Incorporating historical information in biosimilar trials: challenges and a hybrid Bayesian-frequentist approach

Johanna Mielke

Heinz Schmidli

Byron Jones

Novartis Pharma AG, Basel, Switzerland

Abstract

For the approval of biosimilars, it is, in most cases, necessary to conduct large Phase III clinical trials in patients to convince the regulatory authorities that the product is comparable in terms of efficacy and safety to the originator product. As the originator product has already been studied in several trials beforehand, it seems natural to include this historical information into the showing of equivalent efficacy. Since all studies for the regulatory approval of biosimilars are confirmatory studies, it is required that the statistical approach has reasonable frequentist properties, most importantly, that the Type I error rate is controlled – at least in all scenarios that are realistic in practice. However, it is well known that the incorporation of historical information can lead to an inflation of the Type I error rate in the case of a conflict between the distribution of the historical data and the distribution of the trial data. We illustrate this issue and confirm, using the Bayesian robustified meta-analytic-predictive (MAP) approach as an example, that simultaneously controlling the Type I error rate over the complete parameter space and gaining power in comparison to a standard frequentist approach that only considers the data in the new study, is not possible. We propose a hybrid Bayesian-frequentist approach for binary endpoints that controls the Type I error rate in the neighbourhood of the center of the prior distribution, while improving the power. We study the properties of this approach in an extensive simulation study and provide a real-world example.

Keywords: Biosimilarity; Bayesian approach; Equivalence tests; Frequentist approach; Humira; Type I error rate

[This is the peer reviewed version of the following article: Mielke J, Schmidli H, Jones B (2018). *Incorporating historical information in biosimilar trials: challenges and a hybrid Bayesian-frequentist approach*. *Biometrical Journal*. which has been published in final form at <http://onlinelibrary.wiley.com/doi/10.1002/bimj.201700152/full>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.]

1 Introduction

A biosimilar (the test product) is approved as a copy of an already marketed biological product (the reference product) if it can be shown that the biosimilar and the reference product are *highly similar* and that "there are no clinically meaningful differences between the biological product and the reference product in terms of the safety, purity, and potency of the product" (FDA 2009). Biosimilar development has gained much attention in the last ten years because the patents of several biological blockbusters have recently expired (e.g., etanercept, infliximab) which makes it possible for competing companies to bring their own version of the drug onto the market. Currently, there are 36 approved biosimilars in Europe (Generics and Biosimilars Initiative 2017*a*) and 7 approved biosimilars in the US (Generics and Biosimilars Initiative 2017*b*).

In order to demonstrate that a proposed biosimilar has the same efficacy and safety as its corresponding reference product, large Phase III trials in patients are usually requested. For previously approved biosimilars in Europe, these were mostly parallel groups trials (Mielke et al. 2016) with 89-759 subjects. It is well known that drug development is a costly and time-consuming process (DiMasi et al. 2003) and because biosimilars are supposed to be sold at a lower price than the originator product, it is essential to keep the development costs as low as possible. Consequently, it is desirable that all available information is used to determine if the biosimilar and reference products are highly similar. In terms of available information, the development of a biosimilar differs from the development of a new drug in the sense that the reference product is already an established product at the point in time at which the studies for the biosimilar product are conducted. The efficacy of the reference product has already been investigated in several studies and, at least, summary statistics from these trials are publicly available. It seems natural to incorporate the historical information on the reference product into the assessment of biosimilarity since this may lead to a reduction in sample size. The Bayesian framework provides an intuitive way of combining the historical information with the observed data in the new study. Neuenschwander et al. (2010) developed, for example, a meta-analytic-predictive (MAP) approach that can combine several historical trials into a prior distribution taking into account the between-trial variation. The MAP approach assumes that the parameters

corresponding to different trials are not identical but similar. A random-effects meta-analytical model is used to quantify the degree of similarity.

However, Bayesian methods are mostly used in exploratory analyses in early stage development. For biosimilars, in contrast, all clinical studies are confirmatory studies in which the frequentist operating characteristics of the approaches are, in most cases, expected to fulfil certain criteria; most importantly, that the Type I error rate is controlled over the complete parameter space or, if that is not feasible, at least in all scenarios that are realistic in practice. It is well known that the use of historical information can lead to an inflation of the Type I error rate in the case where the observed data in the study do not match the historical information (e.g., Schmidli et al. 2014). A moderate inflation of the Type I error rate might be acceptable if, for example, a rare disease is studied and the development of the drug would not be feasible otherwise (e.g., Hampson et al. 2014). For biosimilars, however, the incorporation of historical information is motivated by savings in terms of resources and time. This might speed up the development and therefore bring the biosimilar earlier to the patients which might reduce the costs of treatment and allow more patients to be treated with the biologic. In addition, as fewer patients need to be enrolled in the trial, the burden on the patients is reduced. However, since the reference product is already approved and there is no unmet medical need, Type I error rate inflation should not occur in scenarios that are relevant in practice. Consequently, a methodology for incorporating historical information that controls the Type I error rate in all realistic scenarios, while still providing an advantage in terms of power, is needed. Here, we present such a methodology.

Bayesian approaches for biosimilar development have already been discussed previously, for example by Chiu et al. (2014). Also, Tsou et al. (2013) proposed a consistency approach and Hsieh et al. (2013) discussed the use of reproducibility probabilities. However, none of these authors focussed on a methodology which can control the Type I error rate in realistic scenarios which we consider highly important.

The goal of our paper is, on the one hand, to illustrate the challenges with strict Type I rate error rate control, and on the other hand to propose a methodology that can control the Type I error rate in the neighbourhood of the mean value of the prior distribution,

while still providing an advantage in terms of power or reduction of the sample size. We focus on binary endpoints, but the general conclusions and ideas are also valid for other types of endpoint.

The rest of the paper is structured as follows: we first show in Section 2 that it is not possible to achieve strict Type I error rate control while gaining power with the Bayesian robustified MAP approach (Schmidli et al. 2014) in comparison to the frequentist two-one-sided-test (TOST) approach (Schuirmann 1987). The TOST approach is commonly used in this setting and considers only the data in the new study. Afterwards, in Section 3, we propose a hybrid Bayesian-frequentist approach that can achieve the desired control of the Type I error rate in an area around the center of the prior distribution while still gaining power in that region. Next, the operating characteristics of the proposed methodology are discussed in an extensive simulation study (Section 4) before we show how to use the proposed methodology in practice by illustrating the planning of a study for a biosimilar with the active substance adalimumab (Section 5). We finish with some conclusions and comments in Section 6. Source code to reproduce all results presented in this paper is available as Supporting Information on the journal’s web page (<http://onlinelibrary.wiley.com/doi/10.1002/bimj.201700152/supinfo>).

2 Strict Type I error rate control and gain in power is incompatible with inclusion of historical data

Before we propose our hybrid Bayesian-frequentist approach, we illustrate the challenges with strict Type I error rate control if historical information is incorporated using the robustified meta-analytic-predictive (MAP) approach that was introduced by Schmidli et al. (2014). It is important to note that this methodology only serves as an example for the challenges related to Type I error rate inflation due to the use of historical information and other approaches share the same main issues. In this paper, we focus on binary endpoints, but the general approach is transferable to other types of endpoints. The necessary notation, the standard two-one-sided-test (TOST) approach that only considers the data in the new study (Schuirmann 1987) and the robustified MAP approach are introduced in Sub-

section 2.1. The operating characteristics of the robustified MAP approach are described in Subsection 2.2.

2.1 Notation & Methodology

In this paper, we assume that a proposed biosimilar (the test product, T) is compared to the authorised product (the reference product, R) in a parallel groups design with equal allocation of n patients to the test and reference groups (total sample size for the study: $N = 2n$ patients). We focus on binary endpoints which is a typical level of measurement in biosimilar development and was, for example, used in the applications of biosimilars with the active substance etanercept and infliximab in Europe (Mielke et al. 2016). Let p_R be the response rate for the reference treatment and p_T be the response rate for the test treatment. The goal is to show equivalent response rates for T and R . More precisely, the following hypotheses are considered:

$$H_0 : |p_R - p_T| \geq \Delta \text{ vs. } H_1 : |p_R - p_T| < \Delta,$$

where Δ is a pre-specified value (the equivalence margin) which is the maximum difference in response rates such that the products would not be considered different from a clinical point of view. We denote the number of observed responders in the reference group as r_R and as r_T in the test group. The observed response rates are given, respectively, by $\hat{p}_R$ and $\hat{p}_T$.

The operating characteristics of the approach that incorporates the historical information are compared to an approach that considers the new data only and can therefore serve as the benchmark. The commonly used frequentist approach that only considers the data from the new study is the two-one-sided-test (TOST) approach (Schuirmann 1987). For that, the null hypothesis is split into two parts:

$$\begin{aligned} H_0^{(1)} : p_R - p_T \leq -\Delta \text{ vs. } H_1^{(1)} : p_R - p_T > -\Delta; \\ H_0^{(2)} : p_R - p_T \geq \Delta \text{ vs. } H_1^{(2)} : p_R - p_T < \Delta. \end{aligned}$$

The standard TOST approach for binary data tests the previously described pair of hypotheses by use of an approximation to the normal distribution. For that, the variability

needs to be estimated and the standard error for the difference in response rates is given by

$$\hat{s} = \sqrt{\frac{\hat{p}_R(1 - \hat{p}_R) + \hat{p}_T(1 - \hat{p}_T)}{n}}.$$

The test statistics are defined as

$$Z_1 = \frac{\hat{p}_T - \hat{p}_R + \Delta}{\hat{s}} \quad \text{and} \quad Z_2 = \frac{\hat{p}_R - \hat{p}_T + \Delta}{\hat{s}}.$$

Since both test statistics follow asymptotically a standard normal distribution, the null hypothesis is rejected if both Z_1 and Z_2 are larger than the $(1 - \alpha)$ -quantile of the standard normal distribution.

This frequentist approach is compared to the robustified MAP approach which serves as a representative of a Bayesian approach. It is important to note that historical information is available for the reference product only. Although some authors (e.g., Chiu et al. 2014) proposed using an informative prior based on the historical information of the reference product both for the test and reference product, we believe that this introduces a bias towards the decision for equality, which is supposed to be shown and not assumed. Therefore, we use informative priors for the reference product only and for the test product, we assume a non-informative prior (uniform distribution). In the following, it is described how the posterior distribution is derived for the reference product. The posterior distribution for the test product is obtained directly by applying Bayes' theorem in the same way.

The robustified MAP approach is a Bayesian, stepwise, procedure that requires the derivation of a so-called MAP prior based on the historical data, to robustify it, and finally combines this prior with the observed data in the new study to give a posterior distribution using Bayes' theorem (Schmidli et al. 2014). The MAP approach is based on a hierarchical model. In the case of a binary endpoint, we assume that the data in the new study follow a binomial distribution with parameter p_R and sample size n and the data in the h historical studies also follow a binomial model with parameters $p_{R,H,i}$ and sample sizes $n_{H,i}$ ($i = 1, \dots, h$). The parameters of the historical data and the new study are linked using the so-called exchangeability assumption, i.e., the joint distribution of the parameters does not change when using a different ordering of the trials. In the following, we make the stronger assumption that

$$\theta_R, \theta_{R,H,1}, \dots, \theta_{R,H,h} \sim N(\mu, \tau^2),$$

where

$$\theta_R = \text{logit}(p_R) \text{ and } \theta_{R,H,i} = \text{logit}(p_{R,H,i}).$$

The parameter μ represents the population mean whereas τ is the between-trial standard deviation. As priors for the between-trial standard deviation in hierarchical models, half-normal or half-t distributions are recommended; see Spiegelhalter et al. (2004) and Gelman (2006). For μ , for example $\mu \sim N(0, 10^2)$ is a reasonable choice. For the reference product, we assume that a MAP prior (f_H) is already given and details on deriving MAP priors can be found in Schmidli et al. (2014). The MAP prior f_H is robustified by combining it with a vague prior f_v , so that the robust MAP prior f_{HR} is given by

$$f_{HR} = (1 - w_R)f_H + w_Rf_v,$$

where $w_R \in [0, 1]$ is a chosen weight that represents the prior probability that the new data and the historical data differ and may be based on clinical judgment about the relevance of the historical data. Higher values of w_R put more weight on the vague part of the prior and lead to a prior distribution with heavier tails. In the case of prior-data conflicts, the historical information is more quickly discounted. However, the gain in power is smaller for higher weights w_R . For binary endpoints, f_v may be chosen to be a uniform distribution. After the trial is finished, the robustified MAP prior is combined with the data that were observed during the trial using Bayes' theorem:

$$f_{R,post}(p_R) = f(p_R|\hat{p}_R, \hat{p}_{R,H}) \propto f_{HR}(p_R)f(\hat{p}_R|p_R),$$

where f_{HR} is the robustified prior as described above and $f(\hat{p}_R|p_R)$ represents the likelihood for the data in the new trial (Schmidli et al. 2014).

In a way similar to the frequentist hypothesis testing criterion, we need to define a Bayesian success criterion that corresponds to the described null and alternative hypotheses so that Type I error rate and power can be assessed. For that, we assume that X_T and X_R are random variables that follow the posterior distributions $f_{T,post}$ and $f_{R,post}$ for the test and reference product, respectively. Then, we claim equivalence of test and reference if

$$B := P(|X_R - X_T| < \Delta) > c, \tag{1}$$

where c has to be chosen such that the desired Type I error rate profile is achieved. Choosing a higher value for c leads to a smaller Type I error rate, but also to a lower power. The impact of this choice is discussed in more detail in Subsection 2.2.

2.2 Operating characteristics of the robustified MAP approach

For an illustration of the properties of the robustified MAP approach, we assume that the derived MAP prior for the reference product is a $Beta(50, 50)$ distribution. We consider different weights for the robustification,

$$w_R = 0, 0.1, 0.5, 0.9,$$

in order to illustrate the impact of this choice. The value $w_R = 0$ corresponds to the situation without robustification. As described before, the prior for the test product is chosen to be non-informative (uniform distribution, $Beta(1, 1)$). We assume that the new study has a sample size of $n = 150$ subjects per group, the critical value for the Bayesian approach is set to $c = 0.95$ and we use $\alpha = 0.05$ as the significance level for the TOST approach. We set the equivalence margin to $\Delta = 0.15$ and calculate the rejection rates under the null hypothesis (to evaluate the Type I error rate) and under the alternative (to evaluate power when $p_R = p_T$) for several true response rates p_R of the reference product. It is important to note that there are two limiting scenarios under the null hypothesis to consider: first, $p_R = p_T - \Delta$ which we label as Situation (a) (response rate of the reference product is smaller than the response rate of the test product) and second $p_R = p_T + \Delta$ (Situation (b), response rate of the reference product is larger than the response rate of the test product). Considering several true response rates for the reference product corresponds to considering different levels of prior-data conflict. The mean value of the chosen MAP prior is $\mu = 0.5$. Therefore, a response rate of $p_R = 0.5$ is a scenario in which the data in the new study exactly match the prior belief. A response rate of $p_R = 0.48$, for example, relates to a scenario in which the response rate differs slightly from the prior belief and a response rate of $p_R = 0.3$ indicates a clear prior-data conflict.

Figure 1 shows the results. We first focus on the left and middle panel which shows the empirical Type I error rates for the Situations (a) and (b). The nominal significance level

is $\alpha = 0.05$. The figure shows that the Type I error rate is controlled if the true response rate in the new trial matches the mean value of the prior distribution ($p_R = 0.5$), but can be substantially higher than the nominal significance level when the data from the new study do not match the prior belief. For the following discussion, we focus on Situation (a); however, Situation (b) is symmetric and all conclusions made for Situation (a) are also valid for Situation (b). In Situation (a), the approach is conservative if the response rate of the reference product is larger than the mean value of the prior (vertical line). On the other hand, we observe an inflation of the Type I error rate for response rates smaller than the mean value of the prior. The Type I error rate inflation is moderate if a high weight w_R is used (maximal Type I error rate for $w_R = 0.9$: 0.0643), but increases up to 0.4654 if no robustification is used (solid line). The Type I error rate for the approach without robustification increases monotonically if the prior-data conflict increases. In contrast, for the robustified approaches, if the response rate increases, the Type I error rate increases first, then reaches a maximum value and afterwards decreases again. Thus, if the prior-data conflict is extreme, the robustified approach discounts the prior information and this reduces the Type I error rate in these situations. However, our concerns are not with the situations in which the prior-data conflict is extreme, since we consider these situations to be avoidable in a biosimilar setting, but with the more realistic situation of a minor difference between historical studies and the new study in which MAP priors do not guarantee a reduction of the Type I error rate. For example, for $w_R = 0.9$ and a true response rate of $p_R = 0.4655$, we get a Type I error rate of 0.0564 which is higher than the nominal significance level of $\alpha = 0.05$. Therefore, it is important to emphasize that while robustification helps to limit the overall maximum Type I error rate, it does not control the Type I error rate at its nominal level in the region which we consider most relevant.

Figure 1: Type I error rates (left and middle panel) and power (right panel) for the robustified MAP approach. The panel labeled with (a) refers to the situation in which $p_R = p_T - \Delta$ and the panel labeled with (b) refers to $p_R = p_T + \Delta$. The vertical line gives the mean value of the prior distribution.

The panel on the right shows the empirical power. Here, we see that all approaches have higher power than the TOST approach in the area around the mean value of the prior distribution. However, the power is highest if no robustification is used ($w_R = 0$) and lowest if the highest weight is chosen ($w_R = 0.9$). This shows that there is a direct trade-off between the power and the Type I error rate profile: a more desirable Type I error rate profile comes with a smaller gain in power.

Obviously, it is possible to achieve Type I error rate control for all Bayesian approaches by increasing the critical value c (see Equation (1)). We calibrated the MAP approaches with the different weights w_R to keep the Type I error rate at the nominal level over the shown parameter space (response rates of the reference product between 0.3 and 0.7). The Type I error rate and power are shown in Figure 2. The Type I error rate is now controlled for all approaches (left and middle panel). However, since the Type I error rate is not constant for all response rates p_R in the new study, the approaches are very conservative for some response rates, specifically in the region around $p_R = 0.5$ which is the mean value

of the prior distribution and therefore also the area in which we expect the response rate in the new study to lie. This conservative behaviour directly transfers into lower power (right panel): the power of the Bayesian approaches is now lower than the power of the TOST approach. Therefore, if strict Type I error rate control is required, the MAP approach (or any other approach using informative priors) cannot give any advantages in terms of power.

Figure 2: Type I error rates (left and middle panel) and power (right panel) for the robustified MAP approach with calibrated critical value c such that the Type I error rate is controlled. The panel labeled with (a) refers to the situation in which $p_R = p_T - \Delta$ and the panel labeled with (b) refers to $p_R = p_T + \Delta$. The vertical line gives the mean value of the prior distribution.

3 The hybrid Bayesian-frequentist approach

The proposed approach combines Bayesian and frequentist ideas and uses two switching rules and a response rate-dependent critical value to improve the performance under the alternative hypotheses (higher power) while maintaining the required Type I error rate under the null hypothesis in the neighbourhood of the mean value of the prior distribution. Both the switching rules and the response rate-dependent critical values are used at the

analysis stage of the study. No interim analysis is required. We only aim to control the Type I error rate in a subset of the parameter space because it was shown in Subsection 2.2 that strict Type I error rate control and a gain in power are incompatible. The chosen subset of the parameter space is the region around the mean value of the prior distribution which is the area in which we expect the true response rate of the new study to lie. We consider it to be less important to control the Type I error rate for response rates in the tails of the prior distribution because, due to extensive experience with the reference product, it is realistic to assume that a new study can be set up which is similar to the historical trials. Profiles of the Type I error rate and power that we would consider acceptable are shown in Figure 3: the Type I error rate is controlled in a neighbourhood of the mean value of the prior distribution that is indicated by the dotted vertical lines. Outside of this region, we accept an inflation of the Type I error rate because we are highly confident that the true response rate will not lie within that region. More formally, let $\bar{p}_H$ be the mean value of the prior distribution. We aim to control the Type I error rate for all response rates p_R in the new study in the interval C ,

$$C = [\bar{p}_H - \delta, \bar{p}_H + \delta], \quad (2)$$

where δ is the parameter that controls the width of the controlled interval. This partial control of the Type I error rate comes with a price: the gain in power in comparison to the MAP approach is expected to be lower due to the partial control of the Type I error rate. The goal is therefore not to develop a test with comparable power to the MAP approach, but to develop a test that achieves a higher power than the TOST approach. The approach is described in Subsection 3.1 and some computational issues and the choice of tuning parameters are discussed in Subsection 3.2.

Figure 3: Desired Type I error rate (maximum of Situations (a) and (b)) and power profile. The horizontal line shows the (hypothetical) Type I error rate and power of an approach that is used as a benchmark. The solid vertical line shows the mean value of the prior distribution. The dotted vertical lines indicate the interval C . The marked points are important positions that are used for the optimisation (see Subsection 3.2).

3.1 Description of the approach

The starting point of our approach is a moment-matched Beta prior for the reference product that can be derived, for example, with the MAP approach (Schmidli et al. 2014). In this context, "moment-matched" means that the estimated first two moments of the generated sample from the distribution of the MAP prior matches the first two moments of the derived Beta distribution. We do not use the concept of robustification for the following two reasons: (1) as shown in Subsection 2.2, there is a decrease in power without reducing the Type I error rate in a relevant way in the regions that are relevant for us. Therefore, the concept of robustification does not seem to bring us closer to the desired operating characteristics (shown in Figure 3), (2) the assumption of a monotonic Type I error rate simplifies the determination of the response rate-dependent critical values (see Subsection 3.2). For the test product, we assume a uniform prior distribution. Afterwards,

the new study is conducted using a parallel groups design with n subjects per group. The information from the new study is combined with the prior distribution using Bayes' theorem as stated in Subsection 2.1. However, we introduce some modifications (response rate-dependent critical values and switching rules) for achieving the power profile that is shown in Figure 3. These modifications are summarized in a flow chart in Figure 4 and are described and motivated in the following. It is important to note that while the approach consists of several tuning parameters, all of these values can be pre-defined during the planning stage of the new study by simulating the operating characteristics. The complete algorithm including all critical values and decision points is therefore chosen independently from the data of the new study and can be pre-specified in the study protocol. We illustrate in Section 5 the steps that have to be conducted using an example and give in Subsection 3.2 an algorithm that tunes the proposed hybrid approach automatically.

The response rate-dependent critical values are motivated by Figure 1 which showed that for the MAP approach without robustification (solid line), the rejection rate under the null hypothesis highly depends on the true response rate of the reference treatment and has a monotonic shape. Therefore, we propose to use instead of a constant critical value a flexible, response rate-dependent, critical value. This allows the use of higher critical values for areas in which the standard MAP approach was too liberal and smaller critical values in areas in which the standard MAP approach was too conservative. Also, the critical value depends on the ordering of the response rates of T and R : the conservative situations are located, for example, in Situation (a) right of the mean value of the prior distribution whereas they are left of the mean value of the prior in Situation (b). Therefore, different critical values shall be defined for Situations (a) and (b). More formally, let

$$c_1, c_2 : \mathbb{R} \rightarrow [0, 1]$$

be two functions which map the response rate of the reference product to the critical value.

Figure 4: Flow chart of the hybrid approach.

Function c_1 relates to the situation in which the response rate for reference is equal to or larger than that for test (Situation (b)) and c_2 to the situation in which the response rate for test is smaller than that for reference (Situation (a)). The true response rates for test and reference are not known in practice and need to be estimated and the functions c_1 or c_2 are evaluated at the estimated response rate of the reference product. Also, the decision

if c_1 or c_2 is used, is based on the estimated response rates. If the true response rate for R would be known, optimal critical value could easily be identified using the quantiles of the distribution of the test statistic for a specific response rate. The additional uncertainty of using estimated response rates makes the determination of valid critical values more difficult because this additional uncertainty has to be incorporated into the choice of critical values. We discuss the identification of the functions c_1 and c_2 in Subsection 3.2.

In addition, the proposed approach uses two switching rules that are based on the estimated response rates. First, the historical information should only be used if the observed data are *similar* to the historical data (Bayes-frequentist switching rule, Switching rule I, see Figure 4). Similarity is here measured as the difference between the mean response rate, $\bar{p}_H$, of the derived MAP prior f_H and the observed response rate, $\hat{p}_R$, in the new study. If

$$|\bar{p}_H - \hat{p}_R| > \gamma_1, \quad (3)$$

the historical information is ignored and the standard TOST approach as described in Subsection 2.1 is applied. Otherwise, the Bayesian approach is used. The motivation for this switching rule is that in the case of a prior-data conflict, the historical data does not provide any useful information on the reference treatment and should be ignored. The negative impact of incorporating historical information in the case of a prior-data conflict is reflected in the operating characteristics of the Bayesian approach which are not desirable (see the solid line in Figure 1) – neither in terms of power nor in terms of the Type I error rate. Therefore, it is reasonable to switch back to the TOST approach in this case. The constant γ_1 has to be chosen beforehand. It reflects the acceptable degree of prior-data conflict and influences the operating characteristics of the approach. We give some advice on its choice in Subsection 3.2.

The second switching rule is used in order to increase the chance of rejection of the null hypothesis in the case where the study shows that the test and the reference products are similar, but the data of the reference product do not match the historical data. If the data do not match the prior, even if in the new study test and reference show perfect similarity, the prior knowledge for the reference product will pull the posterior mean of the reference product away from the mean value of the test product making it difficult to reject the null

hypothesis. Therefore, if

$$|\hat{p}_T - \hat{p}_R| < \gamma_2, \quad (4)$$

we use the constant $\bar{c}$ instead of the response rate-dependent critical values $c_1(\hat{p}_R), c_2(\hat{p}_R)$ as the critical value. Both γ_2 and $\bar{c}$ are tuning parameters and need to be chosen to fit the desired operating characteristics. Some advice on the choice is provided in Subsection 3.2.

Figure 4 gives a summary of the algorithm: in the first step, priors are chosen for test (T) and reference (R) (e.g., MAP prior for R , uniform distribution for T). Afterwards, the tuning parameters and functions c_1, c_2 are chosen (see Subsection 3.2) before the new study is conducted and the response rates for reference and test are estimated ($\hat{p}_R, \hat{p}_T$). Using the first switching rule (Equation (3)), it is decided if the historical data should be ignored or not. In the case where the historical data are not used, the test decision can be made directly using the TOST approach. If not, the Bayesian success criterion B (Equation (1)), that was introduced in Subsection 2.1, is used. If the historical data should be incorporated, the second switching rule (Equation (4)) is applied that decides which critical values are used. In the case where the point estimates for the response rate in test and reference are very similar, $\bar{c}$ is used as the critical value. Otherwise, $\hat{p}_R$ is compared with $\hat{p}_T$. If $\hat{p}_R$ is smaller than $\hat{p}_T$, then $c_2(\hat{p}_R)$ is the critical value, otherwise $c_1(\hat{p}_R)$ is used.

3.2 Optimal choice of response rate-dependent critical values and tuning parameters

In this subsection, we discuss the derivation of optimal response rate-dependent critical values and the choice of tuning parameters. We provide supplementary online material with further details concerning the implementation of the proposed methodology and a step-by-step example with R-code.

Before discussing the optimisation, it is important to note that due to the discrete nature of the problem, it is possible to calculate the exact Type I error rate and exact values for power without using simulations. For that, it is necessary to calculate the test decision for all combinations of number of responders, r_R and r_T , that can be observed in the new study. We denote the test decision for observed values r_R and r_T by d_{r_T, r_R} . This

is a binary variable with the value 1 when the test decision is for the alternative and 0 otherwise. For example, for $n = 150$, it is necessary to evaluate the test decision d_{r_T, r_R} for $151^2 = 22801$ scenarios. Combining the test decision with the probabilities to observe r_R and r_T , which are denoted by $P(X = r_R)$ and $P(Y = r_T)$, respectively, leads to the exact rejection rates:

$$r = \sum_{r_R=0}^n \sum_{r_T=0}^n P(X = r_T)P(Y = r_R)d_{r_T, r_R}.$$

The probabilities $P(X = r_T)$ and $P(Y = r_R)$ for a specific scenario can be derived using the binomial distribution with parameters p_T and p_R , respectively, and a sample size of n subjects per group.

Both the functions that map the estimated response rates to the critical values (c_1, c_2) and several tuning parameters have to be chosen before the hybrid approach can be applied. Choosing the functions c_1, c_2 without any assumptions on the functional form is very difficult. Since the Type I error rate is a monotonic function, we also use monotonic functions for c_1 and c_2 . More specific, we assume a logistic function with the parameters L (the minimal value of the function), U (the difference between the minimal and maximal value of the function), x_0 (the sigmoid's midpoint) and k (the steepness of the curve). In addition, we assume that c_1 and c_2 are complements of each other, i.e.,

$$c_1(x) = L + \frac{U}{1 + \exp(-k(x - x_0))}, \quad c_2(x) = L + \frac{U}{1 + \exp(k(x - x_0))}.$$

Also, γ_1 , γ_2 and $\bar{c}$ have to be chosen. The parameter γ_1 describes the boundary at which it is claimed that the data-prior conflict is so extreme that the historical data should be ignored (see Equation (3)). If γ_1 is 0, the TOST approach would always be used. If $\gamma_1 = 1$, only the Bayesian approach would be used. Therefore, a higher choice of γ_1 leads to higher power, but a less desirable Type I error rate. We found that

$$\gamma_1 = 3\sigma = 3 \cdot \sqrt{\frac{\bar{p}_H \cdot (1 - \bar{p}_H)}{n}} \quad (5)$$

is a reasonable compromise. The parameter γ_2 is the boundary at which the critical value $\bar{c}$ is used instead of $c_1(\hat{p}_R)$ or $c_2(\hat{p}_R)$ which is supposed to make a rejection of the null hypothesis easier in case the response rates of T and R are very similar (see Equation (4)). For γ_2 , similar arguments as for γ_1 and experience gained with simulation studies leads to

a recommended choice of

$$\gamma_2 = \sigma = \sqrt{\frac{\bar{p}_H \cdot (1 - \bar{p}_H)}{n}}. \quad (6)$$

The parameter $\bar{c}$ defines the value to which the Bayesian success criterion B (see Equation (1)) is compared to if the absolute difference in response rates of T and R is smaller than γ_2 in the new study. Generally, a lower value for $\bar{c}$ leads to higher power, but if $\bar{c}$ is chosen too small, it might not be possible to control the Type I error rate in the interval C – even if c_1 and c_2 are chosen to be a constant value of 1, making a rejection of the null hypothesis impossible if the difference in response rates is not smaller than γ_2 . Therefore, it is difficult to give a general recommendation for this parameter, but a suitable choice can be made during the optimisation procedure. In total, five parameters (four related to the response rate-dependent critical values and one tuning parameter) need to be chosen.

The choice of the five parameters can be made automatic by an optimisation procedure which is described in the following. We optimise the power under the alternative that $p_R = p_T = \bar{p}_H$ under the constraint that the Type I error rate range is controlled in the interval C , however, other target functions are also possible (e.g., optimise the power under the alternative that $p_R = p_T \pm \log(1.05)$) and might lead to slightly different results. The chosen target function can be visualised in Figure 3: the Type I error rate is to be controlled at the points in the left panel while the power is optimised at the dot in the right panel. It is important to note that it is sufficient to control the Type I error rate at the boundaries of the interval C because the Type I error rate has a monotonic shape in Situations (a) and (b) for response rates in the center of the prior distribution.

Due to the complexity of the optimisation problem, we use a step-wise procedure which does not guarantee the identification of the global optimum, but experience (e.g., see simulation study in Section 4) has confirmed that this strategy leads to a good solution nonetheless. First, we identify the parameter $\bar{c}$, then, we determine suitable starting values for optimising c_1 and c_2 before we identify the parameters of the optimal functions $c_1, c_2(U, L, k, x_0)$. The procedure is summarised in the flow chart in Figure 5. For the start of the optimization procedure (Step 1), we set

$$\gamma_1 = 3\sigma, \gamma_2 = \sigma, \bar{c} = 0.9, c_1 = c_2 = 0.99.$$

Figure 5: Flow chart of the automatic tuning of the parameters for the hybrid approach.

Then, we calculate the Type I error rate for $p_R = \bar{p}_H \pm \delta$ in Situation (a) and Situation (b) with these parameters (Step 2). While the Type I error rate is larger than 0.05 in Situation (a) or Situation (b), we set

$$\bar{c} := \bar{c} + 0.005$$

and repeat the calculation of the Type I error rate. Once the Type I error rate is controlled, we aim to identify good starting values for the optimisation of the functions c_1, c_2 (Step 3). For that, we determine the constant critical value c_{opt} with $c_1 = c_2 = c_{opt}$ that maximises the power at the mean value of the prior while controlling the Type I error rate in the interval C . This optimisation is performed with the function `crs21m` (Price 1983, Kaelo & Ali 2006) in the R-package `nloptr` (Johnson 2014). Afterwards, the found constant critical value is

used as the starting value for identifying the optimal values of the parameters U, L, k, x_0 for the functions c_1, c_2 (Step 4, see the supplementary material for more details). This optimisation is also first performed with the function `crs2lm`. The Nelder-Mead algorithm (function `optim`) is applied on the found solution in order to improve the performance. Since `crs2lm` is a random search algorithm, we apply both steps several times in order to avoid stopping at a local minimum.

It should be emphasised that this approach leads in general to good results and is especially useful if several scenarios need to be evaluated in a systematic way. However, if a study is planned, it is recommended to evaluate the operating characteristics carefully and to adjust the tuning parameters manually if necessary (Step 5). This is illustrated in the case study in Section 5.

4 Simulation study

We investigate the properties of the proposed approach in a simulation study. All simulations were performed with R version 3.2.3 (R Core Team 2015). Before showing any general results, we would like to illustrate the Type I error rate and power profile for one chosen setting. For that, we fix the sample size per group to $n = 200$, the parameter for the controlled region is $\delta = 0.04$, the mean value of the prior distribution is $\bar{p}_H = 0.667$ and the effective sample size (ESS) is set to $ESS = 100$. The effective sample size describes how informative the prior is (Morita et al. 2008). For a Beta prior with parameters a and b , the ESS is, for example, given by

$$ESS = a + b.$$

With these choices of $\bar{p}_H$ and δ , we aim to control the Type I error rate in the interval

$$C = [0.627, 0.707].$$

We use

$$\gamma_1 = 3 \cdot \sqrt{\frac{0.667 \cdot 0.333}{200}} = 0.1000, \quad \gamma_2 = \sqrt{\frac{0.667 \cdot 0.333}{200}} = 0.0333, \quad \bar{c} = 0.9,$$

as recommended in Subsection 3.2, and apply the strategy described in Subsection 3.2 to determine the optimal response rate-dependent critical values which are displayed in Figure 6.

Figure 6: Example for the functions c_1 and c_2 : optimal response rate-dependent critical values for the described scenario.

Figure 7 shows the empirical power (right panel) and Type I error rate (left panel) for different true response rates in the new study to illustrate different levels of commensurability of prior and data. The proposed hybrid approach is compared to the TOST approach that only considers the new data and the robustified MAP approach with 90% weight on the non-informative part that is calibrated to control the Type I error rate in the interval C (see Subsection 2.1, $c = 0.96$). The proposed hybrid approach keeps the Type I error rate within the defined interval C . The Type I error rate is substantially inflated outside of the controlled interval, but as we only aim to control the Type I error rate in the defined interval C , this is acceptable. In contrast, power is gained in the interval C in comparison to the TOST approach that only considers the observed data of the new study: for example, at the mean value of the response rate of the reference product in the historical trials (see vertical solid line in Figure 7) the TOST approach has a power of 0.8776, but the proposed approach has a power of 0.9363. The power of the TOST approach can be

increased by using a higher number of subjects n . It would have been necessary to include $n = 241$ instead of $n = 200$ subjects to achieve a comparable level of power with the TOST approach. In comparison to the robustified MAP approach, the Type I error rate inflation of the hybrid approach is much higher outside of the interval C . In contrast, there is not a relevant gain in power for the robustified MAP approach in comparison to the TOST approach.

Figure 7: Operating characteristics for the robustified MAP approach with 90% weight on the non-informative part ($w_R = 0.9, c = 0.96$), the proposed hybrid approach and the TOST approach that only considers the data in the new study. The vertical solid line shows the mean value of the prior distribution, the vertical dotted lines indicate the boundaries for the interval C . The displayed Type I error rate is the maximum of Situations (a) and (b).

Next, we discuss the operating characteristics of the hybrid approach for a broader range of scenarios. Table 1 shows the considered settings. We consider three different sample sizes for the new study, two effective sample sizes (ESS) of the prior and ten mean values of the prior distribution for the reference product, $\bar{p}_H$, between 0.5 to 0.8. Due to symmetry, the chosen response rates for the historical data are representative also for response rates from 0.2 to 0.5. The equivalence margin is set to $\Delta = 0.15$. We aim to control the Type I

Table 1: Overview of considered parameter settings (n : sample size per group, ESS: effective sample size (ESS= $a + b$ if prior is $Beta(a, b)$), $\bar{p}_H$: mean value of the prior, Δ : equivalence margins, δ : parameter that controls the width of interval C (Equation (2)).

Parameters	Settings
n	100, 150, 200
ESS of prior	20, 100
$\bar{p}_H$	0.5000, 0.5333, 0.5667, 0.6000, 0.6333, 0.6667, 0.7000, 0.7333, 0.7667, 0.8000
Δ	0.15
δ	0.02, 0.04, 0.06

error in the interval

$$C = [\bar{p}_H - \delta, \bar{p}_H + \delta],$$

and use three different values for δ corresponding to a width of the controlled interval of 0.04, 0.08, 0.12. In total, we consider 180 scenarios. The R-package **BatchJobs** (Bischl et al. 2015) is used for parallel computing. The optimal response rate-dependent critical values and tuning parameters are chosen using the recommendations made in Subsection 3.2. In order to compare the results so that the impact of the different scenarios can be analysed, we display in Figure 8 the empirical power at

$$\bar{p}_H = p_R = p_T,$$

which relates to the scenario of a perfect match of historical data and new data and complete equivalence between response rates of T and R . This value can be found in Figure 7 at the vertical solid line. It is important to keep in mind that the Type I error rate is controlled in the interval C which makes the Figure meaningful even though only a part of the information is shown. The rejection rates are shown for several different sample sizes, different values δ , different mean values of the prior distribution and an ESS of 100. The results for an ESS of 20 were comparable (not shown), except that an ESS of 100 gave us a small advantage in terms of power, but led to a higher inflation of the Type I error rate outside of the interval C . However, as we only aim to control the Type I error rate

within the interval C , this is of course acceptable. The feature that the power increases with a higher ESS is clearly desirable because having a more informative prior should be rewarded.

Figure 8 displays the rejection rates of the proposed hybrid approach (solid line) in comparison to the TOST approach (dotted line). The hybrid approach leads to a gain in power for the three values of δ that are explored. However, if a higher value for δ is used, the advantage of the hybrid approach decreases. This is expected because a higher value of δ leads to a larger interval in which the Type I error rate has to be controlled. This more desirable Type I error rate leads to a less desirable power profile. This was already discussed for the robustified MAP priors in Subsection 2.2. However, it should be emphasized that even if a fairly large interval with a width of 0.12 is controlled, still up to 5.5% power can be gained. Comparing the different sample sizes ($n = 100, 150, 200$), we see that a higher sample size in general leads, as expected, to a higher power. The gain in power seems to be rather independent of the sample size.

The optimisation and choice of tuning parameters can be, as discussed in Subsection 3.2, difficult. However, the results show that, apart from the situation with $\delta = 0.02$ and $n = 100$, the power profiles look smooth, indicating that the optimisation strategy that was proposed in Subsection 3.2 is stable. However, it should be emphasized that if a study is planned using this approach, it is recommended to study the operating characteristics for the specific parameter setting of the study in order to check if manually tuning the proposed method increases the performance. Manually adjusting of the parameters can also be useful in case a specific profile of power and Type I error rate is required. We discuss this in a case study in Section 5.

5 Case study

In this section, we illustrate the proposed method by planning a hypothetical Phase III study for a biosimilar with the active substance adalimumab (originator drug: Humira[®] from AbbVie Ltd.). The considered indication is psoriasis. The efficacy of treatment for psoriasis is often measured with the Psoriasis Area and Severity Index (Fredriksson & Pettersson 1978, PASI). This score is the most widely used measurement scale for the severity of pso-

Figure 8: Operating characteristics for the proposed hybrid approach (solid lines) in comparison to the TOST approach (dotted lines) for different values of δ and different sample sizes for the new study. Displayed is the empirical power assuming that the mean value of the prior is identical to the true response rates of reference and test product in the new study.

riasis and lies between 0 (no disease) and 72 (maximal disease). The rate of patients with at least 90% improvement in the PASI score after week 16 (PASI 90 responders) is an important endpoint for clinical trials in psoriasis (Torres & Puig 2015). Through a systematic search we identified five studies with relatively homogeneous study populations, identical treatment regimens (initial treatment with 80mg, afterwards 40mg every other week) and assessment of PASI 90 at week 16. The results of the studies are displayed in Table 2. In total, 1858 subjects were studied of which 886 were classified as PASI 90 responders at week 16.

Table 2: Historical studies for Humira[®] (adalimumab). The shown response rates correspond to PASI 90 at week 16. *: Response rate was given in the paper and number of responders was calculated using the information in the publication.

Study	Publication	Indication	Responders/Sample size (%)
1	Menter et al. (2008)	moderate to severe psoriasis	366/814 (45.0)*
2	Saurat et al. (2008)	moderate to severe plaque psoriasis	55/108 (51.3)*
3	Thaçi et al. (2010)	moderate to severe psoriasis	183/364 (50.0)
4	Blauvelt et al. (2017)	moderate to severe psoriasis	166/334 (49.7)
5	Reich et al. (2017)	moderate to severe psoriasis	116/248 (46.8)
Total		886/1858 (47.7)	

As shown in the flow chart in Figure 4 in Section 3, it is first necessary to derive the MAP prior for the reference treatment. We obtain the MAP prior with the R-package **RBest** (Weber 2017). For that, we chose the priors for the hyper-parameters of the MAP approach (see Subsection 2.1) as recommended in the clinical trial example in Schmidli et al. (2014): for the between-trial standard deviation τ , a weakly informative half-normal prior with a standard deviation of 1 was used. For the mean value of the prior distribution μ , we used a normal distribution with mean value 0 and standard deviation 10 which is also a weakly informative prior distribution. The MAP prior was approximated with a Beta distribution and the best fit was obtained with the parameters $a = 55.0844$ and $b = 59.3647$. The ESS was approximately 114 subjects. The mean value of the prior distribution is $\bar{p}_H = 0.4813$. For the test product, we set a uniform distribution (i.e., $Beta(1, 1)$) as the prior distribution.

The goal of the new study is to demonstrate equivalence with an equivalence margin of $\Delta = 0.15$, i.e., to test the hypotheses:

$$H_0 : |p_R - p_T| \geq 0.15 \text{ vs. } H_1 : |p_R - p_T| < 0.15.$$

We assume that the Type I error rate is to be controlled for response rates in the interval

$$C = [\bar{p}_H - 0.05, \bar{p}_H + 0.05] = [0.43, 0.53].$$

It is noted that the observed response rates for all historical trials that are shown in Table 2 lie clearly within this range. Before any data from the new study become available, it is necessary to choose the tuning parameters and determine the optimal response rate-dependent critical values. We first use the proposed algorithm (see Figure 5) to get initial values for the response rate-dependent critical values and tuning parameters and we will adjust this choice afterwards according to the desired Type I error rate profile. We assume a sample size of $n = 175$ subjects per group for the new study.

In the first step of the algorithm (see flow chart in Figure 5), we set

$$\gamma_1 = 3 \cdot \sqrt{\frac{0.4813 \cdot 0.5187}{175}} = 0.1133, \gamma_2 = \sqrt{\frac{0.4813 \cdot 0.5187}{175}} = 0.0378, \bar{c} = 0.9, c_1 = c_2 = 0.99.$$

We then calculate the Type I error rate under Situation (a) and Situation (b) (Step 2) for $p_R = \bar{p}_H \pm 0.05$ (p_R at the boundaries of interval C) which has a maximum value of 0.0226 and is therefore smaller than the nominal significance level of $\alpha = 0.05$. That is why we can proceed to Step 3 in which we search for starting values for the target function. The optimal constant critical value is $c_1 = c_2 = 0.9748$. The constant critical value is used as the starting value for the optimisation of the response rate-dependent critical values (Step 4). Since the chosen optimiser (`crs2lm` in the package `nloptr`) is a random optimiser, we repeat the optimisation several times with the chosen starting value and get for the optimal solution a power of 0.8153 which clearly exceeds the power of the TOST approach (0.7414). One could now fix the functions c_1 and c_2 , run the study and get results with the hybrid approach as explained in the flow chart in Figure 4. However, we will analyse the Type I error rate and power profile in order to adjust the tuning parameters manually to our desired profile and to check if further increases in the performance are possible. For that, the leftmost panels in Figure 9 shows the operating characteristics of the hybrid approach (automatic optimised parameters, referred to as Choice 1 in the following) in comparison to the TOST approach. The upper panel gives the Type I error rate. It is shown that the Type I error rate is controlled in the chosen region C (indicated with the vertical dotted lines). Outside of the region, the Type I error rate first increases, reaches a maximum and decreases again if the data do not match the prior at all. This property is comparable to the operating characteristics of the robustified MAP approach (see Figure 1).

Figure 9: Operating characteristics (Type I error rate: upper panels, empirical power: lower panels) of the proposed hybrid approach for three different choices of tuning parameters in comparison to the TOST approach for a hypothetical Phase III study for adalimumab. Whereas for Choice 1 (left), the automatic optimisation was used, the parameters γ_1 and γ_2 were changed for Choice 2 (middle) and Choice 3 (right; see description in the text). The dotted vertical lines indicate the boundaries for interval C (the controlled region), the solid vertical line represents the mean value of the prior distribution. The displayed Type I error rate is the maximum error rate for Situations (a) and (b).

The lower panel shows the empirical power. It confirms that there is a high gain in the center of the prior distribution. However, if we move closer to the boundaries of the controlled interval C (vertical lines), the power of the hybrid approach converges to the TOST approach. We would prefer a relevant gain in power in the complete interval C . The ability to gain power in situations in which the historical data and the data in the new study are not completely commensurate is controlled with the parameter γ_2 : if γ_2 is set to

a higher value, then more often the easier rejection rate $\bar{c}$ is used. We increase in Choice 2 the parameter γ_2 to

$$\gamma_2 = 1.5 \cdot \sqrt{\frac{0.4813 \cdot 0.5187}{175}} = 0.0567.$$

The other tuning parameters ($\gamma_1, \bar{c}$) are kept fixed. Again, we apply the algorithm outlined in Figure 5. The operating characteristics are displayed in Figure 9 in the middle panels. Comparing the empirical power of Choice 2 (lower middle panel) with the empirical power of Choice 1 (lower left panel), we can see that this choice of γ_2 brings the desired change: for this setting, the power in the complete interval C is higher than that of the TOST approach. The Type I error rate outside of the interval is comparable. Therefore, we prefer this set of tuning parameters.

Lastly, we consider decreasing the Type I error rate outside of the controlled region to provide some certainty that even if anything unexpected happens, the Type I error rate would not be extreme. This can be controlled by adjusting γ_1 : if this value is decreased, the hybrid approach more often switches to the TOST approach and ignores the historical data. We set γ_1 to

$$\gamma_1 = 2.5 \cdot \sqrt{\frac{0.4813 \cdot 0.5187}{175}} = 0.0944.$$

The operating characteristics are displayed in the rightmost panels of Figure 9. The Type I error rate (upper panel) is clearly reduced. However, the power peak (lower panel) is much narrower than for Choice 2. Therefore, we prefer Choice 2 and propose using

$$\gamma_1 = 0.1133 \text{ and } \gamma_2 = 0.0567,$$

and to optimise the response rate-dependent critical values with the method proposed in Subsection 3.2.

It is important to note that a formal inclusion of γ_1 and γ_2 in the algorithm that automatically tunes the hybrid approach is not possible with the chosen target function (Subsection 3.2). From the three choices of critical values and tuning parameters shown in this paper, we preferred the operating characteristics achieved with the tuning parameters of Choice 2. However, this choice has a value of the target function of 0.8079 whereas the first choice had a value of 0.8153 and would be preferred in the automatic tuning that is described in Section 3.2. The preference for a specific profile of Type I error rate and power

is too complex to be included in a completely automatic choice of tuning parameters. We provide an algorithm for a first automatic choice, but it is highly recommended that the operating characteristics are checked and the tuning parameters are adjusted if necessary as shown in this case study. For now, we fix the parameters obtained in Choice 2.

Next, it is demonstrated how a test decision is made with the proposed hybrid approach using the information provided in the European public assessment report (EPAR) for the application of Amgevita[®] (Amgen, CHMP 2017). Amgevita[®] is a biosimilar with the active substance adalimumab and was approved in March 2017 in Europe. It is the first approved biosimilar to Humira[®] in Europe. The sponsor conducted a Phase III study in patients with stable moderate to severe plaque psoriasis with the same treatment regimen as in the historical trials (see Table 2). PASI 90 was evaluated at week 16. At that point in time 81 of 172 subjects in the test group (47.1%) and 82 of 173 subjects in the reference group (47.4%) were classified as responders. We will now illustrate how a test decision is made with the hybrid approach by following the algorithm displayed in Figure 4. After estimating the response rates under test and reference treatment, it is first necessary to decide if the historical data should be used or not (Switching rule I, see Equation (3)). For that, we calculate the difference between the mean value of the prior distribution and the observed response rate:

$$|\bar{p}_H - \hat{p}_R| = |0.4813 - 0.474| = 0.0073.$$

Since this value is smaller than $\gamma_1 = 0.0944$, we proceed with the Bayesian approach. Next, we compare the observed response rates in the test and reference groups (Switching rule II, see Equation (4)):

$$|\hat{p}_T - \hat{p}_R| = |0.471 - 0.474| = 0.003.$$

This difference is smaller than γ_2 , therefore, the Bayesian success criterion B (Equation (1)) is compared to $\bar{c} = 0.9$

The parameters for the posterior distribution of the test product are given by

$$a_T = 1 + 81 = 82 \text{ and } b_T = 1 + 172 - 81 = 92$$

and by

$$a_R = 55.0844 + 82 = 137.0844 \text{ and } b_R = 59.3647 + 173 - 82 = 150.3647$$

for the reference product. The Bayesian success criterion (Equation (1)) is $B = 0.9983$ which is clearly larger than $\bar{c} = 0.9$. We therefore reject the null hypothesis and claim equivalence of test and reference. It is important to note that the TOST approach leads to the same conclusion in this example.

6 Conclusion

When a biosimilar is developed, historical information on the reference product, which has been on the market for several years, is available. Including this information into the confirmation of equivalent efficacy might lead to a reduction in terms of sample size, but can also inflate the Type I error rate. In this paper, we first showed in Section 2, using the MAP approach (Neuenschwander et al. 2010), that even with the robustified version of this approach (Schmidli et al. 2014), the Type I error rate cannot be controlled if power is to be gained in comparison to the two-one-sided-test (TOST) approach (Schuirmann 1987). The TOST approach is the standard, frequentist approach that considers only the data in the new study. Strict control of the Type I error rate seems to be incompatible with a gain in power. Therefore, we propose to control the Type I error rate not in the complete parameter space, but only in an interval that is centered around the mean value of the prior distribution. This region is the area we consider to be most important because we expect the true value of the new study to be in that region. For example, for binary endpoints and in the case where the response rate of the historical data is 0.25, we might aim to control the Type I error rate for true response rates in the new study between 0.2 and 0.3. We showed that the MAP approach does not provide any relevant advantage even if the Type I error rate has to be controlled in an interval in the center of the prior distribution only.

Therefore, we proposed in Section 3 a hybrid Bayesian-frequentist approach for binary endpoints that borrows strength from historical data while controlling the Type I error rate in the center of the prior distribution. This is achieved by combining two switching rules and response rate-dependent critical values. We analysed the operating characteristics of the proposed approach in a simulation study (Section 4). It was shown that power can be gained in comparison to the TOST approach while controlling the Type I error rate in the pre-specified interval as long as the controlled interval is not too wide. If, for example, the

interval is chosen to represent the complete parameter space, no gain in power is possible with the proposed hybrid approach.

The decision on the width of the region in which the Type I error rate needs to be controlled depends on how confident the sponsor of the new study is of matching the study result from the historical trials. If, for example, the sponsor was involved in one of the historical trials, it might be easy to set up an identical study using the same research centers and exactly the same inclusion/exclusion criteria. In that case, the sponsor might be able to justify that it is sufficient to control the Type I error rate in a narrow interval. On the other hand, if not much knowledge about the historical trials is publicly available, it might be difficult to set up a comparable study and therefore a wider interval would be necessary. It is important to emphasize that there are also situations in which the proposed approach is not recommended (e.g., in cases in which there is a high uncertainty about the response rate in the new study). However, using simulations, it is easy to investigate the usefulness of applying the hybrid approach at the planning stage of the new trial.

The Type I error rate and power profile can be adjusted by the choice of tuning parameters. We have given general advice on the choice of the parameters and provided an algorithm that chooses the parameters automatically. In addition, we explained the fine tuning of the parameters in Section 5 using historical data for the active substance adalimumab (originator drug name: Humira[®] from AbbVie Ltd.) as a case study.

We only aim to control the Type I error rate in the region surrounding the mean value of the prior distribution. Clearly, this may lead to discussions with regulatory agencies as to whether the evidence obtained with this method is reliable enough to be included in the showing of equivalence between a biosimilar and a reference product. We believe, however, that our hybrid approach is an improvement of current practice where weaker evidence has been used to gain regulatory approval in the past. For example, in the area of new antiepileptic drugs (monotherapy), a single-arm study was accepted in which the efficacy comparison was based on comparing to historical control only (Jacobson et al. 2015, Sperling et al. 2015). Also in the context of biosimilar development, single-arm studies with a comparison to historical data have already been used, for example for the application of Zarzio in Europe (Sandoz) (CHMP 2009). Another example is non-inferiority (NI)

trials for which the FDA guideline states that the non-inferiority study "is dependent on knowing something that is not measured in the study, namely, that the active control had its expected effect in the NI study" (FDA 2016). In all these examples the Type I error rate is uncontrolled and can lie anywhere between 0% and 100%. It is acknowledged that the single-arm study for Zarzio was considered supportive only and the main confirmation of equivalence was based on a pharmacodynamic study which involved a control group. Also, biosimilar trials differ from antiepileptic trials in the sense that using placebo in patients for which an effective treatment is already well-established is ethically questionable, whereas a control group in a biosimilar trial is less controversial. However, we would like to emphasise that the Type I error rate profile obtained with our method is, from a statistical point of view, much more preferable than some approaches already used in practice because, on the one hand, the statistical properties of our method are known and can be taken into consideration when making the decision if a biosimilar is approved or not and, on the other hand, we guarantee partial control of the Type I error rate.

This paper has focussed on a binary endpoint. However, the general idea can be adapted to other types of endpoints, e.g., normally distributed endpoints. In that case, using concepts like Bayes factors (e.g., Lavine & Schervish 1999) might be useful for combining multiple parameter estimates (e.g., mean value and variance for normally distributed endpoints) into a measure for similarity which is required for the proposed switching rules. It should also be noted that the approach was developed within the biosimilar framework, but can be used in all situations in which strict Type I error rate control for scenarios that are realistic in practice is required. Finally, we would like to point out that there several options for modifying standard methodology to achieve the desired Type I error rate and power profile. We found that the proposed approach has reasonable and acceptable operating characteristics, however, we acknowledge that further improvement might be possible and could be an area of further research.

Acknowledgment We would like to thank two anonymous reviewers for helpful and constructive comments. We also gratefully acknowledge the funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 633567 and from the Swiss State Secretariat for Education, Research and In-

novation (SERI) under contract number 999754557. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Swiss Government.

References

- Bischi, B., Lang, M., Mersmann, O., Rahnenführer, J. & Weihs, C. (2015), ‘BatchJobs and BatchExperiments: Abstraction mechanisms for using R in batch environments’, *Journal of Statistical Software* **64**(1), 1–25.
- Blaauvelt, A., Papp, K. A., Griffiths, C. E., Randazzo, B., Wasfi, Y., Shen, Y.-K., Li, S. & Kimball, A. B. (2017), ‘Efficacy and safety of guselkumab, an anti-interleukin-23 monoclonal antibody, compared with adalimumab for the continuous treatment of patients with moderate to severe psoriasis: Results from the phase III, double-blinded, placebo- and active comparator-controlled VOYAGE 1 trial’, *Journal of the American Academy of Dermatology* **76**(3), 405 – 417.
- Chiu, S.-T., Liu, J.-P. & Chow, S.-C. (2014), ‘Applications of the bayesian prior information to evaluation of equivalence of similar biological medicinal products’, *Journal of Biopharmaceutical Statistics* **24**(6), 1254–1263.
- CHMP (2009), ‘Zarzio: EPAR - public assessment report’. Available at http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-/_Public/_assessment/_report/human/WC500046528.pdf (accessed 22 Nov 2017).
- CHMP (2017), ‘Amgevita: EPAR - public assessment report’. Available at http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-/_Public/_assessment/_report/human/004212/WC500225231.pdf (accessed 22 Nov 2017).
- DiMasi, J. A., Hansen, R. W. & Grabowski, H. G. (2003), ‘The price of innovation: new estimates of drug development costs’, *Journal of Health Economics* **22**(2), 151 – 185.
- FDA (2009), ‘Biologics Price Competition and Innovation Act.’. Available at <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/ucm216146.pdf> (accessed 29 Sep 2016).

- FDA (2016), ‘Non-inferiority clinical trials to establish effectiveness’. Available at <https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf> (accessed 28 July 2017).
- Fredriksson, T. & Pettersson, U. (1978), ‘Severe psoriasis—oral therapy with a new retinoid’, *Dermatology* **157**(4), 238–244.
- Gelman, A. (2006), ‘Prior distributions for variance parameters in hierarchical models’, *Bayesian Analysis* **1**(3), 515–534.
- Generics and Biosimilars Initiative (2017*a*), ‘Biosimilars approved in eu-rope’. Available at <http://www.gabionline.net/Biosimilars/General/Biosimilars-approved-in-Europe> (accessed 02 Aug 2017).
- Generics and Biosimilars Initiative (2017*b*), ‘Biosimilars approved in the us’. Available at <http://gabionline.net/Biosimilars/General/Biosimilars-approved-in-the-US> (accessed 02 Aug 2017).
- Hampson, L. V., Whitehead, J., Eleftheriou, D. & Brogan, P. (2014), ‘Bayesian methods for the design and interpretation of clinical trials in very rare diseases’, *Statistics in Medicine* **33**(4), 4186–4201.
- Hsieh, T.-C., Chow, S.-C., Yang, L.-Y. & Chi, E. (2013), ‘The evaluation of biosimilarity index based on reproducibility probability for assessing follow-on biologics’, *Statistics in Medicine* **32**(3), 406–414.
- Jacobson, M. P., Pazdera, L., Bhatia, P., Grinnell, T., Cheng, H. & Blum, D. (2015), ‘Efficacy and safety of conversion to monotherapy with eslicarbazepine acetate in adults with uncontrolled partial-onset seizures: a historical-control phase III study’, *BMC Neurology* **15**, 46–59.
- Johnson, S. G. (2014), ‘The NLOpt nonlinear-optimization package’. R package version 1.0.4.
- Kaelo, P. & Ali, M. M. (2006), ‘Some variants of the controlled random search algorithm for global optimization’, *Journal of Optimization Theory and Applications* **130**(2), 253–264.
- Lavine, M. & Schervish, M. J. (1999), ‘Bayes factors: What they are and what they are not’, *The American Statistician* **53**(2), 119–122.

- Menter, A., Tying, S. K., Gordon, K., Kimball, A. B., Leonardi, C. L., Langley, R. G., Strober, B. E., Kaul, M., Gu, Y., Okun, M. & Papp, K. (2008), ‘Adalimumab therapy for moderate to severe psoriasis: A randomized, controlled phase III trial’, *Journal of the American Academy of Dermatology* **58**(1), 106 – 115.
- Mielke, J., Jilma, B., Koenig, F. & Jones, B. (2016), ‘Clinical trials for authorized biosimilars in the European Union: a systematic review’, *British Journal of Clinical Pharmacology* **82**(6), 1444–1457.
- Morita, S., Thall, P. F. & Mueller, P. (2008), ‘Determining the effective sample size of a parametric prior’, *Biometrics* **64**(2), 595–602.
- Neuenschwander, B., Capkun-Niggli, G., Branson, M. & Spiegelhalter, D. J. (2010), ‘Summarizing historical information on controls in clinical trials’, *Clinical Trials* **7**(1), 5–18.
- Price, W. L. (1983), ‘Global optimization by controlled random search’, *Journal of Optimization Theory and Applications* **40**(3), 333–348.
- R Core Team (2015), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Reich, K., Armstrong, A. W., Foley, P., Song, M., Wasfi, Y., Randazzo, B., Li, S., Shen, Y.-K. & Gordon, K. B. (2017), ‘Efficacy and safety of guselkumab, an anti-interleukin-23 monoclonal antibody, compared with adalimumab for the treatment of patients with moderate to severe psoriasis with randomized withdrawal and retreatment: Results from the phase III, double-blind, placebo- and active comparator-controlled VOYAGE 2 trial’, *Journal of the American Academy of Dermatology* **76**(3), 418 – 431.
- Saurat, J.-H., Stingl, G., Dubertret, L., Papp, K., Langley, R., Ortonne, J.-P., Unnebrink, K., Kaul, M., Camez, A. & for the CHAMPION Study Investigators (2008), ‘Efficacy and safety results from the randomized controlled comparative study of adalimumab vs. methotrexate vs. placebo in patients with psoriasis (CHAMPION)’, *British Journal of Dermatology* **158**(3), 558–566.
- Schmidli, H., Gsteiger, S., Roychoudhury, S., O’Hagan, A., Spiegelhalter, D. & Neuenschwander, B. (2014), ‘Robust meta-analytic-predictive priors in clinical trials with historical control information’, *Biometrics* **70**(4), 1023–1032.
- Schuirman, D. (1987), ‘A comparison of the Two One-Sided Tests procedure and the

- Power Approach for assessing the equivalence of average bioavailability’, *Journal of Pharmacokinetics and Biopharmaceutics* **15**(6), 657–680.
- Sperling, M. R., Harvey, J., Grinnell, T., Cheng, H., Blum, D. & the 045 Study Team (2015), ‘Efficacy and safety of conversion to monotherapy with eslicarbazepine acetate in adults with uncontrolled partial-onset seizures: A randomized historical-control phase III study based in north america’, *Epilepsia* **56**(4), 546–555.
- Spiegelhalter, D., Abrams, K. & Myles, J. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley.
- Thaçi, D., Ortonne, J.-P., Chimenti, S., Ghislain, P.-D., Arenberger, P., Kragballe, K., Saurat, J.-H., Khemis, A., Sprøgel, P., Esslinger, H.-U., Unnebrink, K. & Kupper, H. (2010), ‘A phase IIb, multicentre, randomized, double-blind, vehicle-controlled study of the efficacy and safety of adalimumab with and without calcipotriol/betamethasone topical treatment in patients with moderate to severe psoriasis: the BELIEVE study’, *British Journal of Dermatology* **163**(2), 402–411.
- Torres, T. & Puig, L. (2015), ‘Treatment goals for psoriasis: Should PASI 90 become the standard of care’, *Actas Dermosifiliogr* **106**(3), 155–157.
- Tsou, H.-H., Chang, W.-J., Hwang, W.-S. & Lai, Y.-H. (2013), ‘A consistency approach for evaluation of biosimilar products’, *Journal of Biopharmaceutical Statistics* **23**(5), 1054–1066.
- Weber, S. (2017), *RBesT: R Bayesian Evidence Synthesis Tools*. R package version 1.1-0.