Contemporary Clinical Trials 67 (2018) 100–108

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Periodic benefit-risk assessment using Bayesian stochastic multi-criteria acceptability analysis

Kan Li[a], Shuai Sammy Yuan[b,*], William Wang[b,*], Shuyan Sabrina Wan[b], Paulette Ceesay[b], Joseph F. Heyse[b], Shahrul Mt-Isa[b], Sheng Luo[c]

[a] Department of Biostatistics, The University of Texas Health Science Center at Houston, Houston, TX, USA
[b] Merck Research Lab, Merck & Co, PA, USA
[c] Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA

ARTICLE INFO

Keywords:
Clinical trials
Structured benefit-risk assessment
Periodic benefit-risk evaluation report
Bayesian meta-analysis
Stochastic multi-criteria acceptability analysis
Multi-criteria decision analysis

ABSTRACT

Benefit-risk (BR) assessment is essential to ensure the best decisions are made for a medical product in the clinical development process, regulatory marketing authorization, post-market surveillance, and coverage and reimbursement decisions. One challenge of BR assessment in practice is that the benefit and risk profile may keep evolving while new evidence is accumulating. Regulators and the International Conference on Harmonization (ICH) recommend performing periodic benefit-risk evaluation report (PBRER) through the product's lifecycle. In this paper, we propose a general statistical framework for periodic benefit-risk assessment, in which Bayesian meta-analysis and stochastic multi-criteria acceptability analysis (SMAA) will be combined to synthesize the accumulating evidence. The proposed approach allows us to compare the acceptability of different drugs dynamically and effectively and accounts for the uncertainty of clinical measurements and imprecise or incomplete preference information of decision makers. We apply our approaches to two real examples in a post-hoc way for illustration purpose. The proposed method may easily be modified for other pre and post market settings, and thus be an important complement to the current structured benefit-risk assessment (sBRA) framework to improve the transparent and consistency of the decision-making process.

1. Introduction

Benefit-risk (BR) assessment is essential to ensure the best decisions are made for a medical product in the clinical development process, regulatory marketing authorization, post-market surveillance, and coverage and reimbursement decisions. However, to reach a consensus on the evaluation of benefit and risk is a challenging and complex process as it involves various stakeholders, different data sources, and exhibits dynamic nature as new information continues to emerge. To this end, the structured BR assessment framework has evolved rapidly in the last few years, as evidenced by a large number of regulatory and industry-wide initiatives on structured BR assessment [1–3]. Although it has been widely acknowledged that structured BR assessment is mainly based on a qualitative and descriptive BR framework, quantitative approaches play an important role in order to complement qualitative frameworks by providing objectivity and transparency on the impact of weighting and uncertainty when assessing the BR profile of a medical product [2].

2. Literature review and motivations

A recent systematic review by Mt-Isa et al. [4] identified multi-criteria decision analysis (MCDA) to be among the most promising methods for conducting a quantitative BR assessment, and the method was also highlighted by regulators [1]. The debut of MCDA for the BR assessment of new drugs was provided by Mussen et al. [5]. The principle of the method is to compare drugs using utility scores calculated from multiple criteria of benefits and risks, taking into account their relative importance according to the preferences (a.k.a. weights) of the decision makers. However, in its standard implementation, the impact of uncertainty of criteria and preference on the choice of optimal decision was not considered. Tervonen et al. [6] proposed to use a stochastic multi-criteria acceptability analysis (SMAA) approach for evaluating a drug benefit-risk profile which can account for both variations inherent in criterion measurements and lack of preferences information. Waddingham et al. [7] proposed a Bayesian MCDA model to estimate the distribution of the criterion via synthesizing the evidence observed in previous studies. To mitigate the high degree of uncertainty in the

* Corresponding authors.
E-mail addresses: sammy.yuan@merck.com (S.S. Yuan), william_wang@merck.com (W. Wang).

https://doi.org/10.1016/j.cct.2018.02.016
Received 7 December 2017; Received in revised form 21 February 2018; Accepted 27 February 2018
Available online 02 March 2018
1551-7144/

results of SMAA, Saint-Hilary et al. [8] proposed a simple way to control the weight space of SMAA in benefit-risk assessment. Each of the above approaches has its own advantage of incorporating different sources of uncertainty in BR assessment. However, how to combine these advantages in a real working procedure of BR assessment in the pharmaceutical industry still needs more researches.

One additional challenge of BRA in practice is that the benefit and risk profile may keep evolving while new evidence is accumulating. For example, it is recommended by regulators and the International Conference on Harmonization (ICH) to perform periodic benefit-risk evaluation report (PBRER) through the product's lifecycle for such consideration [13–16]. In (ICH) E2C (R1) guideline [16], it is stated that meta-analyses or pooled analyses could be performed to summarize all available information from any other clinical trial sources in addition to those clinical trials or no-interventional studies completed or still ongoing during the reporting period, such as randomized clinical trials, and safety information from co-development partners or investigator-initiated trials. However, methods and examples of implementing an integrated benefit-risk analysis are rare in literatures, especially those in a quantitative way. How to integrate all the cumulative clinical trial data sources together in a quantitative way to perform a comprehensive BRA remains an open question. Therefore, a quantitative assessment method that can be applied in practice to satisfy the needs of benefit risk assessment with accumulating information during the drug development is much needed.

In this paper, we propose a general statistical framework that could be used for the quantitative benefit risk assessment in which Bayesian meta-analysis and stochastic multi-criteria acceptability analysis (SMAA) will be combined to synthesize the accumulating evidence from early stages of the clinical development to late stages. Specifically, we first adopt a Bayesian approach to conduct a cumulative meta-analysis (CMA) based on the summary level data to get the posterior distribution of the criteria values from the selected benefit and risk endpoints across multiple studies. Then the SMAA approach is used to perform the BR assessment based on the synthesized benefit and risk evidence from the cumulative meta-analysis. The proposed framework is a dynamic process in which the posterior distribution is updated whenever a new clinical trial or another new data source regarding the medical product becomes available for inclusion. The proposed approach aims to systematically assess the benefit-risk balance across the lifecycle of a medical product. Therefore, the approach is ready to be modified as needed to address all stakeholders' requirements in both pre and post market setting.

The remaining of the paper is organized as follows. We first introduce a two-step approach for periodic BRA based on Bayesian evidence synthesis and the SMAA method in Section 3. The details of the proposed method are then illustrated in Sections 4 and 5. Section 4 describes the Bayesian meta-analysis as the first step and Section 5 describes the SMAA as the second step. Next, we show how to apply the proposed method for periodic BRA using two case studies in Section 6. Concluding remarks and discussions are presented in Section 7.

3. A general framework for periodic BRA

Fig. 1 shows the flows of the proposed framework of the periodic BRA. In a drug development program, we usually have multiple studies conducted at different stages (labeled as study 1, 2, …, K + 1). Some studies provide pivotal information on both efficacy and safety for registration (e.g. study 1 and 2); some studies may only provide the long-term safety data (e.g., study K and K + 1). Other data sources could also be included in the framework. As the first step in the process, the Bayesian meta-analysis approach will be used to synthesize different data sources together in a temporal sequence which will give the posterior distribution of the selected key efficacy and safety endpoints (a.k.a., criteria). It is repeated whenever a new data source is available, and respectively for each endpoint if the endpoints are independent. After we get the summary of each endpoint across studies, those summary data will be put into the so-called stochastic multi-criterion acceptability assessment (SMAA) framework as criteria values for ranking of the different treatments included in the drug development program. One thing worth mentioning here is that endpoints could be correlated. However, such cases may need patient-level data or correlation information, and are beyond the scope of this paper.

4. Bayesian meta-analysis for evidence synthesis

To perform the periodic benefit-risk assessment, the first step is to identify the endpoints to be included in the value tree or effect table of the sBRA as in the usual benefit-risk analysis. Thereafter, we can integrate the information from different studies to produce across-study summaries of the data for those endpoints selected. In this paper, we focus on the scenarios that the summary level data are available for each treatment arm in the comparison. In scenarios where pairwise comparisons between treatment arms are available, and where there is a need for indirect treatment comparison, different techniques such as network meta-analysis may be used.

We propose using Bayesian hierarchical models to synthesize the evidence across all trials for each treatment group respectively, for its flexibility and ease to implement [12]. Without loss generality, we consider the count type data first. For treatment arm i in study k, and the selected endpoint j, the number of patients having an event is denoted by Y_{ijk}, where $i = 1, \ldots, I$; $j = 1, \ldots, J$; and $k = 1, \ldots, K$. The outcomes Y_{ijk} are assumed to follow independent binomial distributions

$$Yijk \sim \mathrm{Bin}(n_{ik}, p_{ijk}) \tag{1}$$

where the total number of patient n_{ik} in arm i and study k is known. For different treatment arm i and endpoint j, the probabilities of the event p_{ijk}'s are assumed to be independent and come from the same prior distribution across studies. This is equivalent to assuming that each study has its own independent population which is a sample of the overall population. In the hierarchical model, we use a link function to transfer the probabilities p_{ijk} onto the *logit* scale as

$$g(p_{ijk}) = logit(p_{ijk}) = log\left(\frac{p_{ijk}}{1 - p_{ijk}}\right) = \theta_{ijk}$$

After transformation, the parameter of log odds ratio is assumed to follow a normal prior distribution as

$$\theta_{ijk} \sim N(\mu_{ij}, \sigma_{ij}^2) \tag{2}$$

Note that u_{ij} and $\sigma_{ij}{}^2$ are parameters across the K studies for treatment i and endpoint j. We use a non-informative or weak hyper-prior distribution $p(\mu_{ij}, \sigma_{ij}{}^2)$ for these parameters, to represents the lack of information about the effect at the (overall) population level before the current data are available. For example, $u_{ij} \sim N(0, 10^4)$, and $\sigma_{ij}{}^2 \sim Gamma$ (0.001,0.001), where $i = 1, \ldots, I$; *and* $j = 1, \ldots, J$.

In this paper, we consider the three most common data types and their corresponding metrics: continuous data (e.g., change from baseline for an efficacy endpoint), binary data (the metric is percentage, e.g., the proportion of subject meeting an efficacy endpoint or having a safety event among the population of interest), and Poisson count data (the metric is exposure adjusted incidence, e.g., the rate of subjects experiencing a safety event in one patient year). The link functions allow us to model different data types with minimal changes of the above model. Table 1 lists the likelihood and link functions to be used in the Bayesian hierarchical model for different data types.

When data are accumulating after the completion of each relevant study, cumulative meta-analysis can be used to update the posterior distribution of the criteria measurement for each endpoint. In the Bayesian framework, cumulative meta-analysis is a natural process for updating across study summaries in a chronical way. Essentially, the information from earlier studies is utilized to form a prior distribution

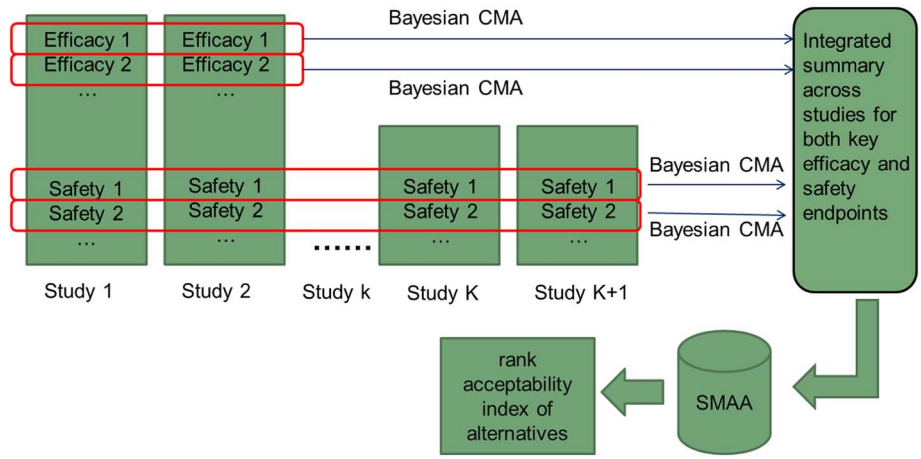

Fig. 1. The general framework of the proposed two-step approach.

and the inclusion of data from each new study will produce an updated posterior distribution and so on.

5. Stochastic multi-criteria acceptability analysis

After the evidence across studies is integrated together, we will then use the SMAA method to compare the benefit-risk profile of different drugs. SMAA [17,18] considers a multi-criteria decision problem consisting of a set of I treatments that are assessed on J criteria. The vector of criteria measurements corresponding to alternative i is denoted by $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iJ}) \in \Xi$, where ξ_{ij} represents the outcome of alternative i on criterion j, and Ξ is the feasible space of criteria values. In our scenario, the alternatives are the I treatments to be compared, and criteria are the J input variables in SMAA model that quantify the selected key efficacy and safety endpoints. The considerations on how to select the efficacy and safety endpoints has been thoroughly discussed by many researchers [9], and is beyond the scope of the present paper. The benefit-risk tradeoff of a specific treatment is represented by the utility score $u(\boldsymbol{\xi}_i, \boldsymbol{w})$, with higher value indicating a more preferable benefit-risk profile. It is generally assumed that the criteria satisfy the independence conditions and an additive value function is applied:

$$u(\boldsymbol{\xi}_i, \boldsymbol{w}) = w_1 u_1(\xi_{i1}) + \ldots + w_J u_J(\xi_{iJ}) \tag{3}$$

where $u_j(\xi_{ij})$ are partial value functions for the j-th criterion of i-th alternative and w_j is the weight representing relative importance (or preference) of the criteria as compared with other criteria. Denote $\boldsymbol{w} = [w_1, w_2, \ldots, w_J] \in \boldsymbol{\Omega}$, and the weight space $\boldsymbol{\Omega} = \{\boldsymbol{w} : \boldsymbol{w} \geq \boldsymbol{0}$ and $\Sigma_{j=1}{}^{J} w_j = 1\}$. Note w is the same for all alternative treatments. In the next several subsection, we are going to introduce the different components in the SMAA method.

5.1. Criteria

To assess the uncertainty in criteria measurements, the criteria values $\boldsymbol{\xi}_i's$ are assumed to be random variables with joint density functions $f(\boldsymbol{\xi}_i)$. Assuming the criterion value are independent, then $f(\boldsymbol{\xi}_i) = \Pi_{j=1}{}^{J} f(\boldsymbol{\xi}_{ij})$. In our proposed framework, $f(\xi_{ij})$ is the marginal posterior distribution $p(\xi_{ij} \mid Y_{ij1}, \ldots, Y_{ijK})$ resulted from the Bayesian cumulative meta-analysis. The relationships between criteria ξ_{ij} and transformed population level mean parameter u_{ij} are listed in the last column of Table 1. For example, in the case of event rate with a binary distribution, the percentage of event is defined as $\xi_{ij} = inv.\ logit(\mu_{ij}) = \frac{exp(\mu_{ij})}{1 + exp(\mu_{ij})}$.

For simplicity, we assumed that the criteria satisfy the independence condition, and the information of each criterion is summarized independently. It is important to acknowledge that such independence may not be realistic in some cases, and not all studies had every efficacy and safety endpoint collected. However, the proposed model can be extended to account for correlation between criteria as data permits, by using multivariate distributions. In our real examples, we have the study summaries for every endpoint and each drug of interest. So the posterior distribution for the criteria value of each endpoint across all treatment arms was computed respectively. If the data are not available for all endpoints in every study, we could still use the studies with data for some endpoints and update the posterior separately for each corresponding endpoint. In other cases where summary data are from literature, especially when multiple treatments were involved without a common comparator, network meta-analysis may be needed to produce estimates for relative effects. One disadvantage for such method is that the relative effects are not suitable for specifying absolute benefit and risk trade-offs of a single drug, as they do not contain information on the baseline effect [10]. Therefore, more assumptions such as those on baseline effect have to be made to rank all the treatments.

Table 1
Criteria types and corresponding likelihood and link functions.

Data type of Y_{ijk}	Metric of endpoints	Likelihood of Y_{ijk}	Link function $g(\cdot) = \theta_{ijk}$	Prior distribution for θ_{ijk}	Criteria value in SMAA model $\xi_{ij} = g^{-1}(\mu_{ij})$
Binary data	Percentage	$Y_{ijk} \sim \mathrm{Bin}(n_{ik}, p_{ijk})$	$log\left(\frac{p_{ijk}}{1 - p_{ijk}}\right)$	$\theta_{ijk} \sim N(\mu_{ij}, \sigma_{ij}{}^2)$	$\frac{exp(\mu_{ij})}{1 + exp(\mu_{ij})}$
Count	Exposure adjusted event rate	$Y_{ijk} \sim \mathrm{Poisson}(\lambda_{ijk} E_{ik})$ E_{ik}: exposure λ_{ijk}: the incidence rate per person-year	$log(\lambda_{ijk})$	$\theta_{ijk} \sim N(\mu_{ij}, \sigma_{ij}{}^2)$	$\exp(\mu_{ij})$
Continuous data		$Y_{ijk} \sim \mathrm{N}(\theta_{ijk}, se_{ijk}{}^2)$, $se_{ijk}{}^2$: standard error	Identity	$\theta_{ijk} \sim N(\mu_{ij}, \sigma_{ij}{}^2)$	μ_{ij}

5.2. Partial value functions

To put criteria on a comparable basis, the partial value functions (PVF) $u_j(\xi_{ij})$ are used to normalize the criterion measurements ξ_{ij} by mapping them into the same scale, e.g. 0 to 1. The PVF is often linear, monotonically increasing, but may also be nonlinear. The linear PVF is defined as $u_j(\xi_{ij}) = (\xi_{ij} - \xi_j^{'})/(\xi^{''}{}_j - \xi_j^{'})$ if the preference direction is increasing; and $u_j(\xi_{ij}) = (\xi_j^{'} - \xi_{ij})/(\xi^{'}{}_j - \xi^{''}{}_j)$ if the preference direction is decreasing, where $\xi^{'}{}_j$ and $\xi^{''}{}_j$ are the least and most preferable values of criteria j. Thus, the best case corresponds to $u_j(\xi_{ij}) = 1$ and the worst case corresponds to $u_j(\xi_{ij}) = 0$. This is consistent among all benefit and risk criteria. Moreover, because using scale ranges that are too large causes imprecision for the preference elicitation, it is recommended that the value $[\xi_j^{'}, \xi^{''}{}_j]$ could be defined based on interval of the 95% credible intervals from the posterior distribution of the criteria j [6]. The measurements out of the range are truncated by the values $\xi^{'}{}_j$ and $\xi^{''}{}_j$ at each direction respectively.

5.3. Choice of weight

The preference information in SMAA is expressed as weights for different criteria. This reflects the consideration of decision makers regarding the relative importance of a criterion comparing with other criteria. A greater weight means to give more importance to the criterion when assessing the benefit-risk profile. Denote such weight vector as $\boldsymbol{w} = (w_1, \ldots, w_J)$ with restriction that $\Sigma_{j=1}{}^J w_j = 1$. Very often, such weights cannot be fully extracted from decision makers because either decision makers are not sure, or reluctant to tell. Also, the weights from different decision makers can be simply different due to different perspectives. Therefore, instead of using an elicited weight vector $\boldsymbol{w}$, the SMAA approach considered the weights as random variables with a joint density function $f(\boldsymbol{w})$ in the feasible weight space $\boldsymbol{\Omega}$.

The weight vector reflects the relative importance of the different criteria from the decision makers and thus is not driven by the observed data from clinical trials. In a fully Bayesian framework, the distribution of the weight vector $f(\boldsymbol{w})$ can be viewed as the prior distribution for the parameter $\boldsymbol{w}$. Saint-Hilary et al. [8] proposed a Dirichlet distribution for $f(\boldsymbol{w})$, which considers both the relative magnitudes and uncertainties in $\boldsymbol{w}$. Specifically, the weight is assumed to follow $Dirichlet(r \cdot (w_1{}^0, \ldots, w_J{}^0))$, where $0 \le w_1{}^0, \ldots, w_J{}^0 \le 1$ and $\Sigma_{j=1}{}^J w_j{}^0 = 1$. Note, now the value of $w_j{}^0$ reflects the relative importance of criteria j, and the mean of the weight $E(w_j) = w_j{}^0$. The constant r varies from 0 to $+\infty$ and controls the variance of the weight which goes to infinity when $r = 0$ and to 0 when $r = +\infty$. This precision parameter r reflects the confidence level of the decision makers in the elicitation of their preferences. In practice, $r > 50$ usually represents a very strong confidence in the weight elicitation [8]. Changing r is analogous to conduct a sensitivity analysis on weight parameters $\boldsymbol{w}$ based on certain prior knowledge $\boldsymbol{w^0} = (w_1{}^0, \ldots, w_J{}^0)$, and further impose uncertainty on this prior knowledge via r. Since $\boldsymbol{w^0}$ can be subjective, the impact of $\boldsymbol{w^0}$ can be evaluated by assigning an additional level of hyper-distribution on it if desired, or by using weight distribution elicited from patient preference data which will inherit uncertainty. However, the relative importance of the elements in w^0 should be determined by clinical considerations, rather than being driven by the study results. The regular SMAA without preferences (equal weights) corresponds to a Dirichlet model with the weight of $w_j{}^0 = 1/J$ and $r = J$.

5.4. Utility scores and ranking

The benefit-risk balance is measured by the utility scores $u(\boldsymbol{\xi_i}, \boldsymbol{w})$ which are also random variables. The BR assessment of the treatments is conducted by comparing the distributions of the utility scores, which can be implemented via computing the distribution of the difference between the utility scores of two treatments i and $i^{'}$ as $\Delta u(\boldsymbol{\xi_i}, \boldsymbol{\xi_{i'}}, \boldsymbol{w}) = u(\boldsymbol{\xi_i}, \boldsymbol{w}) - u(\boldsymbol{\xi_{i'}}, \boldsymbol{w})$. The key idea in original SMAA is that, the ranking of the treatment alternatives is based on the expected percentage of the weight vector that gives the best rank for a treatment over the space of the criteria values. In SMAA, such expected percentage is called rank acceptability index. SMAA defines the favorable rank weights $W_i^r(\xi_i)$ as

$$W_i^r(\boldsymbol{\xi_i}) = \{\boldsymbol{w} \in \boldsymbol{W}, rank(\boldsymbol{\xi_i}\boldsymbol{w}) = r\},$$

where r is the rank ranging from 1 to I. $W_i^r(\boldsymbol{\xi_i})$ can be interpreted as the set of weights that gives rank r to the treatment i when its criteria values is $\boldsymbol{\xi_i}$. The r-th rank acceptability index is then defined by

$$b_i^r = \int_{\boldsymbol{\xi_i} \in \Xi} f(\boldsymbol{\xi_i}) \int_{W_i^r(\boldsymbol{\xi_i})} f(\boldsymbol{w}) d\boldsymbol{w} d\boldsymbol{\xi_i} \tag{4}$$

i.e., b_i^r is the expected volume of weights that gives treatment i a rank r over all possible ξ_i.

Calculating the acceptability index involves high dimensional integrals of criteria and weight distributions on their combined feasible space. In practice, Monte Carlo simulation is applied to obtain approximations of those integrals. The samples of criteria measurements and weights are drawn from their distributions, and total utility scores are calculated based on random samples using formula (3). Suppose we draw totally $d = 1, \ldots, D$ samples of criteria $\boldsymbol{\xi_i}^{(d)}$ for treatment i and weights $\boldsymbol{w}^{(d)}$, and thus obtain D utility scores $u^{(d)}(\boldsymbol{\xi_i}, \boldsymbol{w})$ for treatment i. The acceptability index b_i^r is computed by counting how many times $u^{(d)}(\boldsymbol{\xi_i}, \boldsymbol{w})$ are ranked at order r and then divide that number by D.

In this paper, a Bayesian approach based on a Markov-chain Monte Carlo method is performed in Stan, a probabilistic programming language for Bayesian inference, by specifying the full likelihood function and the prior distributions of all unknown parameters. Stan adopts a No-U-Turn sampler [11], which offers faster convergence and parameter space exploration compared with other MCMC algorithms such as Gibbs sampler. The random samples from the posterior distribution of the effect estimates of each criterion were fed directly into the SMAA model which was coded in R.

6. Application to real clinical trials

In this section, we use two real clinical trial examples to illustrate how to use the proposed framework for periodic benefit-risk assessment. For confidentiality issue, we will mask the drug names and use random samples from the original data.

6.1. Periodic BRA for drug T1

The first example is for the treatment of acute migraine. Four placebo or active-controlled phase III studies are chosen for our purpose. We compare the benefit-risk profile of two different doses of drug T1 (Dose 1 and Dose 2) vs another drug T0. Based on the clinical inputs, we choose five efficacy endpoints: pain freedom, pain relief, the absence of phonophobia, the absence of photophobia, and the absence of nausea all at 2 h postdose. The adverse events of concerns are dry mouth, dizziness, somnolence, nausea, and fatigue. For each endpoint, the criterion values used in SMAA is the percentages of patients who had an event of the endpoints. All four studies collected the data for each of the 10 endpoints.

The steps for the periodic BRA are as follows.

1) Using the data from the first study, conduct the meta-analysis using the Bayesian hierarchical model for binary data as described in formula (1) and (2) in Section 3. The meta-analysis is performed for each endpoint respectively. Then we get the posterior distribution of the criteria value $f(\xi_{ij1} | Y_{ij1})$, where $i = 1, 2, 3$, and $j = 1, \ldots, 10$.
2) Get the initial weight vector $\boldsymbol{w}^0 = (w_1{}^0, \ldots, w_{10}{}^0)$ that will be used in the Dirichlet prior for weights in the SMAA model. This is done by the weight swing method as used in multi criterion decision assessment [19]. The experts are provided a table with the smallest

and biggest value for each criterion. Then they are asked to rank criteria according to the importance of their swing from worst to best performance assuming all the other criteria remain at their worst value. After that, the criterion with rank 1 is assigned an initial value of 100. The initial values of all other criteria are assigned a number between 0 and 100, in comparison to the initial value of the most important criterion. The final weights will be the normalized value by the total sum of initial values and be used as prior knowledge about the weight in the Dirichlet distribution $\boldsymbol{w}^0 = (w_1{}^0, \ldots, w_{10}{}^0)$. The weight elicitation are presented in Appendix A Table A.1. We use $r = 1$ as baseline, which indicates little confidence on the weight elicitation and conducted the sensitivity analysis on set of $r \in [1, 50]$.

3) The SMAA method is used to rank the different treatments. To compute rank acceptability index b_i^1 in Eq. (4) and differences of utility $\Delta u(\xi_i, \xi_{i'}, \boldsymbol{w})$, random samples of the criteria value are drawn from the posterior using MCMC from step (1) and the random samples of $\boldsymbol{w}$ are drawn from the Dirichlet distribution with parameters in step (2). Monte Carlo integration is then used to compute the value of b_i^r. The ranks are given based on the relative magnitude of the overall utility value for each treatment arm.
4) Step 1 and 3 are repeated 4 times whenever a new trial became available for inclusion. Since weight is not driven by the data of clinical trials, Step 2 for weight elicitation is only performed once and the same elicited weights are used across repetitions.

The rank acceptability indices resulting from the analysis with $r = 1$ are visualized as a bar chart in Fig. 2. From the left to the right panel in the figure, the acceptability for drug T1 Dose 2 to be the best treatment increases from 0.56 to 0.72 as more evidence from clinical trial studies are included. The Dose 1 of drug T1 is ranked as the second best treatment across the scenario, and drug T0 is always the least preferable alternative. Fig. 3 displays the distributions of the pairwise differences in utility score. To ensure fair comparisons, we use the same set of initial weight samples w^0 in calculating $\Delta u(\xi_i, \xi_{i'}, \boldsymbol{w})$ when comparing each two treatments. Both doses of drug T1 have a better benefit-risk balance than drug T0 as the majority of the distribution curves are on the right side of 0. The utility score for Dose 1 is smaller than Dose 2 as most of the difference distribution is on the left side of 0. As expected, the distributions become denser as more study results are included, which indicates that the precision of the differences of utilities between treatments increases. The results of sensitivity analysis of the confidence factor based on all studies are presented in Fig. 4. As the strength of the weight elicitation confidence r increases, the first rank acceptability for each alternative converges quickly, remaining stable for $r > 20$. We also present the results for the no preference case for reference. The results are close to those obtained with a high degree of uncertainty in weight elicitation.

In summary, the preference sequence is T1 Dose 2 > T1 Dose 1 > T0 in terms of the BR balance. When more studies are included, the evidence is stronger. And the conclusion is very stable for all different levels of confidence of the decision maker's preference.

6.2. Periodic BRA for drug R1

The second example of drug R1 is for treatment of HIV disease for treatment experienced patients. Two studies are chosen from this program. We compare the benefit-risk profile based on antiretroviral activity and safety of R1vs another drug R0. The efficacy endpoints in this example are: the proportion of patients achieving HIV RNA < 400 copies/mL, the change from baseline in HIV RNA (log10 copies/mL); and the change from baseline in CD4 cell count. The AEs of interest are diarrhea, injection site reaction, headache, nausea, alanine aminotransferase increased, aspartate aminotransferase increased, and creatine phosphokinase increase. In this program, the two studies were conducted at around the same time period as the pivotal studies and lasted a total of 240 weeks. The initial filing was performed at Week 24 with the primary time point at Week 16, and follow-up analyses were performed at Week 48, 96, 156 and 240. After Week 156, all patients in the placebo arm were allowed to switch to the treatment arms. Therefore, we omit the data after Week 156. Starting from Week 48, the pooled data showed a few but unbalanced number of malignancy cases between treatment arms and placebo. Therefore, malignancies should be added to the effect table of BRA after Week 48. To mimic the finding history in our post-hoc analysis, we add malignancy to the list of endpoints in the BRA and do a retrospective analysis by adding the malignancy into the Week 24 analysis.

In this example, we conduct the periodic BR assessment at each time when a new data lock is available to reflect the change of BR profile over the whole 156 weeks. For the two proportion metrics, we use the Bayesian hierarchical model with binomial likelihood; for the two change from baseline metrics, we use the Bayesian hierarchical model

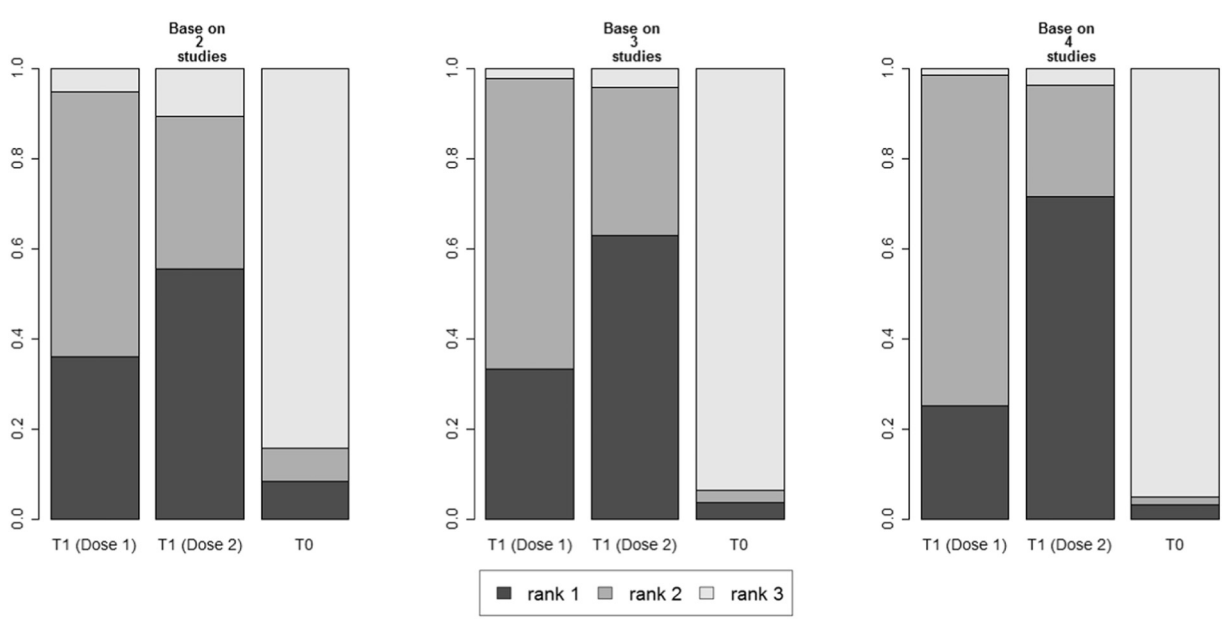

Fig. 2. The rank acceptability indices resulting from the cumulative data with weight confidence $r = 1$.

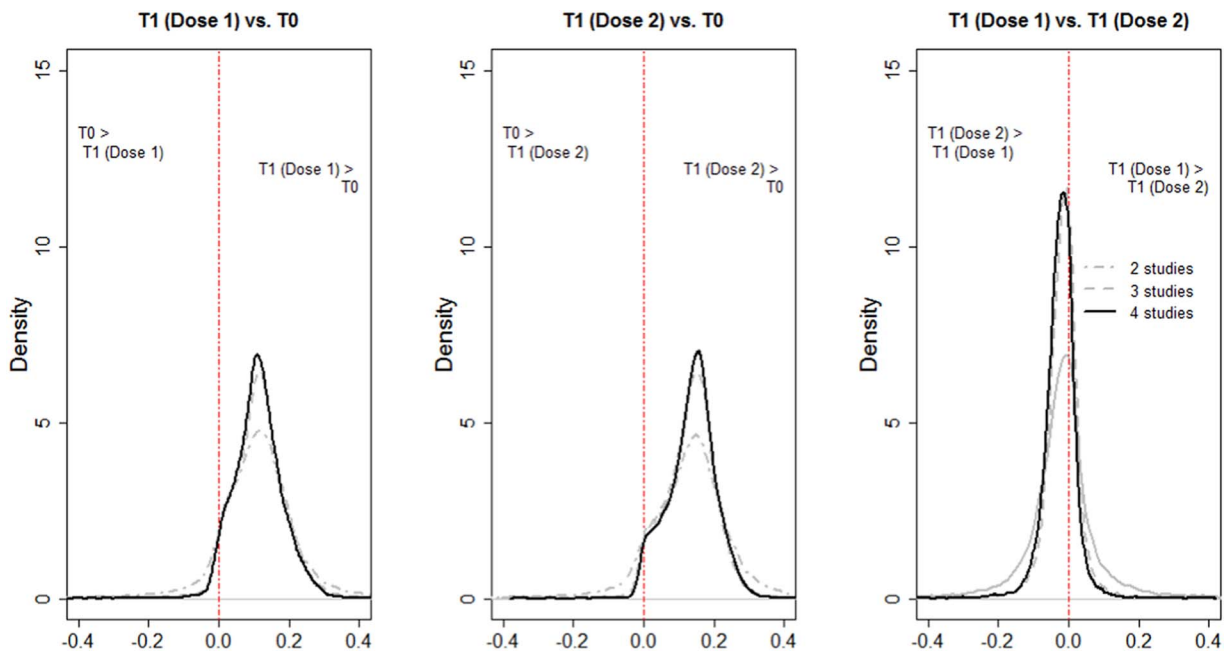

Fig. 3. Distributions of the pairwise differences in utility scores when using the same vector of parameters with $r = 1$, The distribution is updated with increasing number of studies.

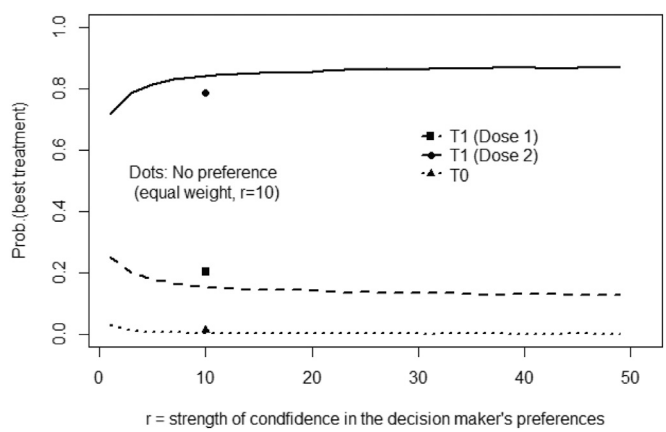

Fig. 4. Probabilities of being the best treatment with confidence r ranging from 1 to 50.

with normal likelihood; for the AE counts, we use the exposure adjusted incidence rate (number of AE per person year) as the metric, and the Bayesian hierarchical model with Poisson likelihood for the meta-analysis. Therefore, we have three different metrics which are not on the same scale. We used the partial value function to scale them into 0 to 1.

We first performed an analysis based on data at Week 24 without considering malignancies as a criterion in the model. We then added a post-hoc analysis using data at 24 week considering malignancies to reflect such finding. The weight elicitation $\boldsymbol{w}^0$ with and without taking into consideration malignancies are presented in Appendix A Table A.2. The rank acceptability indices resulting from the analysis with $r = 1$ are visualized as a bar chart in Fig. 5. We can see that the first rank acceptability for drug R1 has a slight drop from 0.88 to 0.82 when adding malignancy as a risk factor at Week 24. However, that does not impact the fact that R1 has a greater chance to be better than drug R0 over the course of the trials. The distribution of the differences in utility scores between the two alternatives over time are plotted in Fig. 6. R1 becomes more preferable as the density shift to the right as time goes.

In summary, as longer periods of data are available, the drug R1 shows greater benefit-risk balance over the drug R0 and the evidence is stronger when the follow-up time is longer as seen from the fact the PDF is shifting more to the right as the follow-up time is longer. When the malignancy was added to the effect table at Week 48, the BR balance was dropping by a certain amount for the post-hoc analysis of Week 24, but the BR balance difference is still getting greater between the treatment arm and control arm as time moves forward.

7. Discussion

Benefit-risk assessment (BRA) is important to pharmaceutical companies throughout the drug development stages. It is important to initiate the process early in the product development lifecycle in order to understand the benefit and risk profile of the new drug and make a correct Go-and-No-Go decision in the early stages. The probability of success of phase III studies is not only dependent on the efficacy of the drugs but also the adverse effect the new drugs may have. Hence, a rigorous benefit-risk assessment before Phase III is warranted for a comprehensive decision to be made. The evidence from randomized clinical trials is critical to the BRA for marketing application of new drugs and forms the basis of an approval decision by the regulatory agencies and the basis of a reimbursement policy by government or insurance companies. Collecting information on the patients who took the drug once it is marketed is also important to understand the long-term benefit-risk profile during the drug lifecycle, and to provide more evidence for regulatory agencies when evaluating the benefit and risk of the drug to the society as the drug could be exposed to a much broader population than those in the clinical trials.

In this paper, we proposed a general framework for periodical benefit-risk assessment by combining Bayesian meta-analysis and SMAA together and illustrated how to use the proposed methods to update the benefit-risk profile of a drug periodically during the development process using case studies. The quantitative BR framework proposed can contribute to making the assessment transparent, consistent, and rational, e.g., how different sources of data are integrated, how the uncertainty of information impacts an assessment, and how the conclusion is derived. Bayesian approach is especially suitable to tackle these problems. The data from early studies or literature can be used to form an informative prior (instead of non-informative prior when no historical data) in the very first meta-analysis, and posterior distribution from an early analysis will become priors for the next coming Bayesian meta-analysis when new study data is available. The uncertainty in criteria value is incorporated into the posterior distribution

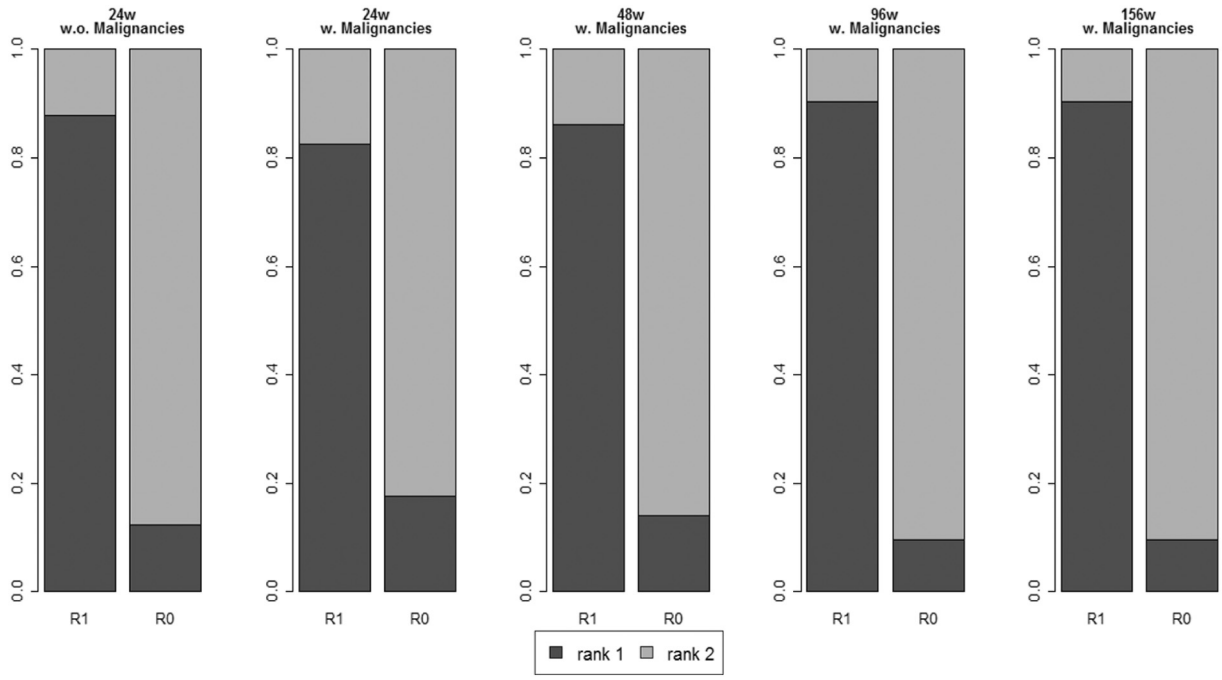

Fig. 5. The rank acceptability indices at different time points as disease evolving.

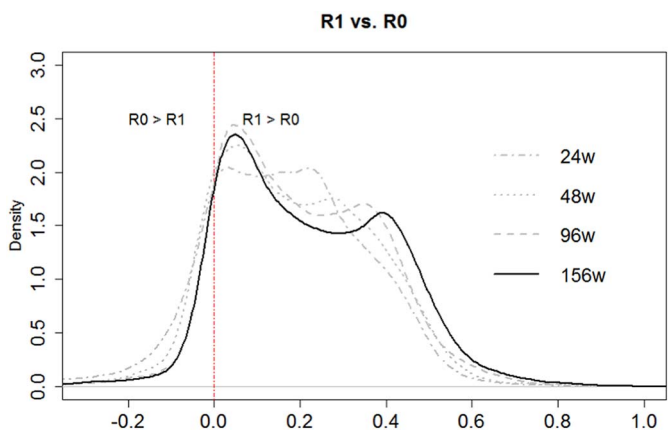

Fig. 6. Distributions of the differences in utility scores between drug R1 and drug R0. The distribution is updated over the course of the trials.

obtained from the Bayesian meta-analysis which is then inputted to SMAA model. The uncertainty of the weights is expressed through Dirichlet process with hyper parameters. Hence, it is a convenient method to integrate information from multiple sources, evaluate uncertainty of the benefit risk profile, and naturally follow the chronological structure of cumulative knowledge acquired on drugs, especially on their safety. The readily available sampling software based on a Markov chain Monte Carlo approach made it easy to implement the proposed method in terms of technique. In addition, the graphical tools are visually intuitive to interpret the result and facilitate communicating with decision makers.

The method proposed in the paper is based on summary level data from clinical trials. Whenever possible, individual-level data should be favored because it provides more information on the variability of patients' responses and the inter-correlations among all effects, thereby enabling more accurate estimates and predictions of benefit-risk profiles. In addition, the importance of including patients in regulatory BR assessment has been increasingly recognized. Patients who live with a disease are directly impacted by approval decisions and so are in a unique position to contribute to BR decisions. The efficacy and safety of a given treatment can differ from what was concluded on the basis of a population of patients in a clinical trial. And the perspective of a patient could also be different from that of regulatory perspective. The extension of proposed framework to the individual level is crucial to demonstrating the BR profile of a drug to individual patients and realization of Patient-Focused Drug Development [2].

There are several limitations in our study. First, the general assumption of independence among criteria in the SMAA may not be easily met. For example, numbers of adverse events for some selected terms could be correlated to some extent if they are in the same System Organ Class (SOC). In our examples, the correlations between criteria were small. In practice, careful thought are needed when selecting criteria for BR assessment and should take place at the planning stage of BR assessment. Multivariate MCDA may be a potential solution for such problem. Also, Bayesian methods can explicitly account for observed correlation among the different BR criteria, either through the prior distribution or through the likelihood model. For example, if the individual-level data are available, a joint posterior distribution of criteria, e.g., a multivariate normal distribution with an unstructured covariance matrix can be estimated to account for dependency among the different criteria (parallel research is ongoing for this issue). Second, although the proposed method controls the uncertainty of weight with a natural interpretation, weights elicitation still requires a substantial input from clinicians and decision makers. Weights can be elicited with the swing method, in which the decision maker is asked to judge the relative importance of the worst-best scale swings. Third, the proposed quantitative BR framework is suitable for the periodic BR reviews of medical products from an early stage of development to post-approval safety monitoring as information continues to accumulate. However, the criteria are generally not characterized in post-approval studies in as much fine-grained detail as in randomized clinical trials. The challenge of integrating the greater variety of data sources may not have a one-size-fits-all solution, but need further methodology development. As one reviewer pointed out, real world evidence (RWE) should also be included for more comprehensive assessment of the benefit risk profile of a product in post marketing setting. Although we did not include RWE in our examples, in principle, the proposed framework could be used for incorporating the RWE. One caveat is that, such data usually has potential bias and confounding. How to deal with

those biases and confounding issue in order to have a clean benefit risk assessment warrants more research.

In summary, the proposed approach could be an important complement to the qualitative BR framework. It presents an opportunity to guide and inform decisions for medical products rather than to dictate them. The actual decision making process by various stakeholders is usually more complex, and is affected by other considerations such as social and economic aspects. These factors may not directly influence a product's benefit-risk profile for regulatory decision-making, but may still have a profound impact on the acceptance of the decision by the patient, public and other stakeholders like the HTA agencies.

Disclaimer

All authors declare no conflict of interest.

Acknowledgements

Sheng Luo's research was supported by the National Institute of Neurological Disorders and Stroke under Award Number R01NS091307.

Appendix A

Table A.1
Weight vector $\boldsymbol{w}^0$ elicited by swing weighting method in drug T1 example.

	Criteria	Weight
Efficacy	Pain free	0.18
	Pain relief	0.16
	Absence of phonophobia	0.15
	Absence of photophobia	0.15
	Absence of nausea	0.16
Safety	Dry mouth	0.03
	Dizziness	0.05
	Fatigue	0.05
	Somnolence	0.03
	Nausea	0.04

Table A.2
Weight vector $\boldsymbol{w}^0$ elicited by swing weighting method in drug R1 example.

	Criteria	Weight (without Malignancies)	Weight (with Malignancies)
Efficacy	Proportion of patients achieving HIV RNA < 400 copies/mL at Week 48	0.25	0.21
	Change from baseline in HIV RNA (log10 copies/mL)	0.07	0.06
	Change from baseline in CD4 cell count	0.24	0.20
Safety	Diarrhea	0.10	0.08
	Injection site reaction	0.09	0.07
	Headache	0.08	0.06
	Nausea	0.10	0.08
	Alanine aminotransferase increased	0.03	0.02
	Aspartate aminotransferase increased	0.02	0.02
	Creatine phosphokinase Increase	0.02	0.02
	Malignancies	–	0.19

References

[1] European Medicines Agency, European Medicines Agency Benefit-risk Methodology Project, Work Package 4 [Internet], Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Report/2012/03/WC500123819.pdf, .
[2] Food and Drug Administration, Structured Approach to Benefit-Risk Assessment in Drug Regulatory Decision-Making [Internet], Available from: https://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm329758.pdf, .
[3] International Conference of Harmonization, Common Technical Document: Revision of M4E Guideline on Enhancing the Format and Structure of Benefit-Risk Information in ICH [Internet], Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/CTD/M4E_R2_Efficacy/M4E_R2_Step_4.pdf, .
[4] S. Mt-Isa, C.E. Hallgreen, N. Wang, T. Callréus, G. Genov, I. Hirsch, et al., Balancing benefit and risk of medicines: a systematic review and classification of available methodologies, Pharmacoepidemiol. Drug Saf. 23 (7) (2014 Jul 1) 667–678.
[5] F. Mussen, S. Salek, S. Walker, A quantitative approach to benefit-risk assessment of medicines – part 1: the development of a new model using multi-criteria decision analysis, Pharmacoepidemiol. Drug Saf. 16 (S1) (2007 Jul 1) S2–15.
[6] T. Tervonen, G. van Valkenhoef, E. Buskens, H.L. Hillege, D.A. Postmus, Stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis, Stat. Med. 30 (12) (2011 May 30) 1419–1428.
[7] E. Waddingham, S. Mt-Isa, R. Nixon, D. Ashby, A Bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment, Biom. J. 58 (1) (2016 Jan 1) 28–42.
[8] G. Saint-Hilary, S. Cadour, V. Robert, M. Gasparini, A simple way to unify multi-criteria decision analysis (MCDA) and stochastic multicriteria acceptability analysis (SMAA) using a Dirichlet distribution in benefit–risk assessment, Biom. J. 59 (3) (2017 May 1) 567–578.
[9] H. Ma, Q. Jiang, C. Chuang-Stein, S.R. Evans, W. He, G. Quartey, et al., Considerations on endpoint selection, weighting determination, and uncertainty evaluation in the benefit-risk assessment of medical product, Stat. Biopharm. Res. 8 (4) (2016 Oct 1) 417–425.
[10] T. Tervonen, H. Naci, G. van Valkenhoef, A.E. Ades, A. Angelis, H.L. Hillege, et al., Applying multiple criteria decision analysis to comparative benefit-risk assessment: choosing among statins in primary prevention, Med. Decis. Mak. 35 (7) (2015 Oct 1) 859–871.
[11] M.D. Hoffman, A. Gelman, The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo, J. Mach. Learn. Res. 15 (1) (2014 Apr 1) 1593–1623.
[12] D. Ashby, A.F. Smith, Evidence-based medicine as Bayesian decision-making, Stat. Med. 19 (23) (2000 Dec 15) 3291–3305.
[13] P. Arlett, R. Postigo, H. Janssen, A. Spooner, Periodic benefit-risk evaluation report:

a European Union regulatory perspective, Pharm. Med. 28 (6) (2014 Dec 1) 309–315.
[14] M.R. Warner, A.M. Wolka, R.A. Noel, Implementing benefit-risk assessment for the periodic benefit-risk evaluation report, Ther. Innov. Reg. Sci. 50 (3) (2016 May) 342–346.
[15] Food and Drug Administration (FDA), E2C(R2) Periodic Benefit-Risk Evaluation Report (PBRER) Guidance for Industry, https://www.fda.gov/downloads/drugs/guidances/ucm299513.pdf dated July 2016, Accessed July 24, 2017.
[16] International Conference on Harmonisation, ICH Harmonised Tripartite Guideline, Periodic Benefit-Risk Evaluation Report (PBRER) E2C (R2), http://www.ich.org/fileadmin/PublicWeb_Site/ICH_Products/Guidelines/Efficacy/E2C/E2C_R2 Step4.pdf, (2012) dated 17 December 2012, Accessed July 24, 2017.
[17] R. Lahdelma, J. Hokkanen, P. Salminen, SMAA-stochastic multiobjective acceptability analysis, Eur. J. Oper. Res. 106 (1) (1998 Apr 1) 137–143.
[18] R. Lahdelma, P. Salminen, SMAA-2: stochastic multicriteria acceptability analysis for group decision making, Oper. Res. 49 (3) (2001 Jun) 444–454.
[19] V. Belton, T. Stewart, Multiple Criteria Decision Analysis: An Integrated Approach, Springer Science & Business Media, 2002.