Biostatistics (2021) **00**, 00, *pp.* 1–16
doi:10.1093/biostatistics/kxab008

Bayesian adaptive design for concurrent trials involving biologically related diseases

MATTHEW. A. PSIODA*

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

matt_psioda@unc.edu

H. AMY XIA, XUN JIANG

Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA

JIAWEI XU, JOSEPH G. IBRAHIM

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

Summary

We develop a Bayesian design method for a clinical program where an investigational product is to be studied concurrently in a set of clinical trials involving related diseases with the goal of demonstrating superiority to a control in each. The approach borrows information on treatment effectiveness using correlated mixture priors using an analysis procedure that is closely related Bayesian model averaging. Mixture priors are constructed by eliciting conjugate priors based on pessimistic and enthusiastic predictions for the data to be observed for each disease and then by eliciting mixture weights for all possible configurations of the pessimistic and enthusiastic priors across the diseases to be studied. The proposed approach provides a robust framework for information borrowing in settings where the diseases may have endpoints based on different data types. We show via simulation that operating characteristics based on the proposed design framework are favorable compared to those based on information borrowing designs using the Bayesian hierarchical model which is poorly suited for information borrowing when there are different data types underpinning the endpoints across which information is to be borrowed.

Keywords: Bayesian model averaging; Clinical trial design; Conjugate prior; Mixture prior.

1. Introduction

Targeted drug development focused on particular molecular targets has led to the development and implementation of so-called basket trials which can be used to evaluate an investigational product (IP) simultaneously in a variety of disease settings and has caused a shift towards increased use of master protocols (Woodcock and LaVange, 2017). The use of basket trials and master protocols has occurred mostly

*To whom correspondence should be addressed.

in the area of oncology (Park *and others*, 2019) but increasingly they are being used in nononcology settings (e.g., REMAP-CAP for community-acquired pneumonia [NCT02735707]). Targeted treatments for inflammatory diseases present another opportunity for trial designs that borrow information on treatment effectiveness across related diseases. Several products have been developed and proven to be effective for multiple inflammatory diseases. For example, Ustekinumab has received United States Food and Drug Administration (FDA) approval for the treatment of Crohn's disease, plaque psoriasis, and psoriatic arthritis. Another example, Etanercept, has approval for the treatment of rheumatoid arthritis, juvenile idiopathic arthritis, psoriatic arthritis, ankylosing spondylitis, and plaque psoriasis. Innovative clinical trial designs are needed for concurrently evaluating an IP in multiple related disease conditions to increase the efficiency of drug development programs.

In this article, we develop a Bayesian adaptive design framework for a clinical program where an IP is to be studied in a set of clinical trials corresponding to a set of biologically related diseases with the goal of demonstrating superiority to a control in each disease. The approach borrows information on treatment effectiveness using correlated mixture priors. Mixture priors are constructed by eliciting conjugate priors based on pessimistic and enthusiastic predictions for the data to be observed for each disease and then by eliciting mixture weights for all possible configurations of the pessimistic and enthusiastic priors across the diseases to be studied. Inference with the correlated mixture prior is operationalized by combining inferences from different configurations of pessimistic and enthusiastic priors using a procedure that is closely related to Bayesian model averaging (BMA) (Madigan and Raftery, 1994; Draper, 1995), where instead of averaging over models for the data as is typical in BMA, one instead averages over the different prior configurations taking into account the degree to which observed data support each configuration and the configuration's prior mixture weight. The proposed approach avoids making unjustified assumptions regarding how treatment effect parameters relate across the diseases being studied. In particular, the method is developed for the situation where the endpoints for each disease are potentially different (e.g., some binary and some continuous). Information borrowing is induced through elicited mixture weights that satisfy a dependence criterion. We propose an elicitation process for the pessimistic and enthusiastic priors using the conjugate prior framework of Chen and Ibrahim (2003).

The design problem we consider in this article is closely related to basket trial design, an area for which there has been an explosion of methodological innovation in recent years. Methods include a frequentist pool/no-pool, two-stage design for binary data (Cunanan *and others*, 2017), a hierarchical modeling approach that allows for one or more baskets to be nonexchangeable with all others (Neuenschwander *and others*, 2016), designs based on multisource (e.g., pairwise) exchangeability models for normally distributed (Kaizer *and others*, 2018) and binary data (Hobbs and Landin, 2018), an approach for binary data which averages inferences from all possible models based on classifications of baskets into sets having common response probabilities (Psioda *and others*, 2021), a calibrated hierarchical modeling approach for binary data (Chu and Yuan, 2018), an approach based on commensurate predictive priors for normally distributed data (Zheng and Wason, 2020), and other approaches using the BHM (Thall *and others*, 2003), including adaptations that incorporate offsets to allow for different treatment effects across baskets (Berry *and others*, 2013). Several of the aforementioned methods use BMA (Kaizer *and others*, 2018; Hobbs and Landin, 2018; Psioda *and others*, 2021) but do so in the traditional sense. For the proposed approach, we simply exploit the BMA framework to efficiently perform inference using a complex mixture of priors.

A crucial point regarding all of the above methods is that they were developed to provide robust information borrowing when treatment effects differ across subgroups, but under the assumption that a common data model applies to all of them. Indeed, many of the above methods cannot be applied when data types are not the same across diseases (e.g., Psioda *and others*, 2021; Chu and Yuan, 2018; Cunanan *and others*, 2017). For those methods that theoretically can, there is not a strong rationale for applying them and new customizations would be needed to do so. For example, the approaches of Neuenschwander *and others* (2016) and Hobbs and Landin (2018) are designed to provide robustness

when the assumption of exchangeability of treatment effects does not hold. However, in the setting of interest, that assumption is not tenable to begin with due to having different models for the data. There are a dearth of methods that attempt to address the challenges of information borrowing in settings where data types differ across diseases studied.

The remainder of the article is organized as follows: in Section 2, we present an overview of the design problem. The clinical trial design methodology is developed in Section 3. In Section 3.1, we provide a description of the class of models for which the design framework is developed. In Section 3.2, we specify the superiority hypothesis test to be performed and describe the Bayesian testing procedure. In Sections 3.3 and 3.4, we introduce the conjugate prior framework of Chen and Ibrahim (2003) and describe how such priors can be elicited. In Sections 3.5 and 3.6, we describe the BMA inference procedure, elicitation of mixing weights, and computational methods. In Section 4, we present two sets of simulation studies that compare information borrowing designs based on the proposed approach to designs based on the Bayesian hierarchical model (BHM). We close the article with some discussion in Section 5.

2. Overview of design problem

In this section, we describe the general design problem for which the subsequent methodology is developed. We consider a scenario in which an IP is to be studied in a set of trials corresponding to a set of biologically related diseases with the goal of demonstrating superiority to control for each disease. Typical phase 2 development programs for inflammatory diseases include evaluating IPs in multiple diseases. For example, IL-17 inhibitors (e.g., Secukinumab, Brodalumab, and Ixekizumab) have been evaluated in phase 2 settings in psoriasis, psoriatic arthritis, and rheumatoid arthritis, and the first two (Secukinumab and Brodalumab) have also been evaluated in asthma. Due to the targeted mechanism of action of these treatments, it is reasonable to expect that patients with inflammatory diseases where IL-17 is implicated will benefit from treatment. Evidence of efficacy for one disease may increase confidence that there is efficacy in the others. This is the precise rationale that drives innovative basket trial design in oncology.

The goal of our design approach is to efficiently perform a set of phase II trials using a design framework that provides the following: (i) a flexible mechanism for incorporating prior information on treatment effectiveness into the analysis for each disease (henceforth, indication), (ii) a robust mechanism for borrowing information on treatment effectiveness across indications that does not assume a common treatment effect parameter or even exchangeability of treatment effect parameters, and (iii) an analysis strategy that does not require all trials to complete enrollment and outcome ascertainment prior to performing the analysis for any one indication. To satisfy condition (ii), we require a method that allows for the possibility that the primary efficacy endpoints for the indication will be different (i.e., some binary and some continuous) so that one cannot assume that treatment effect parameters are shared across indications or even that they are exchangeable. This problem is obvious when the primary endpoints are based on different data types but is just as applicable when all the primary endpoints have the same data type (i.e., binary) but correspond to different treatment effect magnitudes. To satisfy condition (iii), we require a method that does not impose the unrealistic burden that all indications must fully enroll before an analysis can be performed on any of them. A practical design would allow for the analysis of each disease indication to be performed at the time outcome ascertainment completes for that indication.

3. Methods

3.1. *Probability models for the data*

3.1.1. *Exponential family models.* We assume that the distribution for the outcome y_{ij} for subject i from indication j comes from the exponential family of distributions but that the particular member of the family can differ across indications. The general probability distribution for y_{ij} is then given by

$p\left(y_{ij}|\theta_{ij},\tau_{ij}\right)=\exp\left\{a_{ij}\left(\tau_{ij}\right)\left(y_{ij}\theta_{ij}-b(\theta_{ij})\right)+c(y_{ij},\tau_{ij})\right\}$, indexed by the natural parameter θ_{ij} and dispersion parameter τ_{ij}. Typically $a_{ij}\left(\tau_{ij}\right)=\tau_{ij}\omega_{ij}$, where ω_{ij} is a known weight. In the remainder, we take $\omega_{ij}=1$. In most applications, τ_{ij} is taken to be equal for all subjects or equal within treatment groups. The functions $b(\cdot)$ and $c(\cdot)$ determine a particular member of the class of distributions (e.g., normal, Bernoulli, or Poisson).

3.1.2. *Generalized linear models.* Often θ_{ij} is modeled as a function of covariates. Suppose θ_{ij} satisfies $\theta_{ij}=\theta_j(\eta_{ij})$ with the linear predictor $\eta_{ij}=\alpha_{0j}+\alpha_{1j}z_{ij}+\mathbf{x}_{ij}^T\boldsymbol{\beta}_j$, where α_{0j} is an intercept parameter for indication j, α_{1j} is the treatment effect for indication j, z_{ij} is the corresponding indicator of treatment for subject i in indication j, $\mathbf{x}_{ij}$ is a $p_j\times 1$ vector of baseline covariates for subject i in indication j, $\boldsymbol{\beta}_j=\left(\beta_{j,1},...,\beta_{j,p_j}\right)$ is the corresponding $p_j\times 1$ vector of parameters, and $\theta_j(\cdot)$ is a monotone differentiable function. Models of the form described above are known as generalized linear models (GLMs). In the following, we focus on a development without covariates (i.e., $\eta_{ij}=\alpha_{0j}+\alpha_{1j}z_{ij}$) but discuss applications with covariates in depth in Appendix A of the Supplementary material available at *Biostatistics* online.

3.1.3. *Mean parameterization for the no covariate case.* In the no covariate case, it is convenient to parametrize the model for each indication by specifying a separate exponential family model for each treatment group. For control subjects (i.e., $z_{ij}=0$), $\eta_{ij}=\alpha_{0j}$ and $\tau_{ij}=\tau_{0j}$, and for treated subjects, $\eta_{ij}=\alpha_{0j}+\alpha_{1j}$ and $\tau_{ij}=\tau_{1j}$. Rather than work with the natural parametrization in this setting, it is often preferred to use the mean parametrization given by $\mu_{ij}=\partial b\left(\theta_{ij}\right)/\partial\theta_{ij}$. For example, in the case of a logistic model, $b\left(\theta_{ij}\right)=\log\left(1+e^{\theta_{ij}}\right)$ so that $\mu_{ij}=\text{expit}\left(\theta_{ij}\right)$, or equivalently, $\theta_{ij}=\text{logit}\left(\mu_{ij}\right)$, where $\text{expit}(x)=e^x/(1+e^x)$ and $\text{logit}(x)=\log\left(x/[1-x]\right)$. After some elementary algebra and noting that $\tau_{ij}=1$ for the logistic model, this leads to the model reformulation given as $p\left(y_{ij}|\mu_{ij}\right)=\exp\left\{y_{ij}\log\left(\mu_{ij}/1-\mu_{ij}\right)-\log\left(1/1-\mu_{ij}\right)\right\}$.

To emphasize group-level means, we let $\mu_{ij}=\mu_{0j}$ when $z_{ij}=0$ and $\mu_{ij}=\mu_{1j}$ when $z_{ij}=1$. Let $\mathbf{y}_{hj}$ represent the vector of outcomes for treatment group h for indication j and Γ_{hj} be the collection of n_{hj} indices for which $z_{ij}=h$. Following through with the logistic example, the likelihood for treatment group h can then be written as

$$\mathcal{L}\left(\mu_{hj}|\mathbf{y}_{hj}\right)=\exp\left\{\left(\sum_{i\in\Gamma_h}y_{ij}\right)\log\left(\frac{\mu_{hj}}{1-\mu_{hj}}\right)-n_{hj}\log\left(\frac{1}{1-\mu_{hj}}\right)\right\} \tag{3.1}$$

$$=\mu_{hj}^{\sum_{i\in\Gamma_{hj}}y_{ij}}\left(1-\mu_{hj}\right)^{n_{hj}-\sum_{i\in\Gamma_{hj}}y_{ij}}, \tag{3.2}$$

where the form of (3.1) is analogous to that given above for an individual subject, and the more common representation in (3.2) is easily seen to be a kernel of a beta density in μ_{hj} and is therefore conjugate with the family of beta densities. We will exploit this feature to simplify posterior inference for design simulations.

3.2. *Hypothesis testing and decision criteria*

The primary inferential goal for each indication is to prove superiority of the IP to the indication-specific control. In the case of an arbitrary GLM, hypotheses can be formulated as follows: $H_0:\alpha_{1j}\leq\delta_{0j}$ versus $H_1:\alpha_{1j}>\delta_{0j}$, where $\delta_{0j}\geq 0$ and larger values indicate better outcomes. When there are no covariates, one can alternatively formulate the hypotheses in terms of the difference in means or response

probabilities given as $H_0 : \mu_{1j} - \mu_{0j} \leq \delta_{0j}$ versus $H_1 : \mu_{1j} - \mu_{0j} > \delta_{0j}$. This latter formulation will be the primary focus in our subsequent discussion in the article but we discuss applications with covariates in depth in Appendix A of the Supplementary material available at *Biostatistics* online.

The proposed design incorporates a single efficacy analysis for each indication (to occur at the point in time when the indication completes enrollment and outcome ascertainment) and multiple futility analyses. An extension of the design to include multiple efficacy analyses for each indication would be straightforward but may not be advantageous. This is because borrowing information across indications will generally result in a smaller overall sample size per indication compared to an approach where separate trials are conducted in each indication. It may be undesirable to further reduce the sample size in an indication when the treatment is superior to the control as doing so may provide safety information that is too limited.

Let $A_j = 1\left[\mu_{1j} - \mu_{0j} > \delta_{0j}\right]$ (where $1\,[\cdot]$ denotes the indicator function) represent the event defining the alternative hypothesis for indication j and $P\left(A_j \middle| \mathbf{D}\right)$ represent its posterior probability given the observed data $\mathbf{D}$. Note that $P\left(A_j \middle| \mathbf{D}\right)$ will depend not only on $\mathbf{D}$ but also on the priors used for analysis. This notation will be made more explicit subsequently. One may claim superiority of the IP to the control in indication j when $P\left(A_j \middle| \mathbf{D}\right) \geq \phi_{1j}$, where ϕ_{1j} is some pre-specified evidence threshold. Since the data for all indications are needed to compute $P\left(A_j \middle| \mathbf{D}\right)$, it is natural to also evaluate whether stopping indication $j' \neq j$ for futility is warranted for any indication j' where data collection has not already been completed. At such times, one can terminate enrollment when $P\left(A_{j'} \middle| \mathbf{D}\right) \leq \phi_{0j'}$.

3.3. *Conjugate pessimistic and enthusiastic priors*

The design approach we propose in this article is based on elicited *pessimistic* and *enthusiastic* conjugate priors (Chen and Ibrahim, 2003) for each of the J indications to be concurrently studied. Here, our idea of pessimism and enthusiasm follows closely with that described by Spiegelhalter *and others* (1994). Pessimistic priors are centered on a null value of the treatment effect and suggest the hypothesized alternative is unlikely. Similarly, enthusiastic priors are centered on the hypothesized alternative treatment effect and suggest the null is unlikely. Priors are elicited by *predicting* the response vector (or sufficient statistics) for the future trial using the likelihood-based conjugate prior framework of Chen and Ibrahim (2003) but could also be informed by historical data, if such data were available. We will use the term *predicted data* instead of historical data in subsequent discussion to highlight that priors are elicited through making predictions about future data to be observed. For each indication, we elicit both a pessimistic and enthusiastic prior based on predicting future data. This results in 2^J possible configurations for the pessimistic and enthusiastic priors across the J indications. Analyses are performed by combining inferences from all 2^J prior configurations taking into account the degree to which the observed data support each configuration and the configuration's prior mixing weight. If we define a *model* as a set of sampling distributions for the data from the J indications, indexed by their respective parameters and coupled with a particular conjugate prior configuration for those parameters, then we may view this as a setting with 2^J models which vary only in their priors. The proposed analytic approach is equivalent to performing BMA over those 2^J models.

For now, we omit the index for indication and develop the conjugate prior for the no covariate case with mean parameterization as previously described. A development with covariates is given in Appendix A of the Supplementary material available at *Biostatistics* online. The conjugate prior (conditional on $\boldsymbol{\tau}$) is given by $\pi_0\left(\boldsymbol{\mu} \middle| \mathbf{D}^{(p)}, a_0, \boldsymbol{\tau}\right) = c\left(\mathbf{D}^{(p)}, a_0, \boldsymbol{\tau}\right)^{-1} \exp\left\{a_0 \sum_{i=1}^{n} \tau_{z_i}\left[y_{0i}\theta(\mu_{z_i}) - b(\theta(\mu_{z_i}))\right]\right\} \pi_0\left(\boldsymbol{\mu}\right)$, where $\boldsymbol{\mu} = (\mu_0, \mu_1)$, the natural parameter $\theta(\mu_{z_i})$ is represented as a function of the mean for patient i, and $\mathbf{D}^{(p)} = \{\mathbf{y}_0, \mathbf{z}, \}$. Here, $\mathbf{z} = \{z_1, ..., z_n\}$ is a vector of treatment indicators and $\mathbf{y}_0 = \{y_{01}, ..., y_{0n}\}$ is a *predicted* response vector based on $\mathbf{z}$. We refer to $\pi_0\left(\boldsymbol{\mu}\right)$ as the initial prior and obtain the conjugate prior framework of Chen and Ibrahim (2003) upon taking $\pi_0\left(\boldsymbol{\mu}\right) \propto 1$. We incorporate an initial prior in this

development so that the conjugate priors have the familiar appearance of a power prior (Ibrahim and Chen, 2000). The scalar $a_0 \geq 0$ controls the informativeness of the prior with larger values resulting in a more informative prior. In this setting, we need not require $a_0 \leq 1$ as is commonly done with actual historical data. In what follows, we are careful to use the superscript (p) to indicate quantities based on *predicted* data with quantities based on observed data (i.e., actual data) being represented similarly but without a superscript.

The normalizing constant $c(\mathbf{D}^{(p)}, a_0, \boldsymbol{\tau})$ must be computed in our setting. Having $c(\mathbf{D}^{(p)}, a_0, \boldsymbol{\tau})$ is critical for two reasons. First, our approach is to use the prior factorization $\pi_0\left(\boldsymbol{\mu}\middle|\mathbf{D}^{(p)}, a_0, \boldsymbol{\tau}\right) \times \pi_0(\boldsymbol{\tau})$ which necessitates having the normalizing constant so that $\pi_0(\boldsymbol{\tau})$ can be interpreted as a marginal distribution for $\boldsymbol{\tau}$. Second, it is critical that each prior configuration be properly normalized as these constants enter directly into posterior calculations. In regression settings outside the linear model, $c\left(\mathbf{D}^{(p)}, a_0, \boldsymbol{\tau}\right)$ does not exist in closed form. Calculation of $c\left(\mathbf{D}^{(p)}, a_0, \boldsymbol{\tau}\right)$ for regression models is described in Appendix A of the Supplementary material available at *Biostatistics* online.

3.4. *Eliciting pessimistic and enthusiastic priors*

In this section, we discuss elicitation of priors using the mean parameterization for the no covariate case, but the case with covariates is described in Appendix A of the Supplementary material available at *Biostatistics* online. Under the mean parameterization and for $\pi_0(\boldsymbol{\mu}) \propto 1$ (or when $\pi_0(\boldsymbol{\mu})$ is taken to be a member of the appropriate conjugate family of priors), the resulting distribution $\pi_0\left(\boldsymbol{\mu}\middle|\mathbf{D}^{(p)}, a_0, \boldsymbol{\tau}\right)$ is easily recognized to be a product of normal, gamma, and beta distributions for the normal, Poisson, and Bernoulli models, respectively (the latter two not depending on $\boldsymbol{\tau}$). Using the mean parametrization affords one the ability to derive closed-forms for the conjugate prior and leads to posterior computations that do not require MCMC as will be described subsequently.

Our approach to elicitation of pessimistic and enthusiastic priors is designed to be semi-automatic given the standard set of inputs required for sample size determination. For indication j, three quantities are required for specification of the priors: (i) an estimate of the control group mean μ_{0j}, (ii) the hypothesized mean difference, denoted by δ_{1j}, and (iii) a minimum clinically significant mean difference δ_{0j}. The choice for (i) guides elicitation of a conjugate prior for the control group. The choice for (i) and (iii) combine to guide elicitation of the pessimistic conjugate prior for the treated group. Lastly, the choice for (i) and (ii) combine to guide elicitation of the enthusiastic prior for the treated group. Let $\gamma_j = \mu_{1j} - \mu_{0j}$. A pessimistic prior for the treatment group in indication j is defined as one where the prior mean difference is equal to δ_{0j} and $P\left(\gamma_j > \delta_{1j}\right)$ is small whereas an enthusiastic prior is defined as one where the prior mean difference is equal to δ_{1j} and $P\left(\gamma_j < \delta_{0j}\right)$ is small.

To fix ideas, we consider an example for a binary outcome. Assume that available data suggest a control group probability of response $\mu_{0j} = 0.23$, that any degree of efficacy would be viewed as clinically meaningful (i.e., $\delta_{0j} = 0$), and that investigators hypothesize the treatment is likely to improve the probability of response by $\delta_{1j} = 0.27$ so that the treatment group has response probability $\mu_{1j} = 0.50$. To achieve 80% power and a 2.5% nominal one-sided type I error rate, a sample size of $n = 58$ subjects per group would be required based on a standard power calculation using the likelihood ratio test. A pessimistic prediction for the sample proportions to be observed in the study would be $\bar{y}_{00} = \bar{y}_{01} = 0.23$ where $\bar{y}_{0h} = 1/n_h \sum_{i=1}^{n_h} y_{hi}$. An enthusiastic prediction would be $\bar{y}_{00} = 0.23$ and $\bar{y}_{01} = 0.5$. Figure S1 of the Supplementary material available at *Biostatistics* online shows the pessimistic and enthusiastic conjugate power priors obtained by taking $a_0 = 0.5$ and $n = 58$ (top row). The *induced* priors for the treatment effects (i.e., response probability differences) are also provided (bottom row). One can see that the induced pessimistic and enthusiastic priors are both tightly massed about δ_{0j} and δ_{1j}, respectively, and that they place little mass near δ_{1j} and δ_{0j}, respectively.

3.5. *Inference averaged over prior configurations*

In this section, we describe the inference procedure using general representations for the sampling model for each indication and corresponding prior. We will denote a pessimistic prior of the type described in Section 3.3 as $\pi_0\left(\boldsymbol{\psi}_j\right)$ and an enthusiastic prior as $\pi_1\left(\boldsymbol{\psi}_j\right)$, where $\boldsymbol{\psi}_j$ is a generic label for the parameters for indication j. A natural *fully enthusiastic* prior for all indications can be constructed as $\pi_1\left(\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_J\right) = \prod_{j=1}^{J} \pi_1\left(\boldsymbol{\psi}_j\right)$. The term *fully enthusiastic* is intended to imply broad enthusiasm across all indications. Alternatively, a *fully pessimistic* prior could be defined as $\pi_0\left(\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_J\right) = \prod_{j=1}^{J} \pi_0\left(\boldsymbol{\psi}_j\right)$. These two priors represent the extremes of belief regarding the efficacy of the IP across the indications. Rather than simply choosing one of the two extremes above or another conjugate prior based other predictions, we perform inference by averaging analysis results from *all* possible prior configurations for the J indications constructed using the set of enthusiastic priors $\left\{\pi_1\left(\boldsymbol{\psi}_j\right) : j = 1, ..., J\right\}$ and pessimistic priors $\left\{\pi_0\left(\boldsymbol{\psi}_j\right) : j = 1, ..., J\right\}$.

For illustration, consider the case where $J = 3$ indications are to be studied. Table S1 of the Supplementary material available at *Biostatistics* online enumerates all $2^J = 8$ possible prior configurations. To highlight the connection to BMA, we let prior configuration k be denoted by M_k. Our approach assigns prior configuration M_k a prior weight, denoted by $p(M_k)$, that satisfies $p(M_k) > 0$ and $\sum_{k=1}^{2^J} p(M_k) = 1$. In the parlance of BMA, we would refer to $p(M_k)$ as the *prior model probability* for model M_k. Though the model fitting strategy described below has the familiar look of BMA and offers a number of computational advantages, it is completely equivalent to analysis using a 2^J component mixture prior. In particular, if we let $\pi_{0k}\left(\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_J \middle| M_k\right)$ represent the joint prior based on conjugate prior configuration M_k, then the procedure below is equivalent to performing inference using the mixture prior $\sum_{k=1}^{2^J} p(M_k)\pi_{0k}\left(\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_J \middle| M_k\right)$.

Let A_j represent the event that the alternative hypothesis is true for indication j (e.g., $A_j = 1\left[\gamma_j > \delta_{0j}\right]$). Inference for indication j is based on $P\left(A_j \middle| \mathbf{D}, \mathbf{D}^{(p)}, a_0\right)$ where $\mathbf{D}$ is the observed data and $\mathbf{D}^{(p)}$ is *all* predicted data (i.e., both enthusiastic and pessimistic predictions). Note that $P\left(A_j \middle| \mathbf{D}, \mathbf{D}^{(p)}, a_0\right)$ can be represented as

$$P\left(A_j \middle| \mathbf{D}, \mathbf{D}^{(p)}, a_0\right) = \sum_{k=1}^{2^J} P\left(A_j \middle| \mathbf{D}_j, \mathbf{D}_{jk}^{(p)}, a_0\right) P\left(M_k \middle| \mathbf{D}, \mathbf{D}^{(p)}, a_0\right), \tag{3.3}$$

where $\mathbf{D}_j$ is the observed data for indication j and $\mathbf{D}_{jk}^{(p)}$ is the predicted data for indication j under prior configuration M_k. The posterior weight given to the analysis based on prior configuration M_k is denoted by $p\left(M_k \middle| \mathbf{D}, \mathbf{D}^{(p)}, a_0\right)$ and is calculated as

$$p\left(M_k \middle| \mathbf{D}, \mathbf{D}^{(p)}, a_0\right) = \frac{\left[\prod_{j=1}^{J} P\left(\mathbf{D}_j \middle| \mathbf{D}_{jk}^{(p)}, a_0\right)\right] p(M_k)}{\sum_{k'}\left[\prod_{j=1}^{J} P\left(\mathbf{D}_j \middle| \mathbf{D}_{jk'}^{(p)}, a_0\right)\right] p(M_{k'})}, \tag{3.4}$$

where $P\left(\mathbf{D}_j \middle| \mathbf{D}_{jk}^{(p)}, a_0\right)$ is the marginal likelihood for indication j under prior configuration M_k.

Note that in the no covariate case for the normal, Bernoulli, and Poisson models, both (3.3) and (3.4) can be computed without the use of MCMC using standard univariate numerical integration routines. For example, $P\left(A_j \middle| \mathbf{D}_j, \mathbf{D}_{jk}^{(p)}, a_0\right)$ in (3.3) is equal to

$$\int \pi\left(\mu_{0j} \middle| \mathbf{D}_{0j}, \mathbf{D}_{0jk}^{(p)}, a_0\right)\left[1 - F\left(\mu_{0j} + \delta_{0j} \middle| \mathbf{D}_{1j}, \mathbf{D}_{1jk}^{(p)}, a_0\right)\right] \mathrm{d}\mu_{0j}, \tag{3.5}$$

where $\pi(\mu_{0j}|\mathbf{D}_{0j}, \mathbf{D}_{0jk}^{(p)}, a_0)$ is the posterior density for μ_{0j} and $F\left(\cdot|\mathbf{D}_{1j}, \mathbf{D}_{1jk}^{(p)}, a_0\right)$ is the posterior cumulative distribution function (CDF) for μ_{1j}. The integral in (3.5) cannot generally be computed in closed-form. However, standard software such as R (R Core Team, 2016) has built-in functions for densities and CDFs for members of the exponential family (e.g., `pbeta` and `dbeta` functions) and standard numerical integration routines (e.g., `integrate` function) which can perform such computations with great precision and speed. Explicit forms for the posterior distributions and marginal likelihoods in the no covariate case are given in Appendix B of the Supplementary material available at *Biostatistics* online for the normal, Bernoulli, and Poisson models. Model fitting for the general regression setting is discussed in Appendix A of the Supplementary material available at *Biostatistics* online.

3.6. *Eliciting weights for prior configurations*

When $p(M_k) \propto 1$ for each k (i.e., uniform mixing weights), the data for indication j are used to choose between $\pi_1\left(\boldsymbol{\psi}_j\right)$ and $\pi_0\left(\boldsymbol{\psi}_j\right)$ without influence by the data from the other indications. In this case, analysis for indication j is effectively performed using the two-part mixture prior $\frac{1}{2}\pi_1\left(\boldsymbol{\psi}_j\right)+\frac{1}{2}\pi_0\left(\boldsymbol{\psi}_j\right)$. This extends more generally to any indication for which the weights assigned to the prior configurations satisfy an independence criterion. A formal statement of this result is given in Appendix C of the Supplementary material available at *Biostatistics* online, where a proof is also provided. Here, we give a brief summary to convey the general concept.

Without loss of generality, define M_{0d} and M_{1d} to be the dth pair of prior configurations that only differ with respect to the conjugate prior for indication j. Note there are 2^{J-1} such pairs. Here, we assume $\pi_0\left(\boldsymbol{\psi}_j\right)$ is a component of M_{0d} and $\pi_1\left(\boldsymbol{\psi}_j\right)$ is a component of M_{1d}. Analysis for indication j will be independent (i.e., uninfluenced) by data from other indications if and only if $p(M_{0d}) = \pi_{0j}m_d$ and $p(M_{1d}) = (1-\pi_{0j})m_d$ for every d. Here, m_d is the prior weight for the dth prior configuration of the conjugate priors for the *other* indications. When prior model probabilities deviate from the independence assumption, information is borrowed to some degree across the indications through the inference process. In this case, the resulting mixture prior for the J indications cannot be decomposed into a product independent two-part mixture priors for each indication as described above.

One of the benefits of the proposed approach over an approach like borrowing information using the BHM is in the flexibility of the borrowing mechanism. To illustrate this point, Figure 1 presents three

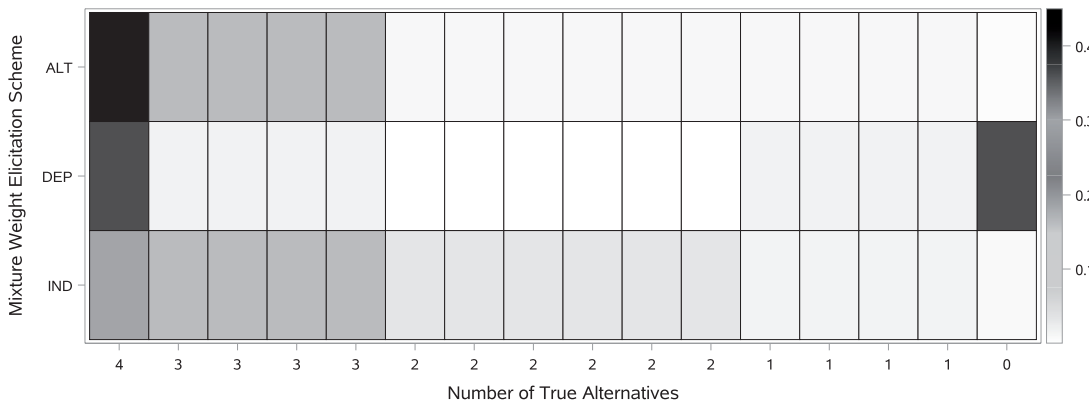

Fig. 1. Example mixture weight elicitation schemes for $J = 4$ indications corresponding to $2^J = 16$ models. Darker shading indicates larger weights.

different weight elicitation schemes: independence (IND), positive dependence (DEP), and a scheme consistent with a relatively strong belief that at least three alternatives are true (ALT). For the IND scheme, we took $\pi_{0j} = 0.675$ for each $j = 1, ..., 4$. Thus, the weight given to prior configurations with k optimistic components is simply $0.675^k (1 - 0.675)^{4-k}$ for $k = 0, ..., 4$. In this case, the fully optimistic prior configuration is given weight 0.208 and the fully pessimistic prior configuration is given weight 0.011 and inference in each indication is based on an independent two-part mixture prior. In contrast, the DEP scheme gives the most weight (and equal weight) to the fully optimistic and fully pessimistic prior configurations. The weight given to each is approximately 0.4. Such a weighting scheme asserts that the effectiveness of the IP is likely to be consistent for all indications. That is to say, the DEP scheme suggests it is most likely that the IP is effective for all indications or ineffective for all indications but does not reflect a preference between those two extremes. This scheme encourages all indications to have analyses that reach the same conclusion (i.e., all null) and induces the type of borrowing that the BHM provides. For the ALT scheme, very high weight is given to the fully optimistic prior configuration (0.45) and prior configurations with only one pessimistic component (0.10 weight to each of those four configurations). All other prior configurations receive minimal weight (i.e., <0.02). Thus, this scheme is consistent with the a relatively strong belief that the investigational product will be effective for most indications studied, a scenario that has occurred in a number of inflammatory disease treatment development programs.

Lastly, we note that the choice of mixture weights drives information borrowing for the proposed method through inducing a correlated joint prior distribution on the J treatment effects. In fact, one can obtain identical *marginal* treatment effect priors using the uniform weighting scheme or the DEP weighting scheme. Moreover, these marginal priors will be reasonably flat over the interval $\left[\delta_{0j}, \delta_{1j}\right]$. However, under the DEP weighting scheme, treatment effects will be correlated and this will result in information borrowing across indications. This prior correlation is illustrated in Figure S2 of the Supplementary material available at *Biostatistics* online for a simple case with $J = 2$ indications having pessimistic and enthusiastic priors matching those described in Section 3.4.

4. Simulation studies

In this section, we perform two sets of simulation studies, each based on designs with $J = 4$ disease indications to be investigated concurrently. We consider a scenario where an IP is to be studied in a collection of inflammatory diseases: psoriasis, psoriatic arthritis, rheumatoid arthritis, and asthma. As described in Section 2, these indications were evaluated in the phase II development programs for both Secukinumab and Brodalumab and our choice of endpoints in each indication align with primary and/or secondary endpoints used in actual phase II trials. In particular, 75% or greater reduction from baseline in psoriasis area-and-severity index score (PASI 75), American College of Rheumatology (ACR) response criteria for 20% improvement (ACR20), ACR response criteria for 50% improvement (ACR50), and Asthma Control Questionnaire (ACQ) total score change from baseline were chosen as the endpoints of interest for psoriasis, psoriatic arthritis, rheumatoid arthritis, and asthma, respectively.

In each simulation study, we compare the proposed-based design to a design that uses the BHM for information borrowing. For simulations presented in Section 4.5, we assume all four indications are replicates of the psoriatic arthritis indication (identical endpoint scenario). Such a case is favorable to the BHM and offers a fair comparison of the proposed approach to an approach using the BHM in situations where the BHM is appropriate. For simulations in Section 4.6, we explore the performance of the two approaches when the indications have distinctly different endpoints (different endpoint scenario). This second scenario is of direct interest to the authors as BHM-based borrowing is not ideal here due to the untenable assumption that trials (i.e., treatment effects) are exchangeable.

Table 1. *Design inputs for inflammatory disease simulations*

Disease indication	Endpoint	Data type	μ_{0j}	μ_{1j}	δ_{1j}	σ_j	N	N_R	% Red.
Psoriasis	$PASI_{75}$	Binary	0.10	0.47	0.37	–	48	24	50.0
Psoriatic arthritis	ACR_{20}	Binary	0.23	0.50	0.27	–	96	58	39.6
Rheumatoid arthritis	ACR_{50}	Binary	0.10	0.40	0.30	–	64	34	46.9
Asthma	ACQ	Continuous	0.00	0.50	0.50	1.00	128	80	37.5

4.1. *Simulation setup details*

Details regarding parameter assumptions and the standard sample size, denoted by N, required to achieve 80% power with a nominal one-sided type I error rate of 2.5% are given in Table 1. The associated power analyses assume a single analysis per indication, 1:1 randomization in each indication, no use of prior information, and no information borrowing. The sample sizes are based on a likelihood ratio test (LRT) for binary endpoints and a t-test for the continuous endpoint.

In both the identical and different endpoint scenarios, we assume enrollment in each indication follows an independent Poisson process with rate parameter equal to 2 so that approximately 2 subjects enroll per month. We further assume that outcome ascertainment takes approximately 4 months for each indication. Therefore, in the results presented in Section 4.5, all indications complete outcome ascertainment at approximately the same time resulting in final analyses for each indication that incorporate near full data for the other indications being concurrently investigated. However, in Section 4.6, due to the varied sample size required for each indication, the assumptions on accrual and outcome ascertainment implies that indications will complete outcome ascertainment at different times, resulting in potential for meaningful early stoppage in some indications (e.g., the asthma and psoriatic arthritis indications).

4.2. *Operating characteristics evaluated*

To compared the designs, we computed several operating characteristics, including (i) the bias of the posterior mean defined as $\mathrm{E}_h\left[\mathrm{E}\left[\gamma_j \middle| \mathbf{D}_\mathrm{F}, \mathbf{D}^{(p)}, a_0\right] - \delta^*_{hj}\right]$, where $\mathbf{D}_\mathrm{F}$ is the final observed data for all indications and δ^*_{hj} is the true value of the estimand, (ii) the null hypothesis rejection rate $\mathrm{E}_h\left[1\left\{P\left(A_j \middle| \mathbf{D}_{\mathrm{F},j}, \mathbf{D}^{(p)}, a_0\right) \geq \phi_{1j}\right\} \times \prod_{j'} 1\left\{P\left(A_j \middle| \mathbf{D}_{\mathrm{F},j'}, \mathbf{D}^{(p)}, a_0\right) \geq \phi_{0j}\right\}\right]$, where $\mathbf{D}_{\mathrm{F},j}$ is the data at the point in time where indication j completes outcome ascertainment and $\mathbf{D}_{\mathrm{F},j'}$ is the data at the point in time where indication j' completes outcome ascertainment for all indications that complete earlier, and (iii) the expected sample size. In the above notation, $\mathrm{E}_h[\cdot]$ is an expectation taken with respect to the *prior predictive distribution* for the data to be observed which is defined according to a null ($h = 0$) or alternative ($h = 1$) sampling prior for the parameters using the approach proposed by Psioda and Ibrahim (2019). For simplicity, we only consider point-mass sampling priors in this article. In the case of point-mass sampling priors, Bayesian versions of power and the type I error rate, whose definitions are based on user-specified sampling prior distributions for the parameters (which need not be degenerate), closely align with the corresponding frequentist constructs. All operating characteristics were estimated using $\geq$ 100 000 simulation studies. For efficacy evaluations, we used posterior probability threshold of $\phi_{1j} = 0.975$ and for futility evaluations, we used $\phi_{0j} = 0.60$ for each j. Futility evaluations for an indication were performed at the time of another indication's efficacy analysis if 50% of the planned outcomes were ascertained for the indication with ongoing data collection.

Table 2. *Design operating characteristics based on correct predictions*

Stat.	# H_1 True	H_1 BHM	H_1 IND	H_1 DEP	H_1 ALT	H_1 REF	H_0 BHM	H_0 IND	H_0 DEP	H_0 ALT	H_0 REF
Pr(H_1)	0						0.016	0.025	0.004	0.028	0.101
	1	0.374	0.798	0.492	0.809	0.801	0.055	0.025	0.011	0.031	0.101
	2	0.626	0.797	0.611	0.817	0.800	0.122	0.025	0.054	0.047	0.101
	3	0.782	0.798	0.842	0.862	0.801	0.231	0.026	0.086	0.045	0.101
	4	0.898	0.797	0.903	0.860	0.801					
Bias	0						−0.002	0.032	0.002	0.034	−0.003
	1	−0.101	−0.012	−0.072	−0.012	−0.002	0.031	0.033	0.011	0.036	−0.003
	2	−0.059	−0.012	−0.048	−0.011	−0.002	0.058	0.033	0.051	0.051	−0.003
	3	−0.031	−0.012	−0.010	−0.006	−0.002	0.091	0.033	0.074	0.049	−0.003
	4	−0.001	−0.012	−0.003	−0.007	−0.002					

Stat.: statistic; P(H_1): probability of concluding H_1 true; BHM: Bayesian hierarchical model; IND: independence; DEP: positive dependence; ALT: belief in $\geq$ 3 true alternatives; REF: reference prior analysis using posterior probability critical value 0.91.

4.3. *Specification of the designs evaluated*

For the proposed designs, we implemented conjugate prior distributions where the treatment group means matched those in Table 1 for the enthusiastic priors and where the means were both equal to the value μ_{0j} for the pessimistic prior (i.e., $\delta_{0j} = 0$). Beta initial priors with shape parameters equal to 0.1 were used for the treatment group means for the binary endpoint indications. For the continuous endpoint indication, a uniform improper prior was used for the mean parameters and a weakly informative gamma prior with mean equal to 1.0 was used for τ. We compared the IND, DEP, and ALT weight elicitation schemes as described in Section 3.6. The BHM prior is specified in detail in Appendix D of the Supplementary material available at *Biostatistics* online.

4.4. *Reference design based on noninformative priors*

To aid in understanding the efficiency gains afforded by information borrowing, for results presented in Section 4.5, we also evaluated a design based on using a non-informative prior for each indication. The priors were taken to equal the initial priors described for the proposed method in Section 4.3. The design, denoted as REF, uses a decreased threshold for the posterior probability critical value ($\phi_{1j} = 0.91$) so that the design achieves 80% power given the reduced sample sizes used. This is done so that one can see the degree of type of type I error rate inflation relative to the nominal 2.5% rate that would be required to maintain 80% power *without* information borrowing. The extent to which information borrowing designs can achieve $\geq$80% power under plausible alternatives while maintaining adequate type I error control under plausible null scenarios and acceptable bias for the posterior mean point estimator provides insight into the extent to which the information borrowing designs truly provide increased efficiency.

4.5. *Simulation study—identical endpoints*

For the simulations presented in this section, each indication was taken to be a replicate of the psoriatic arthritis indication described in Table 1. Table 2 presents estimates of the power, type I error rate, and bias of the posterior mean under the assumption that one of the two elicited predictions for each indication is correct. Here, the term *correct predictions* implies that when the null or alternative hypothesis is true,

Table 3. *Design operating characteristics based on partially correct predictions*

				– Hypothesized –				– Low –				– None –			
Stat.	H	L	N	BHM	IND	DEP	ALT	BHM	IND	DEP	ALT	BHM	IND	DEP	ALT
Pr(H_1)	1	1	2	0.502	0.798	0.542	0.815	0.270	0.316	0.174	0.343	0.091	0.026	0.028	0.041
	1	2	1	0.626	0.798	0.643	0.835	0.369	0.316	0.291	0.391	0.149	0.026	0.051	0.044
	2	1	1	0.723	0.798	0.744	0.847	0.437	0.315	0.413	0.417	0.187	0.026	0.073	0.046
Bias	1	1	2	−0.083	−0.012	−0.061	−0.011	−0.023	0.014	−0.031	0.018	0.048	0.033	0.029	0.045
	1	2	1	−0.070	−0.012	−0.042	−0.009	−0.006	0.014	0.000	0.027	0.067	0.033	0.049	0.048
	2	1	1	−0.048	−0.012	−0.024	−0.008	0.011	0.014	0.026	0.031	0.079	0.033	0.066	0.050

Note: Low effects are equal to half their hypothesized value.
Stat.: statistic; P(H_1): probability of concluding H_1 true; H: hypothesized effect; L: low effect; N: no effect; BHM: Bayesian hierarchical model; IND: independence; DEP: positive dependence; ALT: belief in ≥ 3 true alternatives.

the treatment effect is equal to δ_{0j} or δ_{1j}, respectively. Such a case is favorable to the proposed approach. In this case, the IND design provides approximately 80% power and a type I error rate of approximately 2.5%. This is no coincidence as the sample sizes used for the design simulations were identified so that the IND design had this property, thus providing a reference for comparison to other designs.

The BHM and DEP designs are most powerful when all four indications are true alternatives. In this case, each provide approximately 90% power. This point is important as it illustrates why one may wish to borrow information given that the IND design does not do so and still affords sample size reduction. Under the IND design, the chance of producing substantial evidence of efficacy in all four indications is approximately 41% (100×0.8^4) when all four have the hypothesized level of efficacy. The chance is approximately 66% for the BHM and DEP designs. Thus, in this scenario and for these designs, there is a 25% greater chance of the most favorable outcome for the development program. Note that the chance is also approximately 15% greater for the ALT design compared to the IND design.

There is of course a tradeoff. For the BHM, DEP, and ALT designs, power for true alternative indications decreases as their number decreases. Correspondingly, type I error rates for true null indications increase. Comparatively, the DEP design outperforms the BHM design with regards to power, type I error control, and bias. Compared to both the DEP and BHM designs, the ALT design provides much more stable performance with power never dropping below 80% and a type I error rate never reaching 5%. In situations where sponsors are relatively confident that a treatment will be efficacious in most indications (e.g., all or all but one), the ALT and DEP designs provide worthwhile efficiency gains over the IND design in terms of power, and the tradeoff in terms of type I error inflation (relevant if ≥ 1 indications are truly null) may be desirable.

Table 3 presents estimates of the power, type I error rate, and bias of the posterior mean under the assumption that the actual treatment effect is heterogeneous across the indications. For ease of exposition, the REF design is not included in the table. In the presence of effect heterogeneity, the power for the BHM design is lower than for the IND, DEP, and ALT designs for indications having the hypothesized level of efficacy, and its type I error control is notably worse. For indications with lower than expected efficacy, the relative power comparisons for the designs are mixed. Of note, the ALT design arguably provides the best performance over the range of scenarios. In particular, the power for the low efficacy indications is nearly as high or higher for the ALT design compared to the BHM, IND, and DEP designs across all scenarios, and the type I error rate consistently stays beneath 5%.

Tables 2 and 3 present several of a multitude of scenarios that should be investigated when evaluating the performance of a complex innovative design that borrows information in some way. More results and

Table 4. *Power and type I error rates based on correct predictions*

#	– Indication # 1 –				– Indication # 2 –				– Indication # 3 –				– Indication # 4 –			
H_1	BHM	IND	DEP	ALT	BHM	IND	DEP	ALT	BHM	IND	DEP	ALT	BHM	IND	DEP	ALT
0	0.004	0.029	0.006	0.034	0.010	0.025	0.005	0.028	0.004	0.028	0.005	0.034	0.020	0.024	0.004	0.024
1	0.027	0.029	0.015	0.039	0.057	0.025	0.011	0.031	0.032	0.028	0.013	0.037	0.530	0.793	0.447	0.802
1	0.019	0.029	0.022	0.042	0.025	0.026	0.012	0.031	0.266	0.805	0.548	0.819	0.037	0.024	0.010	0.028
1	0.016	0.028	0.018	0.041	0.339	0.797	0.496	0.808	0.024	0.028	0.015	0.038	0.039	0.023	0.010	0.027
1	0.257	0.815	0.585	0.831	0.027	0.026	0.012	0.031	0.021	0.027	0.018	0.039	0.035	0.024	0.010	0.027
2	0.067	0.029	0.069	0.057	0.084	0.025	0.055	0.046	0.534	0.807	0.656	0.827	0.620	0.794	0.580	0.810
2	0.083	0.029	0.066	0.056	0.641	0.799	0.619	0.817	0.090	0.027	0.063	0.053	0.648	0.794	0.586	0.810
2	0.042	0.029	0.071	0.057	0.528	0.797	0.622	0.818	0.484	0.807	0.671	0.828	0.054	0.023	0.047	0.041
2	0.519	0.814	0.687	0.837	0.080	0.025	0.057	0.046	0.065	0.028	0.066	0.053	0.612	0.796	0.588	0.808
2	0.458	0.814	0.724	0.844	0.043	0.026	0.056	0.048	0.459	0.807	0.690	0.833	0.043	0.023	0.048	0.042
2	0.479	0.812	0.704	0.841	0.516	0.796	0.627	0.818	0.051	0.028	0.067	0.054	0.049	0.023	0.048	0.043
3	0.144	0.029	0.099	0.056	0.747	0.798	0.843	0.860	0.738	0.807	0.846	0.860	0.696	0.794	0.833	0.855
3	0.680	0.813	0.862	0.871	0.111	0.025	0.087	0.046	0.680	0.806	0.851	0.861	0.664	0.793	0.835	0.856
3	0.714	0.813	0.858	0.873	0.734	0.798	0.847	0.862	0.136	0.028	0.096	0.049	0.683	0.795	0.836	0.856
3	0.632	0.812	0.864	0.873	0.638	0.798	0.844	0.861	0.626	0.806	0.854	0.862	0.057	0.023	0.075	0.040
4	0.824	0.814	0.900	0.868	0.794	0.798	0.902	0.859	0.820	0.807	0.897	0.857	0.716	0.795	0.900	0.853

Note: Light gray indicates null and dark gray indicates alternative.
BHM: Bayesian hierarchical model; IND: independence; DEP: Positive dependence; ALT: belief in ≥ 3 true alternatives.

scenarios are presented in Appendix E of the Supplementary material available on *Biostatistics* online. In that appendix, we present similar tables describing operating characteristics of the designs considered when the efficacy level in each indication is uniformly lower than expected (Table E1 of the Supplementary material available on *Biostatistics* online) and when the control and treatment group response probabilities are incorrectly specified (Table E2 of the Supplementary material available on *Biostatistics* online). The results presented in Table E1 of the Supplementary material available on *Biostatistics* online illustrate that no method considered can overcome an overly optimistic assumption regarding the treatment effect in each indication, reinforcing the importance of powering studies using realistic assumptions regarding treatment effectiveness. The results presented in Table E2 of the Supplementary material available on *Biostatistics* online illustrate the performance in a worst-case setting for the proposed approach where control response probabilities are badly misrepresented by the conjugate priors. In that appendix, we describe how the method can be modified to increase robustness.

4.6. *Simulation study—different endpoints*

Table 4 presents estimates of the power and type I error rates under the assumption that one of the two elicited predictions for each indication are correct in the same sense as described in Section 4.5. Here, since the meaning of a true alternative or null is indication-specific, we present estimates of power and type I error rates for all $2^J = 16$ possible scenarios. As noted in Section 4.5, the sample size chosen for each indication was identified so that the IND design provided approximately 80% power for each indication and type I error control near the nominal level of 2.5%. We omit results from the IND design here because they are substantially similar to those from the identical endpoint case.

For the asthma indication (indication # 4), the BHM design only provides 72% power even when all four indications are truly alternative. Contrasted with the results from Section 4.5 where all four indications

Table 5. *Power and type I error rates based on partially correct predictions*

#	– Indication # 1 –				– Indication # 2 –				– Indication # 3 –				– Indication # 4 –			
H_1	BHM	IND	DEP	ALT	BHM	IND	DEP	ALT	BHM	IND	DEP	ALT	BHM	IND	DEP	ALT
0	0.004	0.029	0.006	0.034	0.010	0.025	0.005	0.028	0.004	0.028	0.005	0.034	0.020	0.024	0.004	0.024
1	0.011	0.029	0.009	0.037	0.023	0.026	0.006	0.030	0.015	0.028	0.008	0.036	0.171	0.282	0.079	0.291
1	0.019	0.029	0.022	0.042	0.025	0.026	0.012	0.031	0.266	0.805	0.548	0.819	0.037	0.024	0.010	0.028
1	0.010	0.029	0.011	0.039	0.088	0.314	0.104	0.333	0.011	0.027	0.009	0.036	0.033	0.024	0.006	0.026
1	0.257	0.815	0.585	0.831	0.027	0.026	0.012	0.031	0.021	0.027	0.018	0.039	0.035	0.024	0.010	0.027
2	0.035	0.029	0.041	0.050	0.050	0.026	0.028	0.041	0.393	0.805	0.594	0.823	0.230	0.281	0.143	0.303
2	0.025	0.029	0.021	0.045	0.174	0.314	0.133	0.340	0.028	0.028	0.018	0.042	0.216	0.282	0.103	0.300
2	0.029	0.029	0.043	0.052	0.182	0.316	0.181	0.346	0.368	0.806	0.604	0.826	0.047	0.023	0.025	0.036
2	0.381	0.813	0.629	0.835	0.048	0.025	0.029	0.041	0.035	0.027	0.038	0.047	0.225	0.283	0.144	0.304
2	0.458	0.814	0.724	0.844	0.043	0.026	0.056	0.048	0.459	0.807	0.690	0.833	0.043	0.023	0.048	0.042
2	0.358	0.813	0.638	0.839	0.180	0.317	0.184	0.348	0.033	0.027	0.040	0.047	0.043	0.023	0.025	0.035
3	0.060	0.030	0.066	0.054	0.279	0.316	0.290	0.388	0.499	0.807	0.682	0.839	0.273	0.282	0.250	0.347
3	0.562	0.814	0.801	0.860	0.067	0.025	0.075	0.046	0.563	0.807	0.780	0.849	0.263	0.281	0.374	0.373
3	0.484	0.816	0.710	0.849	0.260	0.316	0.293	0.391	0.062	0.028	0.061	0.049	0.274	0.281	0.253	0.347
3	0.548	0.814	0.808	0.861	0.263	0.316	0.418	0.414	0.545	0.806	0.785	0.852	0.049	0.023	0.064	0.041
4	0.663	0.812	0.856	0.867	0.341	0.315	0.484	0.416	0.660	0.805	0.843	0.857	0.297	0.283	0.443	0.370

Note: Indications 2 and 4 have half the hypothesized effect for scenarios where they are non-null. Light gray indicates null and dark gray indicates alternative.
BHM: Bayesian hierarchical model; IND: Independence; DEP: positive dependence; ALT: belief in $\geq$ 3 true alternatives.

were identical, this illustrates the difficulty in using the BHM to borrow information in the different endpoint setting. Of note, incorporating offsets so that the BHM shrinks deviations from hypothesized effects (Berry *and others*, 2013) instead of the effects themselves does not correct this issue. The results presented do not incorporate offsets as doing so resulted in even poorer performance. Contrasted with the approach based on the BHM, the proposed method using either the IND, DEP, or ALT weighting schemes produces results that are highly similar to the identical endpoint setting. The consistency of performance regardless of whether the sampling models for the data from each indication are different is quite apparent for the proposed method. Table 5 presents estimates of the power and type I error rates under the assumption that the efficacy levels for indications two and four are half what is hypothesized in scenarios where their respective alternative hypotheses are true. One can see in this case that the power for indications one and three are robust for the proposed method but significantly worse for the BHM compared to the results from Table 4.

Appendix F of the Supplementary material available on *Biostatistics* online contains additional tables similar in format to Tables 4 and 5. These supplementary tables present the bias of the posterior mean (Table F1 of the Supplementary material available on *Biostatistics* online) and expected sample size (Table F2 of the Supplementary material available on *Biostatistics* online) under the assumptions matching Table 4 and bias of the posterior mean (Table F3 of the Supplementary material available on *Biostatistics* online) under the assumptions matching Table 5. Tables F1 and F3 of the Supplementary material available on *Biostatistics* online illustrate that all information borrowing designs yield posterior means having some degree of bias but that the bias in the alternative setting is most pronounced for the BHM design for indication 1 where it is higher than for the proposed designs regardless of weighting scheme. Bias in the null setting is the greatest for indication 4 with the IND and ALT designs having slightly more bias than the BHM and DEP designs when viewed broadly across all scenarios for that indication. Table F2 of the

Supplementary material available on *Biostatistics* online illustrates that all designs lead to modest expected sample size reductions in the indications that take longest to complete enrollment (e.g., indication # 4) when futility criteria are met.

5. Discussion

The proposed framework for information borrowing designs provides an innovative mechanism for clinical trialists to increase the efficiency of early phase trials. Both the 21st Century Cures Act and the Prescription Drug User Fee Amendments of 2017 contain provisions that facilitate the use of complex innovative designs in drug development and regulatory decision-making. It is the authors' hope that the proposed method provides an avenue to broaden the use of information borrowing designs beyond trials involving medical devices and in oncology where such approaches are more commonly considered. Concurrent evaluation of an IP in multiple disease indications has potential value beyond information borrowing. Early phase development programs designed using the proposed approach may be nested within a clinical trial master protocol (Woodcock and LaVange, 2017) leading to efficiencies and cost savings that go beyond sample size reduction.

Software

Software to reproduce the simulations for this article can be found on GitHub at https://github.com/psioda/Basket-Hetero.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors wish to thank the associate editor and referees for helpful comments and suggestions which led to improvements of this article.

Conflict of Interest: We have no conflicts of interest to declare for this paper.

Funding

This work was supported by the US National Institutes of Health grant [P30 CA016086/CA/NCI, R01 GM070335/GM/NIGMS].

References

Berry, S. M., Broglio, K. R., Groshen, S. and Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: efficient designs of Phase II oncology clinical trials. *Clinical Trials* **10**, 720–734.

Chen, M.-H. and Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica* **13**, 461–476.

Chu, Y. and Yuan, Y. (2018). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials* **15**, 149–158.

Cunanan, K. M., Iasonos, A., Shen, R., Begg, C. B. and Gönen, M. (2017). An efficient basket trial design. *Statistics in Medicine* **36**, 1568–1579.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Source Journal of the Royal Statistical Society. Series B (Methodological) B* **57**, 45–97.

HOBBS, B. P. AND LANDIN, R. (2018). Bayesian basket trial design with exchangeability monitoring. *Statistics in Medicine* **37**, 3557–3572.

IBRAHIM, J. G. AND CHEN, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15**, 46–60.

KAIZER, A. M., KOOPMEINERS, J. S. AND HOBBS, B. P. (2018). Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics* **19**, 169–184.

MADIGAN, D. AND RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**, 1535.

NEUENSCHWANDER, B., WANDEL, S., ROYCHOUDHURY, S. AND BAILEY, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics* **15**, 123–134.

PARK, J. J. H., SIDEN, E., ZORATTI, M. J., DRON, L., HARARI, O., SINGER, J., LESTER, R. T., THORLUND, K. AND MILLS, E. J. (2019). Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials* **20**, 572.

PSIODA, M. A. AND IBRAHIM, J. G. (2019). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics* **20**, 400–415.

PSIODA, M. A., XU, J., JIANG, Q., KE, C., YANG, Z. AND IBRAHIM, J. G. (2021). Bayesian adaptive basket trial design using model averaging. *Biostatistics* **22**, 19–34.

R CORE TEAM. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

SPIEGELHALTER, D. J., FREEDMAN, L. S. AND PARMAR, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **157**, 357.

THALL, P. F., WATHEN, J. K., BEKELE, B. N., CHAMPLIN, R. E., BAKER, L. H. AND BENJAMIN, R. S. (2003). Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine* **22**, 763–780.

WOODCOCK, J. AND LAVANGE, L. M. (2017). Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine* **377**, 62–70.

ZHENG, H. AND WASON, J. M. S. (2020). Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy. *Biostatistics*. kxaa019.

[*Received June 29, 2020; revised November 30, 2020; accepted for publication February 25, 2021*]