior is data-driven, the vague prior needs to be pre-specified to reflect some lack of knowledge about θ. For the sake of avoiding subjectivity in the analysis, one may be tempted to choose an extremely large variance σ_0^2. However, it should be noted that $\tilde{w} \to 0$ as $\sigma_0^2 \to +\infty$, meaning that choosing a very weakly informative distribution will result in a very low posterior probability for the vague prior (and respectively, a very high posterior probability for the surrogate prior), with no advantage of using this approach compared to the simple general model. The choice of σ_0^2 is therefore critical in the analysis, and could be based on operational characteristics of different simulated scenarios. Nevertheless, Mutsvari *et al.* [95] point out that 'the rationale for using a mixture prior will be strongest when there is some justification for each of the two components' σ_0^2 and w.

5.2.4 PPoS based on multiple surrogate endpoints and the final endpoint

Again with the purpose of making the most of the data collected in the past trials, it may be of interest to combine the information coming from multiple surrogate endpoints assessed in early phase. As stated before, we call surrogate any marker assumed to be predictive of the treatment effect, but which is not necessarily accepted for regulatory purpose.

The relationship between the treatment effects on the final endpoint and M surrogate endpoints could be established in a similar way to the one used in Section 5.2.2, from a meta-analysis of N external clinical trials. Let $\hat{\gamma}_{1i}, ..., \hat{\gamma}_{Mi}$ denote the estimated treatment differences on the M surrogate endpoints in the trial i, and $\delta_{1i}^2, ..., \delta_{Mi}^2$ their sampling variance. The within-trial model in Equation (5.5) could be extended to [105, 10]:

$$\begin{pmatrix} \hat{\theta}_i \\ \hat{\gamma}_{1i} \\ \vdots \\ \hat{\gamma}_{Mi} \end{pmatrix} \Bigg| \begin{pmatrix} \theta_i \\ \gamma_{1i} \\ \vdots \\ \gamma_{Mi} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_i \\ \gamma_{1i} \\ \vdots \\ \gamma_{Mi} \end{pmatrix}, \begin{pmatrix} \sigma_i^2 & \rho_{01i}\sigma_i\delta_{1i} & \cdots & \cdots & \rho_{0Mi}\sigma_i\delta_{Mi} \\ \rho_{01i}\sigma_i\delta_{1i} & \delta_{1i}^2 & \rho_{12i}\delta_{1i}\delta_{2i} & \cdots & \rho_{1Mi}\delta_{1i}\delta_{Mi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{0Mi}\sigma_i\delta_{Mi} & \cdots & \cdots & \cdots & \delta_{Mi}^2 \end{pmatrix} \right).$$

where, in the trial i, ρ_{0mi} is the correlation between the surrogate endpoint m and the final endpoint, and ρ_{mqi} is the correlation between the two surrogate endpoints m and q. As indicated before, it is very unlikely that these values are reported in the publications, but plausible values and range of values could be assumed [26, 108, 105]. Using vectorized notations $G_i = (\gamma_{1i}, ..., \gamma_{Mi})$ and $B = (b_1, ..., b_M)$, and assuming fixed effects for G_i, the linear model to describe the relationship between the endpoints becomes

$$\theta_i \mid G_i, a, B, \tau \sim N(a + B^T G_i, \tau^2).$$

The generalization where G_i are random effects is described in Bujkiewicz *et al.*, 2016 [10]. Some additional assumptions could also be made on the structural relationships between the endpoints, for example that the treatment effect on the final endpoint is independent from the effect on some surrogate endpoints conditionally on some others, see Pozzi *et al.*, 2016 [105] and Bujkiewicz *et al.*, 2016 [10] for more details.

Let $G = (\gamma_1, ..., \gamma_M)$ denote the vector of true treatment differences on the M surrogate endpoints in our drug development, for which the multivariate normal posterior joint distribution obtained from a prior and the data collected in previous trials is

$$G \sim N(G_p, D_p),$$

where G_p is the vector of posterior means and D_p the posterior variance-covariance matrix. The distribution of the treatment effect conditional on the regression parameters becomes

$$\theta \mid a, B, D \sim N(a + B^T G_p, \tau^2 + B^T D_p B).$$

It could be used as a surrogate prior distribution for θ, and the computation of the PPoS follows the same principle as the methods described in the previous sections, depending on the availability of data on the final endpoint and on the chosen approach for handling prior-data conflicts.

5.3 Simulation study

A simulation study was carried out to explore the patterns of behavior of the PPoS, using one surrogate endpoint, according to four components: the precision of the association between the endpoints estimated from the meta-analysis; the validity of the surrogate endpoint; the amount of information available in past trials on each endpoint; the extent of a potential prior-data conflict.

5.3.1 Setting

Data were simulated using a fictive base-case scenario inspired by real clinical trials, and the parameters were varied one at a time to assess their impact on the PPoS. In this base-case scenario, we suppose that the data of one past trial comparing two treatments on the surrogate endpoint are available on $N^\gamma = 86$ patients per arm, with an estimate of the treatment difference $\hat{\gamma} = 4$ and a standard error $\delta = 7\sqrt{2/N^\gamma}$. We also presume that the final endpoint was evaluated in the trial, but is available for only $N^\theta = 30$ patients per arm at the time of the analysis, with an estimated treatment effect $\hat{\theta} = 4$ and a standard error $\sigma = 13\sqrt{2/N^\theta}$. We wish to calculate the PPoS of one future confirmatory trial which primary endpoint is the final endpoint. A sample size of 444 patients ($N_f^\theta = 222$ per arm) is planned in this trial to reach a power of 90%, based on an assumed true difference $\theta = 4$ points, a standard deviation of 13 and a one-sided α of 2.5%.

We suppose that the relationship between the surrogate endpoint and the final endpoint was estimated based on a meta-analysis of external clinical trials using the joint model described in Equation (5.7). The regression coefficients a and b were simulated using a bivariate normal model with means $E(a) = 0$ and $E(b) = 1$, variances $Var(a) = Var(b) = 0.5^2$, and a correlation coefficient 0.5. The standard error τ was simulated using an inverse-gamma distribution with shape parameter $[E(\tau)^2 + 2Var(\tau)]/Var(\tau)$ and scale parameter $E(\tau)[E(\tau)^2 + Var(\tau)]/Var(\tau)$, with $E(\tau) = 1$ and $Var(\tau) = 0.1^2$.

The questions of interest and the corresponding simulation scenarios are presented in Table 5.1. In each scenario, the PPoS is calculated as described in the above sections using either:

- The data on the final endpoint and a weakly informative prior $\theta \sim N(0, 10^6)$;
- The surrogate prior and no data on the final endpoint;
- The surrogate prior and data on the final endpoint, without handling prior-data conflict;
- The surrogate prior and data on the final endpoint, with the testing approach and a testing level $\xi = 0.05$;
- The surrogate prior and data on the final endpoint, with the mixture prior approach, a prior probability $w = 0.1$ and a vague prior component $N(0, 100\sigma^2)$, corresponding to an effective sample size 100 times lower than the sample size of the available data on the final endpoint.

TABLE 5.1: Simulation scenarios

Question of interest	Simulation scenario
How does PPoS vary according to the precision of the association between endpoints estimated from the meta-analysis (Equation (5.12))?	(a) $Var(a)$ varying from 0.1^2 to 8^2 (b) $Var(b)$ varying from 0.1^2 to 8^2 (c) $Corr(a, b)$ varying from -1 to 1
How does PPoS vary according to the dependence between the endpoints?	(d) $E(\tau)$ varying from 0.1 to 8
How does PPoS vary according the amount of available information in the past trial?	(e) N^γ varying from 10 to 150 (f) N^θ varying from 10 to $N^\gamma = 86$
How does PPoS vary according to the extent of a prior-data conflict?	(g) $\hat{\theta} = 0$ and $\hat{\gamma}$ varying from -10 to 20 (h) $\hat{\gamma} = 0$ and $\hat{\theta}$ varying from -10 to 20

The sensitivity of the prior-data conflict approaches to the testing level ξ, the prior probability w and the vague prior component was assessed in [95] and is not repeated here.

All the simulation scenarios above were also performed using a similar base-case scenario but with $E(a) = 1$ and $\hat{\gamma} = 3$, hence assuming that a part of the treatment effect on the final endpoint remains unexplained by the treatment effect on the surrogate endpoint. The resulting patterns of behavior of the PPoS were very similar and are not presented here.

5.3.2 Results

The analyses were conducted using R [106], and 100,000 simulations were run to estimate the parameter distributions.

Figure 5.2 displays the predictive distributions of the estimated treatment difference on the final endpoint in the next study, $\hat{\theta}_f$, in the base-case scenario. Since this scenario was defined under prior-data consistency ($\hat{\theta} = E(a) + E(b)\hat{\gamma}$), the predictive distributions obtained from the surrogate prior and data on the final endpoint are superimposed, whatever the chosen approach for handling or not the prior-data conflicts. For the same reason, all the predictive distributions are centered on the same mean $\hat{\theta}_f = 4$.

The number of patients having the final endpoint assessed in the past trial gives a low power ($N^\theta = 30$, retrospective power of 21.1% to detect a difference $\theta = 4$ with a standard deviation of 13 and a one-sided α of 2.5%), so the predictive distribution obtained using this information only is the least precise, with a PPoS=67% for the next trial. The predictive distribution obtained using the surrogate prior only,

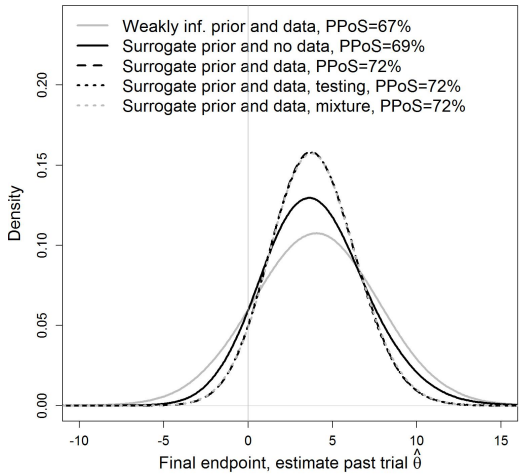

FIGURE 5.2: Base-case scenario. Predictive distributions of $\hat{\theta}_f$ in the next trial.

disregarding the data on the final endpoint, is more precise, with a PPoS=69%, because the number of patients having an evaluation of the surrogate endpoint gives a substantial power ($N^{\gamma} = 86$, retrospective power of 96.3% to detect a difference $\gamma = 4$ with a standard deviation of 7 and a one-sided α of 2.5%) and the relationship between the endpoints is well established from the meta-analysis. Finally, the posterior distributions obtained by combining the surrogate prior with the final endpoint data are the most precise since they use all the available information, leading to a PPoS=72% for the next trial.

The results of the simulation scenarios are presented in Figure 5.3. The scenarios (a)-(f) are under prior-data consistency, so the methods to handle prior-data conflicts have no effect and the PPoS obtained using the surrogate prior and data on the final endpoint are superimposed.

Given $\hat{\gamma}$, the variance of the association between the endpoints estimated from the meta-analysis could be written

$$Var(a + b\hat{\gamma}) \;=\; Var(a) \;+\; \hat{\gamma}^2 Var(b) \;+\; 2\hat{\gamma} Corr(a,b)\sqrt{Var(b)Var(b)}, \qquad (5.12)$$

therefore it increases with $Var(a)$, $Var(b)$ and $Corr(a,b)$, which corresponds to scenarios (a), (b) and (c). As expected, the PPoS calculated with the weakly informative prior is not impacted by the variation. In the other cases, the precision of the surrogate prior becomes smaller as $Var(a + b\hat{\gamma})$ increases, which is reflected by PPoS becoming closer to 50% (corresponding to the maximum uncertainty). The variation of $Corr(a,b)$ has a minimal impact on the results. When final endpoint data are combined with the surrogate prior, the PPoS tends to the one obtained with a weakly informative prior, since the surrogate prior tends to become weakly informative itself. Similar trends are observed in scenario (d), where $E(\tau)$ varies. The validity of the surrogate endpoint, and with it the amount of information it provides on the final endpoint, decreases as τ increases.

On the other hand, the precision of the available evidence increases with the number

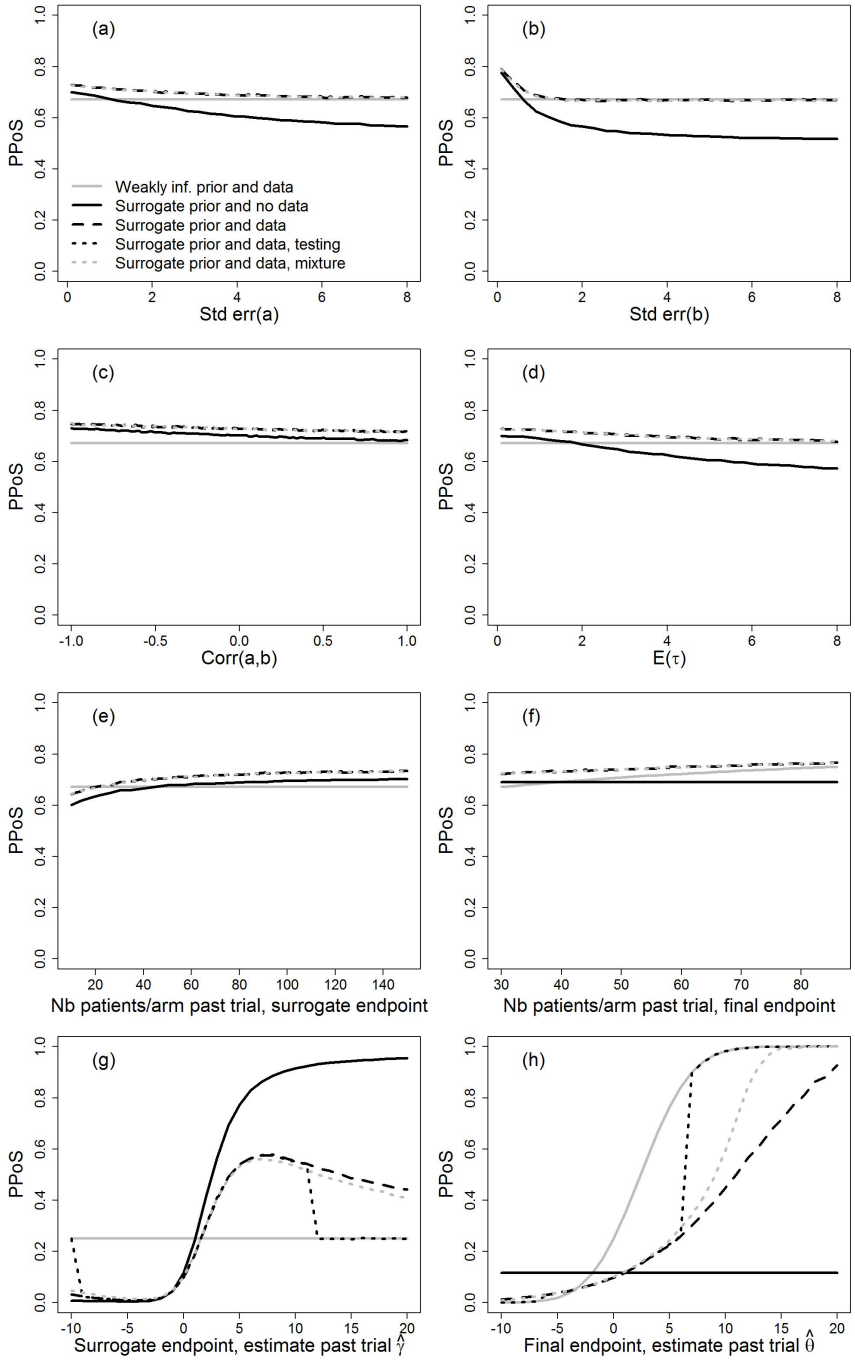

FIGURE 5.3: Results of the simulation scenarios described in Table 5.1.

of patients being evaluated in the past trials, and it slightly impacts the PPoS. When the number of patients with final endpoint assessment remains stable but the number of patients with surrogate endpoint assessment increases (scenario (e)), only the PPoS using the surrogate prior are impacted. In scenario (f), only the PPoS using the final endpoint data are impacted.

Finally, the scenarios with prior-data conflict are presented in (g) and (h), where the surrogate prior and the final endpoint data are consistent when $\hat{\theta} = \hat{\gamma} = 0$, and the prior-data conflict increases as they depart from 0.

In scenario (g), we assume that no treatment effect was observed on the final endpoint in the past trial ($\hat{\theta} = 0$), so the PPoS calculated using these data and a weakly informative prior remains stable and low (25%), indicating a high risk of failure of the next trial. When it is calculated from the surrogate prior only, it increases quickly with the treatment effect observed on the surrogate endpoint, and exceeds 70% for $\hat{\gamma} > 4$. Combining the surrogate prior with the data permits to obtain more realistic PPoS, never exceeding 60%. A prior-data conflict is detected by the testing approach with testing level 0.05 only for extreme values of $\hat{\gamma}$ (approximately < -9 or > 11), because the precision of the data on the final endpoint is quite low. In this case however, the surrogate prior is disregarded and the PPoS is calculated from the data and the weakly informative prior, with a piece-wise curve illustrating the 'all-or-nothing' principle mentioned earlier. Using the mixture prior approach has only a slight impact on the results. However, choosing a smaller prior probability for the surrogate prior ($(1 - w) = 0.9$ in our scenario) or a smaller variance of the vague prior component ($100\sigma^2$ in our scenario) will lead to more flat PPoS curves, closer to 25%.
In scenario (f), the PPoS calculated using the surrogate prior only remains stable and low (12%), because we assume that no treatment effect was observed on the surrogate endpoint in the past trial ($\hat{\gamma} = 0$). The PPoS calculated from the final endpoint data and a vague prior increases with $\hat{\theta}$, exceeding 70% for estimate values greater than 4. The PPoS calculated using the surrogate prior and the data, without handling prior-data conflict, also increases with the treatment effect estimate on the final endpoint but is influenced by the surrogate prior, and reaches 70% only for $\hat{\theta} > 15$. The mixture prior approach lies between the two, increasing until reaching 100% for extreme values of $\hat{\theta}$, but with a lower slope than the PPoS based on the weakly informative prior, indicating that it takes into account that no effect was observed on the surrogate endpoint. In a similar way, the piece-wise curve of the testing approach indicates a prior-data conflict detection for large values of the treatment effect estimate $\hat{\theta} < -6$ and $\hat{\theta} > 6$, approximately.

5.4 Application: Multiple Sclerosis

We illustrate the use of the above methods to support decision-making on a fictive example in Multiple Sclerosis (MS). Consider the development of a new drug with the objective to prevent or delay the accumulation of neurological disability in patients with MS. For regulatory approval, we suppose we need to demonstrate the drug efficacy in reducing the risk of disability progression, defined as an increase of the Expanded Disability Status Scale (EDSS) score sustained for at least 12 weeks, after 2 years of treatment. This requires a clinical trial with a large sample size and a long follow-up, unlikely to be conducted in early development. A rich literature has been published on the surrogacy in MS [118, 122, 121, 123, 120, 119, 124, 117, 116], where the number of brain lesions seen on Magnetic Resonance Imaging (MRI) and the annualized relapse rate have been identified as good predictors of EDSS worsening. Thus, let us assume our development plan is designed with one phase II trial, powered to detect a treatment difference on MRI lesion counts, followed by one phase III trial with the disability progression as primary efficacy endpoint, both versus active control.

The type I error is fixed at 2.5%, one-sided. The phase II trial is conducted with the MRI lesion counts after 1 year of treatment as primary endpoint, with 100 patients per arm to detect a relative risk reduction of 50% with a power of 80%, assuming

a mean number of lesions in the active control arm equal to 4 and a dispersion parameter of 2 (see [105], Appendix 1). The annualized relapse rate and the disability progression at 2 years are collected as secondary endpoints. Given the results summarized in Table 5.2, we wish to calculate the PPoS of the future phase III trial, designed with 337 patients per arm, in order to reach a power of 90% to detect a relative risk reduction of 30%, assuming a proportion of 40% of patients under active control with a disability progression after 2 years of treatment.

TABLE 5.2: Phase II trial. Estimated log risk ratios for the MRI lesion counts, the annualized relapse rate and the disability progression with their standard errors.

	MRI $\hat{\gamma}_1$ (δ_1)	Relapse $\hat{\gamma}_2$ (δ_2)	Disability $\hat{\theta}$ (σ)
N/arm	100	100	10
log(RR) (SE)	-0.511 (0.246)	-0.693 (0.387)	-0.386 (0.646)

N=number of patients. RR=Relative Risk. SE=Standard Error.

We calculated the PPoS of the next trial with and without surrogate priors, which are obtained from either the MRI lesion counts (one surrogate endpoint), the annualized relapse rate (one surrogate endpoint) or both (two surrogate endpoints) to predict the disability progression.

The relationship between the endpoints, using the meta-analytic approach described in Section 5.2.2, has been previously published in Pozzi *et al.*, 2016 [105], based on the data from Sormani *et al.*, 2010 [123]. We translated the `Winbugs` code from [105] in `Stan` [18], in order to use the most up-to-date software for Bayesian inference, but the data of the meta-analysis (Appendix C, Section C.3, Table C.1) are strictly identical to those of the original publication, with a between-endpoint correlation fixed at 0.05 as in their main analysis.

The meta-analytic model was fitted using `R` and `Stan` with the `R` package `RStan`. For three chains, the first 20,000 (40,000 with a thin of 2) Markov Chain Monte Carlo samples were discarded for burn-in, then 10,000 (20,000 with a thin of 2) further simulations were run to estimate the regression parameters. Their posterior means and 95% credible intervals are presented in Table 5.3, for the one-surrogate models and the two-surrogate model. The results of the model using relapse as surrogate endpoint (with those of another model assuming a structural relationship between the endpoints, not used here for simplicity) were already presented in [105], with slight differences due to sampling only. It can be noted that the MRI lesion counts and the annualized relapse rate, taken separately, are good predictors of the disability progression with parameters b significantly different from 0. When both surrogates are used jointly in the model, the mean slope parameter associated to the MRI lesion counts becomes negative, which seems counter-intuitive. However its credibility interval includes 0, and this model is the one providing the smallest error term τ, so we chose to present its results anyway.

Figure 5.4 displays the predictive distributions of the treatment effect on disability progression in the future phase III trial, with the respective PPoS. The prior distributions and the data are consistent, so the methods with and without handling prior-data conflict provide identical results and are presented only once. When the

TABLE 5.3: Posterior means and 95% credible intervals (CrI) of the regression parameters from fitting the meta-analytic models with disability progression as final endpoint.

Model	Parameter		Mean [95% CrI]
MRI as surrogate	Intercept	a	-0.091 [-0.445,0.248]
	MRI	b	0.398 [0.026,0.773]
	Error	τ	0.375 [0.210,0.641]
Relapse as surrogate	Intercept	a	0.078 [-0.099,0.242]
	Relapse	b	0.756 [0.471,1.017]
	Error	τ	0.153 [0.053,0.292]
MRI+Relapse as surrogates	Intercept	a	0.062 [-0.105,0.222]
	MRI	b_1	-0.110 [-0.306,0.082]
	Relapse	b_2	0.749 [0.479,1.011]
	Error	τ	0.099 [0.017,0.236]

PPoS is calculated using a weakly informative prior ($N(0, 10^6)$) and the data of the 20 patients (10 per arm) having a disability progression evaluated in phase II, the predictive distribution of $\hat{\theta}_f$ lacks of precision and the resulting PPoS is equal to 60%, leading to a relatively high uncertainty to take the decision to pursue or not the development. All surrogate priors are more informative, with the greatest precision when the two surrogate endpoints are used jointly in the model, as reflected by the τ in Table 5.3. The observed relative risk on MRI lesion counts is 60%, lower than expected, and provides a PPoS=57% when used as single surrogate, only slightly in favor of the success. On the other hand, the observed risk reduction on relapse (50%), assessed for all 200 patients in phase II, is large and precise enough to provide a PPoS greater than 74%. The analysis using the surrogate prior resulting from the two-surrogate model combined with the available data on the final endpoint gives the greatest precision and provides a reasonable confidence in the study success, with a PPoS for the next trial equal to 72%. In conclusion, these results are all going in the same direction and could support a decision in favor of the continuation of the development. This illustrates how using all the available information on all the potential surrogate endpoints and on the final endpoints strengthen the decision-making process.

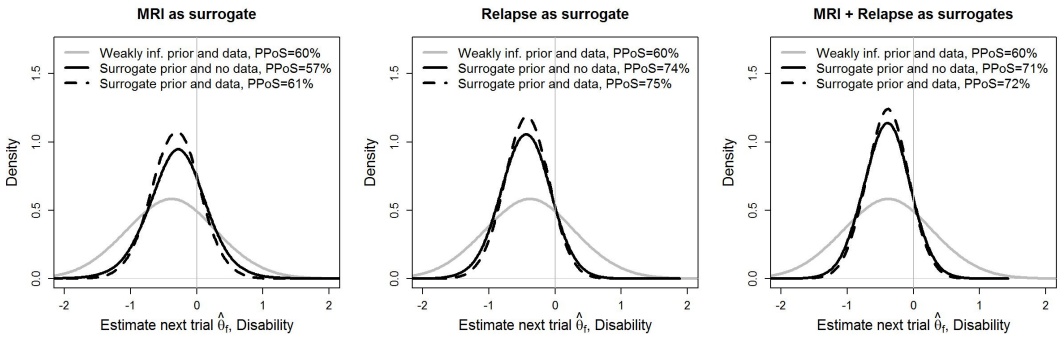

FIGURE 5.4: Predictive distributions of the treatment effect on disability progression, $\hat{\theta}_f$, in the future phase III trial, with the respective PPoS.

5.5 Discussion

We proposed a general, reliable and reproducible methodology to predict the success of a future trial from data on surrogate endpoints, in a way that makes the best use of all the available evidence, in order to enhance decision-making in drug development. The predictions are based on an informative prior, called surrogate prior, derived from the information collected on the surrogate endpoints, combined with the data on the clinical endpoint of interest, if available. The surrogate endpoints considered here are markers used in early phase of development to predict a clinical benefit, but are not necessarily formally accepted for confirmatory phase from a regulatory perspective.

The surrogate prior is built using the past data on the surrogate endpoints during the drug development of interest and the quantitative relationship between the surrogate and the final endpoints, which could be obtained using published methods for surrogacy validation. Approaches based on data from single trials are known to be inadequate as they focus on the patient level association [13, 14, 15], therefore we suggest to use a meta-analytic approach in which the association between the treatment effects on the endpoints is established at the trial level. Since the individual patient data from the published clinical trials are rarely available, we focused on a method using aggregated data [26]. However, this approach usually requires to make assumptions regarding the between-endpoint correlations within each trial, almost never published, and to perform sensitivity analyses accordingly [26, 108, 105]. The availability of individual patient data resolves this problem, because then a two-stage model can be used [15] where the within-trial dependence between the endpoints is first estimated from the patient observations. More generally, meta-analytic approaches require a large amount of external data to estimate the relationship between the endpoints with a sufficient precision. It may happen that these data are not available, or could not be extrapolated to the development of interest. We can argue, however, that designing a drug development using surrogate endpoints in early phases for decision-making should be supported by a certain amount of evidence. Otherwise, we are facing a high uncertainty regarding the ability of the surrogate endpoints to predict the treatment effect on the final endpoint, leading to uncertainty regarding the success of the development itself.

This emphasizes the clear advantage of using both an informative prior based on surrogate endpoints and data collected on the final endpoint. On one hand, if little evidence can be gathered on the relationship between the endpoints, the surrogate prior will be imprecise. On the other hand, if only sparse data on the final endpoint are collected in the past trials, the resulting estimations will be imprecise. Combining both sources of information could increase the precision of the predictions and strengthen the decision-making process.

This will be the case if the surrogate prior and the final endpoint data are consistent. In case of prior-data conflict, more credit should usually be given to the data on the final endpoint, since it will be used in the future development. We considered two approaches addressing this issue: the testing approach and the mixture prior approach. The first is an 'all-or-nothing' approach where the surrogate prior is discarded if a conflict is detected at a predefined testing level. It may be appropriate if there is some biological justification for the mismatch, e.g. that the experimental treatment has a novel mechanism of action, different from the treatments included

in the meta-analysis, leading to a treatment effect on one but not all endpoints. On the other hand, if the surrogacy remains biologically plausible, one may prefer the second approach which discounts (but does not exclude) the informative prior proportionally to the extent of the prior-data conflict by adding a weakly informative component in the prior. A similar approach, the power prior [64, 65], has also been widely used to build informative prior from historical data and could be used in our context. The interpretation of the power parameter, however, may be less straightforward than the weights of the mixture prior approach, which could be naturally interpreted as the probability of compatibility between the data and the surrogate prior. All these methods rely on pre-specified parameters (testing level, weights, precision of the weakly informative component, power parameter), which choice is critical and requires a subjective judgment. While some guidance may be found in the literature [95, 115, 65], we wish to emphasize the importance of assessing the operating characteristics of the approaches before conducting the analyses, and of performing sensitivity analyses for different choices of parameters afterwards.

More generally, sensitivity analyses should be conducted as part of the decision-making process. The robustness of the PPoS calculated using different methods should be evaluated, as illustrated in our simulation study and in the application in Multiple Sclerosis. The impact of the final endpoint data and of the surrogate prior on the PPoS can be assessed by conducting separate analyses considering only one of the two sources of evidence. Of note, one can consider including the past trial(s) of the drug development of interest in the meta-analysis to estimate the relationship between the endpoints, instead of using only external clinical trials. However, in this case, we obtain a data-dependent prior which may lead to difficulties in interpreting the Bayesian model [17]. Also, when several surrogate markers are available, the PPoS may be sensible to the structural assumptions made on the relationships between the endpoints. For example in Multiple Sclerosis, one can assume that the treatment effect on the disability progression is independent from the effect on the MRI lesion counts conditionally on the effect on the annualized relapse rate [105, 10]. Different conclusions obtained from sensitivity analyses can highlight some potential risks associated to the decisions.

Our example in Multiple Sclerosis illustrates the use of the methods in a clinical context, but is not intended to provide a guidance for a drug development in this disease. First, the data used for the meta-analysis were published in 2010 [123] and should be updated with more recent trials, involving again some clinical expertise to select them appropriately. Also, different choices could be made regarding the number of studies in the development, their primary endpoints, the underlying distribution of the parameters and the assumptions for sample calculation.

In most cases in clinical trials, there exists an estimate of the treatment difference reasonably assumed to be normally distributed. Therefore, we focused on normally distributed endpoints, but this may not correspond to the main analysis model. The results are usually asymptotically equivalent and are not expected to have a significant impact on the predictions, but the extension of the methods presented herein to other distributions provides an area for future research.

In conclusion, the proposed methodology is general and can be applied in various contexts. It is expected to be a valuable quantitative tool to support decision-making,

since the use of predictive markers is important to accelerate drug developments and to select promising drug candidates, better and earlier.

Chapter 6

Conclusion and perspectives

The four methodologies presented herein are valuable quantitative tools to support decision-making in the pharmaceutical development. They have strong theoretical foundations and were shown to have soundness in the context of healthcare decisions. They could be used in a wide range of applications throughout the drug life-cycle and their utility has already been demonstrated in very concrete situations.

We focused in this report on two research axes: the drug benefit-risk assessment and the predictive probability of success. The respective methodologies are mainly intended to support decisions at the "study level" and at the "development level". Other topics of research within these decision levels are currently discussed in the scientific literature and the statistical community. Among others we can cite:

- **Decision-making frameworks** for Go/No-Go decisions. Both frequentist and Bayesian frameworks have been developed and used in the pharmaceutical industry to support internal decision-making [49, 76, 34, 24]. These frameworks are based on pre-planned rules regarding some decision parameters, appropriately defined using clinical considerations and operating characteristics.

- **Prior elicitation**. Prior expert knowledge and experience is essential to support decision-making, in particular when the amount of previous data is low or when their mathematical aggregation is challenging. The SHeffield ELicitation Framework (SHELF) [100] was developed to capture expert knowledge about one or more quantities of interest in the form of a probability distribution. This structured approach to prior elicitation has been shown to be beneficial for internal decision-making [25], as it permits not only to homogenize the collection and quantification of expert beliefs, but also to capture their uncertainty. This framework could have a broad range of application in drug development, from the quantification of trial-specific expected quantities (e.g. the mean treatment effect) to more general questions about the success of a development or a portfolio, and there will certainly be further applications and development in the future.

- **Patient preferences elicitation in benefit-risk assessment**. Quantitative methods for drug benefit-risk assessment require value judgments regarding the trade-off between the benefit and risk criteria. Methods for quantifying subjective preferences have been widely studied in the literature [133, 82, 54, 107, 113, 8, 61, 132, 9], and today's research focuses on the incorporation of patient needs and perspectives into drug decision-making. One of the main objectives of the IMI-PREFER consortium [28] is to identify and assess which methodologies are most suitable for eliciting patient preferences on benefits and risks at different decision points in the product life cycle.

More generally, subjectivity is inherent to decision-making analyses: quantitative tools are intended to support the choice between several alternatives based not only on the available information but also on the preferences of the decision-makers. The usefulness of the methods and frameworks already proposed in the literature for preference elicitation should be evaluated, along with potential options for improvement.

Finally, we extended our research to decisions at the "portfolio level", with quantitative methods such as predictions of the number of marketing authorizations over time within a company, portfolios financial risk-value profiles or time-to-milestone analyses. Our research on this topic requires further consideration and is currently unpublished.

In conclusion, quantitative decision-making is increasingly used in the pharmaceutical industry. Given the breadth and the complexity of the context in which decisions are taken, many different questions are raised, and many statistical methods are or could be developed. Evidence-based methods avoid relying on questionable assumptions, and their usefulness is widely recognized by the scientific community. Subjectivity should be incorporated but should also be challenged, and structured quantitative tools permit to guide and to homogenise the decision-making process. Importantly, statistical methods take into account the uncertainty inherent to drug developments and to the use of different sources of evidence. Sensitivity analyses should be conducted as part of the decision-making process, and the confidence in the decisions increases with the robustness of the conclusions.

Bibliography

[1] Alonso, A, Bigirumurame, T, Burzykowski, T, Buyse, M, Molenberghs, G, Muchene, L, Perualila, NJ, Shkedy, Z, and Elst, W Van der. *Applied Surrogate Endpoint Evaluation Methods with SAS and R*. USA: Chapman & Hall/CRC, Taylor & Francis Group, 2017.

[2] Antonijevic, Z. *Optimization of Pharmaceutical R & D Programs and Portfolios*. Switzerland: Springer International Publishing, 2015.

[3] Ashby, D and Smith, A. "Evidence-based medicine as Bayesian decision-making". *Statistics in Medicine* 19 (2000), 3291—3305.

[4] Baker, SG and Kramer, BS. "A perfect correlate does not a surrogate make". *BMC Medical Research Methodology* 3.1 (2003), p. 16. DOI: 10.1186/1471-2288-3-16.

[5] Belton, V and Stewart, TJ. *Multiple Criteria Decision Analysis: An Integrated Approach*. New York, USA: Springer US, 2002. DOI: 10.1007/978-1-4615-1495-4.

[6] Berger, JO. *Statistical decision theory and Bayesian Analysis (2nd ed.)* New York: Springer-Verlag, 1985. DOI: 10.1007/978-1-4757-4286-2.

[7] Bertaina, M, Ferraro, I, Omedè, P, Conrotto, F, Saint-Hilary, G, Cavender, MA, Claessen, BE, Henriques, JPS, Frea, S, Usmiani, T, Grosso Marra, W, Pennone, M, Moretti, C, D'Amico, M, and D'Ascenzo, F. "Meta-analysis Comparing Complete or Culprit Only Revascularization in Patients With Multivessel Disease Presenting With Cardiogenic Shock". *The American Journal of Cardiology* (2018). DOI: 10.1016/j.amjcard.2018.08.003.

[8] Broekhuizen, H, Groothuis-Oudshoorn, CGM, Hauber, AB, Jansen, JP, and IJzerman, MJ. "Estimating the value of medical treatments to patients using probabilistic multi criteria decision analysis". *BMC Medical Informatics and Decision Making* 15.1 (2015), p. 102. DOI: 10.1186/s12911-015-0225-8.

[9] Broekhuizen, H, IJzerman, MJ, Hauber, AB, and Groothuis-Oudshoorn, CGM. "Weighing Clinical Evidence Using Patient Preferences: An Application of Probabilistic Multi-Criteria Decision Analysis". *PharmacoEconomics* 35.3 (2017), pp. 259–269. DOI: 10.1007/s40273-016-0467-z.

[10] Bujkiewicz, S, Thompson, JR, Riley, RD, and Abrams, KR. "Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process". *Statistics in Medicine* 35.7 (2016), pp. 1063–1089. DOI: 10.1002/sim.6776.

[11] Burrello, J, Erhardt, EM, Saint-Hilary, G, Veglio, F, Rabbia, F, Mulatero, P, Monticone, S, and D'Ascenzo, F. "Treatment of Arterial Hypertension in Children and Adolescents: A Network Meta-Analysis". *Hypertension* 72.2 (2018), pp. 306–313. DOI: 10.1161/HYPERTENSIONAHA.118.10862.

[12] Burzykowski, T and Buyse, M. "Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation". *Pharmaceutical Statistics* 5.3 (2006), pp. 173–186. DOI: 10.1002/pst.207.

[13] Burzykowski, T, Molenberghs, G, and Buyse, M. *The Evaluation of Surrogate Endpoints*. USA: Springer Science+Business Media, 2005.

[14] Buyse, M, Molenberghs, G, Paoletti, X, Oba, K, Alonso, A, Elst, W Van der, and Burzykowski, T. "Statistical evaluation of surrogate endpoints with examples from cancer clinical trials". *Biometrical Journal* 58.1 (2016), pp. 104–132. DOI: 10.1002/bimj.201400049.

[15] Buyse, M, Molenberghs, G, Burzykowski, T, Renard, D, and Geys, H. "The validation of surrogate endpoints in meta-analyses of randomized experiments". *Biostatistics* 1.1 (2000), pp. 49–67. DOI: 10.1093/biostatistics/1.1.49.

[16] Cameron, IM, Cardy, A, Crawford, JR, Toit, SW du, Hay, S, Lawton, K, Mitchell, K, Sharma, S, Shivaprasad, S, Winning, S, and Reid, IC. "Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II". *Br J Gen Pract* 61.588 (2011), e419–e426. DOI: 10.3399/bjgp11X583209.

[17] Carlin, BP and Louis, TA. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. USA: Chapman & Hall/CRC, 2000. DOI: 10.1201/9781420057669.

[18] Carpenter, B, Gelman, A, Hoffman, MD, Lee, D, Goodrich, B, Betancourt, M, Brubaker, M, Guo, J, Li, P, and Riddell, A. "Stan: A Probabilistic Programming Language". *Journal of Statistical Software, Articles* 76.1 (2017), pp. 1–32. DOI: 10.18637/jss.v076.i01.

[19] Chen, C and Sun, LZ. "Quantification of PFS Effect for Accelerated Approval of Oncology Drugs". *Statistics in Biopharmaceutical Research* 3.3 (2011), pp. 434–444. DOI: 10.1198/sbr.2011.09046.

[20] Chuang-Stein, C. "Sample size and the probability of a successful trial". *Pharmaceutical Statistics* 5.4 (2006), pp. 305–309. DOI: 10.1002/pst.232.

[21] Chuang-Stein, C, Entsuah, R, and Pritchett, Y. "Measures for Conducting Comparative Benefit:Risk Assessment". *Drug Information Journal* 42 (2008), pp. 223–233. DOI: 10.1177/009286150804200304.

[22] Chuang-Stein, C and Kirby, S. *Quantitative Decisions in Drug Development*. Switzerland: Springer International Publishing, 2017. DOI: 10.1007/978-3-319-46076-5.

[23] Chuang-Stein, C, Kirby, S, French, J, Kowalski, K, Marshall, S, Smith, MK, Bycott, P, and Beltangady, MA. "Quantitative Approach for Making Go/No-Go Decisions in Drug Development". *Drug Information Journal* 45 (2011), pp. 187–202. DOI: 10.1177/009286151104500213.

[24] Crisp, A, Miller, S, Thompson, D, and Best, N. "Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development". *Pharmaceutical Statistics* 17.4 (2018), pp. 317–328. DOI: 10.1002/pst.1856.

[25] Dallow, N, Best, N, and Montague, TH. "Better decision making in drug development through adoption of formal prior elicitation". *Pharmaceutical Statistics* 17.4 (2018), pp. 301–316. DOI: 10.1002/pst.1854.

[26] Daniels, MJ and Hughes, MD. "Meta-analysis for the evaluation of potential surrogate markers". *Statistics in Medicine* 16.17 (1997), pp. 1965–1982. DOI: 10.1002/(SICI)1097-0258(19970915).

[27] D'Ascenzo, F, Iannaccone, M, Saint-Hilary, G, Bertaina, M, Schulz-Schupke, S, Lee, C Wahn, Chieffo, A, Helft, G, Gili, S, Barbero, U, Biondi Zoccai, G, Moretti, C, Ugo, F, D'Amico, M, Garbo, R, Stone, G, Rettegno, S, Omede, P, Conrotto, F, Templin, C, Colombo, A, Park, SJ, Kastrati, A, Hildick-Smith, D, Gasparini, M, and F, Gaita. "Impact of design of coronary stents and length of dual antiplatelet therapies on ischaemic and bleeding events: a network meta-analysis of 64 randomized controlled trials and 102 735 patients". *Eur Heart J* 38.42 (2017), pp. 3160–3172. DOI: 10.1093/eurheartj/ehx437.

[28] de Bekker-GrobEmail, EW, Berlin, C, Levitan, B, Raza, K, Christoforidi, K, Cleemput, I, Pelouchova, J, Enzmann, H, Cook, N, and Hansson, MG. "Giving Patients' Preferences a Voice in Medical Treatment Life Cycle: The PREFER Public–Private Project". *Patient* 10.3 (2017), pp. 263–266. DOI: 10.1007/s40271-017-0222-3. URL: https://www.imi-prefer.eu/.

[29] DiMasi, JA, Hansen, RW, and Grabowski, HG. "The price of innovation, new estimates of drug development costs". *J. Health Econ.* 22.2 (2003), 151–185.

[30] DiMasi, JA, Hermann, JC, Twyman, K, Kondru, RK, Stergiopoulos, S, Getz, KA, and Rackoff, W. "A Tool for Predicting Regulatory Approval After Phase II Testing of New Oncology Compounds". *Clinical Pharmacology & Therapeutics* 98.5 (2015), pp. 506–513. DOI: 10.1002/cpt.194.

[31] DiMasi, JA, Reichert, JM, Feldman, L, and Malins, A. "Clinical Approval Success Rates for Investigational Cancer Drugs". *Clinical Pharmacology & Therapeutics* 94.3 (2013), pp. 329–335. DOI: 10.1038/clpt.2013.117.

[32] DiMasi, JA, Feldman, L, Seckler, A, and Wilson, A. "Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs". *Clinical Pharmacology & Therapeutics* 87.3 (2010), pp. 272–277. DOI: 10.1038/clpt.2009.295.

[33] DiSantostefano, RL, Berlin, JA, Chuang-Stein, C, Quartey, G, Eichenbaum, G, and Levitan, B. "Selecting and Integrating Data Sources in Benefit–Risk Assessment: Considerations and Future Directions". *Statistics in Biopharmaceutical Research* 8.4 (2016), pp. 394–403. DOI: 10.1080/19466315.2016.1225596.

[34] Dunyak, J, Mitchell, P, Hamrén, B, Helmlinger, G, Matcham, J, Stanski, D, and Al-Huniti, N. "Integrating dose estimation into a decision-making framework for model-based drug development". *Pharmaceutical Statistics* 17.2 (2018), pp. 155–168. DOI: 10.1002/pst.1841.

[35] EMA. "Benefit-risk methodology project. Work package 1 Report: description of the current practice of benefit-risk assessment for centralised procedure products in the EU regulatory network" (2011). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/07/WC500109478.pdf.

[36] EMA. "Benefit-risk methodology project. Work package 2 Report: Applicability of current tools and processes for regulatory benefit-risk assessment" (2010). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2010/10/WC500097750.pdf.

[37] EMA. "Benefit-risk methodology project. Work package 3 Report: Field tests" (2011). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/09/WC500112088.pdf.

[38] EMA. "Benefit-risk methodology project. Work package 4 Report: Benefit-risk tools and processes" (2012). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2012/03/WC500123819.pdf.

[39] EMA. "EMEA/H/C/000354-N/0068 - Ketek : EPAR" (2017). URL: http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/human/medicines/000354/human_med_000873.jsp&mid=WC0b01ac058001d124.

[40] EMA. "EMEA/H/C/354/A22/41 - Ketek : EPAR - Scientific discussion - Variation" (2007). URL: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Scientific_Discussion_-_Variation/human/000354/WC500100697.pdf.

[41] EMA. "Guidance document on the content of the <Co->Rapporteur day 80 critical assessment report" (2013). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004800.pdf.

[42] EMA. "Guideline on clinical investigation of medicinal products in the treatment of depression" (2013). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/05/WC500143770.pdf.

[43] EMA. "Guideline on clinical investigation of medicinal products in the treatment of hypertension" (2011). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/12/WC500100191.pdf.

[44] EMA. "Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus" (2012). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129256.pdf.

[45] EMA. "ICH guideline E2C (R2) on periodic benefit-risk evaluation report (PBRER)" (2013). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2012/12/WC500136402.pdf.

[46] EMA. "Point to consider on applications with 1. Meta-analyses and 2. One pivotal study" (2001). URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf.

[47] Figueira, J, Greco, S, and Ehrgott, M. *Multiple Criteria Decision Analysis: State of the Art Surveys*. New York, USA: Springer-Verlag, 2005. DOI: 10.1007/b100605.

[48] Fleming, TR and DeMets, DL. "Surrogate end points in clinical trials: Are we being misled?" *Annals of Internal Medicine* 125 (1996). DOI: 10.7326/0003-4819-125-7-199610010-00011.

[49] Frewer, P, Mitchell, P, Watkins, C, and Matcham, J. "Decision-making in early clinical drug development". *Pharmaceutical Statistics* 15 (2016), pp. 255–263. DOI: 10.1002/pst.1746.

[50] Gail, MG, Pfeiffer, R, Houwelingen, HC Van, and Carroll, RJ. "On meta-analytic assessment of surrogate outcomes". *Biostatistics (Oxford, England)* 1.3 (2000), 231—246. DOI: 10.1093/biostatistics/1.3.231.

[51] Gallo, P, Mao, L, and Shih, VH. "Alternative Views On Setting Clinical Trial Futility Criteria". *Journal of Biopharmaceutical Statistics* 24.5 (2014), pp. 976–993. DOI: 10.1080/10543406.2014.932285.

[52] Gasparini, M, Scala, L Di, Bretz, F, and Racine-Poon, A. "Predictive probability of success in clinical drug development". *Epidemiology Biostatistics and Public Health* 10-1 (2013), e8760–1–14.

[53] Grob, EW deBekker, Juhaeri, J, Kihlbom, U, and B, Levitan. "Giving patients preferences a voice in the medical product lifecycle: why, when and how. The public-private PREFER project: Work package 2". *ISPOR Value & Outcomes Spotlight* 4.3 (2018), 19–21.

[54] Guo, JJ, Pandey, S, Doyle, J, Bian, B, Lis, Y, and Raisch, DW. "A Review of Quantitative Risk-Benefit Methodologies for Assessing Drug Safety and Efficacy - Report of the ISPOR Risk-Benefit Management Working Group". *Value in Health* 13.5 (2010), pp. 657–666. DOI: 10.1111/j.1524-4733.2010.00725.x.

[55] Gupta, U and Verma, M. "Placebo in clinical trials". *Perspectives in Clinical Research* 4.1 (2013), pp. 49–52. DOI: 10.4103/2229-3485.106383.

[56] Hammond, JS, Keeney, RL, and Raiffa, H. *Smart Choices: A Practical Guide to Making Better Decisions*. Boston, MA: Harvard University Press, 1999.

[57] Hauber, AB, Gonzalez, JM, Groothuis-Oudshoorn, CGM, Prior, T, Marshall, D, Cunningham, C, IJzerman, MJ, and Bridges, JFP. "Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force". *Value in Health* 19.4 (2016), pp. 300 –315. DOI: 10.1016/j.jval.2016.04.004.

[58] Hay, M, Thomas, DW, Craighead, JL, Economides, C, and J, Rosenthal. "Clinical development success rates for investigational drugs". *Nat Biotechnol* 32.4 (2014), pp. 40–51. DOI: 10.1038/nbt.2786.

[59] Holden, WL, Juhaeri, J, and Dai, W. "Benefit-risk analysis: a proposal using quantitative methods". *Pharmacoepidemiology & Drug Safety* 12 (2003), 611––616. DOI: 10.1002/pds.887.

[60] Hong, S and Shi, L. "Predictive power to assist phase 3 go/no go decision based on phase 2 data on a different endpoint". *Statistics in Medicine* 31 (2012), pp. 831–843. DOI: 10.1002/sim.4476.

[61] Hughes, D, Waddingham, E, Mt-Isa, S, Goginsky, A, Chan, E, Downey, GF, Hallgreen, CE, Hockley, KS, Juhaeri, J, Lieftucht, A, Metcalf, MA, Noel, RA, Phillips, LD, Ashby, D, A, A Micaleff, and Group, PROTECT Benefit-Risk. "Recommendations for benefit-risk assessment methodologies and visual representations". *Pharmacoepidemiology and Drug Safety* 25.3 (2016), pp. 251–262. DOI: 10.1002/pds.3958.

[62] Hughes, DA, Bayoumi, AM, and Pirmohamed, M. "Current assessment of risk-benefit by regulators: is it time to introduce decision analyses?" *Clinical Pharmacology and Therapeutics* 82 (2007), 123—127.

[63] Iannaccone, M, Saint-Hilary, G, Menardi, D, Vadalà, P, Bernardi, A, Bianco, M, Montefusco, A, Omedè, P, D'Amico, S, Piazza, F, Scacciatella, P, D'Amico, M, Moretti, C, Biondi-Zoccai, G, Gasparini, M, Gaita, F, and F, D'Ascenzo. "Network meta-analysis of studies comparing closure devices for femoral access after percutaneous coronary intervention". *J Cardiovasc Med* 19.10 (2018), pp. 586–596. DOI: do10.2459/JCM.0000000000000697.

[64] Ibrahim, JG and Chen, MH. "Power Prior Distributions for Regression Models". *Statistical Science* 15.1 (2000), pp. 46–60.

[65] Ibrahim, JG, Chen, MH, Gwon, Y, and Chen, F. "The power prior: theory and applications". *Statistics in Medicine* 34.28 (2015), pp. 3724–3749. DOI: 10.1002/sim.6728.

[66] ICH. "ICH guideline E6: Guideline for good clinical practice" (1996). URL: https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdf.

[67] ICH. "ICH guideline E9: Statistical Principles for Clinical Trials" (1998). URL: http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html.

[68] *IMI-PREFER*. 2018. URL: https://www.imi-prefer.eu/.

[69] IMI PROTECT. "IMI PROTECT Work package 5: Benefit-risk integration and representation" (2013). URL: http://www.imi-protect.eu/wp5.shtml.

[70] IMI PROTECT. "IMI PROTECT Work package 5:Report 1:b:ii Benefit – Risk Wave 1 Case Study Report: Ketek® (telithromycin)" (2012). URL: http://www.imi-protect.eu/documents/QuarteyetalBenefitRiskWave1CasestudyReportTelithromycinFeb2012.pdf.

[71] JG, JG Norstrom. "The use of precautionary loss functions in risk analysis". *IEEE Transactions on reliability* 45.3 (1996), pp. 400–403.

[72] Johns, D and Andersen, JS. "Use of predictive probabilities in Phase II and Phase III clinical trials". *Journal of Biopharmaceutical Statistics* 9.1 (1999), pp. 67–79.

[73] Johnson, FR, Lancsar, E, Marshall, D, Kilambi, V, Muhlbacher, A, Regier, DA, Bresnahan, BW, Kanninen, B, and Bridges, JFP. "Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force". *Value in Health* 16.1 (2013), pp. 3 –13. DOI: 10.1016/j.jval.2012.08.2223.

[74] Keeney, RL and Winterfeldt, D von. "Practical value models". *Published Articles and Papers* 36 (2009). URL: //research.create.usc.edu/published_papers/36.

[75] Kelbert, M and Mozgunov, P. "Generalization of Cramér-Rao and Bhattacharyya inequalities for the weighted covariance matrix". *Mathematical Communications* 22.1 (2017), pp. 25–40.

[76] Kirchner, M, Kieser, M, Gotte, H, and Schuler, A. "Utility-based optimization of phase II/III programs". *Statistics in Medicine* 35.2 (2015), pp. 305–316. DOI: 10.1002/sim.6624.

[77] Lahdelma, R, Hokkanen, J, and Salminen, P. "SMAA - stochastic multiobjective acceptability analysis". *European Journal of Operational Research* 106 (1998), 137—143. DOI: 10.1016/S0377-2217(97)00163-X.

[78] Langhans, SD and Lienert, J. "Four Common Simplifications of Multi-Criteria Decision Analysis do not hold for River Rehabilitation". *PLOS One* 11.3 (2016), pp. 1–27. DOI: 10.1371/journal.pone.0150695.

[79] Leong, J, Salek, S, and Walker, S. *Benefit-Risk Assessment of Medicines*. Switzerland: Springer International Publishing, 2015.

[80] Marcelon, L, Verstraeten, T, Dominiak-Felden, G, and Simondon, F. "Quantitative benefit–risk assessment by MCDA of the quadrivalent HPV vaccine for preventing anal cancer in males". *Expert Review of Vaccines* 15.1 (2016), pp. 139–148. DOI: 10.1586/14760584.2016.1107480.

[81] Marsh, K, Lanitis, T, Neasham, D, Orfanos, P, and Caro, J. "Assessing the Value of Healthcare Interventions Using Multi-Criteria Decision Analysis: A Review of the Literature". *PharmacoEconomics* 32.4 (2014), pp. 345–365. DOI: 10.1007/s40273-014-0135-0.

[82] Marsh, K, IJzerman, M, Thokala, P, Baltussen, R, Boysen, M, Kalo, Z, Longrenn, T, Mussen, F, Peacock, S, Watkins, J, and Devlin, N. "Multiple Criteria Decision Analysis for Health Care Decision Making - Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force". *Value in Health* 19 (2016), pp. 125–137.

[83] Marsh, K, Sculpher, M, Caro, JJ, and Tervonen, T. "The Use of MCDA in HTA: Great Potential, but More Effort Needed". *Value in Health* (2017). DOI: 10.1016/j.jval.2017.10.001.

[84] Martin, DM and Mazzotta, M. "Non-monetary valuation using Multi-Criteria Decision Analysis: Sensitivity of additive aggregation methods to scaling and compensation assumptions". *Ecosystem Services* 29 (2018), pp. 13–22. DOI: 10.1016/j.ecoser.2017.10.022.

[85] Mateu-Figueras, G, Pawlowsky-Glahn, V, and Egozcue, JJ. "The normal distribution in some constrained sample spaces". *SORT: statistics and operations research transactions* 37.1 (2013), pp. 29–56.

[86] McIntosh, MW. "The population risk as an explanatory variable in resarch synthesis of clinical trials". *Statistics in Medicine* 15.16 (1996), pp. 1713–1728. DOI: 10.1002/(SICI)1097-0258(19960830).

[87] Morton, A. "Treacle and Smallpox: Two Tests for Multicriteria Decision Analysis Models in Health Technology Assessment". *Value Health* 20 (2017), pp. 512–515. DOI: 10.1016/j.jval.2016.10.005.

[88] Mozgunov, P and Jaki, T. "An information-theoretic approach for selecting arms in clinical trials". *arXiv:1708.02426* (2017).

[89] Mozgunov, P, Jaki, T, and Gasparini, M. "Loss Functions in Restricted Parameter Spaces and Their Bayesian Applications". *arXiv:1706.02104* (2017).

[90] Mt-Isa, S, Peters, R, Phillips, LD, Chan, K, Hockley, KS, Wang, N, Ashby, D, and Tzoulaki, I. "Review of visualisation methods for the representation of benefit-risk assessment of medication: Stage 1 of 2" (2013). URL: //www.imi-protect.eu/documents/ShahruletalReviewofvisualisationmethodsfortherepresentationofBRassessmentofmedicationStage1F.pdf.

[91] Mt-Isa, S, Hallgreen, CE, Asiimwe, A, Downey, G, Genov, G, Hermann, R, Huges, D, Lieftucht, A, Noel, R, Peters, R, Phillips, LD, Shepherd, S, Micaleff, A, Ashby, D, and Tzoulaki, I. "Review of visualisation methods for the representation of benefit-risk assessment of medication: Stage 2 of 2" (2013). URL: http://www.imi-protect.eu/documents/ShahruletalReviewofvisualisationmethodsfortherepresentationofBRassessmentofmedicationStage2A.pdf.

[92] Mt-Isa, S, Ouwens, M, Robert, V, Gebel, M, Schacht, A, and Hirsch, I. "Structured Benefit-risk assessment: a review of key publications and initiatives on frameworks and methodologies". *Pharmaceutical Statistics* 15 (2015), pp. 324–332. DOI: 10.1002/pst.1690.

[93] Mussen, F, Salek, S, and Walker, S. "A quantitative approach to benefit-risk assessment of medicines - Part 1: the development of a new model using multi-criteria decision analysis". *Pharmacoepidemiology and Drug Safety* 16 (2007), S2—S15. DOI: 10.1002/pds.1435.

[94] Mussen, F, Salek, S, and Walker, S. Benefit-Risk Appraisal of Medicines: A Systematic Approach to Decision-making. Chichester, UK: John Wiley & Sons Ltd, 2008. Chap. Development and Application of a Benefit-Risk Assessment Model Based on Multi-Criteria Decision Analysis.

[95] Mutsvari, T, Tytgat, D, and Walley, R. "Addressing potential prior-data conflict when using informative priors in proof-of-concept studies". *Pharmaceutical Statistics* 15.1 (2016), pp. 28–36. DOI: 10.1002/pst.1722.

[96] Nemeroff, CB and Thase, ME. "A double-blind, placebo-controlled comparison of venlafaxine and fluoxetine treatment in depressed outpatients". *Journal of Psychiatric Research* 41 (2007). 10.1016/j.jpsychires.2005.07.009, pp. 351–359.

[97] Neuenschwander, B, Rouyrre, N, Hollaender, N, Zuber, E, and Branson, M. "A proof of concept phase II non-inferiority criterion". *Statistics in Medicine* 30 (2011), pp. 1618–1627. DOI: 10.1002/sim.3997.

[98] NICE Decision Support Unit. "Multi-Criteria Decision Analysis (MCDA)" (2011). URL: http://www.nicedsu.org.uk/MCDAforHTADSU.pdf.

[99] Nixon, R, Dierig, C, Mt-Isa, S, Stöckert, I, Tong, T, Kuhls, S, Hodgson, G, Pears, J, Waddingham, E, Hockley, K, and Thomson, A. "A case study using the PrOACT-URL and BRAT frameworks for structured benefit risk assessment". *Biometrical Journal* 58.1 (2016), pp. 8–27. DOI: 10.1002/bimj.201300248.

[100] O'Hagan, A and Oakley, JE. "The Sheffield elicitation framework (version 3.0)". *School of Mathematics and Statistics, University of Sheffield* (2016). URL: http://www.tonyohagan.co.uk/shelf/SHELF3.html.

[101] OHagan, A, Stevens, JW, and Campbell, MJ. "Assurance in clinical trial design". *Pharmaceutical Statistics* 4 (2005), pp. 187–201. DOI: 10.1002/pst.175.

[102] OHagan, A, Buck, CE, Daneshkhah, A, Eiser, JR, Garthwaite, PH, Jenkinson, DJ, Oakley, JE, and Rakow, T. *Uncertain judgements: Eliciting experts probabilities*. Chichester, UK: John Wiley & Sons Ltd, 2006.

[103] PhRMA. *BRAT Software beta v3.01*. 2011. URL: http://www.cirs-brat.org/.

[104] PhRMA. *The PhRMA BRAT Framework for Benefit-Risk Assessment - User's Guide to the Process*. 2011. URL: http://www.cirs-brat.org/.

[105] Pozzi, L, Schmidli, H, and Ohlssen, DI. "A Bayesian hierarchical surrogate outcome model for multiple sclerosis". *Pharmaceutical Statistics* 15.4 (2016), pp. 341–348. DOI: 10.1002/pst.1749.

[106] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: http://www.R-project.org/.

[107] Riabacke, M, Danielson, M, and Ekenberg, L. "State-of-the-Art Prescriptive Criteria Weight Elicitation". *Advances in Decision Sciences* (2012), pp. 1–24. DOI: 10.1155/2012/276584.

[108] Riley, RD. "Multivariate meta-analysis: the effect of ignoring within-study correlation". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.4 (2009), pp. 789–811. DOI: 10.1111/j.1467-985X.2008.00593.x.

[109] Rufibach, K, Jordan, P, and Abt, M. "Sequentially updating the likelihood of success of a Phase 3 pivotal time-to-event trial based on interim analyses or external information". *Journal of Biopharmaceutical Statistics* 26.2 (2016), pp. 191–201. DOI: 10.1080/10543406.2014.972508.

[110] Saint-Hilary, G, Robert, V, and Gasparini, M. "Decision-making in drug development using a composite definition of success". *Pharmaceutical Statistics* (2018), pp. 1–15. DOI: 10.1002/pst.1870.

[111] Saint-Hilary, G, Robert, V, Gasparini, M, Jaki, T, and Mozgunov, P. "A novel measure of drug benefit–risk assessment based on Scale Loss Score". *Statistical Methods in Medical Research* (2018), pp. 1–16. DOI: 10.1177/0962280218786526.

[112] Saint-Hilary, G, Cadour, S, Robert, V, and Gasparini, M. "A simple way to unify multicriteria decision analysis (MCDA) and stochastic multicriteria acceptability analysis (SMAA) using a Dirichlet distribution in benefit–risk assessment". *Biometrical Journal* 59.3 (2017), pp. 567–578. DOI: 10.1002/bimj.201600113.

[113] Sashegyi, A, Felli, J, and Noel, R. *Benefit-Risk Assessment in Pharmaceutical Research and Development*. USA: Chapman and Hall/CRC, 2013.

[114] Say, RE and Thomson, R. "The importance of patient preferences in treatment decisions - challenges for doctors". *BMJ* 327.7414 (2003), pp. 542–545.

[115] Schmidli, H, Gsteiger, S, Roychoudhury, S, OHagan, A, Spiegelhalter, DJ, and Neuenschwander, B. "Robust meta-analytic-predictive priors in clinical trials with historical control information". *Biometrics* 70.4 (2014), pp. 1023–1032. DOI: 10.1111/biom.12242.

[116] Sormani, MP, Arnold, DL, and Stefano, N De. "Treatment effect on brain atrophy correlates with treatment effect on disability in multiple sclerosis". *Ann Neurol* 75.1 (2014), pp. 43–49. DOI: 10.1002/ana.24018.

[117] Sormani, MP and Bruzzi, P. "MRI lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials". *Lancet Neurol* 12.7 (2013), pp. 669 –676. DOI: 10.1016/S1474-4422(13)70103-0.

[118] Sormani, MP, Comi, P Bruzzi andG, and Filippi, M. "MRI metrics as surrogate markers for clinical relapse rate in relapsing-remitting MS patients". *Neurology* 58 (2002), 417–421.

[119] Sormani, MP, Li, DK, Bruzzi, P, Stubinski, B, Cornelisse, P, Rocak, S, and Stefano, N De. "Combined MRI lesions and relapses as a surrogate for disability in multiple sclerosis". *Neurology* 77 (2011), 1684–1690. DOI: 10.1212/WNL.0b013e31823648b9.

[120] Sormani, MP, Stubinski, B, Cornelisse, P, Rocak, S, Li, D, and Stefano, N De. "Magnetic resonance active lesions as individual-level surrogate for relapses in multiple sclerosis". *Mult Scler* 17.5 (2011), 541–549. DOI: 10.1177/1352458510391837.

[121] Sormani, MP, Bonzano, L, Roccatagliata, L, Cutter, GR, Mancardi, GL, and Bruzzi, P. "Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: a meta-analytic approach". *Ann Neurol* 65.3 (2009), 268–275.

[122] Sormani, MP, Bruzzi, P, Beckmann, K, and al., et. "MRI metrics as surrogate endpoints for EDSS progression in SPMS patients treated with IFN beta-1b". *Neurology* 60 (2003), 1462–1466.

[123] Sormani, MP, Bonzano, L, Roccatagliata, L, Mancardi, GL, Uccelli, A, and Bruzzi, P. "Surrogate endpoints for EDSS worsening in multiple sclerosis. A meta-analytic approach". *Neurology* 75.4 (2010), pp. 302–309. DOI: 10.1212/WNL.0b013e3181ea15aa.

[124] Sormani, MP, Signori, A, Siri, P, and al., et. "Time to first relapse as an endpoint in multiple sclerosis clinical trials". *Mult Scler* 19 (2013), 466–474.

[125] Spiegelhalter, DJ, Abrams, KR, and Myles, JP. *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, UK: John Wiley & Sons Ltd, 2004.

[126] Spiegelhalter, DJ, Reedman, LS, and Blackburn, PR. "Monitoring clinical trials: conditional power or predictive power?" *Control Clin Trials* 7.1 (1986), 8––17. DOI: 10.1016/0197-2456(86)90003-6.

[127] Stallard, N, Whitehead, J, and Cleall, S. "Decision-making in a phase II clinical trial: a new approach combining Bayesian and frequentist concepts". *Pharmaceutical Statistics* 4 (2005), pp. 119–128. DOI: 10.1002/pst.164.

[128] Tang, Z. "Defensive Efficacy Interim Design: dynamic benefit-risk ratio view using probability of success". *Journal of Biopharmaceutical Statistics* 0 (2016), pp. 1–8. DOI: 10.1080/10543406.2016.1198370.

[129] Tang, Z. "Optimal Futility Interim Design: A Predictive Probability of Success Approach with Time-to-Event Endpoint". *Journal of Biopharmaceutical Statistics* 25 (2015), pp. 1312–1319. DOI: 10.1080/10543406.2014.983646.

[130] Tervonen, T and Lahdelma, R. "Implementing stochastic multicriteria acceptability analysis". *European Journal of Operational Research* 178 (2007), pp. 500–513. DOI: 10.1016/j.ejor.2005.12.037.

[131] Tervonen, T, Valkenhoef, G Van, Buskens, E, Hillege, HL, and Postmus, D. "A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis". *Statistics in Medicine* 30 (2007), 1419—1428. DOI: 10.1002/sim.4194.

[132] Tervonen, T, Gelhorn, H, Bhashyam, S Sri, Poon, JL, Gries, KS, Rentz, A, and Marsh, K. "MCDA swing weighting and discrete choice experiments for elicitation of patient benefit-risk preferences: a critical assessment". *Pharmacoepidemiology and Drug Safety* (2017), pp. 1–9. DOI: 10.1002/pds.4255.

[133] Thokaka, P, Devlin, N, Marsh, K, Baltussen, R, Boysen, M, Kalo, Z, Longrenn, T, Mussen, F, Peacock, S, Watkins, J, and IJzerman, M. "Multiple Criteria Decision Analysis for Health Care Decision Making - An Introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force". *Value in Health* 19 (2016), pp. 1–13.

[134] U.S. Food and Drug Administration. "Advancing Use of Patient Preference Information as Scientific Evidence in Medical Product Evaluation" (2017). URL: https://www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/ucm574320.htm.

[135] U.S. Food and Drug Administration. "Approval Date(s) and History, Letters, Labels, Reviews for NDA 021144 - Ketek" (2017). URL: https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm?event=overview.process&ApplNo=021144.

[136] U.S. Food and Drug Administration. "Providing Postmarket Periodic Safety Reports in the ICH E2C (R2) Format (Periodic Benefit-Risk Evaluation Report)" (2013). URL: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm346564.pdf.

[137] Varian, HR. *Intermediate Microeconomics - A Modern Approach (8th ed.)* New York: WW Norton & Co, 2010.

[138] Waddingham, E, Mt-Isa, S, Nixon, R, and Ashby, D. "A Bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment". *Biometrical Journal* 58 (2016), pp. 28–42. DOI: 10.1002/bimj.201300254.

[139] Wang, M, Liu, GF, and Schindler, J. "Evaluation of program success for programs with multiple trials in binary outcomes". *Pharmaceutical Statistics* 14 (2015), pp. 172–179. DOI: 10.1002/pst.1670.

[140] Wang, Y, Fu, H, Kulkarni, P, and Kaiser, C. "Evaluating and utilizing probability of study success in clinical development". *Clinical Trials* 10.3 (2013), pp. 407–413. DOI: 10.1177/1740774513478229.

[141] Welton, NJ, Sutton, AJ, Cooper, NJ, Abrams, KR, and Ades, AE. *Evidence Synthesis for Decision Making in Healthcare*. Chichester, UK: John Wiley & Sons Ltd, 2012.

[142] Whitehead, A. *Meta-Analysis of Controlled Clinical Trials*. Chichester, UK: John Wiley & Sons Ltd, 2002.

[143] Williams, TA, Burrello, J, Sechi, LA, Fardella, CE, Matrozova, J, Adolf, C, Baudrand, R, Bernardi, S, Beuschlein, F, Catena, C, Doumas, M, Fallo, F, Giacchetti, G, Heinrich, DA, Saint-Hilary, G, Jansen, PM, Januszewicz, A, Kocjan, T, Nishikawa, T, Quinkler, M, Satoh, F, Umakoshi, H, Widimsky, J, Hahner, S, Douma, S, Stowasser, M, Mulatero, P, and Reincke, M. "Computed Tomography and Adrenal Venous Sampling in the Diagnosis of Unilateral Primary Aldosteronism". *Hypertension* (2018). DOI: 10.1161/HYPERTENSIONAHA.118.11382.

[144] Wong, CH, Siah, KW, and Lo, AW. "Estimation of clinical trial success rates and related parameters". *Biostatistics* (2018). DOI: 10.1093/biostatistics/kxx069.

[145] Zhang, J and Zhang, JJ. "Joint probability of statistical success of multiple phase III trials". *Pharmaceutical Statistics* 12 (2013), pp. 358–365. DOI: 10.1002/pst.1597.

Appendix A

Scale Loss Score (SLoS): supplemental material

TABLE A.1: Parameters (α, β) of the Beta posterior distributions of the benefit and risk parameters for Telithromycin (Teli.) and Comparator (Comp.)

	CAP		ABS	
	Teli.	Comp.	Teli.	Comp.
Cure rate	$(2286, 233)$	$(814, 114)$	$(607, 126)$	$(176, 52)$
Hepatic AE	$(58, 1264)$	$(47, 1076)$	$(14, 1303)$	$(4, 943)$
Cardiac AE	$(7, 1315)$	$(5, 1118)$	$(2, 1315)$	$(2, 945)$
Visual AE	$(15, 1307)$	$(5, 1118)$	$(17, 1300)$	$(5, 942)$
Syncope	$(3, 1319)$	$(4, 1119)$	$(1, 1316)$	$(2, 945)$

TABLE A.2: Simulation scenarios with four criteria

			Probability of Benefit (θ_{i1} and θ_{i2})								
			0.1			0.5			0.9		
			0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Probability of Risk (θ_{i3} and θ_{i4})	0.9	0.9	◇●	◇	◇	◇	◇	◇	◇	◇	◇●
		0.5	◇	◇●	◇	◇	◇	◇	◇	◇●	◇
		0.1	◇	◇	◇	◇	◇	◇	◇	◇	◇
	0.5	0.9	◇	◇	◇	◇	◇	◇	◇	◇	◇
		0.5	◇	◇	◇	◇	◇●	◇	◇	◇	◇
		0.1	◇	◇	◇	◇	◇	◇	◇	◇	◇
	0.1	0.9	◇	◇	◇	◇	◇	◇	◇	◇	◇
		0.5	◇	◇●	◇	◇	◇	◇	◇	◇●	◇
		0.1	◇●	◇	◇	◇	◇	◇	◇	◇	◇●

● = treatment T1 ; ◇ = treatment T2

Correlation matrices (Figures S12-S15):

- Positive correlations between criteria

$$\Omega = \begin{bmatrix} 1 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1 \end{bmatrix}$$

- Positive correlations between criteria except for the benefit criterion $j = 2$ which is supposed to be negatively correlated with the others

$$\Omega = \begin{bmatrix} 1 & -0.8 & 0.8 & 0.8 \\ -0.8 & 1 & -0.8 & -0.8 \\ 0.8 & -0.8 & 1 & 0.8 \\ 0.8 & -0.8 & 0.8 & 1 \end{bmatrix}$$

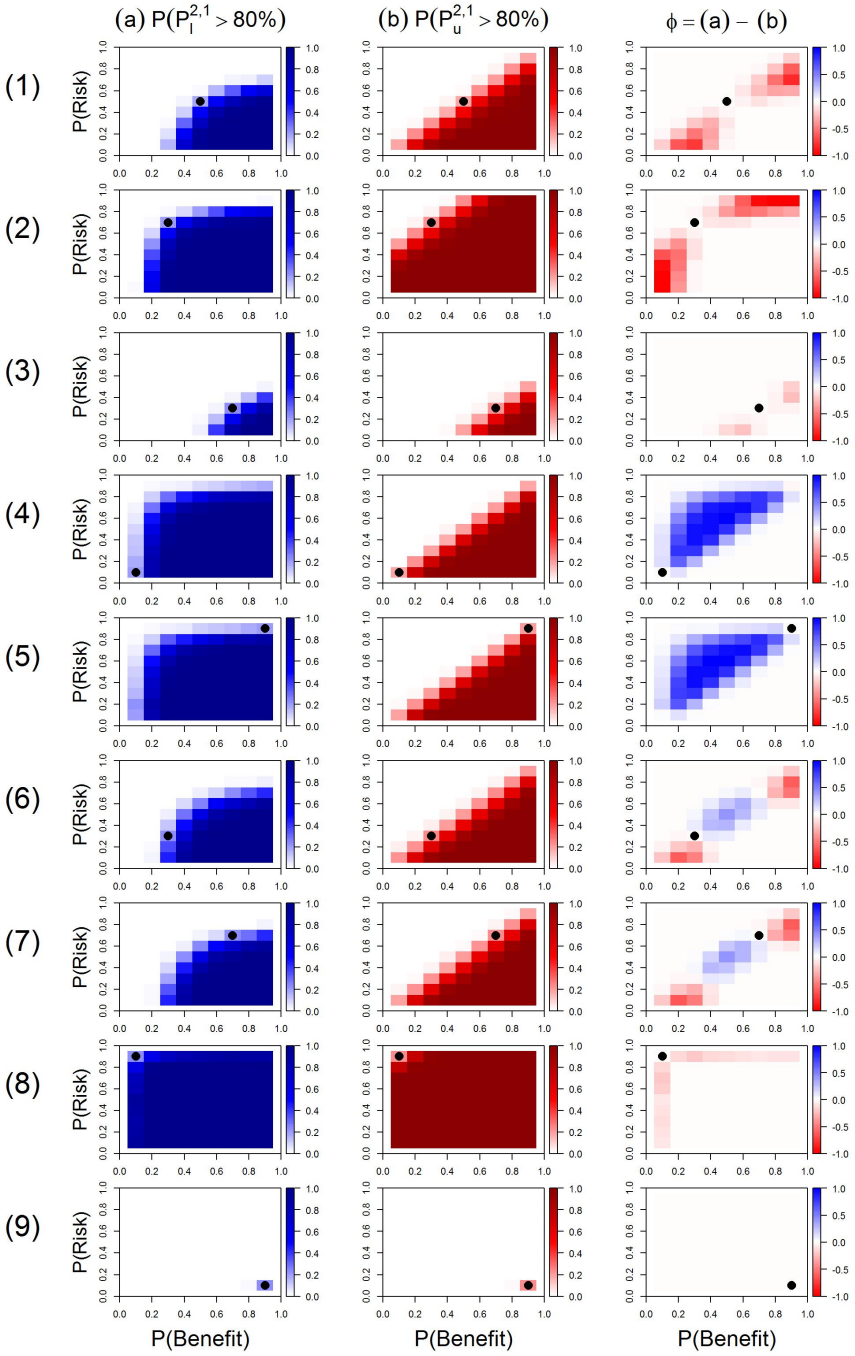

FIGURE A.1: Results of MCDA and SLoS performances in all simulation scenarios for two uncorrelated equally important criteria ($w_j = \tilde{w}_j = 0.5$ for $j = 1, 2$). • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

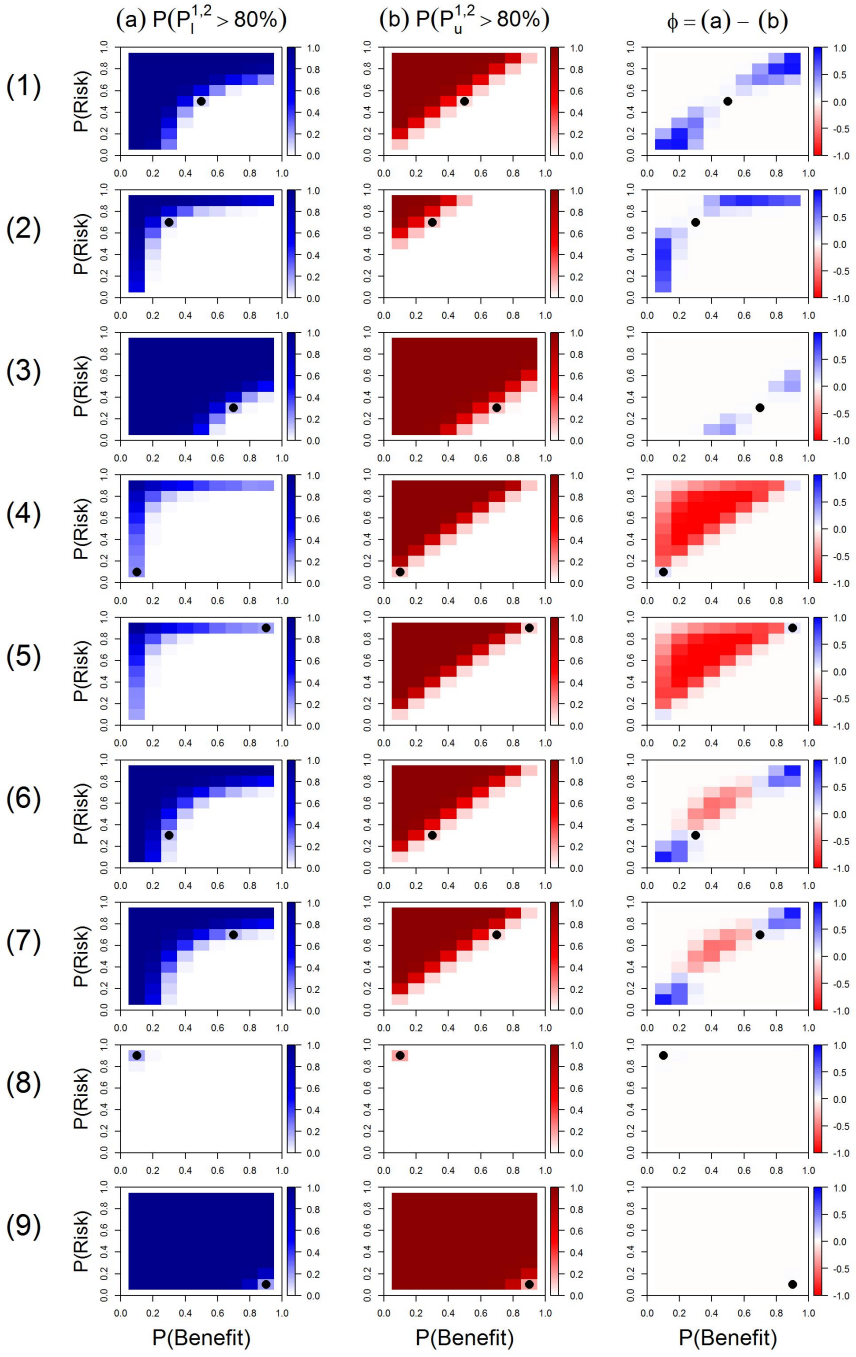

FIGURE A.2: Results of MCDA and SLoS performances in all simulation scenarios for two correlated equally important criteria ($w_j = \tilde{w}_j = 0.5$ for $j = 1,2$) with $\rho = 0.8$ (positive correlation). • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

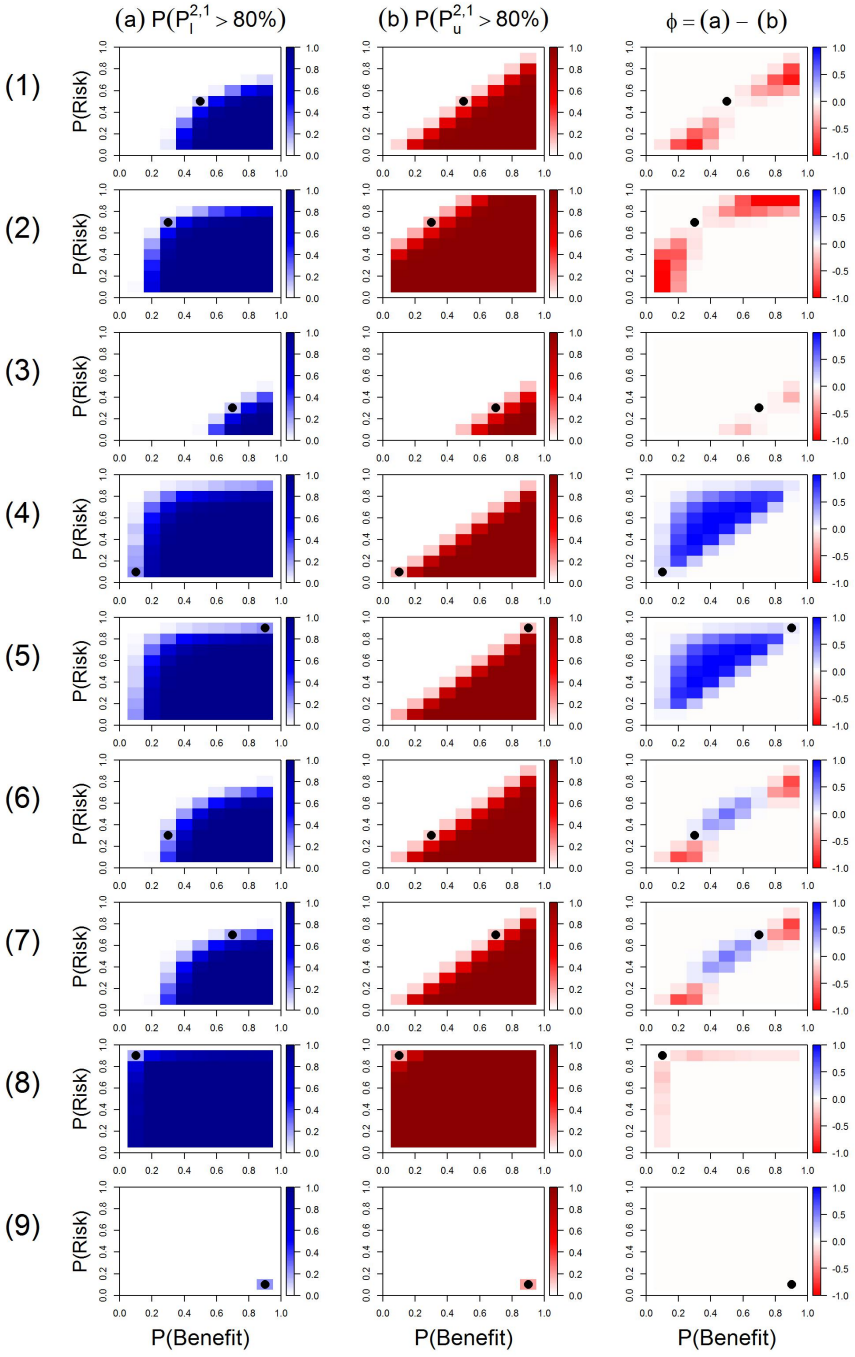

FIGURE A.3: Results of MCDA and SLoS performances in all simulation scenarios for two correlated equally important criteria ($w_j = \tilde{w}_j = 0.5$ for $j = 1, 2$) with $\rho = 0.8$ (positive correlation). • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

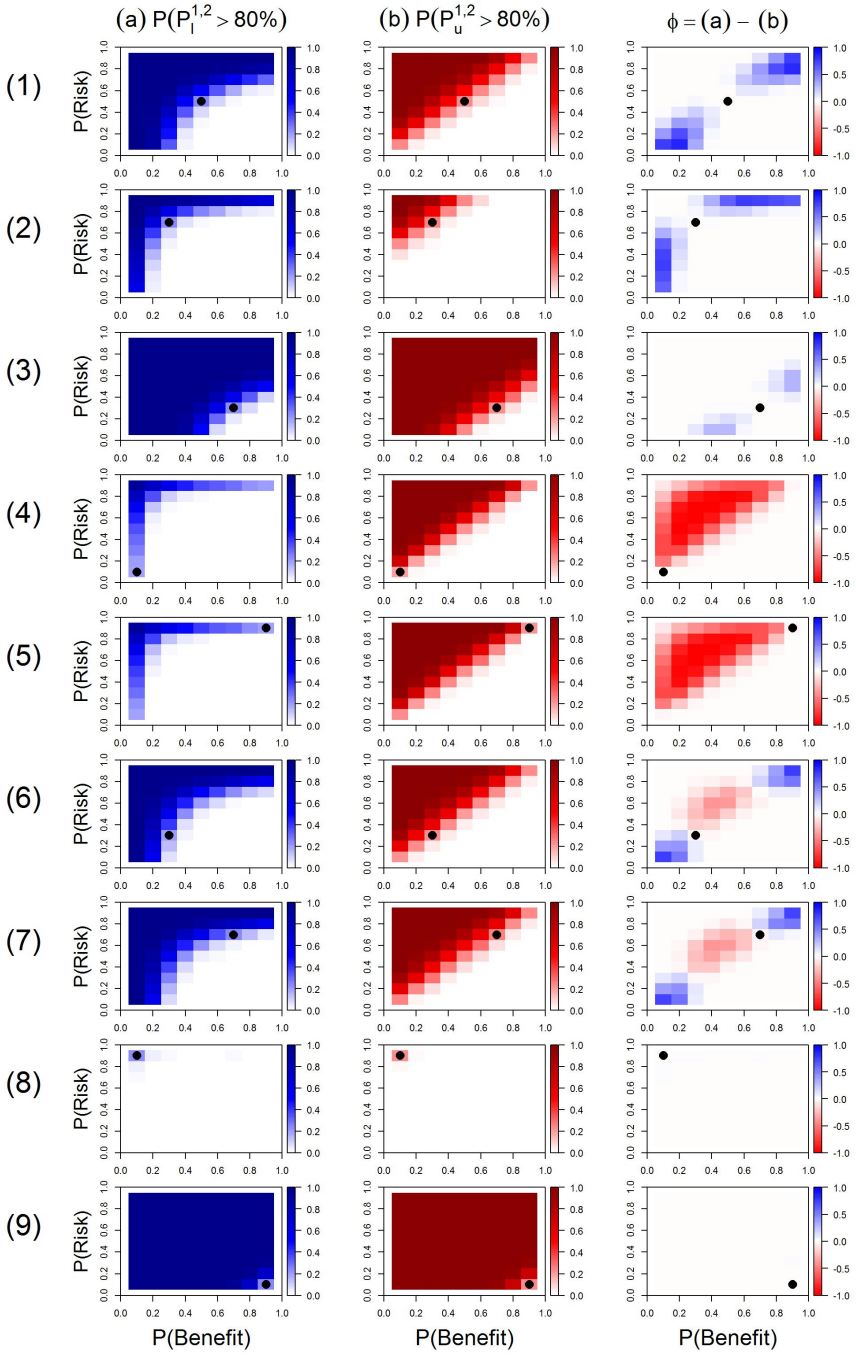

FIGURE A.4: Results of MCDA and SLoS performances in all simulation scenarios for two correlated equally important criteria ($w_j = \tilde{w}_j = 0.5$ for $j = 1, 2$) with $\rho = -0.8$ (negative correlation). • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

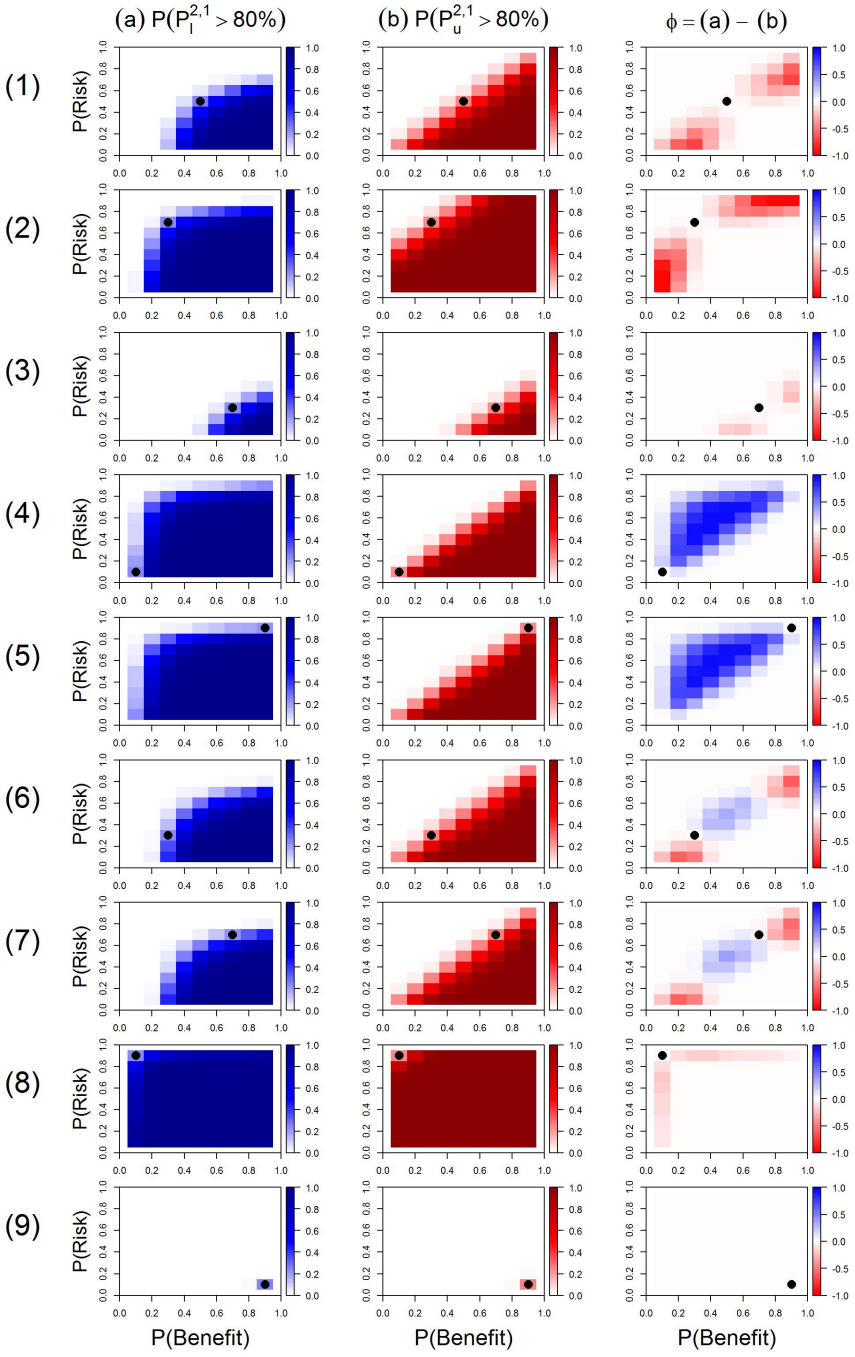

FIGURE A.5: Results of MCDA and SLoS performances in all simulation scenarios for two correlated equally important criteria ($w_j = \tilde{w}_j = 0.5$ for $j = 1, 2$) with $\rho = -0.8$ (negative correlation). • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

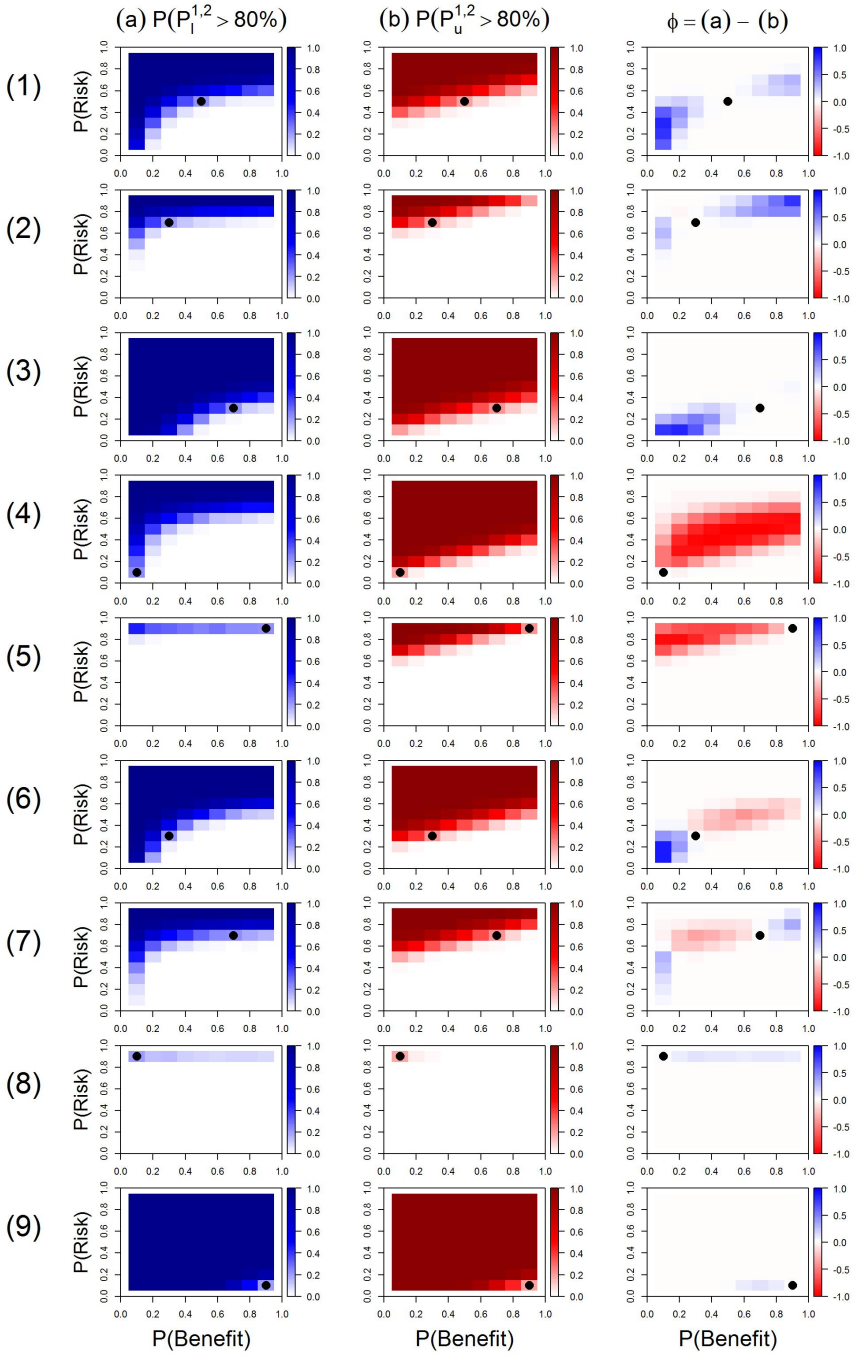

FIGURE A.6: Results of MCDA and SLoS performances in all simulation scenarios for two uncorrelated criteria, MCDA weights $(w_1, w_2) = (0.25, 0.75)$ and mapped SLoS weights $(\tilde{w}_1, \tilde{w}_2) = (0.30, 0.70)$. • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

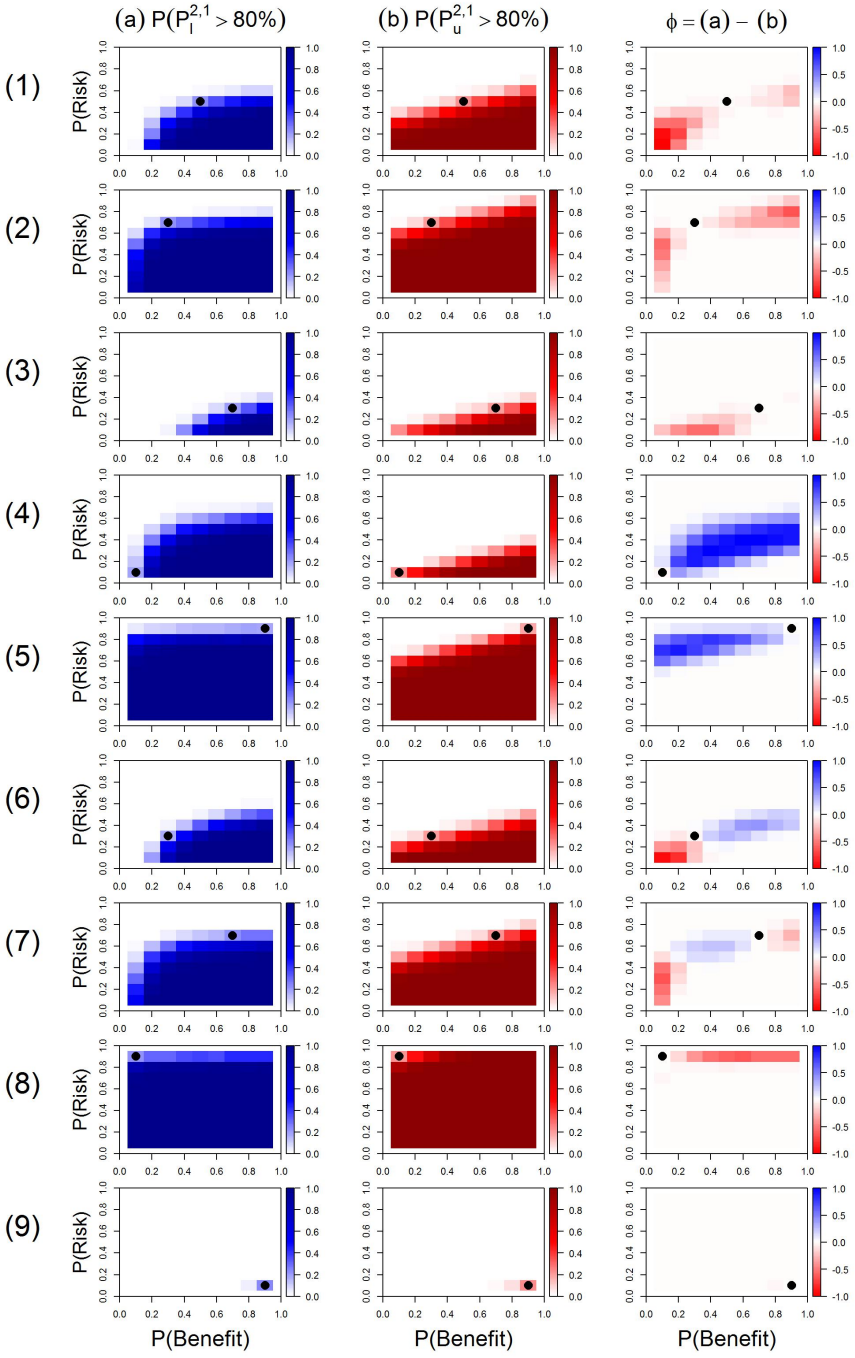

FIGURE A.7: Results of MCDA and SLoS performances in all simulation scenarios for two uncorrelated criteria, MCDA weights $(w_1, w_2) = (0.25, 0.75)$ and mapped SLoS weights $(\tilde{w}_1, \tilde{w}_2) = (0.30, 0.70)$. • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

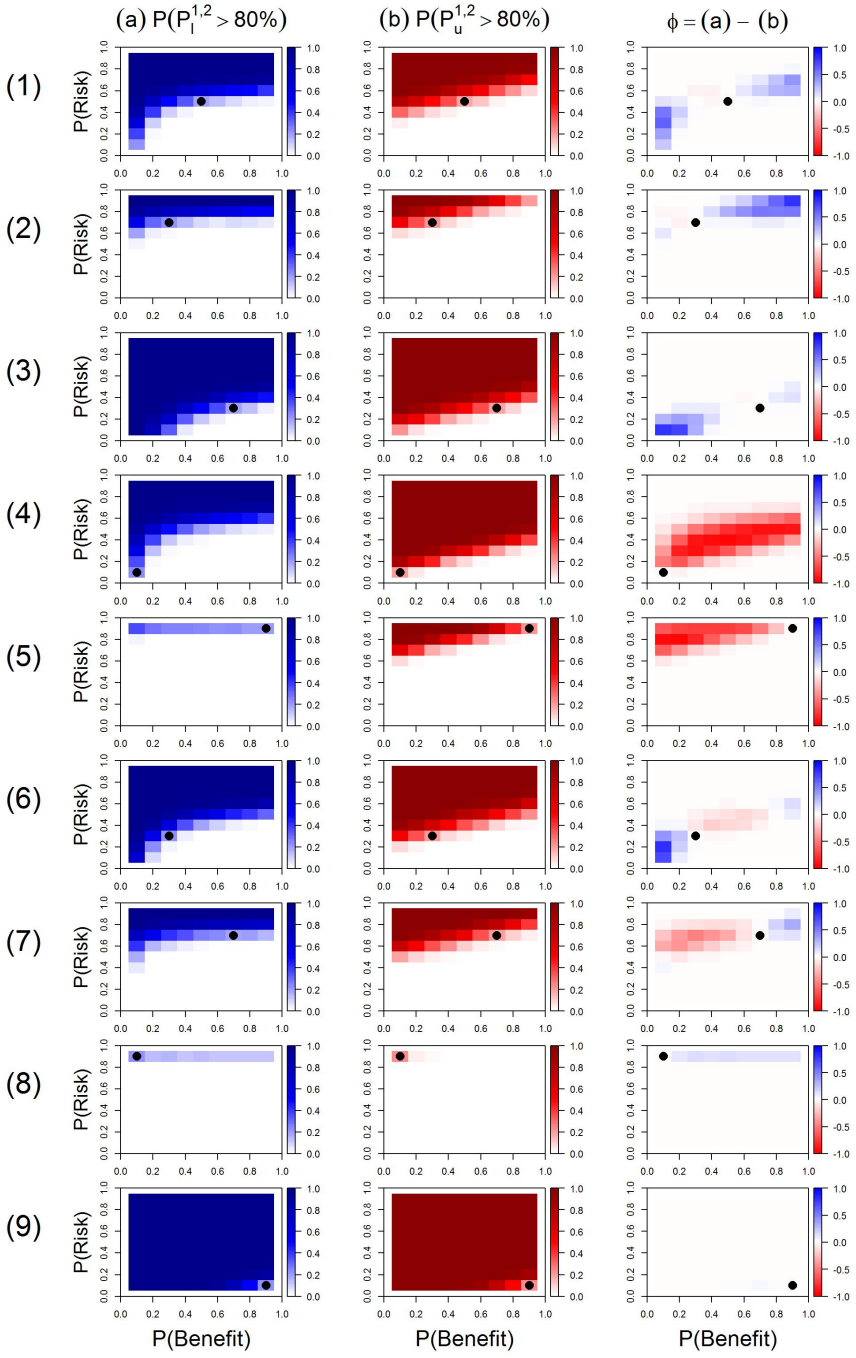

FIGURE A.8: Results of MCDA and SLoS performances in all simulation scenarios for two uncorrelated criteria, MCDA and SLoS weights $(w_1, w_2) = (\tilde{w}_1, \tilde{w}_2) = (0.25, 0.75)$ (no mapping). • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

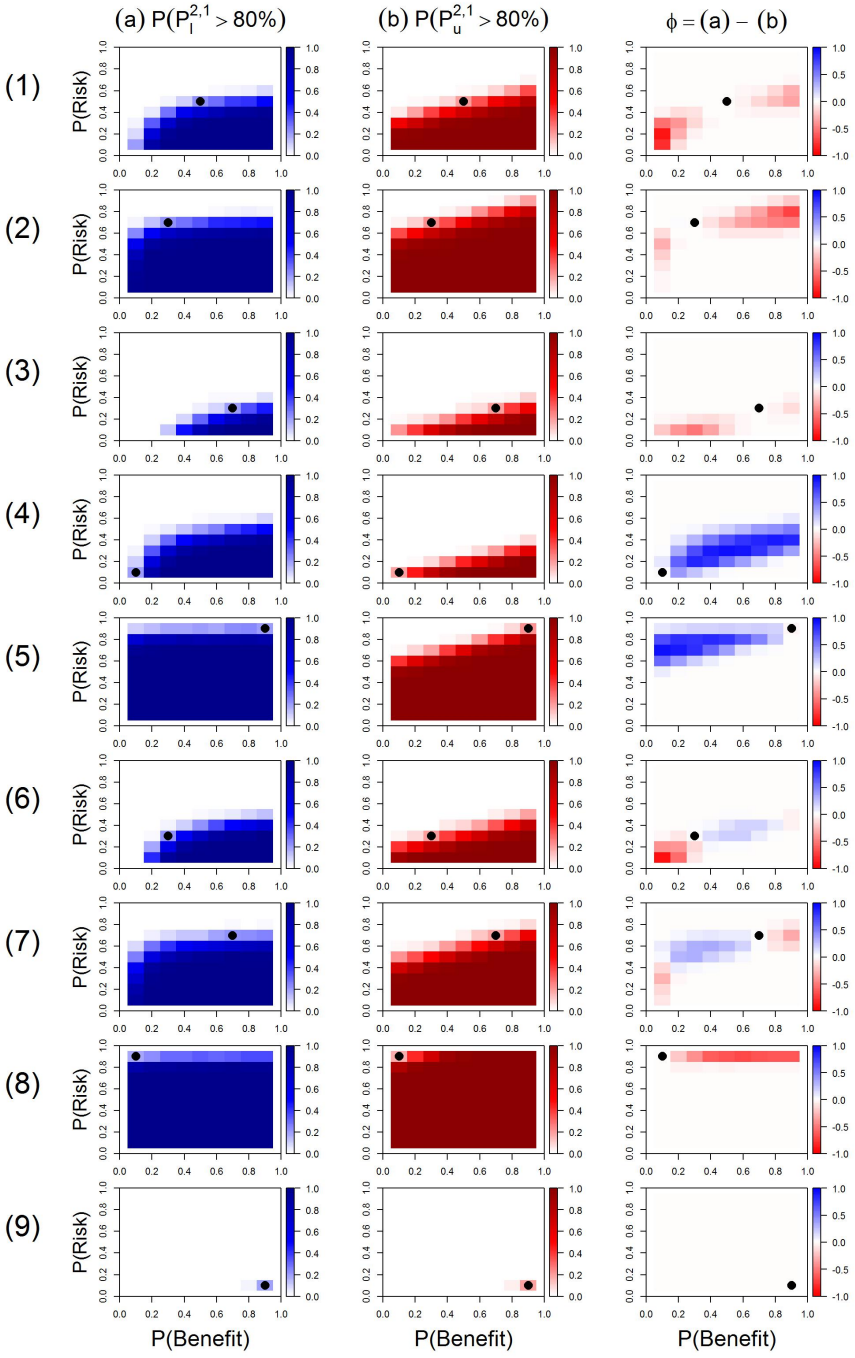

FIGURE A.9: Results of MCDA and SLoS performances in all simulation scenarios for two uncorrelated criteria, MCDA and SLoS weights $(w_1, w_2) = (\tilde{w}_1, \tilde{w}_2) = (0.25, 0.75)$ (no mapping). • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

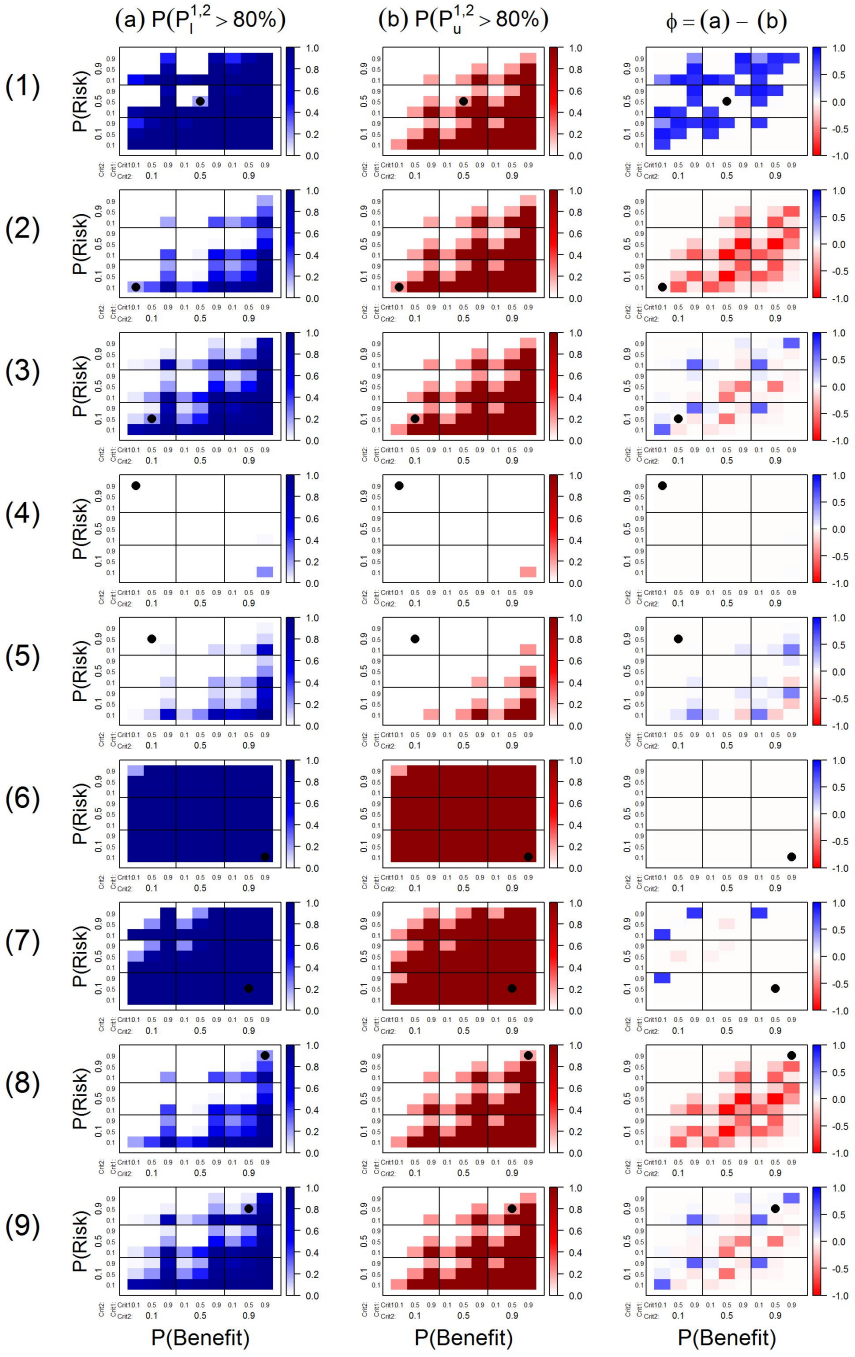

FIGURE A.10: Results of MCDA and SLoS performances in all simulation scenarios for four uncorrelated equally important criteria with weights $w_j = \tilde{w}_j = 0.25$ for $j = 1, ..., 4$. • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

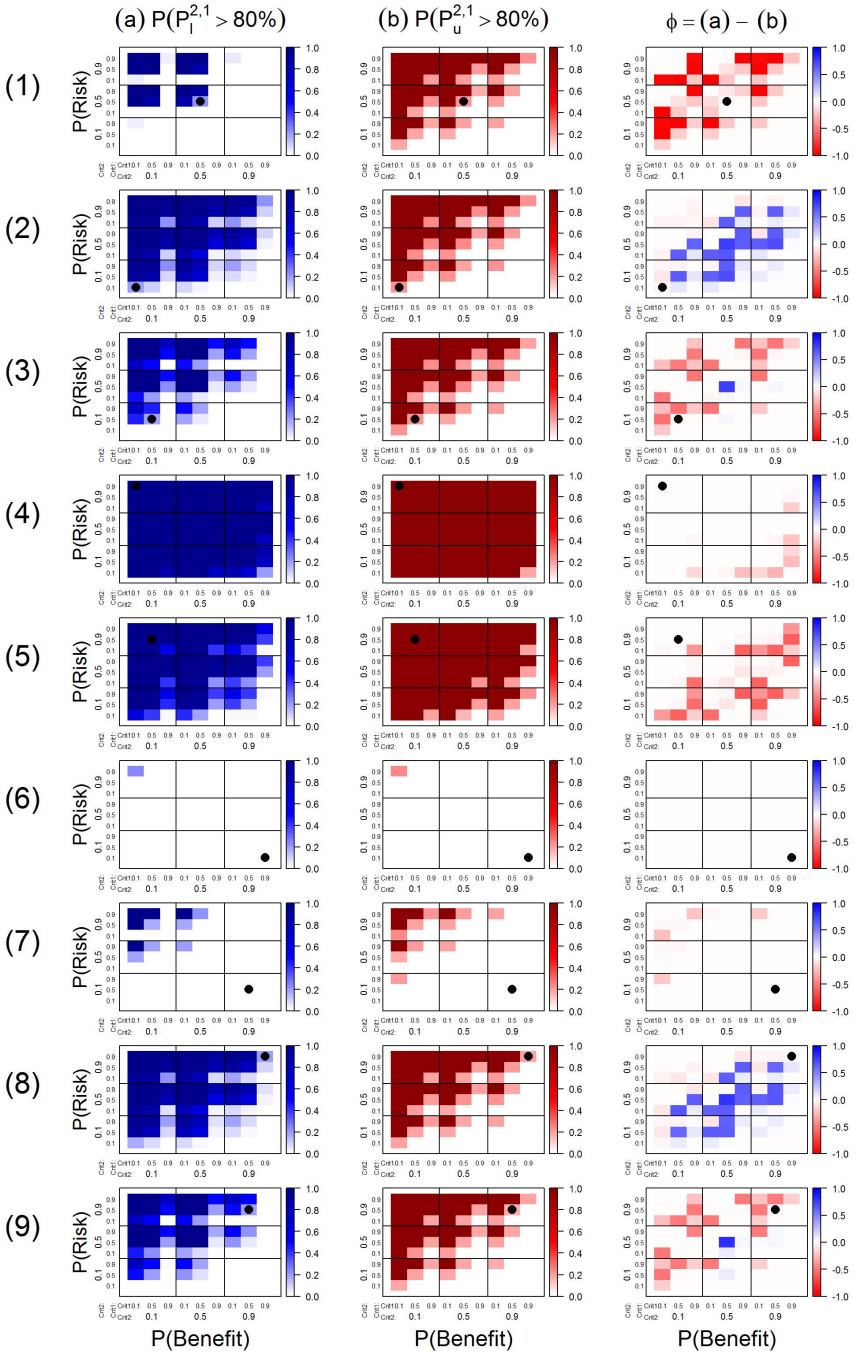

FIGURE A.11: Results of MCDA and SLoS performances in all simulation scenarios for four uncorrelated equally important criteria with weights $w_j = \tilde{w}_j = 0.25$ for $j = 1, ..., 4$. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

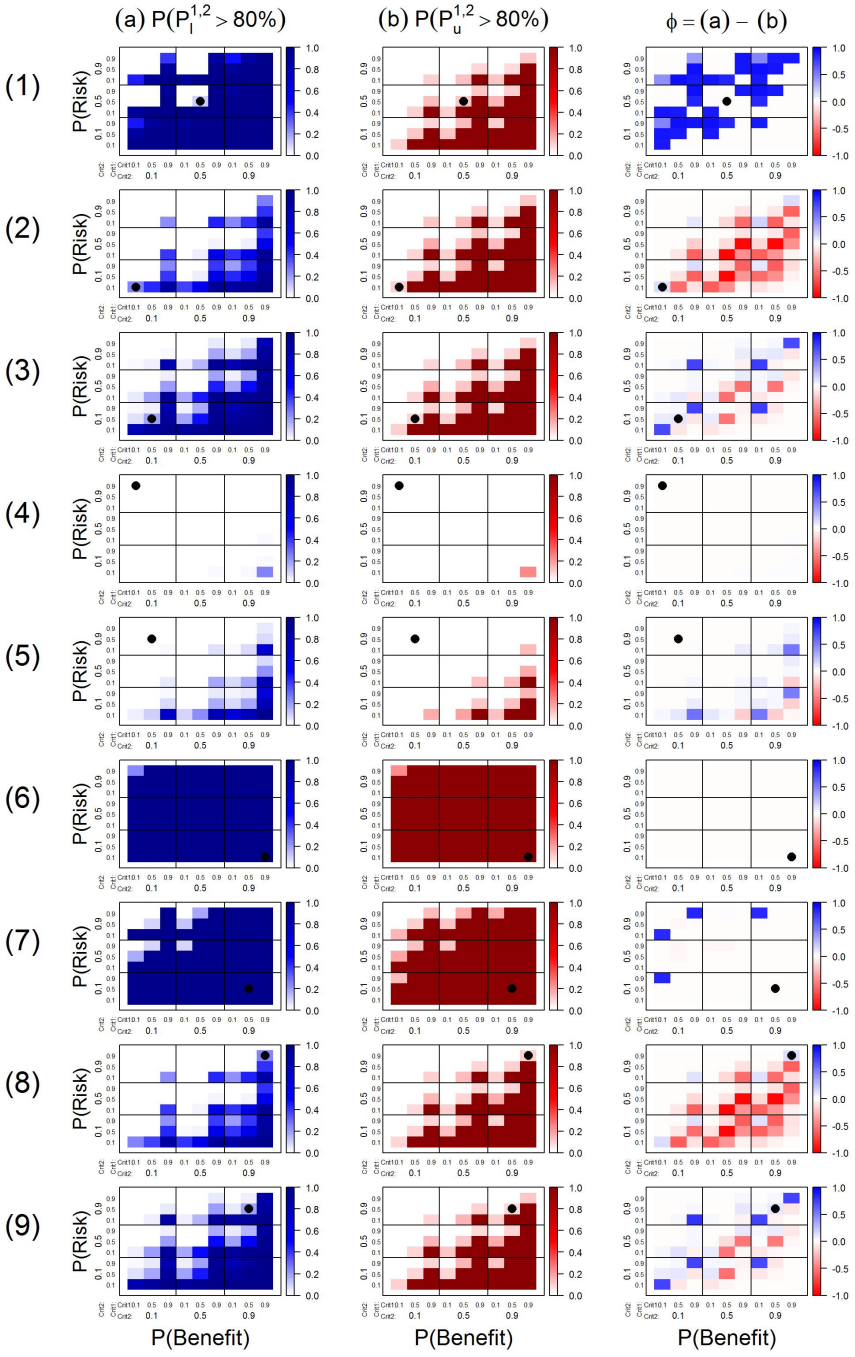

FIGURE A.12: Results of MCDA and SLoS performances in all simulation scenarios for four correlated equally important criteria with weights $w_j = \tilde{w}_j = 0.25$ for $j = 1, ..., 4$, positive correlations. • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

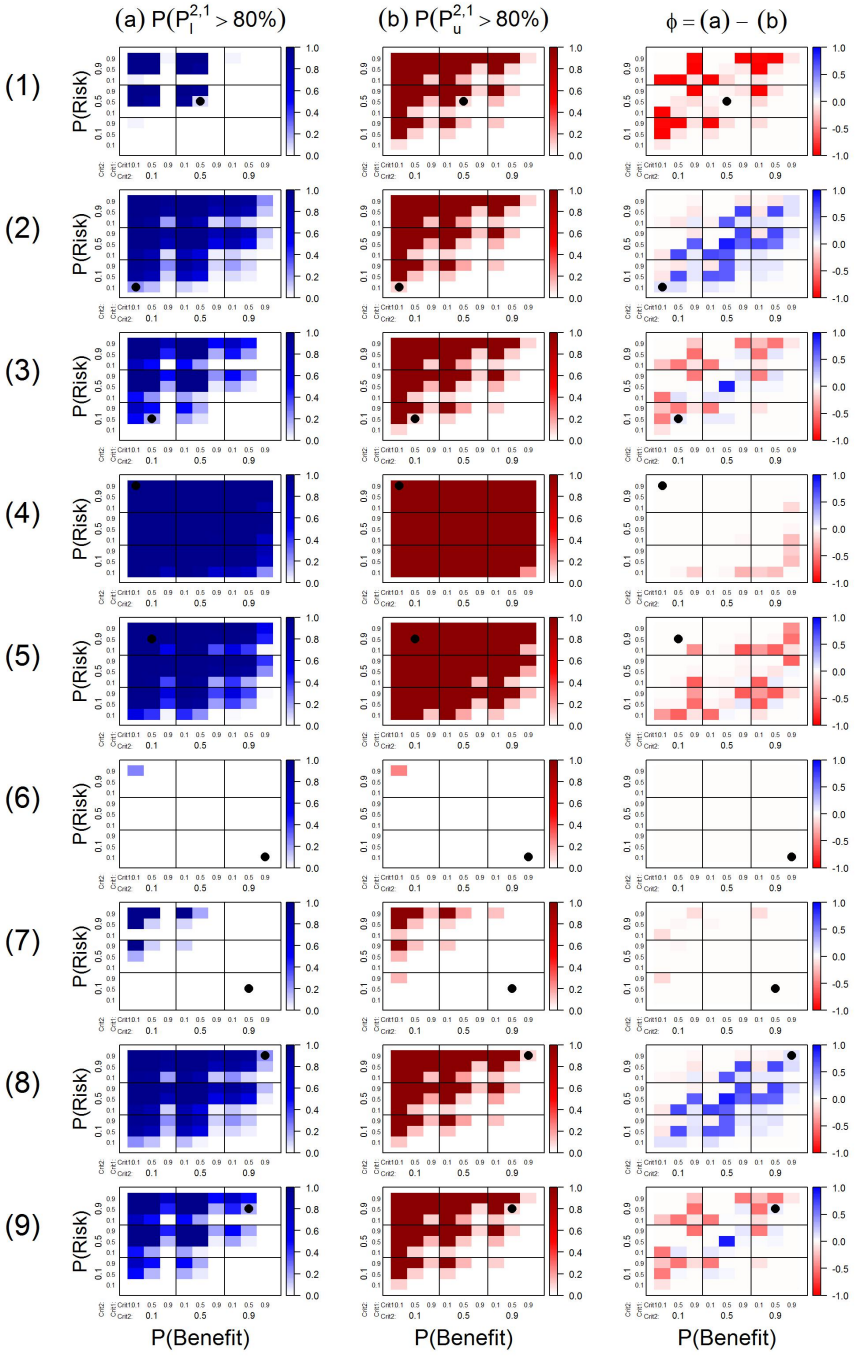

FIGURE A.13: Results of MCDA and SLoS performances in all simulation scenarios for four correlated equally important criteria with weights $w_j = \tilde{w}_j = 0.25$ for $j = 1, ..., 4$, positive correlations. • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

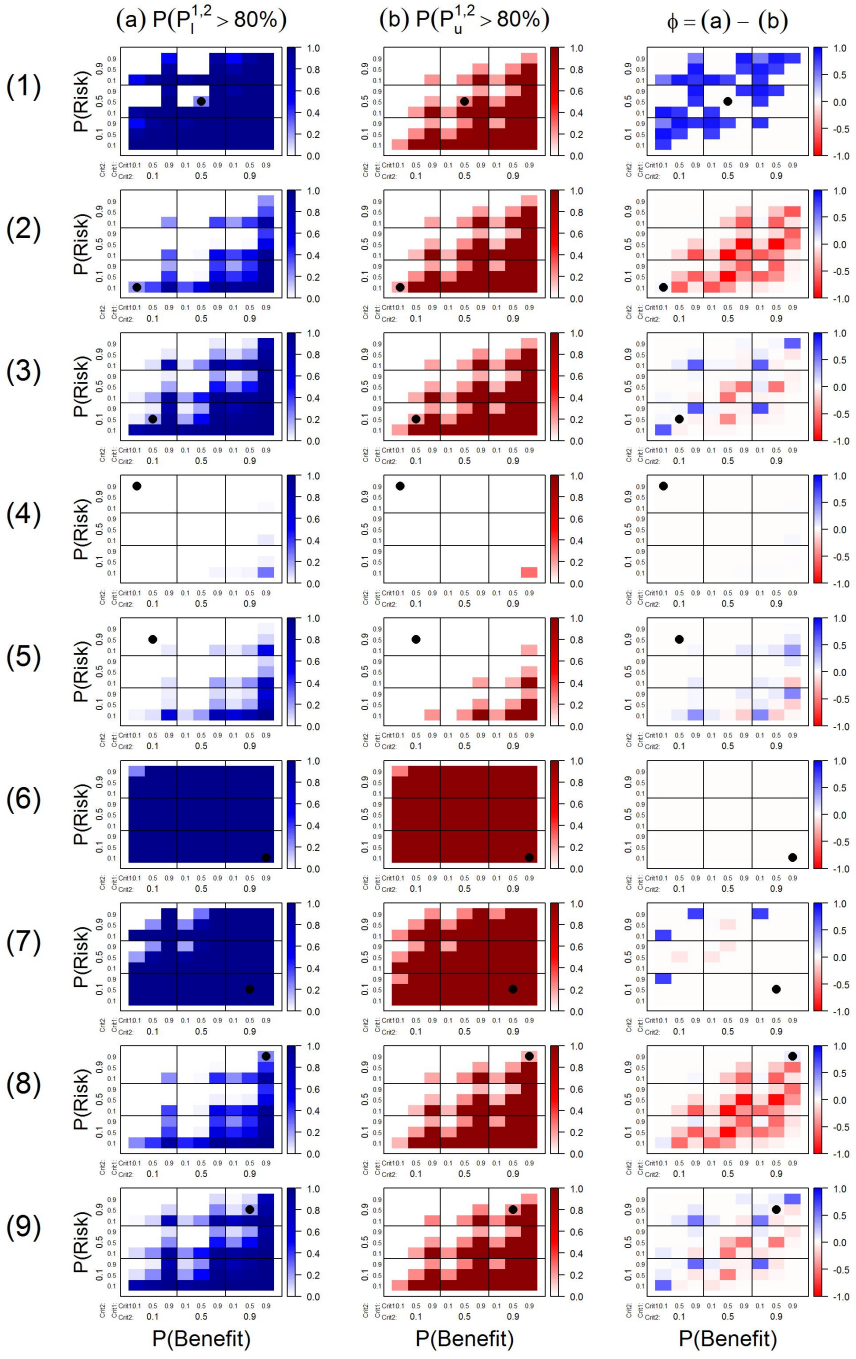

FIGURE A.14: Results of MCDA and SLoS performances in all simulation scenarios for four correlated equally important criteria with weights $w_j = \tilde{w}_j = 0.25$ for $j = 1, ..., 4$, positive correlations between criteria except for criterion $j = 2$, negatively correlated with the others. • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

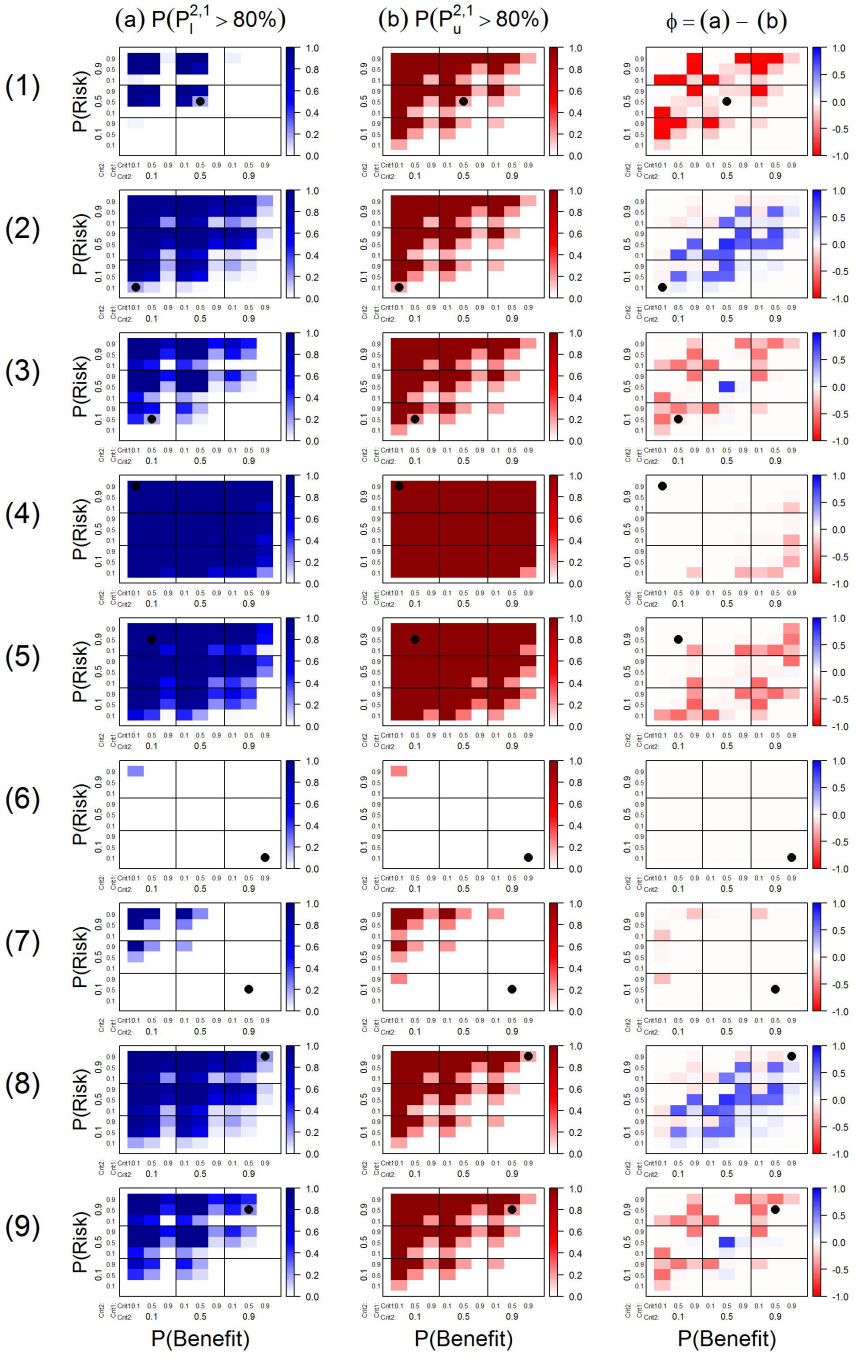

FIGURE A.15: Results of MCDA and SLoS performances in all simulation scenarios for four correlated equally important criteria with weights $w_j = \tilde{w}_j = 0.25$ for $j = 1, ..., 4$, positive correlations between criteria except for criterion $j = 2$, negatively correlated with the others. • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

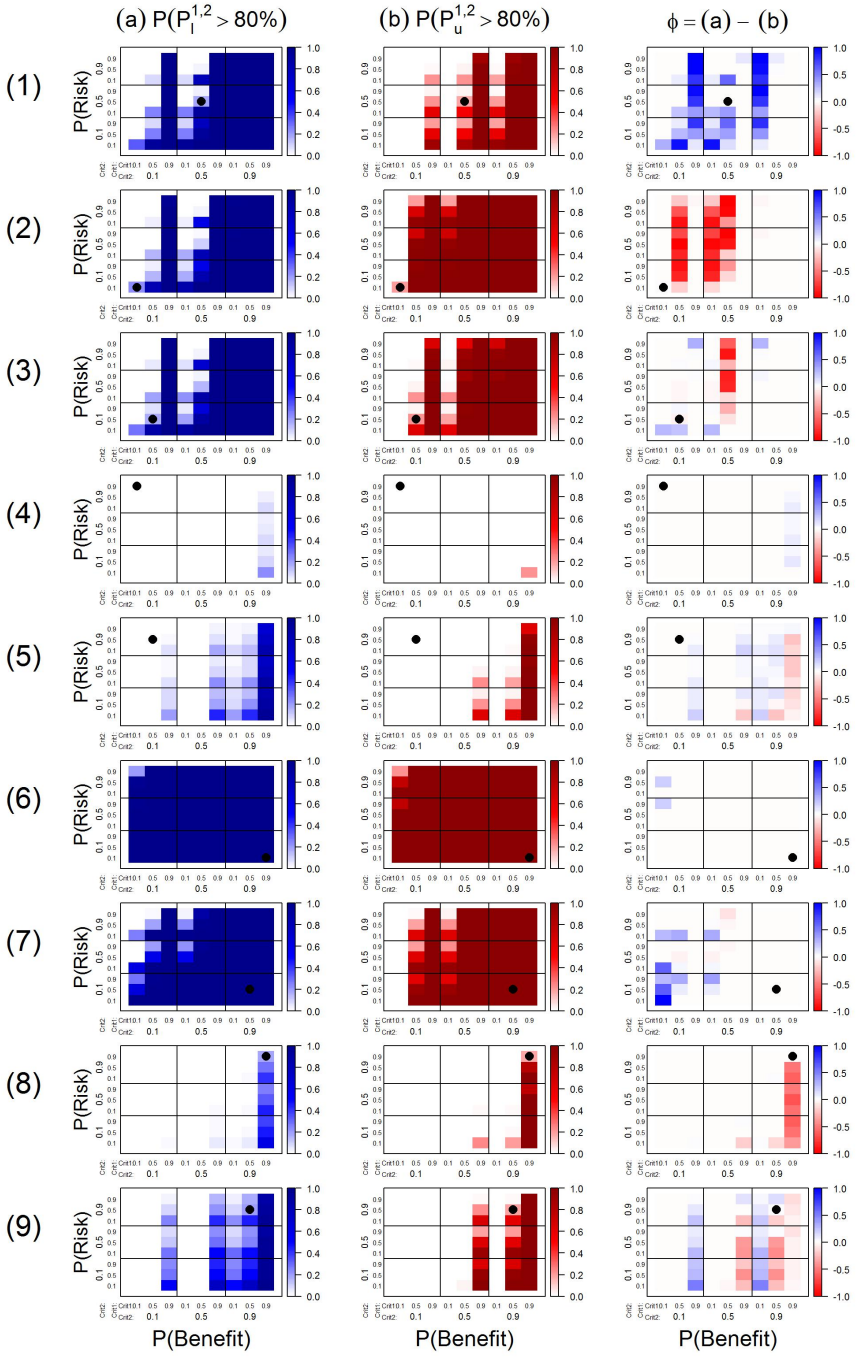

FIGURE A.16: Results of MCDA and SLoS performances in all simulation scenarios for four uncorrelated criteria, with MCDA weights $(w_1, w_2, w_3, w_4) = (0.10, 0.10, 0.40, 0.40)$ and mapped SLoS weights $(\tilde{w}_1, \tilde{w}_2, \tilde{w}_3, \tilde{w}_4) = (0.15, 0.15, 0.43, 0.43)$. • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

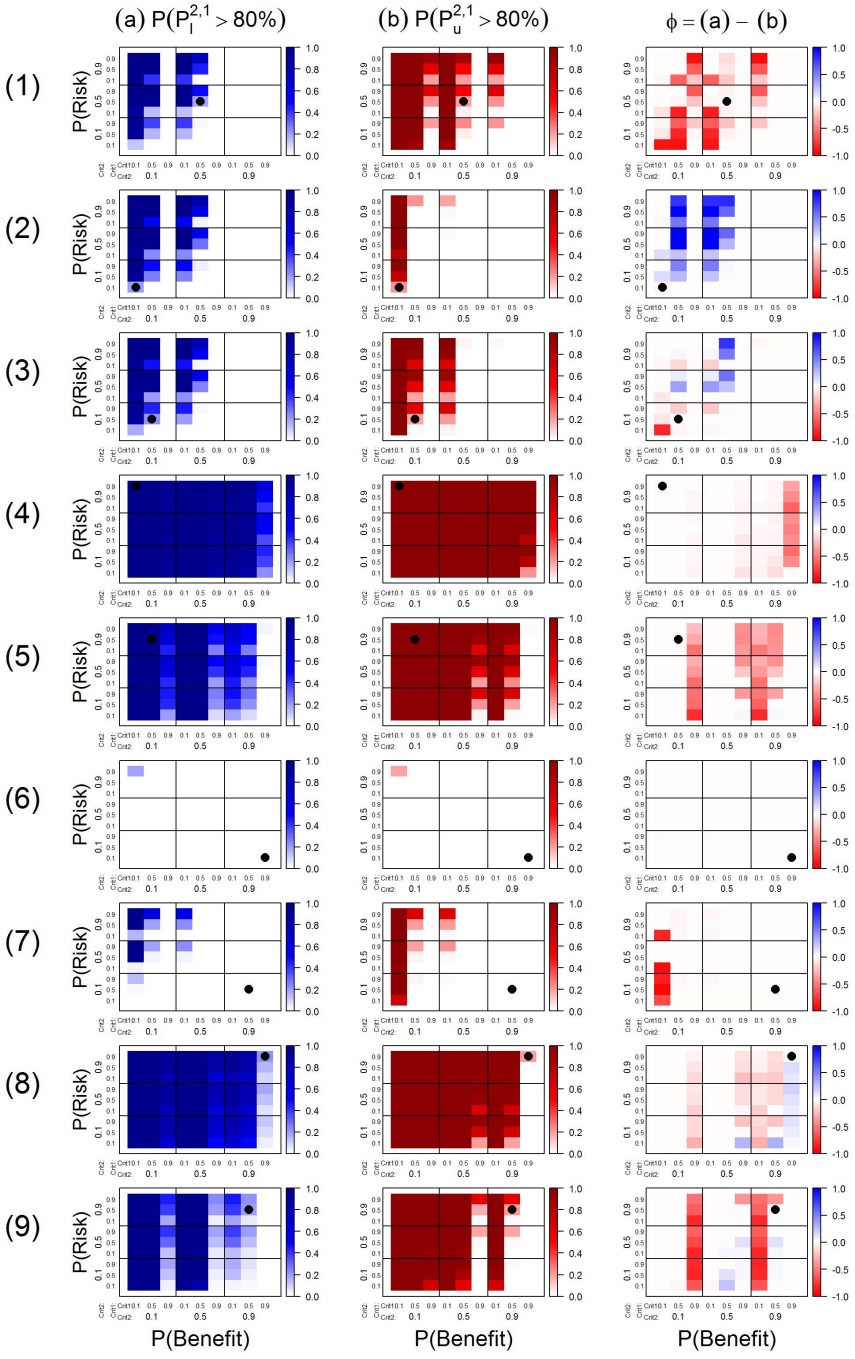

FIGURE A.17: Results of MCDA and SLoS performances in all simulation scenarios for four uncorrelated criteria, with MCDA weights $(w_1, w_2, w_3, w_4) = (0.10, 0.10, 0.40, 0.40)$ and mapped SLoS weights $(\tilde{w}_1, \tilde{w}_2, \tilde{w}_3, \tilde{w}_4) = (0.15, 0.15, 0.43, 0.43)$. • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

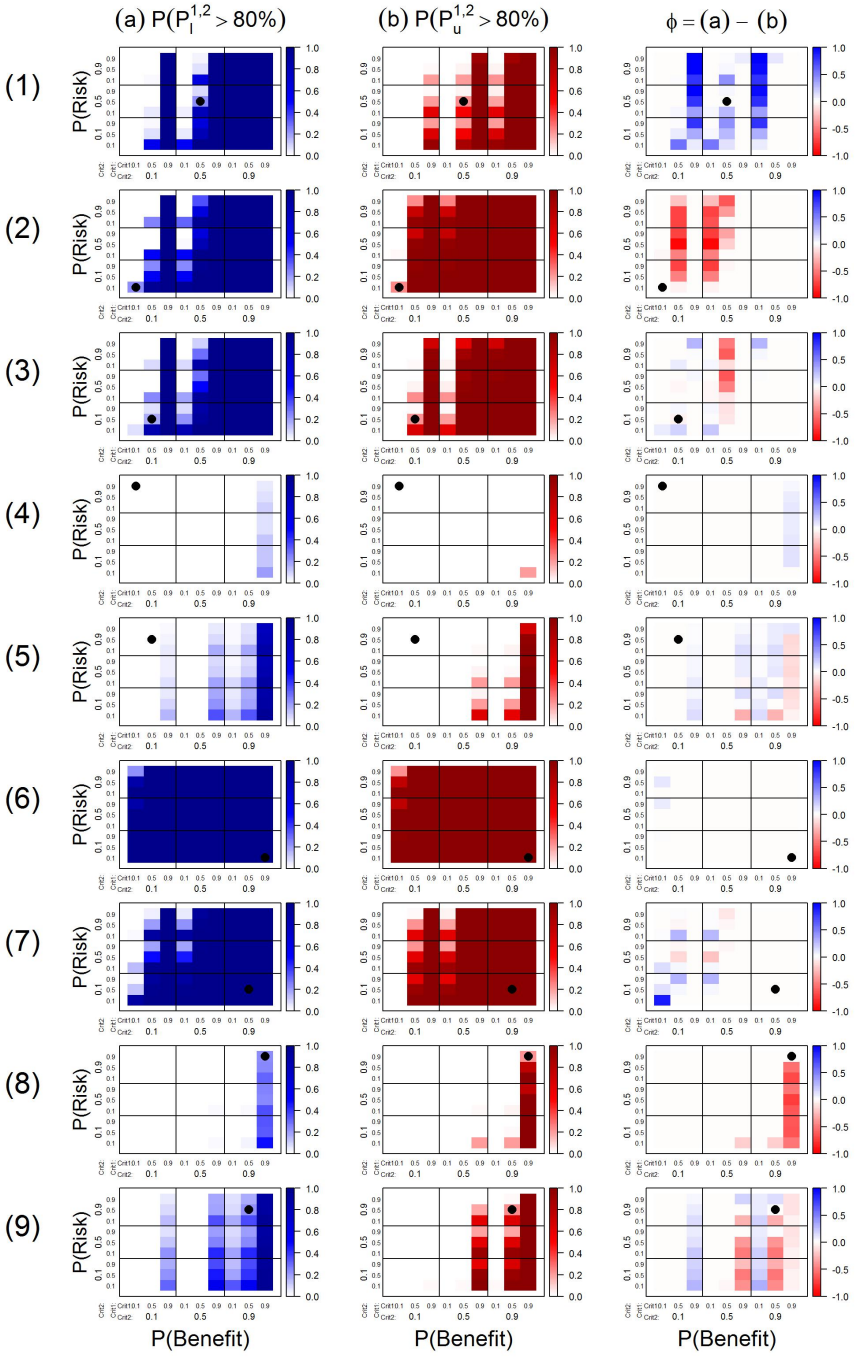

FIGURE A.18: Results of MCDA and SLoS performances in all simulation scenarios for four uncorrelated criteria, with MCDA weights $(w_1, w_2, w_3, w_4) = (\tilde{w}_1, \tilde{w}_2, \tilde{w}_3, \tilde{w}_4) = (0.10, 0.10, 0.40, 0.40)$ (no mapping). • = T_1. Left panel: $P[\mathcal{P}_d^{1,2} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{1,2} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{1,2} > 0.8] - P[\mathcal{P}_u^{1,2} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_1 more often (resp., less often) than MCDA.

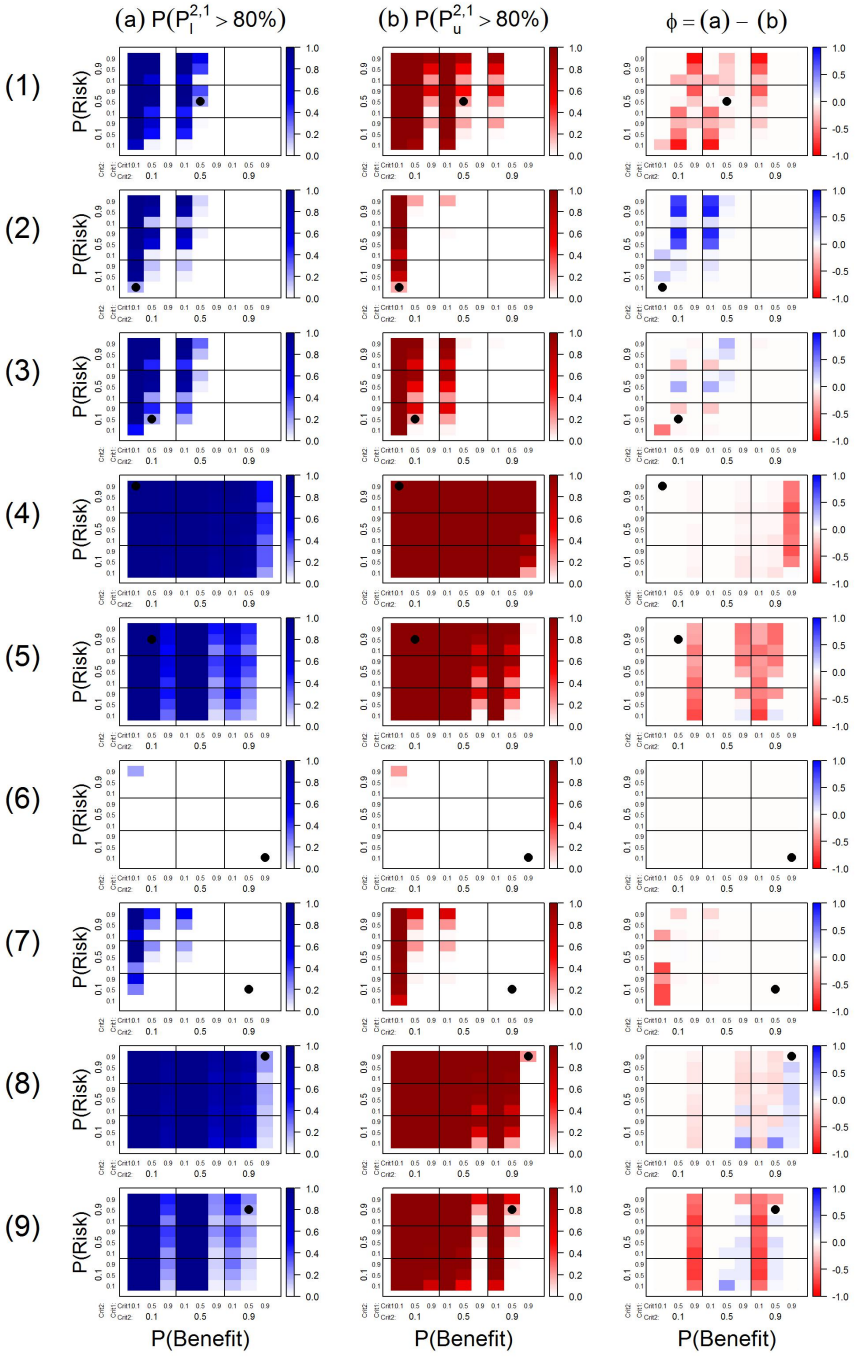

FIGURE A.19: Results of MCDA and SLoS performances in all simulation scenarios for four uncorrelated criteria, with MCDA weights $(w_1, w_2, w_3, w_4) = (\tilde{w}_1, \tilde{w}_2, \tilde{w}_3, \tilde{w}_4) = (0.10, 0.10, 0.40, 0.40)$ (no mapping). • = T_1. Left panel: $P[\mathcal{P}_d^{2,1} > 0.8]$. Middle panel: $P[\mathcal{P}_u^{2,1} > 0.8]$. Right panel: $\phi = P[\mathcal{P}_d^{2,1} > 0.8] - P[\mathcal{P}_u^{2,1} > 0.8]$, for which blue cells (resp., red cells) indicate that SLoS recommends T_2 more often (resp., less often) than MCDA.

Appendix B

Decision-making using a composite definition of success: supplemental material

B.1 Example in Major Depressive Disorder: sensitivity analyses

This section presents the analyses conducted to assess the robustness of the results presented in Section 3.

B.1.1 Weight elicitation

The impact of the the weight elicitation on the benefit-risk assessment is evaluated using a Dirichlet Stochastic Multicriteria Acceptability Analysis (Dirichlet SMAA) model (Supplementary Material, [112]). In this model, the weights are treated as random variables following a Dirichlet distribution:

$$(w_1, ..., w_n) \sim Dirichlet\left(c \times \left(w_1^0, ..., w_n^0\right)\right)\left((w_1, ..., w_n) \in W\right)$$

with

(i) $(w_1^0, ..., w_n^0) = (0.5, 0.2, 0.075, 0.075, 0.075, 0.075)$,

(ii) c the confidence factor reflecting the decision-makers' confidence in their elicitation, varying from 1 to 10^5. The variances of the w_i are inversely proportional to c, and the deterministic weights used in the main MCDA model correspond to an infinite confidence factor $c = +\infty$.

The predictive probabilities of composite success for each regimen according to the confidence factor c are presented in Figure S1. When the decision-makers are much uncertain in their weight elicitation ($1 \leq c \leq 5$), the predictive probabilities of composite success of the Low dose, the High dose and the dose increase are very similar (around 40%). They separate and tend to the results of the main analysis when the decision-makers are more confident in their elicitation. On the other hand, the predictive probability of composite success of the High dose with supplementation ranges between 58% and 78% and always remains above those of the other regimen, confirming that it seems to be the best strategy for further development.

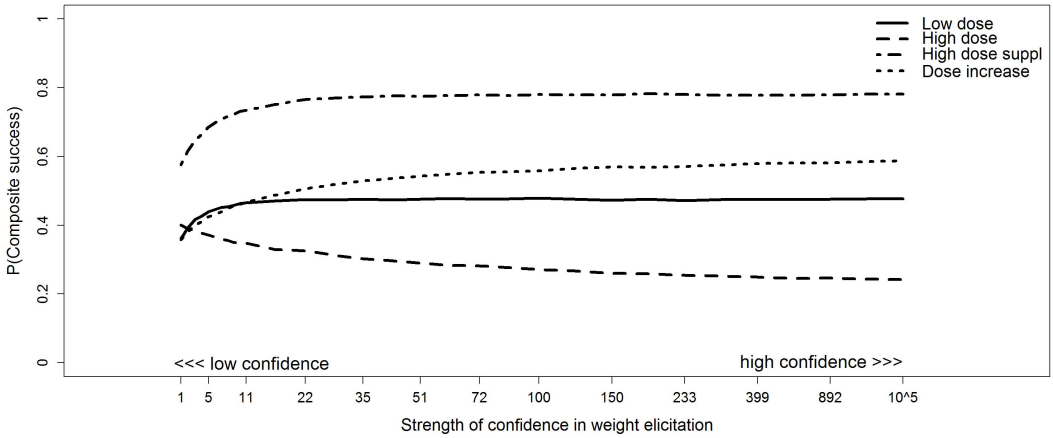

FIGURE B.1: Predictive probabilities of composite success when the benefit-risk assessment is performed using a Dirichlet SMAA model with c ranging from 1 to 10^5.

B.1.2 Correlations between criteria

While the case of uncorrelated criteria is considered in the main MCDA model for benefit-risk assessment, we investigated the robustness of the results in case of correlated criteria:

- Positive correlations between all criteria, with correlation matrix:

$$\Omega = \begin{bmatrix} 1 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 1 & 0.8 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 1 & 0.8 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 1 \end{bmatrix}$$

- Benefit criterion negatively correlated with the risk criteria, and positive correlations between risk criteria, with correlation matrix:

$$\Omega = \begin{bmatrix} 1 & -0.8 & -0.8 & -0.8 & -0.8 & -0.8 \\ -0.8 & 1 & 0.8 & 0.8 & 0.8 & 0.8 \\ -0.8 & 0.8 & 1 & 0.8 & 0.8 & 0.8 \\ -0.8 & 0.8 & 0.8 & 1 & 0.8 & 0.8 \\ -0.8 & 0.8 & 0.8 & 0.8 & 1 & 0.8 \\ -0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 1 \end{bmatrix}$$

The predictive distributions are presented in Figures S2 and S3, and the corresponding predictive probabilities of success in Tables B.1 and B.2.
The predictive probabilities of success $PPoS_1$ and $PPoS_2$ are unchanged since they are based on the benefit criterion only. The correlations only have an impact on the differences in benefit-risk utility scores: because the MCDA model is based on an additive formula, the correlations do not impact the mean of their predictive distributions, but they impact their variance. They are less precise when all criteria

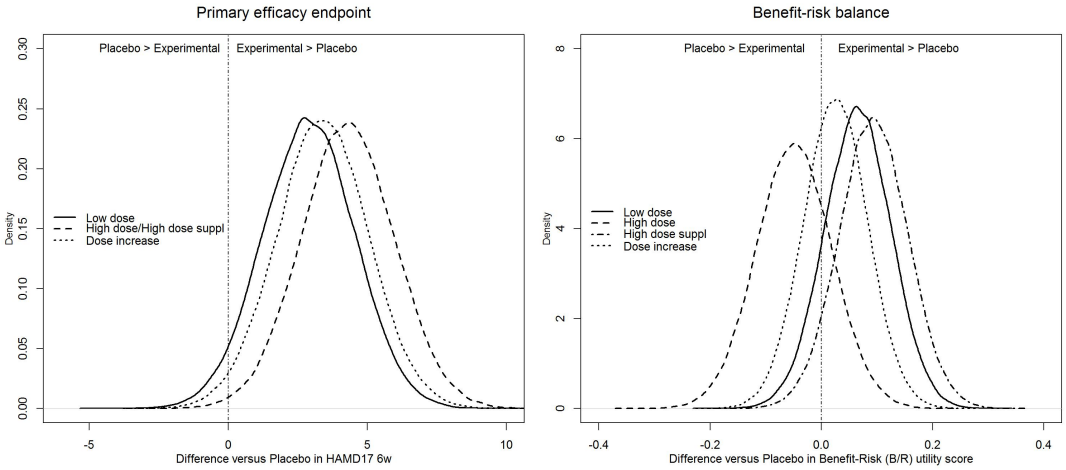

FIGURE B.2: Left: predictive distributions of the differences in HAM-D_{17} mean total score of each regimen versus placebo in the next Phase III study. Right: predictive distributions of the differences in benefit-risk utility scores of each dose versus placebo in the next Phase III study, when all criteria are positively correlated.

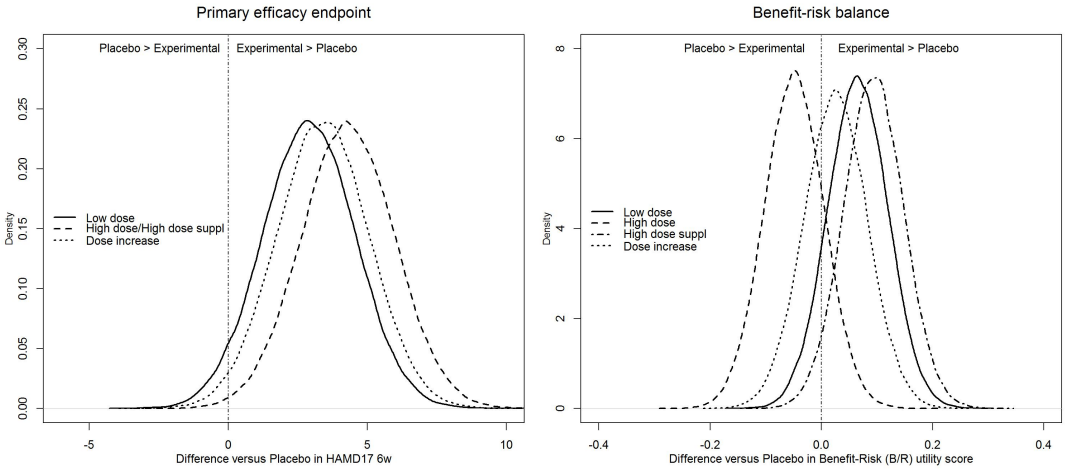

FIGURE B.3: Left: predictive distributions of the differences in HAM-D_{17} mean total score of each regimen versus placebo in the next Phase III study. Right: predictive distributions of the differences in benefit-risk utility scores of each dose versus placebo in the next Phase III study, when the benefit criterion is negatively correlated with the risk criteria, and the risk criteria are positively correlated between themselves.

are positively correlated, and (here) more precise when the benefit criterion is negatively correlated with the risk criteria while risk criteria are correlated between themselves. However, the corresponding predictive probabilities $PPoS_3$ are only slightly impacted and the conclusions are unchanged.

B.1.3 Clinical assumptions for the strategy refinement

Since no data were available on the strategy refinement regimen, their success is predicted using together previous evidence on other regimen and clinical assumptions,

TABLE B.1: Predictive probabilities of success when all criteria are positively correlated

Dose	$PPoS_1$ (statistical significance)	$PPoS_2$ (clinical relevance)	$PPoS_3$ (positive B/R balance)	$PPoS$ (overall)
Low dose	74%	48%	86%	48%
High dose	93%	78%	23%	23%
High dose suppl	93%	78%	94%	78%
Dose increase	83%	59%	67%	58%

TABLE B.2: Predictive probabilities of success when the benefit criterion is negatively correlated with the risk criteria, and the risk criteria are positively correlated between themselves

Dose	$PPoS_1$ (statistical significance)	$PPoS_2$ (clinical relevance)	$PPoS_3$ (positive B/R balance)	$PPoS$ (overall)
Low dose	74%	48%	89%	48%
High dose	93%	78%	18%	18%
High dose suppl	93%	78%	96%	78%
Dose increase	83%	59%	67%	58%

which are translated into priors on some parameters (probability of Hypokalemia, proportion of patients receiving the High dose), based on clinical assumptions (Section 3.4). The following sensitivity analyses were conducted:

- **High dose with potassium supplements.** In the main analysis, the probability of experiencing a Hypokalemia ξ_{42} with the High dose with potassium supplements is assumed to be the same as with the placebo. This may be optimistic since it supposes that the patients are compliant with their supplementation in potassium, and that this one is perfectly adapted to the patient's diet, which may be difficult to achieve in practice. A sensitivity analysis is conducted with the probability of Hypokalemia ξ_{42} varying from 0 to 0.71 (0.71 being the posterior mean for the High dose without supplementation).
- **Dose increase.** According to the clinicians, the proportion of patients increasing to High dose in the Phase III study ζ would vary from 30% to 40%. The reliability of this assumption was assessed in a sensitivity analysis where the mean proportion of patients receiving the High dose z varies from 0.05 to 0.95, with $\zeta \sim U[z - 0.05, z + 0.05]$.

The predictive probabilities of composite success for each regimen according to the varying parameters are presented in Figures S4 and S5.

The probability of composite success of the High dose with potassium supplements remains stable (between 73% and 78%) for probabilities of Hypokalemia lower than 0.3, and then it decreases until reaching the $PPoS$ of the High dose without supplements. It is the regimen with the greatest $PPoS$ for probabilities of Hypokalemia lower than 0.43. Since the potassium supplements are initially expected to prevent the occurrence of Hypokalemia, we conclude that the results are robust even in case

of substantial departure from the clinical assumptions.

The probability of composite success of the dose increase regimen reaches its maximum when the mean proportion of patients receiving the High dose is equal to 38% (close to the clinical assumption 35%), and tends to the $PPoS$ of the High dose when the mean proportion increases. It remains between the $PPoS$ of the Low dose and of the High dose with potassium supplements when the mean proportion of patients receiving the High dose varies between 0% and 64%, and never exceeds the $PPoS$ of the High dose with potassium supplements.

Overall, we conclude that the results are robust to changes in the clinical assumptions.

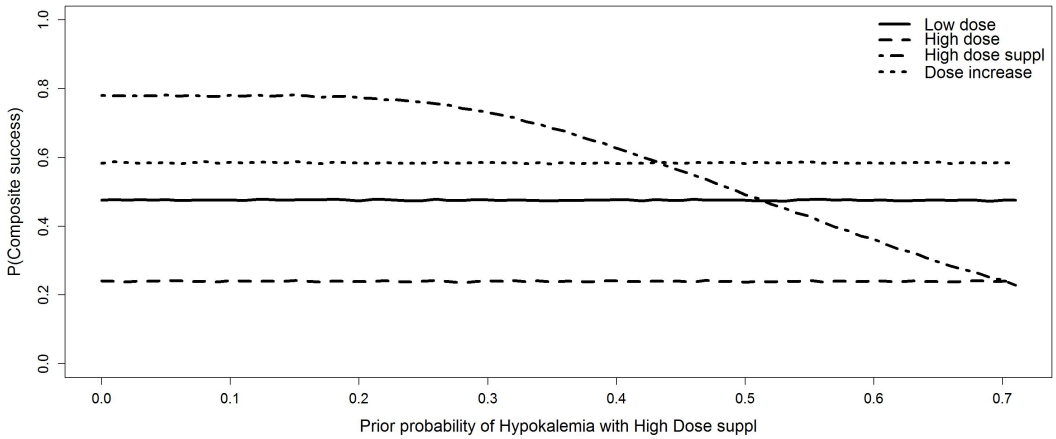

FIGURE B.4: Predictive probabilities of composite success when the probability of Hypokalemia varies from 0 to 0.71 for the High dose with potassium supplements.

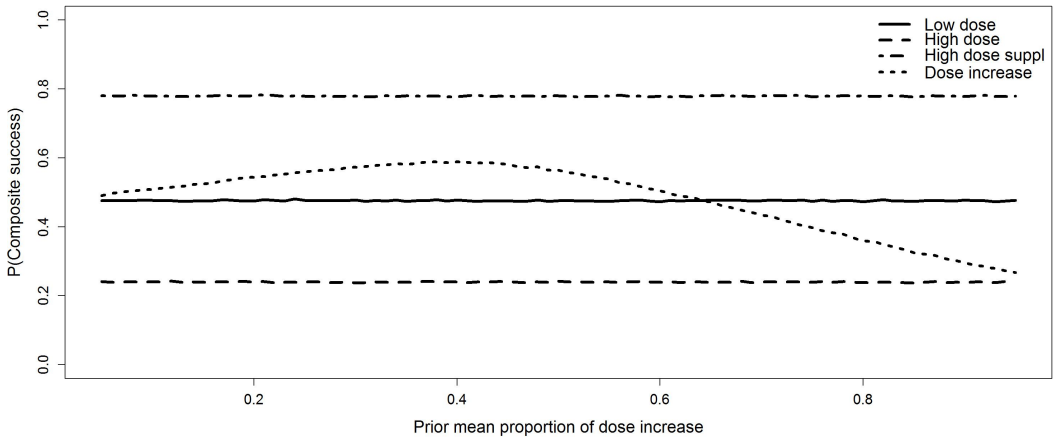

FIGURE B.5: Predictive probabilities of composite success when the mean proportion of patients receiving the High dose varies from 0.05 to 0.95.

B.2 Alternative example in Major Depressive Disorder

B.2.1 Context and data

Let us consider the example in Major Depressive Disorder presented in Section 3, with the two following changes:

- The threshold of minimal clinical relevance is fixed at $d_T = 2$ points. This could be relevant if, for example, the drug is an add-on therapy to be administered on top of a standard therapy, and where the control arm of the trial is the standard therapy plus a placebo. In this case, the difference between the arms often does not need to be as large as for a monotherapy versus placebo alone, if the overall amount of efficacy and safety resulting from the experimental combination are large enough to justify its approval.
- In order to better investigate the performances of the regimens, the project team decides to include three experimental arms in the next Phase III trial, to be selected prior to the study among the four possible regimen (Low dose, High dose, High dose with potassium supplements and Dose increase). The objective is to detect a difference of at least one regimen versus the control arm on the primary endpoint, using a Bonferroni correction to maintain the familywise type I error rate at 2.5% (one-sided).

A sample size of 114 patients per arm (4 arms, 456 patients in total) is planned in the next Phase III to reach a power of 80%, based on an anticipated difference of 3 points (above the threshold of minimal clinical relevance here) on the HAM-D_{17} total score at 6 weeks, a standard-deviation of 7 and a one-sided α of $0.025/3$.

We assume that the same data were observed in Phase II trial, and the same model is used.

B.2.2 Results

The predictive distributions of the differences in HAM-D_{17} mean total score and of the differences in benefit-risk utility score of each regimen versus control in the next Phase III study are presented in Figure S6. The predictive probability of composite success of the development strategies, $PPoS$, along with the predictive probabilities of its components, $PPoS_1$, $PPoS_2$ and $PPoS_3$ are presented in Table B.3.

In this example, for all regimen, $PPoS_2$ is greater than $PPoS_1$ because the clinical relevance criterion ($d^* > d_T = 2$) is easier to reach than the statistical significance criterion ($d^* > c = 2.22$). The predictive probabilities of positive benefit-risk $PPoS_3$ are unchanged compared to the initial example. It follows that the High dose could be excluded from the selected regimen for Phase III, as in the initial example, due to its low probability to show a positive benefit-risk balance versus the control.

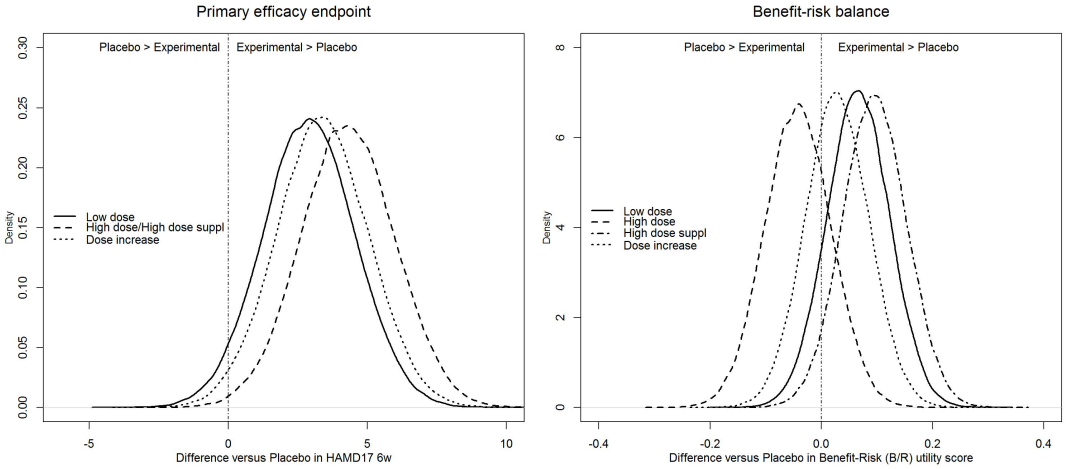

FIGURE B.6: Left: predictive distributions of the differences in HAM-D_{17} mean total score of each regimen versus placebo in the next Phase III study. Right: predictive distributions of the differences in benefit-risk utility scores of each dose versus placebo in the next Phase III study.

.

TABLE B.3: Predictive probabilities of success

Dose	$PPoS_1$ (statistical significance)	$PPoS_2$ (clinical relevance)	$PPoS_3$ (positive B/R balance)	$PPoS$ (overall)
Low dose	66%	71%	88%	66%
High dose	89%	91%	24%	24%
High dose suppl	89%	91%	95%	89%
Dose increase	76%	80%	69%	68%

Appendix C

Predictive probability of success using surrogate endpoints: supplemental material

C.1 Multi-arm trials

We need to account for correlations between treatment effect estimates coming from multi-arm trials. We provide here an example with three-arm trials, which could be simply extended to more arms. The within-trial model becomes:

$$\left.\begin{pmatrix} \hat{\theta}_{(1)i} \\ \hat{\theta}_{(2)i} \\ \hat{\gamma}_{(1)i} \\ \hat{\gamma}_{(2)i} \end{pmatrix}\right| \begin{pmatrix} \theta_{(1)i} \\ \theta_{(2)i} \\ \gamma_{(1)i} \\ \gamma_{(2)i} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{(1)i} \\ \theta_{(2)i} \\ \gamma_{(1)i} \\ \gamma_{(2)i} \end{pmatrix}, \begin{pmatrix} \sigma^2_{(1)i} & \rho_{i\theta}\sigma_{(1)i}\sigma_{(2)i} & \rho_{(11)i}\sigma_{(1)i}\delta_{(1)i} & \rho_{(12)i}\sigma_{(1)i}\delta_{(2)i} \\ \rho_{i\theta}\sigma_{(1)i}\sigma_{(2)i} & \sigma^2_{(2)i} & \rho_{(21)i}\sigma_{(2)i}\delta_{(1)i} & \rho_{(22)i}\sigma_{(2)i}\delta_{(2)i} \\ \rho_{(11)i}\sigma_{(1)i}\delta_{(1)i} & \rho_{(21)i}\sigma_{(2)i}\delta_{(1)i} & \delta^2_{(1)i} & \rho_{i\gamma}\delta_{(1)i}\delta_{(2)i} \\ \rho_{(12)i}\sigma_{(1)i}\delta_{(2)i} & \rho_{(22)i}\sigma_{(2)i}\delta_{(2)i} & \rho_{i\gamma}\delta_{(1)i}\delta_{(2)i} & \delta^2_{(2)i} \end{pmatrix} \right).$$

where the first contrast is indexed by (1) and the second contrast indexed by (2).

C.2 Bayesian linear regression with a normal-inverse-gamma prior

The vector of regression parameters is noted $\beta = \begin{pmatrix} a \\ b \end{pmatrix}$.

C.2.1 Priors

We use the marginal and conditional prior distribution for τ^2 and $\beta \mid \tau^2$:

$$\begin{aligned} \tau^2 &\sim \Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{\nu_0 v_0}{2}\right) \\ \beta \mid \tau^2 &\sim N\left(\beta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tau^2 \Lambda_0^{-1}\right) \end{aligned}$$

with $rank(\Lambda_0) = 2$ (proper prior), leading to a normal-inverse-gamma joint prior distribution:

$$\beta, \tau^2 \sim N\text{-}\Gamma^{-1}\left(\beta_0, \Lambda_0^{-1}, \frac{\nu_0}{2}, \frac{\nu_0 v_0}{2}\right).$$

C.2.2 Model

The regression model is given by

$$\begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_N \end{pmatrix} = \begin{pmatrix} 1 & \gamma_1 \\ \vdots & \vdots \\ 1 & \gamma_N \end{pmatrix} \begin{pmatrix} a & b \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_N \end{pmatrix}$$

with matrix notations $\theta = X\beta + \epsilon$ with $\epsilon \sim N(0, \tau^2 I_n)$.

C.2.3 Likelihood

The likelihood can be written

$$L\left(\beta, \tau^2, \theta, X\right) = \frac{1}{\left(2\pi\tau^2\right)^{-N/2}} \exp\left(-\frac{\left(\theta - X\beta\right)^T \left(\theta - X\beta\right)}{2\tau^2}\right).$$

C.2.4 Posterior

The joint posterior distribution can be expressed as

$$\beta, \tau^2 \mid \theta, X \sim N\text{-}\Gamma^{-1}\left(\beta_N, \Lambda_N^{-1}, \frac{\nu_N}{2}, \frac{\nu_N v_N}{2}\right)$$

with $\Lambda_N = X^T X + \Lambda_0$, $\beta_N = \Lambda_N^{-1}(X^T\theta + \Lambda_0\beta_0)$, $\nu_N = \nu_0 + N + 1$ and

$$v_N = \frac{2}{\nu_N}\left(\beta_0^T\Lambda_0\beta_0 + \theta^T\theta - \beta_N^T\Lambda_N\beta_N + \nu_0 v_0\right).$$

The corresponding marginal posterior distributions are:

$$\tau^2 \mid \theta, X \sim \Gamma^{-1}(\frac{\nu_N}{2}, \frac{\nu_N v}{2})$$
$$\beta \mid \tau^2, \theta, X \sim N(\tau^2\Lambda_N^{-1})$$

C.3 Application in Multiple Sclerosis: data for the meta-analysis

TABLE C.1: Estimated log risk ratios for the MRI lesion counts, the annualized relapse rate and the disability progression with their standard errors.

Trial	N	Treatment arm 1	Treatment arm 2	MRI $\hat{\gamma}_{1i}$ (δ_{1i})	Relapse $\hat{\gamma}_{2i}$ (δ_{2i})	Disability $\hat{\theta}_i$ (σ_i)
2,38	248	IFN-1b 1.6 MIU	Pbo	-0.994 (0.276)	-0.083 (0.150)	0.000 (0.240)
2,38	247	IFN-1b 8 MIU	Pbo	-0.892 (0.274)	-0.416 (0.147)	-0.342 (0.263)
21	26	Methylpred	Pbo	(–)	-0.211 (0.687)	0.122 (0.811)
22	251	GA	Pbo	(–)	-0.342 (0.213)	-0.128 (0.261)
23,39	172	IFN-1a 6 MIU	Pbo	-0.400 (0.208)	-0.386 (0.253)	-0.446 (0.301)
24	150	IVIg	Pbo	(–)	-0.896 (0.269)	-0.357 (0.376)
4	51	Mitoxantrone	Pbo	-0.734 (0.502)	-1.080 (0.484)	-1.660 (0.816)
25	40	IVIg	Pbo	(–)	-0.994 (0.479)	-0.223 (0.807)
5,40	376	IFN-1a 22μg	Pbo	-0.545 (0.227)	-0.342 (0.369)	-0.386 (0.189)
5,40	371	IFN-1a 44μg	Pbo	-1.050 (0.222)	-0.386 (0.178)	-0.315 (0.186)
26	802	IFN-1a 60μg	IFN-1a 30μg	-0.117 (0.125)	0.049 (0.112)	0.000 (0.116)
27	188	IFN-1b	IFN-1a	(–)	-0.342 (0.270)	-0.844 (0.342)
28	306	Hydrolytic enzymes	Pbo	(–)	-0.163 (0.196)	0.077 (0.220)
29	942	Natalizumab	Pbo	-1.770 (0.124)	-1.140 (0.156)	-0.528 (0.141)
30	1171	Natalizumab+IFN-1a	IFN-1a	-1.770 (0.112)	-0.799 (0.121)	-0.236 (0.115)
31	223	Alemtuzumab 12mg	IFN-1a	(–)	-1.170 (0.304)	-1.140 (0.366)
31	221	Alemtuzumab 24mg	IFN-1a	(–)	-1.510 (0.269)	-1.020 (0.354)
32	764	IFN	GA	-0.198 (0.151)	0.030 (0.188)	0.285 (0.226)
33,34	870	Cladribine 3.5mg/kg	Pbo	-1.310 (0.156)	-0.868 (0.164)	-0.363 (0.165)
33,34	893	Cladribine 5.25mg/kg	Pbo	-1.470 (0.149)	-0.799 (0.161)	-0.308 (0.160)
35	118	IFN-1a 30μg+AZA	IFN-1a 30μg	(–)	-0.139 (0.408)	0.207 (0.425)
35	123	IFN-1a 30μg+AZA+pred	IFN-1a 30μg	(–)	-0.357 (0.250)	0.039 (0.431)
36	130	IFN-1a+methylpred	IFN-1a+Pbo	-0.261 (0.320)	-0.994 (0.438)	-0.446 (0.397)
37	1345	IFN-1b 250μg	GA	-0.328 (0.098)	0.058 (0.230)	0.255 (0.112)
37	1347	IFN-1b 500μg	GA	-0.328 (0.082)	-0.030 (0.181)	0.049 (0.118)

N=number of patients. Trial numbers are based on the references numbers provided in [123].