Published in final edited form as:

Stat Methods Med Res. 2018 May ; 27(5): 1547–1558. doi:10.1177/0962280216665418.

Detecting and Accounting for Violations of the Constancy Assumption in Non-Inferiority Clinical Trials

Joseph S. Koopmeiners¹ and Brian P. Hobbs²

¹Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, USA

²Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, USA

Abstract

Randomized, placebo-controlled clinical trials are the gold standard for evaluating a novel therapeutic agent. In some instances, it may not be considered ethical or desirable to complete a placebo-controlled clinical trial and, instead, the placebo is replaced by an active comparator (AC) with the objective of showing either superiority or non-inferiority to the AC. In a non-inferiority trial, the experimental treatment is considered non-inferior if it retains a pre-specified proportion of the effect of the AC as represented by the non-inferiority margin. A key assumption required for valid inference in the non-inferiority setting is the constancy assumption, which requires that the effect of the AC in the non-inferiority trial is consistent with the effect that was observed in previous trials. It has been shown that violations of the constancy assumption can result in a dramatic increase in the rate of incorrectly concluding non-inferiority in the presence of ineffective or even harmful treatment. In this paper, we illustrate how Bayesian hierarchical modeling can be used to facilitate multi-source smoothing of the data from the current trial with the data from historical studies, enabling direct probabilistic evaluation of the constancy assumption. We then show how this result can be used to adapt the non-inferiority margin when the constancy assumption is violated and present simulation results illustrating that our method controls the type-I error rate when the constancy assumption is violated, while retaining the power of the standard approach when the constancy assumption holds. We illustrate our adaptive procedure using a non-inferiority trial of raltegravir, an antiretroviral drug for the treatment of HIV.

Keywords

Non-inferiority trial; constancy assumption; Bayesian hierarchical modeling; multi-source smoothing; HIV

Reprints and permission: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Joseph S. Koopmeiners, Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455, USA, koopm007@umn.edu.

Supplemental material: The supporting materials referenced in Sections 3 and 4 are available at the Statistical Methods in Medical Research Website.

1 Introduction

Randomized, placebo-controlled clinical trials are the gold standard for evaluating a novel therapeutic agent. In some instances, it may not be considered ethical or desirable to complete a placebo-controlled clinical trial and, instead, a non-inferiority trial utilizing an active comparator (AC) is considered. For example, it is generally considered unethical to randomize subjects to placebo if an approved treatment is available. In this case, investigators may attempt to show that a new treatment provides an improvement over placebo by implementing a trial designed to demonstrate that the new treatment is non-inferior to the currently approved treatment. Alternately, a non-inferiority design may be desirable if the new treatment is not expected to result in an efficacy benefit compared to the existing treatment but is expected to have an improved safety profile or some other secondary benefit to the patient. In this case, the new treatment would be viewed as an improvement over the existing treatment as long as it is non-inferior to the existing treatment for key efficacy endpoints.

Statistical methods for non-inferiority trials have been studied extensively in the literature¹⁻³. Statistical inference in a non-inferiority trial requires pre-specification of a non-inferiority margin defined on the domain of the model parameter characterizing the difference in efficacy between treatments. The non-inferiority margin determines the extent to which clinical effectiveness for the experimental treatment (ET) may be diminished in relation to the AC and still qualify as non-inferior. The new treatment is considered non-inferior if the lower limit of the 95% confidence interval for the effect of the new treatment relative to the AC exceeds the non-inferiority margin. For example, in a non-inferiority trial with a binary endpoint, one might summarize the effect of the ET using the difference in the response rate between the ET and AC and set the non-inferiority margin equal to 0.1. In this case, the ET would be declared non-inferior to the AC if the lower limit of the 95% confidence interval for the difference exceeds -0.10. Several approaches have been proposed for specifying the non-inferiority margin⁴⁻⁶ but currently the statistical community is devoid of a consensus approach.

An important assumption required for valid inference in the non-inferiority setting is the assumption of inter-trial constancy. Under the constancy assumption, it is assumed that the effect of the AC in the current trial is consistent with the effect that was observed in previous trials, and thereby enables superiority with respect to placebo to be inferred from the information acquired over the sequence of trials. Results of an extensive simulation study illustrating that violations of the constancy assumption due to inter-trial heterogeneity can result in a dramatic increase in the rate of incorrectly concluding non-inferiority in the presence of ineffective or even harmful treatments have been reported in the literature⁷. These results highlight the need for statistical methods that utilize all available information to detect and account for violations of the constancy assumption in non-inferiority clinical trials.

Two recent papers have discussed statistical approaches that account for violations of the constancy assumption in non-inferiority clinical trials^{8,9}. Odem-Davis and Fleming⁸ develop a bias-adjusted non-inferiority margin that accounts for both bias and uncertainty in the

historical treatment effect of the AC, while Liu, Li and Odem-Davis⁹ develop a robust range that allows investigators to estimate the degree to which the non-inferiority margin is robust to bias in the historical estimate of the treatment effect of the AC. Alternately, one could consider an adaptive approach that evaluates the constancy assumption and adapts the non-inferiority margin accordingly.

In this manuscript, we propose an adaptive approach for detecting and accounting for violations of the constancy assumption in non-inferiority clinical trials. First, we illustrate how Bayesian hierarchical modeling facilitates multi-source smoothing of the data from the current trial with the data from historical studies, enabling direct probabilistic evaluation of the constancy assumption. We then propose an approach to adapt the non-inferiority margin when violations of the constancy assumption are detected. Our approach controls the type-I error rate when the constancy assumption is violated, while achieving power that is equal or greater than the standard approach when the constancy assumption holds. The later goal of maintaining power is achievable using hierarchical priors that facilitate flexible data-dependent partial pooling of the concurrent and historical data^{10–12}, incorporating more information from historical trials when the constancy assumption holds. This results in greater precision that overcomes the additional variability due to adapting the non-inferiority margin.

The remainder of this manuscript proceeds as follows. In Section 2, we present an overview of Bayesian methods for incorporating historical information in clinical trials and present two Bayesian hierarchical models that enable dynamic multi-source smoothing of information pertaining to the clinical effectiveness of the AC. We then illustrate how these models can be used to detect violations of the constancy assumption in Section 3 and propose an approach for adapting the non-inferiority margin when violations of the constancy assumption are detected in Section 4. We illustrate our adaptive procedure using a non-inferiority trial of raltegravir, an antiretroviral drug for the treatment of HIV in Section 5. An extension of our approach to the case of multiple historical trials is discussed in Section 6 and we conclude with a brief discussion in Section 7.

2 Bayesian Hierarchical Models for Multi-Trial Smoothing

2.1 Overview and Notation

Pocock¹³ was the first to discuss the use of Bayesian models for incorporating historical information in clinical trials in settings where an equal randomized design is infeasible due to lack of patient/financial resources or ethical considerations. More recently, it has become a matter of convention that incorporating historical data into trial analysis derives most naturally from the Bayesian perspective^{14–16} through the prior distribution, either explicitly, using the historical data likelihood^{17;18}, or implicitly, through expert elicitation¹⁹.

The aforementioned approaches facilitate partial pooling of concurrent and historical data, wherein the extent to which the model borrows strength from the historical controls is pre-determined by the prior's effective sample size. This “static” approach to incorporating historical data is well known to be sensitive to unforeseen inter-trial effects, yielding biased estimators that may lead to invalid conclusions when the current data are inconsistent with

the historical data. More recently, “dynamic” approaches for incorporating historical controls using hierarchical modeling with sparsity inducing spike-and-slab hyperpriors as well as empirical Bayesian inference have been proposed^{10–12}. For these models, the extent of shrinkage towards the historical information isn't predetermined, but rather estimated from data, yielding a stochastic effective sample size for posterior inference. The resulting dynamic Bayesian estimators borrow more strength in the absence of evidence for trial effects, thereby controlling the extent of bias induced from using the historical information.

In the non-inferiority setting, the dynamic approach provides a natural framework for evaluating the constancy assumption, while also providing the additional benefit of achieving more precise estimates of the treatment effect when the constancy assumption is valid. In this section, we present two hierarchical Bayesian models (also referred to as “commensurate prior” models) that facilitate dynamic borrowing. Thereafter, we will explain how these models can be used to detect violations of the constancy assumption in Section 3 and adapt the non-inferiority margin when violations of the constancy assumption are detected in Section 4.

The remainder of this manuscript will make use of the following assumptions and notation. We assume that data are available from a sequence of two clinical trials. The first study provides the historical data for the AC, which is compared to placebo in a randomized, controlled clinical trial. In the second study (i.e. the current trial) the ET is compared to the AC in a randomized clinical trial. Let $y_{H,1}, y_{H,2}, \dots, y_{H,n_H}$ denote binary outcome variables for n_H subjects randomized to the AC in the historical study with response probability p_H , where $y_{H,j}$ indicates the presence of therapeutic response for the j th patient. Similarly, we let $y_{C,1}, y_{C,2}, \dots, y_{C,n_C}$ denote the presence of response for each of n_C subjects randomized to the AC in the current study with response probability p_C , and let $y_{E,1}, y_{E,2}, \dots, y_{E,n_{ET}}$ denote response for each of n_{ET} subjects randomized to ET in the current study with response probability p_{ET} . We note that data for subjects randomized to placebo in the historical study will not be utilized in our hierarchical models since information pertaining to placebo is not observed in the second study.

2.2 Generalized Linear Model

The first model we consider is a commensurate prior model for generalized linear models (GLMs) using the logit link function¹¹. Let parameter λ_H denote the log-odds of obtaining clinical response for patients randomized to the AC in the historical study,

$\log\left(\frac{p_H}{1-p_H}\right) = \lambda_H$. Using similar notation, we let parameter λ_C denote the log-odds of obtaining clinical response for each patient randomized to AC in the current study,

$\log\left(\frac{p_C}{1-p_C}\right) = \lambda_C$, and let parameter β characterize the treatment effect for patients randomized to ET in the current study, which is modeled as an additive adjustment of the

log-odds, $\log\left(\frac{p_{ET}}{1-p_{ET}}\right) = \lambda_C + \beta$.

Information pertaining to the effectiveness of the AC can be found in both studies. This allows us to formulate a hierarchical model that accounts for the potential for inter-trial

effects that reduce the comparability of the study populations and limit our ability to assess the extent to which the ET may be considered superior to placebo. This requires that we avoid assuming that the data are exchangeable, but rather assume that the historical data are partially informative with respect to the current data. We can estimate the extent of partial informativeness in the presence of the data using a hierarchical model.

Let Δ parameterize the between study difference in the log-odds of response for AC: $\Delta = \lambda_C - \lambda_H$. We assume a minimally informative prior for the historical log-odds, λ_H , and formulate the joint model using two assumptions concerning the prior distribution for Δ . First, we assume that the historical data are non-systematically biased *a priori*, i.e. $E(\Delta) = 0$. This assumption is reasonable because it forms the basis of valid inference in the non-inferiority setting (i.e. constancy) and because systematic bias in the historical data would preclude incorporating historical data into our estimation procedure. Second, we assume that the extent to which the historical data influence estimation of β can be characterized by a latent precision parameter specified in the prior for Δ , $\pi(\Delta) = N(0, \tau_1^{-1})$. The hyperparameter τ_1 has been referred to as the “commensurability parameter” in the prior literature.

A hierarchical model is constructed by assuming a hyperprior distribution for τ_1 . Note that a point mass prior fixed at $\tau_1 = \infty$ assumes that the data are exchangeable, which neglects to account for inter-study heterogeneity and risks biased estimators. By way of contrast, assuming point mass $\tau_1 = 0$ induces independence between λ_H and λ_C , reducing efficiency. Both extremities risk misleading conclusions concerning non-inferiority. More generally, large values of τ_1 imply that the data from study 2 are commensurate with (or reasonably consistent with) the data from study 1, inducing more shrinkage for estimating λ_C towards the information pertaining to λ_H . Conversely, small values for τ_1 reflect a lack of reproducibility with respect to the information acquired in the historical study, limiting the extent of shrinkage.

Estimation of τ_1 from the data is intrinsically difficult, but feasible by inducing sparsity over the precision domain using a spike-and-slab mixture density^{11;20}. Formally, the mixture distribution has two components. It is locally uniform on the “slab,” $[0, u]$, with probability ν , and a point mass at the “spike,” K , with probability $1 - \nu$,

$$\pi(\tau_1) \equiv \nu U(0, u) + (1 - \nu)K, \text{ where } 0 < u \ll K, \text{ and } \nu \in (0, 1). \quad (1)$$

Estimation of τ_1 is to some extent always imprecise in the absence of strong evidence against inter-trial heterogeneity. Yet, this hierarchical model induces a smoothing procedure that has been shown to yield desirable frequentist properties for estimating β using all of the available information. Intuitively, $\tau_1 = K$ represents the condition of commensurability between the two trials, while $0 < \tau_1 < u$ characterizes varying degrees of heterogeneity (or lack of reproducibility). Posterior inference on τ_1 is insensitive to modest shifts in K and u , but not in ν , which represents the prior mixture weight associated with the slab. Generally, a

value for ν needs to be calibrated in the context of the problem in consideration of both the expectation and tolerance for bias.

The joint posterior distribution for the full parameter vector, $\Theta_1 = (\lambda_H, \lambda_C, \beta, \tau_1)$ can be found in the Appendix. Draws from the posterior distribution can be obtained using MCMC sampling implemented in WinBUGS²¹ or JAGS²².

2.3 Beta Commensurate Prior

In the previous section, we formulated a hierarchical model that characterized commensurability by assuming normality on the log-odds domain. Alternatively, a hierarchical model could be formulated to measure commensurability on the probability domain using properties of the beta distribution. Let $Beta(a, b)$ denote the beta distribution under the conventional parameterization, where a and b represent shape parameters. We proceed by assuming the spike-and-slab commensurate prior described in Section 2.1 using the following re-parameterization

$$a = (1 - \gamma^2) \frac{\mu}{\gamma^2} \text{ and } b = (1 - \gamma^2) \frac{(1 - \mu)}{\gamma^2}.$$

Under this parameterization, μ represents the mean of the distribution and γ denotes the variance in terms of the proportion of the maximum variance given μ . To specify a beta commensurate prior for p_C , we set $\mu = p_H$. This approach offers the advantage of characterizing commensurability on the probability scale directly, and thereby avoiding the log-odds transformation. However, the prior for p_C is no longer “centered” about p_H since the beta distribution is not symmetric about the mean.

Finally, we let latent precision parameter $\tau_2 = 1/\gamma$ serve as the commensurability parameter and assume a spike-and-slab mixture prior for τ_2 that differs from (1) only in that the “slab” will be defined on $[1, u]$ rather than $[0, u]$ due to the restricted support of γ . Hyperparameter τ_2 can be interpreted similarly to τ_1 in the previous model, in the sense that large values induce more shrinkage to the historical information. Although, due to the nature of the bounded probability domain, this beta commensurate prior does impose that the mean is a function of the variance, which was absent the previous model. We will compare the performance of this model to the GLM commensurate prior specified above by simulation in Sections 3 and 4.

The joint posterior distribution for the full parameter vector, $\Theta_2 = (p_H, p_C, p_{ET}, \tau_2)$, can be found in the Appendix. As with the GLM, draws from the posterior distribution can be obtained using MCMC sampling implemented in WinBUGS²¹ or JAGS²².

3 Detecting Violations of the Constancy Assumption

We now return to the setting of non-inferiority clinical trials. The constancy assumption states that the effect of the AC is consistent with what was observed in previous studies. It is clear that the presence of inter-trial effects, as discussed in Section 2.1, implicates violation

of the constancy assumption. Moreover, in addition to allowing for the incorporation of historical information, the hierarchical models presented in Sections 2.2 and 2.3 provide a probabilistic basis for testing the constancy assumption in the presence of data observed in the non-inferiority trial. Specifically, the spike and slab prior specified for τ_j enables posterior inference as it pertains to the constancy assumption directly (we note, here, that j indexes the model and $j = 1, 2$). Recall that using the spike and slab prior, $\tau_j = K$ implies commensurability/constancy. In this case, the posterior probability that the constancy assumption is violated is equal to the posterior probability that $\tau_j < K$ or $P(\tau_j < K | \vec{y}_H, \vec{y}_C, \vec{y}_{ET})$.

We completed a small simulation study to evaluate our ability to detect violations of the constancy assumption using the models discussed in Section 2. We simulated a sequence of two clinical trials. In the first clinical trial, outcomes for the AC group were drawn from a Bernoulli distribution with a response probability of 0.80. In the second clinical trial, outcomes for the AC group were simulated from a Bernoulli distribution with response probability of 0.70, 0.75 or 0.80, while outcomes for the ET group were simulated from a Bernoulli distribution with a response probability of 0.80. Simulations were completed with 200 subjects/group and 400 subjects/group in each trial. These sample sizes are consistent with the sample sizes of non-inferiority trials reported in the literature, including our example in Section 5. For the results presented, we set $K = 40$, $u = 4$ and $\nu = 0.5$ for the GLM commensurate prior model and $K = 400$, $u = 4$ and $\nu = 0.50$ for the beta commensurate prior model. In addition, we also completed additional simulations varying these parameters for both models in order to determine the sensitivity of our results to the specification of the “spike-and-slab” prior. These results are presented in the Supplementary Materials and will also be discussed below.

Simulation results can be found in Figure 1. We present our results in the form of an ROC curve for $P(\tau_j < K | \vec{y}_H, \vec{y}_C, \vec{y}_{ET})$. In this case, the x-axis represents the type-I error rate, the y-axis is the power and the ROC curve, $\text{ROC}(t)$, is the power corresponding to a type-I error rate of t . Both models performed similarly well in their ability to detect violations of the constancy assumption. For a difference of 0.1 between the current and historical trial ($P_C = 0.70$ vs. $P_H = 0.80$), both models exhibited approximately 90% power for $N = 400$ per group and 60% power for $N = 200$ per group with a type-I error rate of 0.05. In contrast, neither method exhibited substantial power for detecting moderate violations of the constancy assumption ($P_C = 0.75$ vs. $P_H = 0.80$; the beta commensurate prior exhibited 30% power and the GLM model exhibited 35% power for $N = 400$ per group with a type-I error rate of 0.05). Results obtained from additional simulations evaluating robustness to the specification of the “spike-and-slab” prior can be found in the Supplementary Materials. The resulting ROC curves are very similar when varying the hyperparameters K , u and ν illustrating that our results are robust to specification of the “spike-and-slab” prior.

4 Adapting the Margin

The results presented in Section 3 illustrate that commensurate priors can be used to detect violations of the constancy assumption. We now discuss how this knowledge can be

incorporated into the analysis of non-inferiority clinical trials by adapting the non-inferiority margin when violations of the constancy assumption are detected.

Let Γ denote the non-inferiority margin. In a standard non-inferiority trial, the ET is declared non-inferior to the AC if:

$$P(P_{ET} - P_C > \Gamma | \vec{y}_C, \vec{y}_{ET}) > \psi.$$

A number of authors have discussed approaches for choosing a non-inferiority margin^{4–6} but, for the purposes of this discussion, the method used to determine the non-inferiority margin is not important and we proceed assuming only that an appropriate non-inferiority margin was determined prior to commencing the trial.

The hierarchical models described in Section 2 provide two key pieces of information for adapting the non-inferiority margin: 1) the posterior probability of non-constancy, which can be used to identify when it is appropriate to adapt the margin, and 2) the estimated difference of the effect of the AC in the current trial versus the AC in the historical trial, $\Omega = p_H - p_C$. A natural approach to adapting the non-inferiority margin is to observe the posterior probability of non-constancy and adapt the margin if it exceeds some pre-specified threshold, ϕ . Recall that $P(\tau_j < K | \vec{y}_H, \vec{y}_C, \vec{y}_{ET})$ is the posterior probability of non-constancy for models $j = 1, 2$. In this case, our adapted non-inferiority margin is equal to Γ^* , where:

$$\Gamma^* = \begin{cases} \Gamma & \text{if } P(\tau_j < K | \vec{y}_H, \vec{y}_C, \vec{y}_{ET}) < \phi \text{ or } \hat{\Omega} < 0 \\ \Gamma - \hat{\Omega} & \text{if } P(\tau_j < K | \vec{y}_H, \vec{y}_C, \vec{y}_{ET}) > \phi \text{ and } 0 < \hat{\Omega} < \Gamma \\ 0 & \text{if } P(\tau_j < K | \vec{y}_H, \vec{y}_C, \vec{y}_{ET}) > \phi \text{ and } \Gamma < \hat{\Omega} \end{cases}$$

where $\hat{\Omega}$ represents the posterior mean of the difference in the response probabilities of the AC in the current versus historical trial. Our adaptive procedure has three cases. First, we do not adapt the margin if the posterior probability of non-constancy is less than ϕ or if the response probability of the AC in the current trial is greater than in the historical trial. Second, we adapt the non-inferiority margin by the difference between the response probabilities of the AC in the historical and current trial if this difference is less than Γ and the posterior probability of non-constancy exceeds ϕ . Finally, we set the margin equal to 0 and complete a superiority trial if the difference between the response probabilities of the AC in the historical and current trial is greater than the non-inferiority margin and if the posterior probability of non-constancy exceeds ϕ .

There are two advantages of the proposed adaptive approach. First, our approach allows us to adapt the non-inferiority margin and control the type-I error rate when the constancy assumption is violated. Controlling the type-I error rate is of primary importance to regulatory agencies and, from this perspective, our adaptive approach represents an improvement over the fixed-margin approach, which can inflate the type-I error rate if the constancy assumption is violated⁷. A second, perhaps more subtle advantage, is that the hierarchical models will use the information found in the historical trial to estimate the

response probability of the AC in relation to the extent of evidence supporting the constancy assumption. This will result in a more precise estimate of the difference between the ET and the AC, which will allow us to retain or improve upon the power of the standard approach when the constancy assumption holds.

We completed a small simulation study to evaluate the operating characteristics of our proposed adaptive procedure. We again simulated a sequence of two clinical trials. In the first clinical trial, outcomes for the AC group were drawn from a Bernoulli distribution with a response probability of 0.80. In the second clinical trial, the non-inferiority margin, Γ was set equal to 0.10 and outcomes for the AC group were simulated from a Bernoulli distribution with response probability of 0.70, 0.725, 0.75, 0.775 or 0.80, while outcomes for the ET group were simulated from a Bernoulli distribution with a response probability of either 0.70 (i.e. $0.80 - \Gamma$, to evaluate the type-I error rate) or 0.80 (to evaluate power). We note here that we are defining type-I and type-II errors relative to the historical treatment effect of the AC. That is, we define a type-I error as the case where the treatment effect of the ET is inferior to the historical treatment effect of the AC but we declare non-inferiority and we define a type-II error as the case where the treatment effect of the ET is equal to the historical treatment effect of the AC but we fail to declare non-inferiority. Simulations were completed with 200 subjects/group and 400 subjects/group in each trial and we used the same hyperparameter specifications as in Section 3. We considered values of ϕ determined by identifying thresholds from our simulations in Section 3 such that the model would detect non-constancy and adapt the margin either 60% or 80% of the time when $P_C = 0.75$. For the GLM model, we considered values of ϕ equal to 0.3232 and 0.2800 when $n = 200/\text{group}$ and 0.33465 and 0.26445 when $n = 400/\text{group}$. For the beta commensurate prior, we considered values of ϕ equal to 0.1204 and 0.0995 when $n = 200/\text{group}$ and 0.12575 and 0.0884 when $n = 400/\text{group}$.

For comparison, we also considered two approaches to evaluating non-inferiority where the data from the second trial were fit using reference priors instead of the hierarchical priors described in Section 2. First, we considered the standard approach to evaluating non-inferiority, where non-inferiority was evaluated by considering $P(P_{ET} - P_C > \Gamma | \vec{Y}_G, \vec{Y}_{ET})$. Second, we considered a naive approach that adapted the margin for all trials with the adapted margin equaling Γ if $\hat{\Omega} = 0$, $\Gamma - \hat{\Omega}$ if $\hat{\Omega} < \Gamma$ and 0 if $\Gamma < \hat{\Omega}$.

In addition to evaluating the type-I error and power of our adaptive procedure, we also present the effective historical sample size (EHSS) as a measure of the gain in posterior precision induced by dynamically incorporating historical data through our hierarchical models. Effective sample size (ESS) is a measure of the extent of *a posteriori* information that results from updating the information in the prior with the information in the likelihood^{23;24}. Following the approach of Hobbs, Carlin and Sargent²⁵, we calculate the EHSS as the difference between the ESS and the number of observations randomized to the AC in the second trial, n_C . In a “static” model, the EHSS is determined *a priori* through prior specification but in a dynamic model, like those discussed in Section 2, EHSS is stochastic and will depend on the consistency between the data in the current trial and the historical trial. Therefore, we will report the average EHSS for different values of P_C to illustrate the

dynamic borrowing properties of the proposed hierarchical models. A full introduction of EHSS can be in the Supplementary Material.

Simulation results obtained using $n = 200$ subjects per group can be found in Figure 2. Violations of the constancy assumption inflate the type-I error rate for the standard approach with type-I error rates exceeding 0.5 when $P_C = 0.70$. In contrast, the proposed adaptive approach controls the type-I error rate under even extreme violations of the constancy assumption for both the GLM and beta commensurate prior models. Both models also exhibit higher power than the naive adaptive approach, which is not surprising because the commensurate prior approach only adapts when necessary and incorporates historical information for more precise estimates when the constancy assumption holds. Furthermore, both models also meet or exceed the power of the standard model when constancy holds (represented by the solid black line) for all values of P_C . Finally, we see that the EHSS increases with P_C for both models and reaches a maximum when $P_C = 0.80$ (i.e. when constancy holds). This illustrates the advantage of using our hierarchical approach for inducing dynamic borrowing. Dynamic borrowing allows our model to down-weight the historical data when they are not consistent with the current data, which decreases bias and, along with adapting the margin, allows us to control the type-I error rate. Furthermore, while adapting the margin results in additional variability, we are able to maintain power by dynamically incorporating a substantial amount of historical data when the constancy assumption holds. Additional simulation results for $n = 400$ subjects per group yielded similar results and can be found in the Supplementary Materials.

5 Illustration

In this section, we illustrate the application of our method to a non-inferiority trial for once daily raltegravir²⁶. Raltegravir is an antiretroviral drug used as part of combination therapy for the treatment of HIV. A previous definitive trial of raltegravir illustrated that a twice-daily raltegravir-based regimen was non-inferior to an efavirenz-based regimen with a primary endpoint of virological response at 48 weeks (defined as viral RNA loads < 50 copies per mL)²⁷. In the current trial, researchers are interested in determining whether a once-daily raltegravir-based regimen is non-inferior to the twice-daily regimen using the same primary endpoint. The non-inferiority margin was determined in advance to be -10% and the once-daily regimen will be declared non-inferior to the twice-daily regimen if the lower limit of the 95% credible interval for the difference in the probability of virological response at 48 weeks exceeds -10%.

Table 1 presents the results of the original definitive trial for twice-daily raltegravir (Lennox 2009)²⁷ and the non-inferiority trial for once-daily raltegravir (Eron 2011)²⁶. Here, the standard model refers to a binomial model with a uniform (0,1) prior on the probability of virological response for each group. Using the standard analysis, which is consistent with the original, published results²⁶, the lower-limit for the 95% credible interval for the difference in the virological response rate is less than -10.0 and once-daily raltegravir fails to attain non-inferiority with respect to twice-daily raltegravir.

We next consider the two commensurate prior models fit using the hyperparameters specified in Section 3. We first observe that there is slight evidence for a violation of the constancy assumption but we would not adapt the margin in either case regardless of our value for ϕ because the estimated rate of virological response in the current trial is greater than the estimated rate of virological response in the previous trial. In addition, this example illustrates a secondary benefit of our method, which is that the hierarchical models incorporate a substantial amount of historical information (EHSS = 76.4 and 109.6 for the GLM and beta models, respectively, 27% and 39% of the 280 subjects in the first trial), resulting in a more precise estimate of the probability of virological response in the current trial. As a result, the lower limit of the 95% credible interval for the difference in the rate of virological response now exceeds the non-inferiority margin. This results in a complicated interpretation of the commensurate prior results because the upper limit of the credible interval for the difference is also less than 0, which suggests superiority of the twice-daily regimen but also meets the pre-specified threshold for non-inferiority. Regardless, this does illustrate a secondary benefit of our approach, which yields enhanced efficiency when the constancy assumption holds or is only slightly violated.

6 Multiple Historical Studies

To this point, we have focused on the case where historical data are available for only a single trial. In practice, it will often be the case that historical information will be available from two or more previous clinical trials evaluating the AC. In this case, we would like to evaluate the constancy assumption by comparing the effect of the AC in the current trial to the effect of the AC in all previous trials. We can extend our approach to accommodate multiple historical trials, as follows. For the remainder of this section, we assume M historical trials and let $y_{m,1}, y_{m,2}, \dots, y_{m,n_m}$ denote binary outcomes variables for the n_m subjects randomized to the AC in the m th historical trial with response probability p_m , where $y_{m,j}$ indicates the presence of therapeutic response for the j th patient in the m th historical trial. Notation for the current study will remain as defined in Section 2.1.

We can extend the commensurate prior model for GLMs to account for multiple historical trials by hierarchically modeling the log-odds of obtaining clinical response for the AC in the historical trials. Let λ_m denote the log-odds of obtaining clinical response for patients

randomized to the AC in the m th historical study, $\log\left(\frac{p_m}{1-p_m}\right) = \lambda_m$. The trial-specific log-odds are assumed to follow a normal distribution centered at the overall historical log-odds with precision η_H : $\lambda_m \sim N(\lambda_H, \eta_H)$ for $m = 1, \dots, M$. Assuming that the data from the current trial are modeled as in Section 2.2, we can evaluate the consistency of the log-odds for the AC in the current trial, λ_C , with the overall historical log-odds from the historical trials, λ_H , by defining $\delta = \lambda_C - \lambda_H$ and assuming $\delta \sim N(0, \tau_3)$. The precision, τ_3 , is the commensurability parameter between the current log-odds and overall historical log-odds, for which we assume the spike-and-slab mixture prior defined by (1). The joint posterior distribution for the full parameter vector, $\Theta_3 = (\lambda_1, \dots, \lambda_M, \lambda_H, \lambda_C, \beta, \tau_3)$, can be found in the Appendix. Draws from the posterior distribution can be obtained using MCMC sampling implemented in WinBUGS²¹ or JAGS²².

The methods described in Sections 3 and 4 can be applied directly to the model described above to detect and account for violations of the constancy assumption in the presence of multiple historical trials. The presence of multiple historical trials would result in a more firmly established historical response rate for the AC. This will allow us to better discriminate between situations when the constancy assumption holds and when the constancy assumption is violated, resulting in a decreased probability of adapting the margin when the constancy assumption holds, thus retaining power, and more precisely adapting the margin when the constancy assumption is violated, resulting in better overall performance for our adaptive procedure.

7 Discussion

We have proposed an adaptive procedure for detecting and accounting for violations of the constancy assumption in non-inferiority clinical trials. We illustrated how Bayesian hierarchical models for dynamically incorporating historical information in clinical trials can be used to detect violations of the constancy assumption and that we are able to calculate the posterior probability that the constancy assumption holds by inducing sparsity over the precision domain using a “spike-and-slab” mixture density as the prior distribution for the commensurability parameter. We then proposed a procedure for adapting the non-inferiority margin when there is high posterior probability that the constancy assumption is violated. The proposed adaptive procedure controls the type-I error rate when the constancy assumption is violated and maintains the power of the standard approach when the constancy assumption holds.

The hierarchical models discussed in this manuscript are for binary outcomes but similar models have also been proposed for normally distributed outcomes and censored survival data^{11;12}. An advantage to binary endpoints is that the sample proportion is a sufficient statistic that is typically reported in published results of a clinical trial. As a result, the proposed adaptive method can be applied without having access to the original data. The same is true for normally distributed outcomes, where the sample mean and variance are sufficient and typically reported, but this is not the case for censored survival data, in which case the full data from the historical trial would need to be available in order to apply the adaptive procedure proposed in this manuscript.

A limitation to our proposed method is that we are unable to distinguish between violations of the constancy assumption due to changes in the effect of the AC or changes in the underlying study population. For example, results were recently reported for a non-inferiority trial of nevirapine in patients co-infected with HIV and tuberculosis, which used evafirenz as the AC²⁸. Evafirenz was previously evaluated in a general HIV positive population²⁹. We would expect the virological response rate of evafirenz to be lower in the nevirapine trial than in the historical trial because patients co-infected with HIV and tuberculosis generally have worse clinical outcomes than the general HIV population. Our method would identify this as a violation of the constancy assumption and adapt the margin but it is not clear that this is of interest scientifically. Care should be taken when applying our method, and when designing non-inferiority trials, in general, to determine the expected effect of the AC if the population in the current trial is dramatically different than in

historical trials. A placebo-controlled trial could be considered if there is clinical equipoise regarding the efficacy of the treatment in the new population but, in general, a randomized trial of the ET vs. the AC would be preferable because the AC has been shown superior to placebo in a different population. In this case, propensity scores could be used to determine the expected effect of the AC in the new population and use our proposed adaptive approach to identify violations of the constancy assumption and adapt the non-inferiority margin, when appropriate. Future work is needed to determine how this approach would be implemented.

We considered hierarchical models that characterize commensurability on both the log-odds scale and on the probability scale, directly. GLMs are known to be limited by assumptions of linearity for treatment effects and in the presence of interactions, which can be difficult to specify, in practice. In this manuscript, we consider the setting of randomized trials where covariates are balanced across treatment groups and adjustment to prevent confounding is not typically required. Therefore, one would expect the GLM version of the hierarchical model to be robust to treatment and trial effects for the response rates that are characterized as linear adjustments on the log-odds domain. While, we do not believe that robustness poses a substantial issue for the hierarchical GLM model, our simulation and case studies do support improved performance for the beta commensurate model in the dynamic borrowing procedure as evident from EHSS, suggesting an advantage from characterizing commensurability on the probability scale, rather than the log-odds scale.

Finally, our proposed adaptive procedure is robust to violations of the constancy assumption but other issues related to the design of non-inferiority trials must also be considered. One of the key decisions in the design of non-inferiority trials is the choice of AC¹. There are often multiple appropriate ACs and, while our method will detect violations of the constancy assumption for the chosen AC, this does not alleviate concerns regarding the sponsor choosing an AC that results in the most favorable comparison for the ET. In addition, our work is primarily concerned with violations of the constancy assumption with respect to the primary efficacy endpoint but the safety profile of the AC may also change over time. The hierarchical models discussed in this manuscript could provide a framework for evaluating the constancy assumption for important safety endpoints but further work is needed to determine how this additional information could be used in combination with the efficacy endpoints to evaluate non-inferiority of the ET.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This work was partially supported by a research grant from Medtronic Inc. (JSK) and NIH grants R01CA157458 and P30CA016672 (BPH).

Appendix

A Posterior for Generalized Linear Model

Let S_H , S_C , and S_{ET} denote the numbers of responses observed for patients receiving AC (in historical and current studies) and ET, respectively. The joint posterior distribution for the full parameter vector, $\Theta_1 = (\lambda_H, \lambda_C, \beta, \tau_1)$ is proportional to,

$$\begin{aligned} p(\Theta_1 | \vec{y}_H, \vec{y}_C, \vec{y}_{ET}) &\propto \left(\frac{1}{1+\exp(-\lambda_H)} \right)^{S_H} \left(\frac{\exp(-\lambda_H)}{1+\exp(-\lambda_H)} \right)^{n_H-S_H} \\ &\times \left(\frac{1}{1+\exp(-\lambda_C)} \right)^{S_C} \left(\frac{\exp(-\lambda_C)}{1+\exp(-\lambda_C)} \right)^{n_C-S_C} \\ &\times \left(\frac{1}{1+\exp(-\lambda_C-\beta)} \right)^{S_{ET}} \left(\frac{\exp(-\lambda_C-\beta)}{1+\exp(-\lambda_C-\beta)} \right)^{n_{ET}-S_{ET}} \\ &\times \{(1-\nu)N(\lambda_C | \lambda_H, \tau_1^{-1}) + \nu N(\lambda_C | \lambda_H, K^{-1})\} \pi(\tau_1) \pi(\lambda_H, \beta), \end{aligned} \quad (2)$$

where $\pi(\lambda_H, \beta)$ represents the product of independent normal priors with mean 0 and large variance.

B Posterior for Beta Commensurate Prior

The joint posterior distribution for parameter vector, $\Theta_2 = (p_H, p_C, p_{ET}, \tau_2)$ resulting from this model is proportional to,

$$\begin{aligned} p(\Theta_2 | \vec{y}_H, \vec{y}_C, \vec{y}_{ET}) &\propto (p_H)^{S_H} (1-p_H)^{n_H-S_H} (p_C)^{S_C} (1-p_C)^{n_C-S_C} (p_{ET})^{S_{ET}} (1-p_{ET})^{n_{ET}-S_{ET}} \\ &\times \{(1-\nu)\text{Beta}(p_C | \mu=p_H, \gamma=\tau_2^{-1}) + \nu\text{Beta}(p_C | \mu=p_H, \gamma=K^{-1})\} \\ &\times \pi(\tau_2) \pi(p_H, p_{ET}), \end{aligned} \quad (3)$$

where $\pi(p_H, p_{ET}) \propto 1$, the product of uniform $\text{Beta}(1, 1)$ priors for p_H and p_{ET} .

C Posterior for Generalized Linear Model with Multiple Historical Trials

Let $S_1, \dots, S_M$, S_C , and S_{ET} denote the numbers of responses observed for patients receiving AC (in the M historical studies and the current study) and ET, respectively. The joint posterior distribution for the full parameter vector, $\Theta_3 = (\lambda_1, \dots, \lambda_M, \lambda_H, \lambda_C, \beta, \tau_3)$ is proportional to,

$$\begin{aligned}
p(\Theta_3 | \vec{y}_1, \dots, \vec{y}_M, \vec{y}_C, \vec{y}_{ET}) &\propto \prod_{m=1}^M \left(\frac{1}{1+\exp(-\lambda_m)} \right)^{S_m} \left(\frac{\exp(-\lambda_m)}{1+\exp(-\lambda_m)} \right)^{n_m-S_m} N(\lambda_m | \lambda_H, \eta_H) \\
&\times \left(\frac{1}{1+\exp(-\lambda_C)} \right)^{S_C} \left(\frac{\exp(-\lambda_C)}{1+\exp(-\lambda_C)} \right)^{n_C-S_C} \\
&\times \left(\frac{1}{1+\exp(-\lambda_C - \beta)} \right)^{S_{ET}} \left(\frac{\exp(-\lambda_C - \beta)}{1+\exp(-\lambda_C - \beta)} \right)^{n_{ET}-S_{ET}} \\
&\times \{ (1-\nu) N(\lambda_C | \lambda_H, \tau_3^{-1}) + \nu N(\lambda_C | \lambda_H, K^{-1}) \} \pi(\tau_3) \pi(\lambda_H, \beta),
\end{aligned}
\tag{4}$$

where $\pi(\lambda_H, \beta)$ represents the product of independent normal priors with mean 0 and large variance.

References

1. Fleming TR. Current issues in non-inferiority trials. *Statistics in Medicine*. 2008; 27(3):317–332. [PubMed: 17340597]
2. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues the encounters of academic consultants in statistics. *Statistics in Medicine*. 2003; 22(2):169–186. [PubMed: 12520555]
3. Fleming TR, Odem-Davis K, Rothmann MD, et al. Some essential considerations in the design and conduct of non-inferiority trials. *Clinical Trials*. 2011; 8(4):432–439. [PubMed: 21835862]
4. Wang SJ, Hung H. Assessing treatment efficacy in noninferiority trials. *Controlled Clinical Trials*. 2003; 24(2):147–155. [PubMed: 12689736]
5. Wang SJ, James Hung HM. Tact method for non-inferiority testing in active controlled trials. *Statistics in Medicine*. 2003; 22(2):227–238. [PubMed: 12520559]
6. Rothmann M, Li N, Chen G, et al. Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine*. 2003; 22(2):239–264. [PubMed: 12520560]
7. Everson-Stewart S, Emerson SS. Bio-creep in non-inferiority clinical trials. *Statistics in Medicine*. 2010; 29(27):2769–2780. [PubMed: 20809482]
8. Odem-Davis K, Fleming TR. Adjusting for unknown bias in noninferiority clinical trials. *Statistics in Biopharmaceutical Research*. 2013; 5(3):248–258.
9. Liu Q, Li Y, Odem-Davis K. On robustness of non-inferiority clinical trial designs against bias, variability, and non-constancy. *Journal of Biopharmaceutical Statistics*. 2014 epub ahead of print.
10. Hobbs BP, Carlin BP, Mandrekar SJ, et al. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*. 2011; 67(3):1047–1056. [PubMed: 21361892]
11. Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*. 2012; 7(3):639–674. [PubMed: 24795786]
12. Murray TA, Hobbs BP, Lystig TC, et al. Semiparametric bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. *Biometrics*. 2014; 70(1):185–191. [PubMed: 24308779]
13. Pocock SJ. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*. 1976; 29:175–188. [PubMed: 770493]
14. Spiegelhalter, DJ., Abrams, KR., Myles, JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. London: Wiley; 2004.

15. Pennello G, Thompson L. Experience with reviewing Bayesian medical device trials. *Journal of Biopharmaceutical Statistics*. 2008; 18:81–115. [PubMed: 18161543]
16. Berry, SM., Carlin, BP., Lee, JJ., et al. *Bayesian Adaptive Methods for Clinical Trials*. Chapman & Hall Ltd; 2011.
17. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Statistical Science*. 2000; 15:46–60.
18. Neuenschwander B, Capkun-Niggli G, Branson M, et al. Summarizing historical information on controls in clinical trials. *Clinical Trials*. 2010; 7:5–18. [PubMed: 20156954]
19. Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*. 1994; 50:337–349. [PubMed: 7980801]
20. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*. 1988; 83(404):1023–1032.
21. Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modelling. *The Statistician: Journal of the Institute of Statisticians*. 1994; 43:169–177.
22. Plummer M. rjags: Bayesian graphical models using MCMC. R package version 3-10. 2013
23. Morita S, Thall PF, Müller P. Determining the effective sample size of a parametric prior. *Biometrics*. 2008; 64:595–602. [PubMed: 17764481]
24. Morita S, Thall PF, Müller P. Prior effective sample size in conditionally independent hierarchical models. *Bayesian Analysis*. 2012; 7:591–614.
25. Hobbs BP, Carlin BP, Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*. 2013; 10:430–440. [PubMed: 23690095]
26. Eron JJ, Rockstroh JK, Reynes J, et al. Raltegravir once daily or twice daily in previously untreated patients with HIV-1: a randomised, active-controlled, phase 3 non-inferiority trial. *Lancet Infect Dis*. 2011; 11(12):907–915. [PubMed: 21933752]
27. Lennox JL, DeJesus E, Lazzarin A, et al. Safety and efficacy of raltegravir-based versus efavirenz-based combination therapy in treatment-naïve patients with HIV-1 infection: a multicentre, double-blind randomised controlled trial. *Lancet*. 2009; 374(9692):796–806. [PubMed: 19647866]
28. Bonnet M, Bhatt N, Baudin E, et al. Nevirapine versus efavirenz for patients co-infected with HIV and tuberculosis: a randomised non-inferiority trial. *Lancet Infect Dis*. 2013; 13(4):303–312. [PubMed: 23433590]
29. Riddler SA, Haubrich R, DiRienzo AG, et al. Class-sparing regimens for initial treatment of HIV-1 infection. *N Engl J Med*. 2008; 358(20):2095–2106. [PubMed: 18480202]

Figure 1.

ROC curves illustrating the power to reject the null hypothesis of constancy corresponding to the specified type-I error rate using the GLM with commensurate prior or beta commensurate prior with $p_H = 0.80$ and p_{AC} equal to 0.70 or 0.75.

Figure 2.

Type-I error rate and power for the standard approach (solid line), naive approach (dashed line) and proposed adaptive approach with two different values for ϕ (dotted line, dotted and dashed line) as a function of P_C for a sample size of 200 subjects/group. The bottom panels present the EHSS for the GLM with commensurate prior and beta commensurate prior as a function of P_C .

Table 1

Results of a non-inferiority trial of raltegravir analyzed using the standard approach and our proposed adaptive approach using the GLM with commensurate prior and commensurate Beta prior.

Model	Twice Daily $\hat{P}_{AC}$ 95% CI	Once Daily $\hat{P}_{ET}$ 95% CI	Difference	95% CI	Post Prob Non-Const	EHSS
Lennox 2009						
Standard	85.8(81.5, 89.6)					
Eron 2011						
Standard	88.7(85.4, 91.6)	83.1(79.2, 86.6)	-5.6	(-10.5, -0.8)	-	0.0
GLM Com. Prior	88.3(85.2, 91.1)	83.2(79.2, 86.8)	-5	(-9.9, -0.3)	0.341	76.4
Com. Beta Prior	88.0(85.0, 90.8)	83.1(79.2, 86.7)	-5	(-9.8, -0.3)	0.113	109.6