Published in final edited form as:
J Biopharm Stat. 2015 ; 25(3): 508–524. doi:10.1080/10543406.2014.921514.

A Counterfactual P-value Approach for Benefit-Risk Assessment in Clinical Trials

Donglin Zeng[*], **Ming-Hui Chen**[†], **Joseph G. Ibrahim**[*], **Rachel Wei**[‡], **Beiying Ding**[‡], **Chunlei Ke**[‡], and **Qi Jiang**[‡]

[*]Department of Biostatistics, University of North Carolina, McGavran Greenberg Hall, CB#7420, Chapel Hill, NC 27599, USA

[†]Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, CT 06269, USA

[‡]Global Biostatistical Science, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA

Summary

Clinical trials generally allow various efficacy and safety outcomes to be collected for health interventions. Benefit-risk assessment is an important issue when evaluating a new drug. Currently, there is a lack of standardized and validated benefit-risk assessment approaches in drug development due to various challenges. To quantify benefits and risks, we propose a counterfactual p-value (CP) approach. Our approach considers a spectrum of weights for weighting benefit-risk values and computes the extreme probabilities of observing the weighted benefit-risk value in one treatment group as if patients were treated in the other treatment group. The proposed approach is applicable to single benefit and single risk outcome as well as multiple benefit and risk outcomes assessment. In addition, the prior information in the weight schemes relevant to the importance of outcomes can be incorporated in the approach. The proposed counterfactual p-values plot is intuitive with a visualized weight pattern. The average area under CP (AUCP) and *preferred probability* over time are used for overall treatment comparison and a bootstrap approach is applied for statistical inference. We assess the proposed approach using simulated data with multiple efficacy and safety endpoints and compare its performance with a stochastic multi-criteria acceptability analysis (SMAA) approach.

Keywords

Area under the CP-region; Benefit risk assessment; Counterfactual p-value; Preferred probability; Prior distribution

1 Introduction

Evaluation of balance between benefits and risks is fundamental in development, registration and use of drugs. Risk-benefit assessment (RBA) is generally considered challenging and has received considerable attention from regulatory agencies, governance bodies, patients and industry. The US Food and Drug Administration (FDA) internally piloted a framework with the intention to provide a standard RBA framework. In Europe, the Committee for Medicinal Products for Human Use (CHMP) had a comprehensive review of available qualitative and quantitative methods and processes for regulatory RBA in their working report in 2010 and 2012 [1,2]. Pharmaceutical Research and Manufacturers of America (PhRMA) developed a benefit risk action team (BRAT) framework to enable a structured and transparent approach to Benefit-risk assessment [3] and then transferred the BRAT framework to the Centre for Innovation in Regulatory Science, Ltd. (CIRS) in order to further the technical development and broaden the input from the scientific community.

Multiple methods are available to quantify benefits and risks of new drugs [4]. Drug RBA typically includes multiple benefit and risk criteria. In this setting, multi-criteria decision analysis (MCDA) was proposed to provide a framework for systematic analysis of complex decision problems involving value trade-off [5]. This approach constructs a multi-criteria decision model for benefits and risks, and quantifies them into some summarized risk-benefit scores. However, MCDA only provides the point estimate of the score for combined benefits and risks. Therefore, the uncertainty associated with sampling variation of criteria measurements is not incorporated. In addition, the approach requires specifying the weight for each criterion which involves subjective judgment; the decision makers may not reach a consensus about the weights [5]. To overcome the limitations of MCDA, a stochastic multi-criteria acceptability analysis (SMAA) approach [6-11] has been proposed. The SMAA-2 method [7,8] extends the original SMAA [6] by considering all ranks in the analysis. The SMAA methodology has been applied to risk assessment [9-10] and SMAA-2 has been applied to drug RBA [11]. The SMAA approach [11] considers a multi-criteria decision problem and quantifies the decision uncertainty through descriptive measures calculated as multidimensional integrals over stochastic parameter spaces to aid in decision making. The weights in the SMAA approach are random variables comparing to elicited weights in traditional approaches. The SMAA approach can incorporate the sampling variation in criteria measurements and characterize the benefits and risks using **different** prior distributions for weights. The decision making in the SMAA approach is mainly done through central weight vectors and confidence factors if there is no preference information, and through rank acceptability indices to find the best alternative when preference information is incorporated. However, the ranking approach in SMAA may not be statistically efficient because of disregarding the quantitative values of the benefits and risks. In addition, there is a lack of graphical representation of the risk-benefit assessment over the weight selection in the MCDA approaches including the SMAA.

In this paper, we propose a new concept, called the counterfactual p-value (CP), to quantify benefit-risk balance when comparing two treatment plans. The proposed concept will automatically incorporate prior importance of weighing each benefit or risk endpoint. We will further propose a graphical display of this concept which shows different weighting schemes for benefit-risk analysis. Some summary measures based on the graph will be used for comparison. Finally, the proposed method will be used to analyze a simulated dataset.

2 Method

2.1 Counterfactual P-value

We consider two treatment plans (A vs B). For plan A, we define a multivariate benefit-risk value as $\mathbf{u}_A = (u_{1A}, \ldots, u_{KA})^T$, where larger values are associated with better outcomes; similarly for plan B, we define a multivariate benefit-risk value $\mathbf{u}_B = (u_{1B}, \ldots, u_{KB})^T$. Note that both $\mathbf{u}_A$ and $\mathbf{u}_B$ refer to some concatenations of benefit and risk attributes. From a given study, we can estimate these values, and we let $(\hat{u}_{1A}, \ldots, \hat{u}_{KA})^T$ denote the corresponding random vector for the estimated value in plan A and $(\hat{u}_{1B}, \ldots, \hat{u}_{KB})^T$ is the random vector for the estimated value in plan B. Usually $(\hat{u}_{1A}, \ldots, \hat{u}_{KA})^T$ and $(\hat{u}_{1B}, \ldots, \hat{u}_{KB})^T$ can be transformations of the endpoint outcomes using value functions [3], where the transformations are used to make these concatenated benefit-risk values comparable. We assume $(\hat{u}_{1A}, \ldots, \hat{u}_{KA})^T \sim N(\mathbf{u}_A, \mathbf{\Sigma}_A/n_A)$ and $(\hat{u}_{1B}, \ldots, \hat{u}_{KB})^T \sim N(\mathbf{u}_B, \mathbf{\Sigma}_B/n_B)$, where n_A and n_B denote the group sizes of treatment A and treatment B, respectively. Note that the normality assumption holds approximately in large sample sense but this assumption is not essential in our method and it can be replaced by any parametric distribution.

For any given weight $\mathbf{w} = (w_1, \ldots, w_K)^T$, where $\sum_{k=1}^{K} w_k = 1$, we define $v_A(\mathbf{w}) = w_1 u_{1A} + \ldots + w_K u_{KA}$ and $v_B(\mathbf{w}) = w_1 u_{1B} + \ldots + w_K u_{KB}$ as weighted combinations of benefit-risk values. Correspondingly, we obtain $\hat{v}_A(\mathbf{w}) = w_1 \hat{u}_{1A} + \ldots + w_K \hat{u}_{KA}$ and $\hat{v}_B(\mathbf{w}) = w_1 \hat{u}_{1B} + \ldots + w_K \hat{u}_{KB}$. Then, we define a probability for plan A associated with $\mathbf{w}$ as

$$p_A(\mathbf{w}) = \Pr(\text{Observing value } \hat{\nu}_A(\mathbf{w}) \text{ or larger if subjects were treated with plan B});$$

similarly, we define a probability for plan B associated with $\mathbf{w}$ as

$$p_B(\mathbf{w}) = \Pr(\text{Observing value } \hat{\nu}_B(\mathbf{w}) \text{ or larger if subjects were treated with plan A}).$$

We call both probabilities as *counterfactual p-values* for two reasons: (1) treating subjects already in one plan with the other plan never happens so the scenarios in both definitions are counterfactual; (2) the probability in $p_A(\mathbf{w})$ is similar to a p-value for testing the null hypothesis $H_0 : v_A(\mathbf{w}) = v_B(\mathbf{w})$ against $H_a : v_A(\mathbf{w}) > v_B(\mathbf{w})$. According to the definition of the counterfactual p-values, if plan A is better than plan B, then we would expect $p_A(\mathbf{w})$ to be more likely less than 1/2 and $p_B(\mathbf{w})$ larger than 1/2; if plan A is equivalent to plan B, then both probabilities will be equal to 1/2 on average.

Under the normality assumption, we know that if all n_A subjects in treatment plan A were treated in plan B, the group mean of the weighted values should have an approximate normal distribution with mean $\hat{v}_B(\mathbf{w})$ and variance $\mathbf{w}^T \hat{\mathbf{\Sigma}}_B \mathbf{w}/n_A$. Therefore, we can actually estimate these counterfactual p-values as

$$\begin{aligned}\hat{p}_A(\mathbf{w}) &= 1-\Phi\left((\hat{\nu}_A(\mathbf{w})-\hat{\nu}_B(\mathbf{w}))/\sqrt{\mathbf{w}^T\hat{\mathbf{\Sigma}}_B\mathbf{w}/n_A}\right),\\ \hat{p}_B(\mathbf{w}) &= 1-\Phi\left((\hat{\nu}_B(\mathbf{w})-\hat{\nu}_A(\mathbf{w}))/\sqrt{\mathbf{w}^T\hat{\mathbf{\Sigma}}_A\mathbf{w}/n_B}\right),\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

2.2 Graphical display of counterfactual p-values

For any weight vector $\mathbf{w} \in \Omega$, where Ω is the feasible region of weights incorporating prior weight information, we can calculate the pair $\left(\hat{p}_B(\mathbf{w}), \hat{p}_A(\mathbf{w})\right)$. We then plot $\left(\hat{p}_B(\mathbf{w}), \hat{p}_A(\mathbf{w})\right)$ for all feasible $\mathbf{w}$. We call such a plot a CP region. Therefore, we expect that if plan B is better than plan A, then most of the points should be in the quadrant $\{(x, y) : x < 0.5, y > 0.5\}$.

One major question is how to choose the weight vector. When no prior information on the importance of each endpoint is available, one non-informative choice is to sample the weight vector uniformly from its feasible space. However, if one knows apriori the relative importance and denotes it as a prior weight vector $\mathbf{w}_0$, then one possibility is to sample the weights $\mathbf{w}$ from a Dirichelet distribution with parameter $\mathbf{w}_0$. See [5, 6] for discussions on approaches to determining the prior weight. In another scenario, when the prior importance is not known exactly but the importance order of the endpoints is known, then a uniform distribution from the ordered region can be used to sample weights. The following algorithm is thus given to obtain the CP-region:

1. Using the raw data, we estimate $(\hat{u}_{1A}, \ldots, \hat{u}_{KA})$ and their estimated covariance matrix $\hat{\mathbf{\Sigma}}_A/n_A$; similarly, we estimate $(\hat{u}_{1B}, \ldots, \hat{u}_{KB})$ and their estimated covariance matrix $\hat{\mathbf{\Sigma}}_B/n_B$.
2. We sample a vector $\mathbf{w}$ from a prior distribution $f_W(\mathbf{w})$.
3. We obtain $\hat{\upsilon}_A(w)=\sum_{k=1}^{K} w_k\hat{u}_{kA}$ and $\hat{\upsilon}_B(w)=\sum_{k=1}^{K} w_k\hat{u}_{kB}$.
4. We calculate the CP-values

$$\hat{p}_A(\mathbf{w})=1-\Phi\left((\hat{\nu}_A(\mathbf{w})-\hat{\nu}_B(\mathbf{w}))/\sqrt{\mathbf{w}^T\hat{\mathbf{\Sigma}}_B\mathbf{w}/n_A}\right)$$

   and

$$\hat{p}_B(\mathbf{w})=1-\Phi\left((\hat{\nu}_B(\mathbf{w})-\hat{\nu}_A(\mathbf{w}))/\sqrt{\mathbf{w}^T\hat{\mathbf{\Sigma}}_A\mathbf{w}/n_B}\right).$$

5. Repeat Steps 2-4 many times and then plot the calculated $\hat{p}_A(\mathbf{w})$ versus $\hat{p}_B(\mathbf{w})$.

2.3 Example

Motivated by a phase III clinical trial of an experimental oncology drug, we simulated a dataset including 2 efficacy endpoints, progression free survival (PFS) and overall survival (OS) time, and time to the first occurrence of 2 adverse events of interest (AE1 and AE2) for 500 patients each on two treatment arms, placebo (plan A) or active treatment (plan B). The simulated data assumes a treatment effect on the progression free survival and a similar overall survival outcome between the two treatment arms. Among the two AE's of interest, the treatment increased the incidence of the first adverse event and was associated with the occurrence of another adverse event (ie, 0% incidence for the placebo arm). We use the survival probabilities of these four endpoints at month 10 for illustration. We estimate these probabilities using the Kaplan-Meier estimates and estimate the covariance matrices using the bootstrap. The estimated probabilities and their estimated covariance matrices $\left(\hat{\mathbf{\Sigma}}_A/n_A, \hat{\mathbf{\Sigma}}_B/n_B\right)$ for the two treatment arms are given in Table 1. For example, at month 10, the proportions of the patients who did not have disease progression are similar in the two treatment groups (90.5% for A and 91.4% for B). The mortality rate within 10 months is **significantly** lower in treatment B but the adverse event rates are higher.

In this example, some prior information on the relative importance of these four endpoints is available and the prior weights after consulting clinicians are 0.48, 0.29, 0.01 and 0.22 for PFS, OS, AE1 and AE2 respectively. Therefore, to produce the proposed CP-region, we generate weights from the Dirichlet distribution with parameter vector $\mathbf{w}_0$ = (0.48, 0.29, 0.01, 0.22) so that the average weights are exactly the same as their relative importance. We randomly draw 5,000 weights from this distribution and calculate the CP-values for both treatments. For each draw of weight, we plot the derived CP values and the obtained points are given in Figure 1. To reflect the weight contribution for each plotted point, we add a vertical line from each point to *x*-axis where the vertical line has four colors, with the length of each color segment reflecting the weight of the corresponding endpoint from this particular draw. Furthermore, since the covariances for the endpoints are similar between the two groups, $\hat{p}_A(\mathbf{w}) \approx 1 - \hat{p}_B(\mathbf{w})$ and the points are close to a diagonal line as seen in Figure 1.

In Figure 1, the upper-left square contains all the draws where the counter-factual p-value for treatment B is smaller than the one for treatment A; the bottom-right square contains the opposite. From the expression of the CP values, if $\hat{p}_A(\mathbf{w}) > 0.5$ then $\hat{v}_A(\mathbf{w}) - \hat{v}_B(\mathbf{w}) < 0$ so $\hat{p}_B(\mathbf{w}) < 0.5$. The opposite is also true. So there are no points in either the upper-right and the lower-left regions. From this figure, the general message is that if OS or PFS is weighted most, then the p-value for treatment B is smaller than that of treatment A; in contrast, if adverse event 2 is weighted most, then the p-value for treatment A is smaller. This is consistent with the estimates provided in Table 1. In the plot, we also present some summary statistics including the area under the CP region, the B-preferred probability, etc. Definitions and details of the calculation of these quantities are given in the next section.

2.4 Statistics based on CP-values

The area under the CP region (AUCP) is defined as the signed area from the plotted CP-values, where the measure along the horizontal axis is taken as the probability measure corresponding to p_B's distribution. Formally, we define it as

$$AUCP=\int \left(p_A\left(\mathbf{w}\right)-0.5\right)dF_B\left(\mathbf{w}\right),$$

where $F_B(\mathbf{w})$ denotes the distribution of $p_B(\mathbf{w})$ given the data. Empirically, we can estimate AUCP as

$$\frac{1}{m}\sum_{k=1}^{m}\left[p_A\left(\mathbf{w}_k\right)-0.5\right],$$

where $\mathbf{w}_1, \ldots, \mathbf{w}_m$ are m random draws from the prior distribution of $\mathbf{w}$. The AUCP not only depends on the frequency of one preferring plan A, but also depends on how large the weighted value for plan A is as compared to plan B. The larger the AUCP is than zero, the more likely we will see the plotted p-values to appear in the upper-left region and the more treatment plan B is favored. Another summary quantity, called the B-preferred probability (BPP), is simply the proportion that $p_A(\mathbf{w}) > p_B(\mathbf{w})$, i.e., $\Pr(p_A(\mathbf{w}) > p_B(\mathbf{w})|\text{Data}) = \Pr(p_B(\mathbf{w}) < 0.5|\text{Data})$. BPP can be estimated as the proportion of the pairs within the upper-left region. The larger the BPP is above 0.5, the more likely we will see small counter-factual p-value for treatment B.

Although the color pattern in Figure 1 indicates how importance of each endpoint influences the p-value, one may be interested in the overall or central weight in either B-preferred region or A-preferred region. For this purpose, as indicated in the upper-right corner of Figure 1, we present the average weights in either region to summarizing the weight information in the region of preferring A or preferring B. These weights are closely related to the central weights in the SMAA approach but are not equivalent. This clearly indicates that (a) if we give more weights to OS or PFS, then treatment B tends to have a better benefit-risk balanced value than treatment A; (b) if we weight adverse event 2 the most, then treatment A is preferred (note that treatment A has 0 such events); (c) the endpoint of adverse event 1 seems to have little influence on which treatment to be preferred because minimal prior weight was given to it.

Both AUCP and BPP are calculated conditional on the given data. Statistically, it will be interesting to know whether AUCP is significantly larger than zero or whether BPP is larger than 0.5. In other words, we need to account for data randomness to make appropriate inference based on these statistics. Note that AUCP and BPP both depend on survival estimates, and their covariance estimates which are assumed to be asymptotically normal. The delta method then implies that these statistics will be also asymptotically normal. However, estimating their asymptotic covariance matrix is difficult due to complex variance estimation of the estimated covariance matrices in the counterfactual p-values. Therefore, we suggest to use the bootstrap to obtain the inference for these two statistics. For the example, the bootstrap result indicates a significant preference for treatment B between time 5 to 15 months; that is, AUCP>0 and BPP>0.5 in that interval.

2.5 Generalization to time-dependent endpoints

Many endpoints such as the four endpoints in the above example are temporal so it will be interesting and important to incorporate a time component in a benefit-risk analysis. The definition of the counterfactual p-values can be easily generalized to time-dependent endpoints by allowing both values and their covariance matrices to depend on time *t*. Thus, we obtain a time-dependent AUCP(t) and BPP(t), which can be unsmooth due to the use of the Kaplan-Meier estimates of the survival probabilities. Similarly, the bootstrap method can be used to obtain their pointwise confidence bands. For the example, the estimates of these functions and their confidence bands are plotted in Figure 2. One way to summarize time-dependent AUCP or BPP curves is to do a weighted integration of the curves, where the weights can be chosen to be time-dependent to reflect either clinical importance or study sample variability over time.

3 Sensitivity Analysis

We conduct additional numerical studies to examine the performance of the proposed method using the same data example as before. Specifically, we study the sensitivity of the CP method to the choice of prior distributions for **w**. We also compare the CP method with the SMAA method proposed in [11].

We consider generating the weights from a prior distribution which is a mixture of the original Dirichlet distribution and a uniform distribution with varying mixing probabilities. For each given prior distribution, we calculate the probabilities of preferring treatment B based on the four endpoints at each time point *t* using the proposed counterfactual approach and the SMAA method. The obtained preference probability curves are given in Figure 3.

From Figure 3, it is clear that the preference curves of treatment B have very similar shapes and patterns between our approach and the SMAA method. However, our method tends to give a larger preference probability for treatment B, mainly due to the fact that our approach uses more data information in addition to the ranks in the combined endpoints. These observations are consistent no matter what mixing probabilities are used in the prior distribution for **w**. As indicated in Figure 3 (a)-(d), the preference probability for treatment B appears to be sensitive to the choice of the prior distribution for **w**. Furthermore, we obtain the center weights of preferring treatment B for each endpoint, which are defined as the average of **w** that yields the preference to treatment B. The center weights for the four endpoints are plotted over time in Figure 4 and they look similar between the two approaches.

For illustration, we also carry out the calculation using the uniform prior for all the weights, i.e., there is no preference as to which endpoint is more important. The obtained preference curves over time and center weights are given in Figure 5. The conclusions are similar to Figure 4 but the differences between our approach and the SMAA are very little. The latter fact also indicates that the decision of preferring one treatment plan over the other can be largely influenced by the prior distribution of **w**.

Finally, we perform simulation studies to examine the operating characteristics of the proposed CP-value method. In the simulation study, we generate two survival endpoints T_1 (overall survival time) and T_2 (time to adverse event) from a gamma frailty model where the frailty distribution has mean one and variance σ^2. There are 500 subjects in treatment A group and the same number subjects in treatment B group. Let r_1 denote the hazard ratio between B and A for T_1 and let r_2 denote the hazard ratio for T_2. We consider different scenarios including (1) there is no difference between treatment A and B ($r_1 = r_2 = 1$) ; (2) treatment B moderately prolongs overall survival but increases the risk of adverse event ($r_1 = 0.75$, $r_2 = 1.25$); (3) treatment B significantly prolongs overall survival but increases the risk of adverse event ($r_1 = 0.5$, $r_2 = 1.25$). The prior weights for T_1 and T_2 are set to be (0.78, 0.22). For each simulation, we compare the proposed method to the SMAA method. We also calculate the ratios between the number needed to benefit and the number needed to be harm (NNTB/NNTH) and report their standard deviations in parentheses. The results from 500 replicates are given in Table 2. The tables show that the CP method is comparable to the SMAA method but the preference probability for treatment B is more sensitive to the benefit of treatment B vs A. Both the CP method and the SMAA method are not sensitive to the correlation between the two survival endpoints. Instead, the NNTB/NNTH ratios is very variable.

4 Concluding Remarks

We have proposed a graphical method based on different counterfactual p-values to quantify the benefit-risk assessment among different treatments. The proposed method has the following advantages: (1) it incorporates more quantitative information in data, such as the actual mean and variability of the values from all endpoints, than the rank-based methods; (2) the proposed CP region provides an intuitive summary of the weights associated with the preference of one treatment over the other; (3) compared to existing multi-criteria decisions, the proposed method provides a spectrum of weights in benefit-risk analysis thus will be useful for investigators to understand a complete picture of how weights characterizes benefits and risks balance; (4) the area under the CP (AUCP) and the B preference probability (BPP) have been proposed and they can be easily used for comparing treatments; and (5) we have also proposed a temporal CP region which enables us to examine the change of benefits and risks trade-off over time. As a note, we do not claim the superiority of the proposed approach to existing methods due to the use of a single data set and the lack of gold standard in benefit-risk analysis. Instead, one highlight of the proposed method is its connection to causal inference, the flexibility of incorporating prior weight information, extension to incorporate temporal relationship and graphical visualization.

One limitation of the proposed method is that it is only limited to the comparisons between two treatments. When multiple treatments are present, one possibility is to conduct pair-wise comparison between any two treatments. Inference then should account for multiple comparisons in the procedure. In our numerical studies, the prior distribution for **w** does not affect the temporal pattern of preference but it can change the preference probability of one treatment from 80% to 50%. Therefore, choosing an appropriate prior distribution for **w** remains an important issue, for which obtaining opinion from clinicians or experts is essential. Finally, the proposed CP approach and its associated statistics can easily be implemented in R. We have provided general R code for implementing this procedure in the appendix.

The proposed method can also be generalized to compare two competing treatments from different studies, but **not** without some cautions. The validity of the method relies on the key assumption that the patients in one study would have performed similarly as the patients in the other study if they were given the other study's treatment. Obviously, this assumption may not be valid when the two studies recruit patients from different sources. One possible way to alleviate this issue is to apply the approach similar to inferring causal effect in observational studies, for instance, using propensity scores matching so that the matched patients from these two different studies are comparable.

Acknowledgment

We thank Dr. Steve Snapinn at the Amgen Inc. for encouragement of this joint project and helpful comments. We also thank Ms. Ja-an Lin for helping to program the algorithms and generate all the plots.

References

1. Committee for Medicinal Products for Human Use (CHMP). Report of the CHMP working group on benefit-risk assessment models and methods. 2010. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2010/01/WC500069668.pdf
2. Committee for Medicinal Products for Human Use (CHMP). Report of the CHMP working group on benefit-risk assessment tools and processes. 2012. Available from: http://www.ema.europa.eu/ocs/en_GB/document_library/Report/2012/03/WC500123819.pdf
3. Coplan PM, Noel RA, Levitan BS, Ferguson J, Mussen F. Development of a Framework for Enhancing the Transparency, Reproducibility and Communication of the Benet-Risk Balance of Medicines. Clinical pharmacology & Therapeutics. 2011; 89(2):312–315. [PubMed: 21160469]
4. Guo JJ, Pandey S, Doyle J, Bian B, Lis Y, Raisch D. A review of quantitative risk-benefit methodologies for assessing drug safety and efficacy - Report of the ISPOR Risk-Benefit Management Working Group. Value in Health. 2010; 13(5):657–666. [PubMed: 20412543]
5. Mussen F, Salek S, 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(Suppl. I):S12–S15.
6. Lahdelma R, Hokkanen J, Salminen P. SMAA-stochastic multiobjective acceptability analysis. European Journal of Operational Research. 1998; 106(1):137–143. DOI: 10.1016/S0377-2217(97)00163-X.
7. Lahdelma R, Salminen P. SMAA-2: stochastic multicriteria acceptability analysis for group decision making. Operations Research. 2001; 49(3):444–454. DOI: 10.1287/opre.49.3.444.11220.
8. Tervonen T, Figueira JR. A survey on stochastic multicriteria acceptability analysis methods. Journal of Multi-Criteria Decision Analysis. 2008; 15(1-2):1–14. DOI: 10.1002/mcda.407.
9. Tervonen T, Linkov I, Steevens J, Chappell M, Figueira JR, Merad M. Risk-based classification system of nanomaterials. Journal of Nanoparticle Research. 2009; 11(4):757–766. DOI: 10.1007/s11051-008-9546-1.
10. Tervonen T, Figueira JR, Lahdelma R, Almeida DJ, Salminen P. A stochastic method for robustness analysis in sorting problems. European Journal of Operational Research. 2009; 192(1):236–242. DOI: 10.1016/j.ejor.2007.09.008. Forest Science. 49(6):928–937.

11. Tervonen T, van Valkenhoef G, Buskens E, Hillege HL, Postmus D. A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. Statistics in Medicine. 2011; 30(12):1419–28. DOI: 10.1002/sim.4194. [PubMed: 21268053]

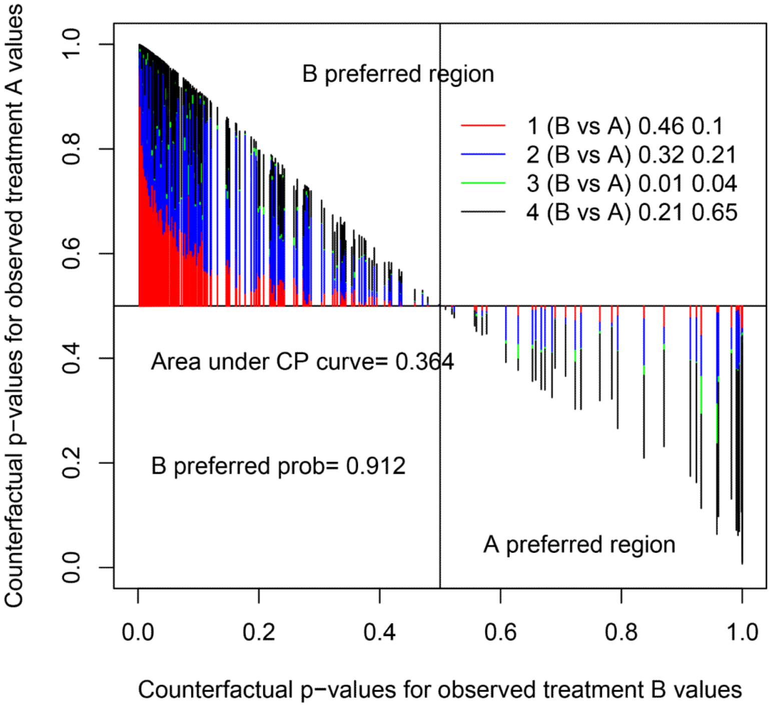

Figure 1.

Counterfactual p-value region at time 10 for the example: 1–progression free survival; 2–overall survival; 3–adverse event 1; 4–adverse event 2.

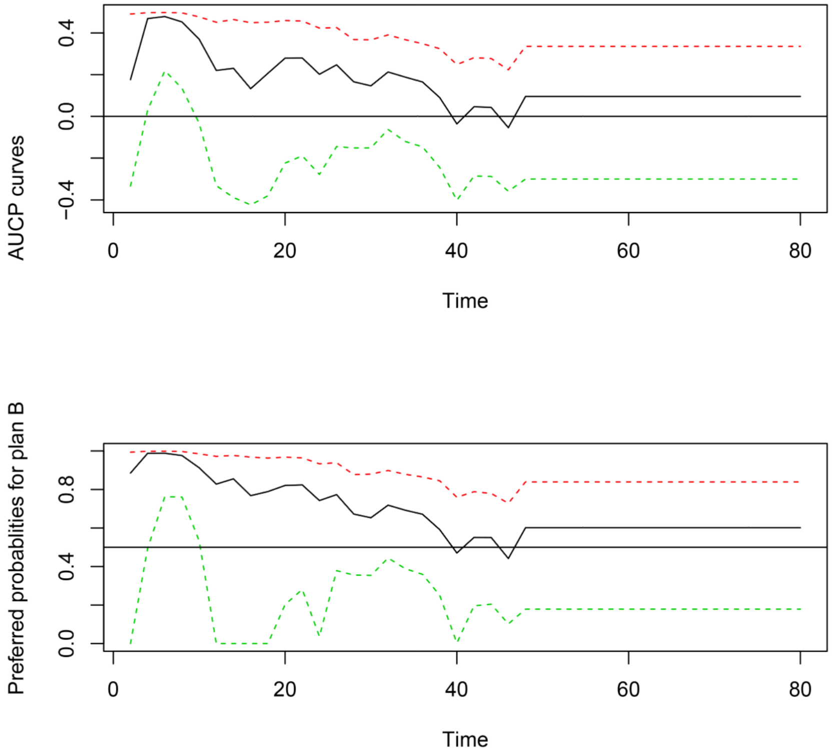

Figure 2.

AUCP and BPP curves based on Counterfactual p-value regions over time in the example: the solid curves are the estimates and the dashed ones are 95% pointwise confidence bands.

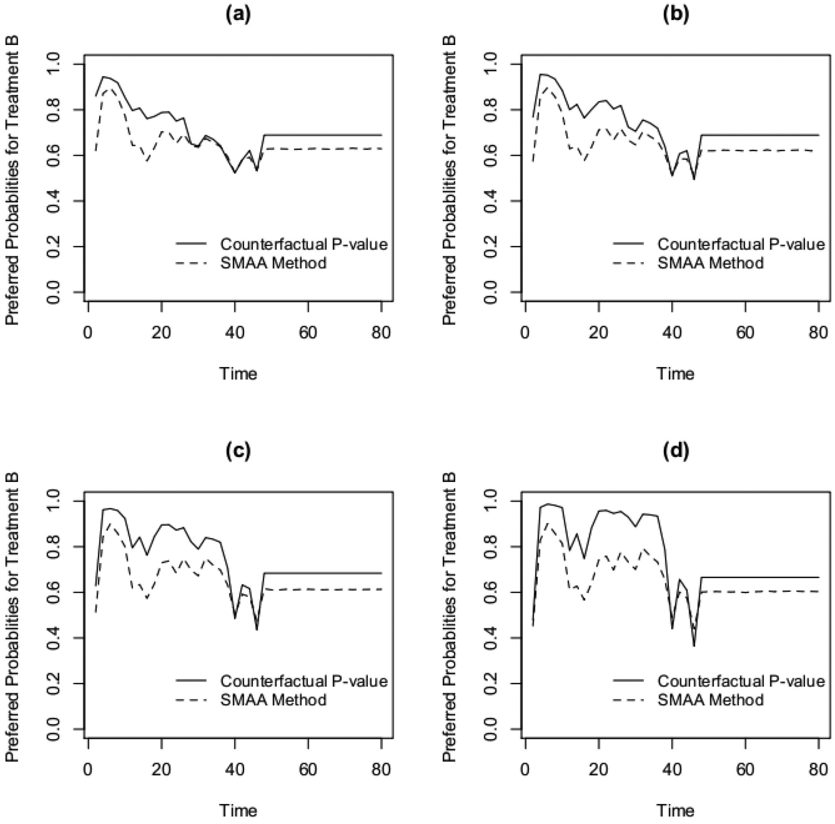

Figure 3.

The curves of preference probabilities for treatment B under different prior distributions for **w**: (a) the prior distribution is the Dirichlet distribution for **w** with parameters (0.48, 0.29, 0.01, 0.22); (b) the prior distribution for **w** is a mixture of the Dirichlet distribution in (a) and the uniform distribution in the space $\left\{(w_1, w_2, w_3, w_4) : \sum_{k=1}^{4} w_k = 1, w_1 \geq w_2 \geq w_4 \geq w_3 \geq 0\right\}$ with the mixing probabilities equal to (0.8, 0.2); (c) the distribution for **w** is the same as (b) except that the mixing probabilities are (0.5, 0.5); (d) the distribution for **w** is the same as (b) except that the mixing probabilities are (0.2, 0.8).

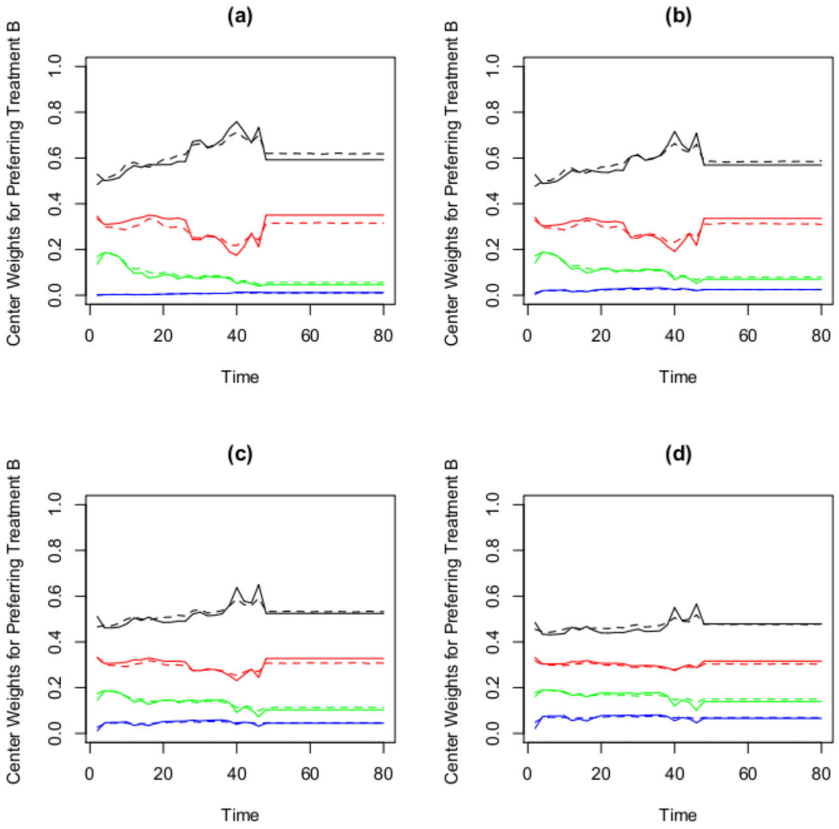

Figure 4.

The central weights of the four endpoints in preferring treatment B: (a)–(d) use the same prior weight distributions as Figure 3: the solid curves are from the CP-value method and the dashed curves are from the SMAA method. The black, red, blue and green curves correspond to each of four endpoints in the order of progression free survival, overall survival, adverse event 1 and adverse event 2.

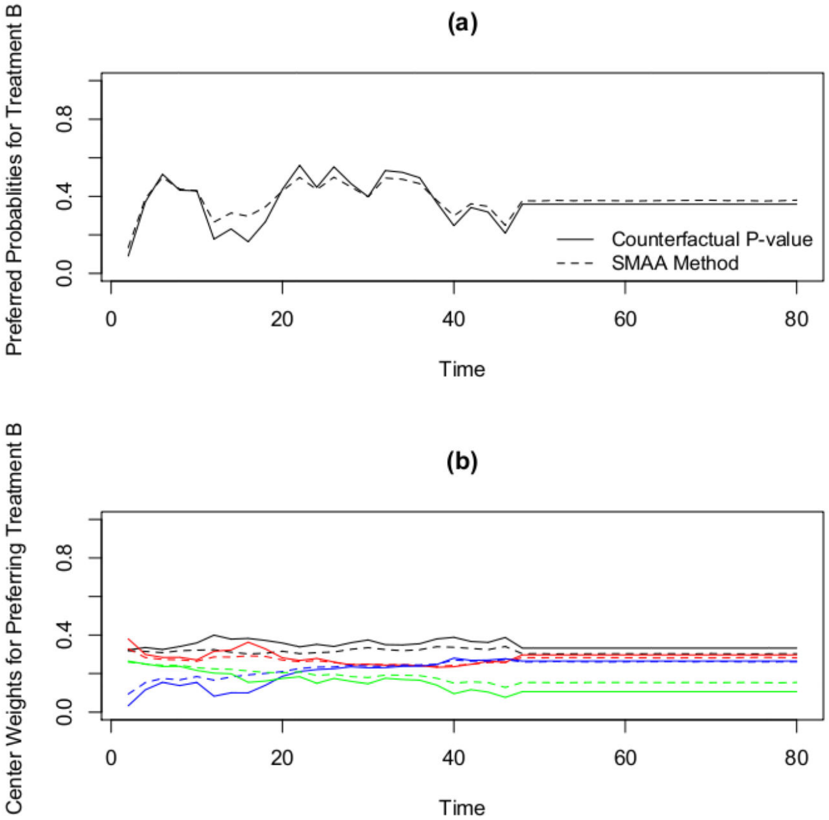

Figure 5.

The comparison between the CP-value method and the SMAA method under the uniform prior distribution for **w**: The prior distribution is the uniform distribution in the space $\left\{(w_1, w_2, w_3, w_4) : \sum_{k=1}^{4} w_k = 1, w_k \ \geq \ 0, \ k=1, \ \ldots, \ 4\right\}$. The curve definitions are the same as Figure 3 and Figure 4.

Table 1

Survival estimates at Month 10 in demo example

Treatment	Endpoint	Survival prob.	Covariance ($\times 10^{-3}$)			
A (Placebo)	overall survival	0.792	0.298	0.043	0.002	0.000
	PFS	0.905	0.043	0.177	−0.009	0.000
	adverse event 1	0.936	0.002	−0.009	0.139	0.000
	adverse event 2	1.000	0.000	0.000	0.000	0.000
B (Active Treatment)	overall survival	0.846	0.279	0.033	−0.017	0.000
	PFS	0.914	0.033	0.167	0.006	0.006
	adverse event 1	0.880	−0.017	0.006	0.245	−0.007
	adverse event 2	0.983	0.000	0.006	−0.007	0.036

Table 2

Simulation results from 500 replicates

		CP method		SMAA	NNTB/NNTH
σ^2	(r_1, r_2)	AUCP	B preference		
0.5	(1,1)	−0.014(0.236)	0.495(0.437)	0.510(0.217)	−9.732(222.581)
	(0.75,1.25)	0.071(0.228)	0.640(0.405)	0.590(0.209)	0.786(53.727)
	(0.50,1.25)	0.170(0.213)	0.778(0.330)	0.689(0.187)	−41.558(1268.20)
1.0	(1,1)	−0.018(0.247)	0.479(0.440)	0.525(0.222)	1.126(33.817)
	(0.75,1.25)	0.074(0.227)	0.650(0.403)	0.585(0.215)	−1.717(120.847)
	(0.50,1.25)	0.164(0.209)	0.781(0.332)	0.698(0.186)	14.089(471.06)
2.0	(1,1)	0.009(0.233)	0.538(0.441)	0.502(0.216)	−1.435(867.67)
	(0.75,1.25)	0.042(0.237)	0.595(0.419)	0.579(0.206)	−1.881(89.940)
	(0.50,1.25)	0.170(0.213)	0.768(0.314)	0.673(0.188)	−2.790(1261.504)