Decision-making in drug development using a composite definition of success

Decision-making in drug development using a composite definition of success

Gaelle Saint-Hilary[a*], Veronique Robert[b] and Mauro Gasparini[a]

[a] Dipartimento di Scienze Matematiche (DISMA) Giuseppe Luigi Lagrange, Politecnico di Torino, Torino, Italy
[b] Department of Biostatistics, Institut de Recherches Internationales Servier (IRIS), Suresnes, France
[*] **Correspondence to:** Gaelle Saint-Hilary, Politecnico di Torino, Dipartimento di Scienze Matematiche (DISMA) Giuseppe Luigi Lagrange, Corso Duca degli Abruzzi 24, 10129 Torino, Italy. E-mail: gaelle.sainthilary@polito.it

Abstract

Evidence-based quantitative methodologies have been proposed to inform decision-making in drug development, such as metrics to make go/no-go decisions or predictions of success, identified with statistical significance of future clinical trials. While these methodologies appropriately address some critical questions on the potential of a drug, they either consider the past evidence without predicting the outcome of the future trials or focus only on efficacy, failing to account for the multifaceted aspects of a successful drug development. As quantitative benefit-risk assessments could enhance decision-making, we propose a more comprehensive approach using a composite definition of success based not only on the statistical significance of the treatment effect on the primary endpoint, but also on its clinical relevance, and on a favorable benefit-risk balance in the next pivotal studies. For one drug, we can thus study several development strategies before starting the pivotal trials by comparing their predictive probability of success. The predictions are based on the available evidence from the previous trials, to which new hypotheses on the future development could be added. The resulting predictive probability of composite success provides a useful summary to support the discussions of the decision-makers. We present a fictive, but realistic, example in Major Depressive Disorder inspired by a real decision-making case.

Keywords: Decision-making; Composite success; Probability of success; Benefit-risk; Bayesian analysis.

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 [1, 2, 3]). 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 [4, 5, 6]. 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 [7, 8], to assess the benefit-risk balance of the treatments [9, 10, 11, 12] or to predict the statistical significance or the futility of clinical trials [13, 14, 15, 16, 17, 18, 19, 20]. 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 2. In Section 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. Additional information including source code to reproduce the results may be found in Supplemental Material.

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 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 2.1 to 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 2.4.
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 [21]. 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.

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 [22, 23, 24] can be calculated according to Bayes theorem as:

$$f(\delta|\mathbf{Y}=\mathbf{y})=\frac{f_{\mathbf{Y}}(\mathbf{y}|\delta)f(\delta)}{f(\mathbf{y})}, \tag{1}$$

where $f_{\mathbf{Y}}$ is the density of $\mathbf{Y}$ conditional on δ and $f(\mathbf{y})=\int f_{\mathbf{Y}}(\mathbf{y}|\delta)f(\delta)d\delta$.
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:

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

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

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

Using Equation (1), it can be re-written as:

$$P(d^*>D|\mathbf{Y}=\mathbf{y})=\frac{\int\int_{z>D} f_{d^*}(z|\delta)f_{\mathbf{Y}}(\mathbf{y}|\delta)f(\delta)dzd\delta}{\int f_{\mathbf{Y}}(\mathbf{y}|\delta)f(\delta)d\delta}. \quad (2)$$

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 $d^*|\delta \sim N(\delta,s^{*2})$, where s^{*2} is its variance in the next trial. From the posterior distribution of δ and the distribution of $d^*|\delta$, we obtain the predictive distribution:

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

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

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

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 (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* [15] or *Bayesian predictive power* [25, 26]. 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 [1, 2, 3]). We define the probability of clinical relevance on the primary endpoint, $PPoS_2$, as the probability in Equation (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 [27]. 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.

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 [9, 10, 11, 12, 28, 29] 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) [30, 31, 32], since the European Medicine Agency Benefit-Risk Methodology Project suggested that it is one of the most comprehensive among the quantitative methodologies they considered [33, 34, 35, 36], and it is also recommended by the IMI PROTECT Work package 5 [37]. 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 [30, 31], 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.* [38] 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 [30]:

(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 [10, 30, 38, 39] 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:

$$u_i=u(\boldsymbol{\xi}_i,\mathbf{w})=w_1u_1(\xi_{i1})+...+w_nu_n(\xi_{in})=\sum_{j=1}^{n}w_ju_j(\xi_{ij}).$$

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.* [38], 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:

$$f(\xi_{ij}|X_{ij}=x_{ij})=\frac{f_{X_{ij}}(x_{ij}|\xi_{ij})f(\xi_{ij})}{f(x_{ij})}, \tag{3}$$

where $f_{X_{ij}}$ is the density of X_{ij} conditional on ξ_{ij} and $f(x_{ij})=\int f_{X_{ij}}(x_{ij}|\xi_{ij})f(\xi_{ij})d\xi_{ij}$. 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

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

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

$$\begin{aligned} PPoS_3 &= P(\Delta^* u_{12}>0|\mathbf{X_1}=\mathbf{x_1},\mathbf{X_2}=\mathbf{x_2}) \\ &= \int\int\int_{v>0} f_{\Delta^* u_{12}}(v|\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)f(\boldsymbol{\xi}_1|\mathbf{X_1}=\mathbf{x_1})f(\boldsymbol{\xi}_2|\mathbf{X_2}=\mathbf{x_2})dvd\boldsymbol{\xi}_1d\boldsymbol{\xi}_2. \end{aligned}$$

Using Equation (3) and the vectorized notations, it can be re-written as:

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

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 $f(\boldsymbol{\xi}_1|\mathbf{X_1}=\mathbf{x_1})$ and $f(\boldsymbol{\xi}_2|\mathbf{X_2}=\mathbf{x_2})$ of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are obtained using classical Bayesian methods [40], either analytically or with Markov Chain Monte Carlo (MCMC) simulations.

(ii) Values $\boldsymbol{\xi}_1^{*(k)}$ and $\boldsymbol{\xi}_2^{*(k)}$ are sampled from $f(\boldsymbol{\xi}_1|\mathbf{X_1}=\mathbf{x_1})$ and $f(\boldsymbol{\xi}_2|\mathbf{X_2}=\mathbf{x_2})$, 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 $\mathbf{x_1}^{*(\mathbf{k})}$ and $\mathbf{x_2}^{*(\mathbf{k})}$ of the performances of the treatments in the next trial are simulated from $f_{\mathbf{X_1}}(\mathbf{x_1}|\boldsymbol{\xi}_1^{*(k)})$ and $f_{\mathbf{X_2}}(\mathbf{x_2}|\boldsymbol{\xi}_2^{*(k)})$, 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 $\Delta^{*(k)}u_{12}=u(\mathbf{x_1}^{*(\mathbf{k})},\mathbf{w})-u(\mathbf{x_2}^{*(\mathbf{k})},\mathbf{w})$.

(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 $\mathbb{1}[false]=0$.

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:

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

It is highly unlikely that δ is independent of $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, since the primary endpoint is almost always one of the criteria considered in the benefit-risk assessment. Therefore,

we consider the joint distribution of $(\delta,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ to write explicitely the formula of the PPoS, following the same principle as in the previous sections:

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

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.

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:

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

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.

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.

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 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.

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.

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 2.

The HAM-D_{17} total score is usually and reasonably assumed to be normally distributed [41]. 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 σ_1=0.98, σ_2=1.02 and σ_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 $\delta_i \sim N(0,2\times10^4)$ (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 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 $Bin(n_{ij},\xi_{ij})$ and uniform conjugate priors $\xi_{ij} \sim Beta(1,1)$.
The partial value functions of all criteria are defined as linear functions as presented in Section 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 3.

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 2, and assuming that 114 patients per arm are included:

(i) Means HAM-D_{17} total score: $m_i^*|\xi_{i1} \sim N(\xi_{i1},\sigma_i^{*2})$ 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 observed in the Phase II study.

(ii) Differences in HAM-D_{17} mean total score versus placebo: $d_i^*|\delta_i \sim N(\delta_i,s_i^{*2})$ 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: $p_{ij}^*|\xi_{ij}=r_{ij}^*/114$ with $r_{ij}^* \sim Bin(114,\xi_{ij})$ for i=1,2,3 and j=2,...,6.

(iv) Benefit-risk utility scores: $u(m_i^*,p_{i2}^*,...,p_{i6}^*,\mathbf{w})=w_1u_1(m_i^*)+w_2u_2(p_{i2}^*)+...+w_6u_6(p_{i6}^*)$ 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.

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 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.

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 $1.96s^* \approx 1.82$ 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 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.

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 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 3, and the predictive probabilities of success are summarized in Table 5. 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.

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 [42], 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 Supplemental Material. Overall, the conclusions are robust to uncertainty in weight elicitation, correlations, and departure from clinical assumptions.

3.6 Alternative example

An alternative example is presented in Supplementary Material, 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 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 [43].

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 [44, 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 [11, 30] are valuable tools to help building such models, and some support could also be found from the general literature on MCDA [45, 46].

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 [12, 37, 50]. 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 [42, 51], 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 [38].

- The sensitivity of the results to the choice of the priors used in the Bayesian analysis should be evaluated [26]. 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 [52]. 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 [47]. 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 [48, 49].

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 [21]. 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.* [54] or Frewer *et al.* [8]. 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 [52]. 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' [55], 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.

Acknowledgements: This research is supported by the Institut de Recherches Internationales Servier (IRIS).

References

[1] EMA. Guideline on clinical investigation of medicinal products in the treatment of hypertension, 2011. Available at `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2016/07/WC500209943.pdf`. Accessed June 14, 2017.

[2] EMA. Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus, 2012. Available at `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129256.pdf`. Accessed June 14, 2017.

[3] EMA. Guideline on clinical investigation of medicinal products in the treatment of depression, 2013. Available at `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/05/WC500143770.pdf`. Accessed June 14, 2017.

[4] EMA. Guidance document on the content of the <Co->Rapporteur day 80 critical assessment report, 2013. Available at `http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/10/WC500004856.pdf`. Accessed June 14, 2017.

[5] EMA. ICH guideline E2C (R2) on periodic benefit-risk evaluation report (PBRER), 2013. Available at `http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/periodic-benefit-risk-evaluation-report.html`. Accessed June 14, 2017.

[6] U.S. Food and Drug Administration. Providing Postmarket Periodic Safety Reports in the ICH E2C (R2) Format (Periodic Benefit-Risk Evaluation Report), 2013. Available at `https://www.fda.gov/downloads/drugs/guidances/ucm346564.pdf`. Accessed June 14, 2017.

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

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

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

[10] 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* 2016; **19**: 1–13.

[11] 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* 2016; **19**: 125–37.

[12] 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 Statististics* 2015; **15**: 324–332. DOI: 10.1002/pst.1690.

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

[14] Johns D and Andersen JS. Use of predictive probabilities in Phase II and Phase III clinical trials. *Journal of Biopharmaceutical Statistics* 1999; **9(1)**: 67–79.

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

[16] 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* 2005; **4**: 119–128. DOI: 10.1002/pst.164.

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

[18] Tang Z. Defensive Efficacy Interim Design: dynamic benefit-risk ratio view using probability of success. *Journal of Biopharmaceutical Statistics* 2016; **27**: 683–690. DOI: 10.1080/10543406.2016.1198370.

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

[20] Zhang J and Zhang JJ. Joint probability of statistical success of multiple phase III trials. *Pharmaceutical Statistics* 2013; **12**: 358–365. DOI: 10.1002/pst.1597.

[21] EMA. Point to consider on applications with 1. Meta-analyses and 2. One pivotal study, 2001. Available at `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf`. Accessed June 14, 2017.

[22] Ashby D and Smith A. Evidence-based medicine as Bayesian decision-making. *Statistics in Medicine* 2000; **19**: 3291-3305.

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

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

[25] Spiegelhalter DJ, Reedman LS, Blackburn PR. Monitoring clinical trials: conditional power or predictive power? *Control Clin Trials* 1986; **7(1)**: 8-17.

[26] Rufibach K., Burger H. U. and Abt M. Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development. *Pharmaceutical Statistics* 2016; **15**: 438-446. DOI: 10.1002/pst.1764.

[27] ICH guideline E9: Statistical Principles for Clinical Trials, 1998. Available at `http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html`. Accessed December 16, 2017.

[28] 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. Available at `http://www.imi-protect.eu/documents/ShahruletalReviewofvisualisationmethodsfortherepresentationofBR assessmentofmedicationStage1F.pdf`. Accessed June 14, 2017.

[29] 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. Available at http://www.imi-protect.eu/documents/ShahruletalReviewofvisualisationmethodsfortherepresentationofBR assessmentofmedicationStage2A.pdf. Accessed June 14, 2017.

[30] 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* 2007; **16**: S2-S15. DOI: 10.1002/pds.1435.

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

[32] NICE Decision Support Unit. Multi-Criteria Decision Analysis (MCDA) for Health Technology Assessment, 2011. Available at http://scharr.dept.shef.ac.uk/nicedsu/wp-content/uploads/sites/7/2016/03/MCDA-for-HTA-DSU.pdf. Accessed June 14, 2017.

[33] 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. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/07/WC500109478.pdf. Accessed June 14, 2017.

[34] EMA. Benefit-risk methodology project. Work package 2 Report: Applicability of current tools and processes for regulatory benefit-risk assessment, 2010. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Report/2010/10/WC500097750.pdf. Accessed June 14, 2017.

[35] EMA. Benefit-risk methodology project. Work package 3 Report: Field tests, 2011. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/09/WC500112088.pdf. Accessed June 14, 2017.

[36] EMA. Benefit-risk methodology project. Work package 4 Report: Benefit-risk tools and processes, 2012. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Report/2012/03/WC500123819.pdf. Accessed June 14, 2017.

[37] IMI PROTECT. IMI PROTECT Work package 5: Benefit-risk integration and representation, 2013. Available at http://www.imi-protect.eu/wp5.shtml. Accessed June 14, 2017.

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

[39] Marcelon L, Verstraeten T, Dominiak-Felden G, Simondon F. Quantitative benefit-risk assessment by MCDA of the quadrivalent HPV vaccine for preventing anal cancer in males. *Expert Rev Vaccines* 2016; **15(1)**:139-48. DOI: 10.1586/14760584.2016.1107480.

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

[41] Cameron IM, Cardy A, Crawford JR, du Toit SW, Hay S, Lawton K, Mitchell K, Sharma S, Shivaprasad S, Winning S, Reid IC. Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II. *Br J Gen Pract* 2011; **61 (588)**: e419-e426. DOI: 10.3399/bjgp11X583209.

[42] Saint-Hilary G, Cadour S, Robert V, 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* 2017; **59(3)**: 567–578. DOI: 10.1002/bimj.201600113.

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

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

[45] Belton V, Stewart TJ. Multiple Criteria Decision Analysis: An Integrated Approach. Kluwer Academic Publishers, MA, 2002.

[46] Figueira J, Greco S and Ehrgott M. Multiple Criteria Decision Analysis: State of the Art Surveys. Springer Science + Business Media, Inc, 2005.

[47] Wang Y, Fu H, Kulkarni P, and Kaiser C. Evaluating and utilizing probability of study success in clinical development. *Clinical Trials*, **10(3)**:407-413, 2013. DOI: 10.1177/1740774513478229.

[48] 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* 2012; **31**: 831–843. DOI: 10.1002/sim.4476.

[49] Chen C and Sun LZ. Quantification of PFS effect for accelerated approval of oncology drugs. *Statistics in Biopharmaceutical Research*, **3(3)**:434-444, 2011. DOI: 10.1198/sbr.2011.09046.

[50] Nixon R., Dierig C., Mt-Isa S., Stockert 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* 2016; **58**, 827. DOI: 10.1002/bimj.201300248.

[51] Tervonen, T., Van Valkenhoef, G., Buskens, E., Hillege, H. L., and Postmus, D. A stochastic multicriteria model for evidence-based decision making in drug benefitrisk analysis. *Statistics in Medicine* 2011; **30**, 1419–1428. DOI: 10.1002/sim.4194.

[52] Gallo P, Mao L, Shih VH. Alternative views on setting clinical trial futility criteria. *J Biopharm Stat* 2014; **24(5)**:976-93. DOI: 10.1080/10543406.2014.932285.

[53] Kennedy EH, Kangovi S, Mitra N. Estimating scaled treatment effects with multiple outcomes. *Statistical Methods in Medical Research* 2017. DOI:10.1177/0962280217747130 (first published online).

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

[55] Chuang-Stein C and Kirby S. *Quantitative Decisions in Drug Development*. Springer International Publishing AG, Switzerland, 2017.

Table 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

Table 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)				
$\delta_i=\xi_{31}-\xi_{i1}$	$d_i=m_3-m_i$	$N(0,s_0^2)$	$N(\delta_i,s_i^2)$	$N\left(\frac{s_0^2}{s_0^2+s_i^2}d_i,\frac{s_0^2s_i^2}{s_0^2+s_i^2}\right)$
Occurrence of adverse events in each arm (i=1,...,3 ; j=2,...,6)				
ξ_{ij}	r_{ij}/n_{ij}	$Beta(1,1)$	$Bin(n_{ij},\xi_{ij})/n_{ij}$	$Beta(r_{ij}+1,n_{ij}-r_{ij}+1)$

$\sigma_0^2=10^4$; $s_0^2=2\times\sigma_0^2=2\times10^4$.

Table 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%

Table 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%

Table 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