Received: 8 September 2018 | Accepted: 8 October 2019

DOI: 10.1111/biom.13163

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Bayesian design of biosimilars clinical programs involving multiple therapeutic indications

Matthew A. Psioda[1] | **Kuolung Hu**[2] | **Yang Zhang**[3] | **Jean Pan**[4] | **Joseph G. Ibrahim**[1]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina

[2]Biometrics, Ionis Pharmaceuticals Inc, Carlsbad, California

[3]Biostatistics, Atara Biotherapeutics Inc, Thousand Oaks, California

[4]Global Development, Biosimilars, Biostatistics, Amgen Inc, Thousand Oaks, California

Correspondence
Matthew A. Psioda, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB#7420, Chapel Hill, NC 27599.
Email: matt_psioda@unc.edu

Funding information
NIH, Grant/Award Numbers: GM 70335, P01CA142538

Abstract

In this paper, we propose a Bayesian design framework for a biosimilars clinical program that entails conducting concurrent trials in multiple therapeutic indications to establish equivalent efficacy for a proposed biologic compared to a reference biologic in each indication to support approval of the proposed biologic as a biosimilar. Our method facilitates information borrowing across indications through the use of a multivariate normal correlated parameter prior (CPP), which is constructed from easily interpretable hyperparameters that represent direct statements about the equivalence hypotheses to be tested. The CPP accommodates different endpoints and data types across indications (eg, binary and continuous) and can, therefore, be used in a wide context of models without having to modify the data (eg, rescaling) to provide reasonable information-borrowing properties. We illustrate how one can evaluate the design using Bayesian versions of the type I error rate and power with the objective of determining the sample size required for each indication such that the design has high power to demonstrate equivalent efficacy in each indication, reasonably high power to demonstrate equivalent efficacy simultaneously in all indications (ie, globally), and reasonable type I error control from a Bayesian perspective. We illustrate the method with several examples, including designing biosimilars trials for follicular lymphoma and rheumatoid arthritis using binary and continuous endpoints, respectively.

KEYWORDS

Bayesian clinical trial design, Bayesian type I error rate, biosimilars, equivalence trial

1 | INTRODUCTION

A biosimilar is a biological product that is highly similar to and has no clinically meaningful differences with an approved reference biologic (U.S. Food and Drug Administration, 2015). There are four major components of biosimilarity that are evaluated in a stepwise manner to establish the totality of evidence required for FDA approval of a biosimilar: analytical characterization (eg, molecular structure and function), nonclinical assessments (eg, animal toxicity), clinical pharmacology, and clinical safety and efficacy. The clinical program for a biosimilar is conducted to demonstrate equivalent pharmacokinetics and efficacy, similar safety, and similar immunogenicity between the proposed and reference biologics with a goal of ultimately demonstrating that there are no clinically meaningful differences between the two products (U.S. Food and Drug Administration, 2015). One or more trials may be required to eliminate residual uncertainty regarding the

similarity of clinical efficacy and safety between the proposed and reference biologic, and these trials are typically designed to provide statistical evidence that the proposed biologic is neither inferior nor superior (in most cases) to the reference biologic based on prespecified equivalence margins.

Due to clinically relevant differences in the mechanism of action for a proposed biologic in different disease conditions (hereafter referred to as *indications*), a biosimilar program may include clinical trials for different indications with the goal of generating the comparative efficacy and safety data needed to support approval as a biosimilar for the indications studied and potentially other indications held by the reference biologic. By way of example, the biologic rituximab (brand name MabThera/Rituxan) has indications for treatment of non-Hodgkin's lymphoma (NHL), chronic lymphocytic leukemia, rheumatoid arthritis (RA), granulomatosis with polyangiitis, and microscopic polyangiitis. One biosimilar program (Deeks, 2017) included separate clinical trials in RA (Yoo *et al*., 2017) and follicular lymphoma (FL, a type of NHL) (Kim *et al*., 2017) in addition to analytical characterization and nonclinical assessments of the proposed biologic. On the basis of these collective data, the product was approved as a biosimilar by the European Medicines Agency (EMA) in 2017 for use in all indications held by rituximab.

In this paper, we develop a Bayesian clinical trial design methodology for the design of a biosimilars program that concurrently investigates several treatment indications with a goal of establishing equivalent clinical efficacy between a proposed and reference biologic in each of them. Our approach incorporates informative priors for each indication and allows for information borrowing on treatment efficacy equivalence across indications, resulting in a biosimilars program that is more efficient (ie, requires fewer subjects) than conducting independent trials in each indication. The use of informative priors may be justified, given the evidence of biosimilarity collected earlier in the development program (eg, analytical characterization and nonclinical studies).

Borrowing information on treatment effectiveness across different indications presents unique challenges. In particular, qualitatively different endpoints may be used for different indications. For example, both binary and continuous endpoints have been used in RA biosimilars trials (Yoo *et al*., 2017) whereas FL trials have used a binary objective response endpoint (Coiffier *et al*., 2016; Kim *et al*., 2017), which is common in oncology trials. For our approach, information borrowing is achieved using an informative multivariate normal prior, which we refer to as a *correlated parameter prior* (CPP), that induces prior correlation between the treatment effects for the different indications. The CPP is constructed from two elicited probabilities that are readily interpretable in the context of biosimilars development: (a) the marginal probability of treatment efficacy equivalence for an indication; and (b) the updated (conditional) probability of treatment efficacy equivalence for an indication given equivalence in another. The overarching goal of our design approach is to identify the minimum sample size required for each indication to be studied such that the analysis has high Bayesian (or average) power to demonstrate equivalent efficacy in each indication, reasonably high power to demonstrate equivalent efficacy simultaneously for all indications (ie, globally), and reasonable Bayesian type I error control. Use of these Bayesian operating characteristics for evaluating clinical trial designs has been recently studied by Psioda and Ibrahim (2018; 2019), Chen *et al*. (2014), Ibrahim *et al*. (2012), and Chen *et al*. (2011). Evaluating type I error rates in settings where multiple indications are investigated to help establish a global claim (ie, that the proposed biologic has equivalent efficacy compared to the reference product) is challenging because of the possibility that treatment equivalence may hold for only a subset of the indications studied. Generalizing the strategy proposed by Psioda and Ibrahim (2019) for evaluating Bayesian type I error rates for a single hypothesis in the presence of pertinent prior information, we develop a procedure for evaluating the Bayesian type I error rate for a composite global hypothesis derived from indication-specific hypotheses when some or all of them may be null.

Bayesian design strategies using the Bayesian hierarchical model (BHM) have been proposed to borrow information across indications in concurrent biosimilars trials by Berry *et al*. (2011) though the authors did not consider an application with endpoints based on different data types. Trial design methods incorporating information borrowing based on the BHM are increasingly common, especially in the area of oncology (Thall *et al*., 2003; Barry *et al*., 2015; Liu *et al*., 2017; Chu and Yuan, 2018). Haitao *et al*. (2017) propose an information-borrowing design for biosimilars trials, but the problem considered relates to borrowing information from historical data in the context of a single trial, which is quite different from our focus. There is a dearth of statistical methods that directly address information borrowing when the groups across which information is borrowed have outcomes of different types (eg, binary and continuous).

The rest of the paper is organized as follows. In Section 2, we give a brief overview of the design

problem and describe the example applications considered here and in the Supplementary Materials. In Section 3, we introduce a mathematical notation for the generalized linear model (GLM). In Section 4, we develop the indication-specific and global equivalence hypotheses defined with respect to the canonical parameter in a GLM as well as for a difference in means or proportions. In Section 5, we define the CPP and discuss prior elicitation. In Section 6, we describe the simulation-based design framework used to perform sample size determination and to evaluate the operating characteristics of a design based on a CPP chosen for analysis. In Section 7, we present a design application using the CPP that involves designing biosimilars trials investigating treatment efficacy equivalence for RA and FL indications where the endpoints are continuous and binary, respectively. We close with some discussion in Section 8.

2 | PRELIMINARIES

The general framework we develop in this paper can be cast a sample size determination method where the goal is to determine the minimum sample size required in each indication, subject to Bayesian type I error and power constraints, as well as other user-specified sample size considerations. Although the methodology is developed to be quite general, we ground the discussion with concrete illustrations based on a biosimilars program with $J = 2$ indications. Many of the concepts that may be unfamiliar to the reader (eg, sampling priors discussed in Section 6.3) can be visualized effectively in the case where $J = 2$, which may help with understanding the methodology. Moreover, it is likely that J will be small for most biosimilars programs (eg, $J = 2$ or $J = 3$); so this choice is also practical.

For our primary example, we consider proportional sample size reduction across the two indications studied, meaning that the ratio of the actual sample size for a trial in a given indication based on the proposed method to the sample size required for an independent trial (having the same power and without use of prior information) is equal for all indications studied. We supplement that example with additional examples and discussion in the Supplementary Materials. In Appendix A of the Supplementary Materials, we consider an application with sample size reduction occurring in one indication only to mirror a situation where one indication is substantially more difficult to enroll than the other. In Appendix B of the Supplementary Materials, we consider an application with $J = 3$ indications.

3 | GENERALIZED LINEAR MODELS

In this section, we describe the general sampling distribution for the data. Let $j = 1, ..., J$ index indication and $i = 1, ..., n_j$ index subject within an indication. Denote the outcome for subject i in indication j by y_{ij}. We assume y_{ij} has a probability distribution in the exponential family given as follows:

$$p(y_{ij}|\theta_{ij}, \tau_j) = \exp\{a_{ij}^{-1}(\tau_j)(y_{ij}\cdot\theta_{ij} - b(\theta_{ij})) + c(y_{ij}, \tau_j)\},$$

indexed by the natural parameter θ_{ij} and the indication-specific scale parameter τ_j. The functions $b(\cdot)$ and $c(\cdot)$ determine a particular family in the class, such as the binomial, normal, Poisson, and so forth. The functions $a_{ij}(\tau_j)$ are commonly of the form $a_{ij}(\tau_j) = \tau_j^{-1}w_{ij}^{-1}$, where the w_{ij}'s are known weights. We assume that $w_{ij} = 1$ throughout.

Now suppose θ_{ij} satisfies $\theta_{ij} = \theta_j(\eta_{ij})$ with $\eta_{ij} = \alpha_j + z_{ij}\gamma_j + \mathbf{x}_{ij}^T\boldsymbol{\beta}_j$, where α_j is an intercept parameter for indication j, γ_j is the treatment effect parameter 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, $\boldsymbol{\beta}_j = (\beta_{j,1}, ..., \beta_{j,p_j})$ is the corresponding $p_j \times 1$ vector of parameters, and $\theta_j(\cdot)$ is a monotone differentiable function. We use the label $\mathbf{D}_j = \{(y_{ij}, z_{ij}, \mathbf{x}_{ij}) : i = 1, ..., n_j\}$ to refer to the data for subjects from indication j and $\mathbf{D}$ to refer to the data for all subjects. Models of the form described above are known as GLMs. The function $\theta_j(\cdot)$ links the linear predictor η_{ij} to the natural parameter θ_{ij}. When $\theta_{ij} = \eta_{ij}$, the link is said to be a *canonical* link.

4 | HYPOTHESIS TESTING

4.1 | Indication-specific equivalence hypotheses for GLMs

The primary inferential goal for each indication is to prove treatment efficacy equivalence of the proposed biologic compared to the reference biologic. In the case of a GLM, the associated hypotheses can be formulated as

$$H_{0j} : |\gamma_j| \geqslant \delta_j \quad \text{vs} \quad H_{1j} : |\gamma_j| < \delta_j, \tag{1}$$

where $\delta_j \geqslant 0$ is the largest absolute value of γ_j that is not clinically meaningful (ie, the equivalence margin). Using the canonical link functions, this is a test of the difference in means for a linear model, the log ratio of means for a Poisson regression model, and the log odds ratio for a logistic regression model.

4.2 | Hypotheses based on differences in means or proportions

Biosimilars trials frequently test equivalence hypotheses based on a difference in means even when the data are not normally distributed (eg, using a difference in proportions for binary data), typically based on a model without covariates. Such a choice presents no challenge to the proposed methodology. One only needs to employ the identity link instead of the canonical link for nonnormal data.

For example, in the case of binary data, the probability of response $\pi_j(z)$ for subjects in treatment group z from indication j is $\pi_j(z) = \exp(\alpha_j + \gamma_j z)/(1 + \exp(\alpha_j + \gamma_j z))$ for the logit link (ie, the canonical link) and is $\pi_j(z) = \alpha_j + \gamma_j z$ for the identity link. For the latter, $\alpha_j \in (0, 1)$ is the control group response probability and γ_j is the difference in response probabilities between the treated and control groups. To simplify model fitting, we consider the alternative formulation

$$\pi_j(z) = \max(\min(\alpha_j + \gamma_j z, 1), 0), \tag{2}$$

which, for fixed α_j and γ_j, is equivalent to $\pi_j(z) = \alpha_j + \gamma_j z$. In practice, inference can be performed using the posterior distribution for $\pi_j(1) - \pi_j(0)$, which is essentially equivalent to that for γ_j provided the sample size in each group is not too small, actual response probabilities are not too extreme, and reasonable priors are employed. Importantly, using formulation (2) is advantageous from a computational perspective as it allows the prior distribution for γ_j to have unbounded support, which helps in constructing a correlated prior distribution for the $\{\gamma_1, ..., \gamma_J\}$ that directly takes into account each indication's equivalence margin on the scale on which it is defined. The prior is developed in detail in Section 5. In a slight abuse in notation, in what follows we will write general probability statements based on γ_j and the associated equivalence margin δ_j while acknowledging that formal inference for binary endpoint indications is based on the posterior distribution for $\pi_j(1) - \pi_j(0)$.

4.3 | Indication-specific evidence evaluation

In the proposed testing framework, one rejects the null hypothesis for indication j when $P(|\gamma_j| < \delta_j|\mathbf{D}) \geqslant p_{0j}$ where p_{0j} is a prespecified posterior probability critical value. Our use of $\mathbf{D}$ as opposed to $\mathbf{D}_j$ in the posterior probability reflects the fact that analysis using the CPP will generally induce information borrowing across indications and, therefore, posterior inference for any one indication will be influenced by the data from all indications. The choice of p_{0j} should correspond to an evidence threshold thought to be compelling to stakeholders. A default choice is to take $p_{0j} = 0.95$ which, under some assumptions (eg, a fixed sample size, a noninformative prior, and a single analysis) will result in a type I error rate of approximately 0.05 when $|\gamma_j| = \delta_j$. As our focus in this paper is not to define what constitutes compelling evidence, we shall simply fix $p_{0j} = 0.95$ for our examples.

4.4 | Global equivalence hypotheses and evidence evaluation

While it is important that each indication has high power to prove treatment efficacy equivalence, it is also important that the program as a whole is reasonably powered to simultaneously demonstrate treatment efficacy equivalence in *all* indications. A program that fails to prove equivalence in every indication studied may not eliminate all residual uncertainty about the biosimilarity of the proposed biologic. We define the *global* (or equivalently, program-level) equivalence hypothesis with this goal in mind. The global alternative hypothesis asserts that the equivalence criteria hold for each of the J indications to be studied. Let Θ be the parameter space for $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_J)$ and define the global alternative space as $\Theta_1 = \{\boldsymbol{\gamma} : |\gamma_j| < \delta_j,\ j = 1, ..., J\}$ with complement Θ_0. The global equivalence hypotheses may then be defined generally as $H_0 : \boldsymbol{\gamma} \in \Theta_0$ vs $H_1 : \boldsymbol{\gamma} \in \Theta_1$ with the decision to reject the global null hypothesis occurring when $P(\boldsymbol{\gamma} \in \Theta_1|\mathbf{D}) \geqslant p_0$, where p_0 is a prespecified posterior probability critical value which we assume to be equal to 0.95 for our examples.

5 | THE CPP

The CPP is a multivariate normal prior for $\boldsymbol{\gamma}$ and can be written as $\boldsymbol{\gamma}|\pi_0, \pi_1 \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with the positive definite covariance matrix $\boldsymbol{\Sigma}$ determined indirectly by elicited scalar hyperparameters π_0 and π_1. The hyperparameters π_0 and π_1 correspond, respectively, to the (prior) marginal probability of treatment equivalence for an indication and the conditional probability of treatment equivalence for an indication given equivalence in another. Formally, we define $\pi_0 = P(|\gamma_j| < \delta_j)$ for all j and $\pi_1 = P(|\gamma_j| < \delta_j||\gamma_k| < \delta_k)$ for all $j \neq k$. Given a choice for π_0 and π_1 and the specified equivalence margins, it is a simple computational problem to determine $\boldsymbol{\Sigma}$. A computational approach is described in Appendix C of the Supplementary Materials.

5.1 | On the elicitation of π_0 and π_1

The intent behind parameterizing the covariance matrix $\mathbf{\Sigma}$ in terms of π_0 and π_1 is to provide interpretable hyperparameters to support prior elicitation. Recall that a key purpose of the clinical program in the multistage biosimilars development process is to eliminate *residual uncertainty* regarding the equivalence of treatment efficacy between the proposed and reference product. For our purposes, we define the residual uncertainty regarding treatment efficacy equivalence as $1 - \pi_0$. An agnostic viewpoint would correspond to $\pi_0 = 1/3$ which suggests that equivalence is no more likely than inferiority (ie, $\gamma_j \leqslant -\delta_j$) or superiority (ie, $\gamma_j \geqslant \delta_j$). A more realistic but still relatively agnostic perspective would correspond to equal prior probabilities on the two hypotheses (ie, $\pi_0 = 0.5$). If we define compelling evidence of treatment efficacy equivalence in an indication as $P(|\gamma_j| < \delta_j|\mathbf{D}) > 0.95$, then choosing $\pi_0 \in [0.5, 0.8]$ would be consistent with having pertinent knowledge from earlier stages in the development program and the presence of a nonnegligible degree of residual uncertainty.

To elicit π_1, one must ask the question, "If a trial were conducted in one indication and equivalence proved, how would that modify π_0 for the indications yet to be studied?" It may also be helpful to frame the question in terms of the percent reduction in residual uncertainty regarding treatment efficacy equivalence in one indication resulting from proving equivalence in another, defined formally as $\Delta_{\text{RU}} = \left(\frac{\pi_1 - \pi_0}{1 - \pi_0}\right) \times 100$. Having determined π_0, eliciting Δ_{RU} is equivalent to eliciting π_1. Biosimilars have been approved for indications not directly studied in clinical efficacy trials by the EMA (Deeks, 2017) suggesting that proof of treatment efficacy equivalence in one indication may provide a substantial reduction in uncertainty about equivalence in other indications. Values of $\Delta_{\text{RU}} \in [25, 50]$ may be appropriate when, for the set of indications studied in a given clinical program, there is a sound scientific justification for information borrowing.

6 | SIMULATION-BASED DESIGN FRAMEWORK

In this section, we develop a simulation-based design procedure that can be used for sample size determination and evaluation of Bayesian operating characteristics based on an elicited CPP to be used for analysis. We extend the simulation-based procedure for characterizing Bayesian versions of the type I error rate and power developed previously (Chen *et al*., 2011; Ibrahim *et al*., 2012; Chen *et al*., 2014; Psioda and Ibrahim, 2018; 2019) to the scenario where one is evaluating several related hypotheses as well as a global composite hypothesis.

6.1 | Sampling priors and Bayesian operating characteristics

To formally define the Bayesian type I error rate and Bayesian power, we first introduce the concept of sampling or design priors (O'Hagan and Stevens, 2001; Wang and Gelfand, 2002). Let $\theta = (\boldsymbol{\gamma}, \boldsymbol{\psi})$ be the collection of all parameters for all indications where $\boldsymbol{\gamma}' = [\gamma_1, ..., \gamma_J]$ is the vector of treatment effect parameters, and $\boldsymbol{\psi}$ is the collection of all nuisance parameters (ie, $\alpha_1, ..., \alpha_J$, $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J$, and $\tau_1, ..., \tau_J$). A sampling prior is simply a probability distribution for θ that reflects a (possibly assumed) state of knowledge about θ.

When there are J indications to be studied simultaneously, one needs to evaluate the operating characteristics of the design for each of the 2^J possible scenarios regarding treatment efficacy equivalence across the respective indications. These scenarios are summarized in Table 1 for the case where $J = 2$. The scenarios are identified by a two-letter designation such as AN where (in this example) "A" indicates a true alternative for indication 1 and "N" indicates a true null for indication 2.

In order to evaluate operating characteristics based on each of the 2^J scenarios, one must specify sampling prior distributions for θ that are consistent with each of them. For example, a valid sampling prior to distribution for the AN scenario in Table 1 would give nonzero mass to values of θ that satisfy both $|\gamma_1| < \delta_1$ and $|\gamma_2| \geqslant \delta_2$.

6.2 | Defining the Bayesian type I error rate and power

In this section, we formally define the Bayesian type I error rate and power using the notation of (Psioda and Ibrahim, 2019) extended for a biosimilars program with multiple clinical efficacy trials, where it is of interest to characterize the power to demonstrate indication-specific equivalence as well as global equivalence.

TABLE 1 Scenarios for equivalence

Scenario label	True indication hypothesis		True global hypothesis
	$j = 1$	$j = 2$	
AA	H_{11}	H_{12}	H_1
AN	H_{11}	H_{02}	H_0
NA	H_{01}	H_{12}	H_0
NN	H_{01}	H_{02}	H_0

For fixed θ, define the null hypothesis rejection rate for indication j as

$$r_j(\theta) = \mathrm{E}[1\{P(|\gamma_j| < \delta_j|\mathbf{D}) \geqslant p_{0j}\}|\theta], \quad (3)$$

where $1\{P(|\gamma_j| < \delta_j|\mathbf{D}) \geqslant p_{0j}\}$ is an indicator that one rejects H_{0j} based on the posterior probability $P(|\gamma_j| < \delta_j|\mathbf{D})$ and prespecified critical value p_{0j}. Note that the expectation in (3) is with respect to the distribution of the data $\mathbf{D}$ given θ. It should be understood that $r_j(\theta)$ implicitly depends on the chosen analysis prior which in our case is the CPP described in Section 5. Further, define the global null hypothesis rejection rate to be

$$r(\theta) = \mathrm{E}[1\{P(\boldsymbol{\gamma} \in \Theta_1|\mathbf{D}) \geqslant p_0\}|\theta]. \quad (4)$$

Now consider one of the 2^J scenarios for which power and/or type I error control will be evaluated, and denote the associated sampling prior generically as $\pi^{(s)}(\theta)$ (here (s) is used to denote sampling prior). The Bayesian null hypothesis rejection rate for indication j is formally defined as

$$r_j^{(s)} = \mathrm{E}_{\pi^{(s)}}[r_j(\theta)] = \int_\theta r_j(\theta)\pi^{(s)}(\theta)d\theta, \quad (5)$$

and can be recognized as a weighted average null hypothesis rejection rate with weights determined by the user-specified sampling prior $\pi^{(s)}(\theta)$. The Bayesian global null hypothesis rejection rate $r^{(s)} = \mathrm{E}_{\pi^{(s)}}[r(\theta)]$ has an analogous interpretation as a weighted average rate.

Thus far, we have been careful to use the term *null hypothesis rejection rate*. This is because whether $r_j^{(s)}$ can be interpreted as a Bayesian type I error rate or Bayesian power depends on the scenario in question (see Table 1). For the AN scenario from Table 1, the value $r_1^{(s)}$ is the indication-specific Bayesian power for indication 1 and $r_2^{(s)}$ is the indication-specific Bayesian type I error rate for indication 2. For the AN, NA, and NN scenarios, the quantity $r^{(s)}$ is the Bayesian type I error rate associated with the global null hypothesis whereas for the AA scenario it is the Bayesian power. An algorithm for simulation-based estimation of $r_j^{(s)}$ and $r^{(s)}$ is given in Appendix D of the Supplementary Materials.

6.3 | Construction of sampling priors

In this section, we discuss the construction of sampling priors corresponding to the scenarios in Table 1. In settings where information is being borrowed across indications, one should consider several different sampling priors when evaluating the performance of a design that uses the CPP (or any informative analysis prior). The purpose of using multiple sampling priors is to characterize type I error rates and power for a coherent set of possibilities for the true value of $\boldsymbol{\gamma}$ to support decision making regarding whether the chosen analysis prior is adequate in the opinion of regulatory and nonregulatory stakeholders.

We will consider two types of sampling priors: (a) point-mass sampling priors and (b) nondegenerate sampling priors. Point-mass sampling priors are trivial to construct and the associated Bayesian type I error rate or power aligns with classical frequentist versions. More generally, one can construct nondegenerate sampling priors for one or more components of $\boldsymbol{\gamma}$ by first defining a plausible distribution for $\boldsymbol{\gamma}$, denoted by $\pi^{(D)}(\boldsymbol{\gamma})$, and then conditioning on an event (eg, $\gamma_j = -\delta_j$) to induce a plausible sampling prior for one of the scenarios in Table 1 *given* the scenario is true. This type of conditioning approach was considered by Psioda and Ibrahim (2018; 2019) when constructing sampling priors for a single set of hypotheses. Herein, we discuss an extension for multiple related hypotheses.

To keep context clear, we will use subscripts such as $1, AA$ in the sampling prior labels (eg, $\pi_{1,AA}^{(s)}(\boldsymbol{\gamma})$) where the first component in the subscript indicates whether the sampling prior is a point-mass prior (1 = point-mass) and the second component indicates the underlying scenario (AA = true alternative for both indications). To fix ideas, consider the design of a biosimilars program evaluating the equivalence of a proposed and reference biologic for the treatment of FL ($j = 1$) and RA ($j = 2$). We consider the same example in the design application in Section 7. We assume that the FL treatment evaluation will be based on a difference in proportions endpoint using objective overall response, an equivalence margin of $\delta_1 = 0.10$, and that prior data suggest the response probability for patients treated with the reference biologic is 0.81. We assume the RA treatment evaluation will be based on a mean change from baseline endpoint (eg, a composite disease activity score using 28-joint counts and C-reactive protein levels—DAS28-CRP), an equivalence margin of 0.6 units per European League Against Rheumatism guidelines, and that prior data suggest a mean change from baseline equal to −2.0 for the reference biologic and a standard deviation for the change equal to 1.4.

For ease of exposition, we will assume point-mass sampling priors on the nuisance parameters and no covariates. Specifically, we assume (based on the identity link)

$$\pi^{(s)}(\alpha_1, \alpha_2, \tau_2) \propto 1(\alpha_1 = 0.81) \times 1(\alpha_2 = -2.0) \times 1(\tau_2 = 1.4^2). \quad (6)$$

These same point-mass sampling priors for the nuisance parameters are used for the design application in Section 7 and for the applications presented in the Supplementary Materials.

Panel A of Figure 1 presents a bivariate contour plot for an example choice for $\pi^{(D)}(\boldsymbol{\gamma})$ based on a CPP with a prior probability of treatment efficacy equivalence equal to $\pi_0 = 0.75$ and $\pi_1 = 0.875$ corresponding to a 50% reduction in residual uncertainty of treatment efficacy equivalence in one indication given equivalence in the other (ie, $\Delta_{\text{RU}} = 50$). These choices induce a CPP covariance matrix with standard deviations equal to 0.087 and 0.522 for the FL and RA indications, respectively, and correlation equal to 0.835. This choice for $\pi^{(D)}(\boldsymbol{\gamma})$ will serve as the reference from which we will construct the nondegenerate sampling priors specific to the scenarios in Table 1 for use in design evaluations.

Panel B of Figure 1 presents the marginal prior distribution $\pi^{(D)}(\gamma_1)$ for the difference in proportions for the FL indication. Panels C and D present, respectively, the prior distributions for the FL difference in proportions conditional on inferiority (panel C) and equivalence (panel D) for the RA indication. Panels E and F, respectively, present the prior distribution conditional on the RA mean difference equaling its boundary null value for inferiority (panel D) and equaling 0 (panel F). These plots help to illustrate how information regarding the treatment effect in one indication influences the distribution for the treatment effect in the other.

6.3.1 | Sampling priors for the AA scenarios

The most optimistic sampling prior for the AA scenario (ie, the sampling prior that only allows for perfect equivalence) is given by $\pi^{(s)}_{1,AA}(\boldsymbol{\gamma}) \propto 1(\gamma_1 = 0, \gamma_2 = 0)$. This point-mass prior places all mass at the mode of the distribution in panel A of Figure 1. When stakeholders of the proposed biologic are highly confident in equivalence, using this AA sampling prior as the sole basis for power analysis is appropriate. However, this approach can be anti-conservative for sample size determination and may leave the study underpowered when the reference biologic is modestly more or less effective than the proposed biologic.

A less optimistic AA sampling prior, denoted by $\pi^{(s)}_{2,AA}(\boldsymbol{\gamma})$, is obtained by conditioning $\pi^{(D)}(\boldsymbol{\gamma})$ on the event $\{|\gamma_j| < c{\cdot}\delta_j : j = 1, 2\}$ for $c \in (0, 1)$. Panel A of Figure 2 presents a contour plot for the truncated alternative sampling prior $\pi^{(s)}_{2,AA}(\gamma_1, \gamma_2)$ based on $c = 0.5$.

Panel B of Figure 2 presents a histogram for the marginal distribution for the FL difference in proportions for the truncated sampling prior. Though the primary power analysis will often be based on the $\pi^{(s)}_{1,AA}$ sampling prior, we recommend, at minimum, performing a supplemental power analysis using the $\pi^{(s)}_{2,AA}$ sampling prior with $c \approx 0.5$ to assess the robustness of power for the chosen sample sizes.

6.3.2 | Sampling priors for the AN and NA scenarios

In this section, we focus discussion on the AN scenario (the NA scenario is analogous). A reasonable worst-case point-mass AN sampling prior (for evaluating type I error rates) is given by $\pi^{(s)}_{1,AN}(\boldsymbol{\gamma}) \propto 1(\gamma_1 = 0, \gamma_2 = -\delta_2)$. This sampling prior places all mass at the dark square labeled "AN" in Panel A of Figure 1. Of course, an even less favorable AN sampling prior is given by $\pi^{(s)}_{1,AN}(\boldsymbol{\gamma}) \propto 1(\gamma_1 = \delta_1(1 - \epsilon), \gamma_2 = -\delta_2)$. for small $\epsilon > 0$. However, such a sampling prior is likely not plausible given the implied belief that the proposed biologic is nearly superior in one indication and inferior in the other.

We would not generally advocate using the Bayesian type I error rate based on the $\pi^{(s)}_{1,AN}(\boldsymbol{\gamma})$ sampling prior as the sole determining factor regarding whether a chosen CPP provides acceptable performance under that scenario. This is because the $\pi^{(s)}_{1,AN}(\boldsymbol{\gamma})$ sampling prior will not generally correspond to what are viewed to be the most likely parameter values under that scenario. If we acknowledge $\pi^{(D)}(\boldsymbol{\gamma})$ as a reasonable probabilistic relationship between the treatment effects, a more plausible AN sampling prior, denoted by $\pi^{(s)}_{2,AN}(\boldsymbol{\gamma})$, is obtained by conditioning $\pi^{(D)}(\boldsymbol{\gamma})$ on the event $\{|\gamma_1| < \delta_1, \gamma_2 = -\delta_2\}$. The induced marginal sampling prior distribution for the FL difference in proportions is presented in panel C of Figure 2. Under the more plausible sampling prior distribution (ie, $\pi^{(s)}_{2,AN}(\boldsymbol{\gamma})$), the type I error inflation associated with information borrowing is much less substantial.

7 | APPLICATION: BIOSIMILARS PROGRAM DESIGN WITH THE CPP

In this section, we compare designs based on analysis CPPs using various choices for π_0 and π_1 as well as designs based on the BHM to evaluate their performance with respect to Bayesian power and type I error rates as defined in Section 6.2 based on the sampling priors discussed in Section 6.3. For CPP designs, a uniform prior was used for the intercept in the binomial model (identity link) and a uniform improper prior was used for the normal model intercept and standard deviation. Priors for the BHM are given in Section 7.3.

7.1 | Comparison of designs based on the CPP

We focus on the design of two biosimilars trials evaluating treatment efficacy equivalence for FL and RA indications as described in Section 6.3. A traditional approach would be to conduct independent trials in each of the two indications to establish equivalence in both to support

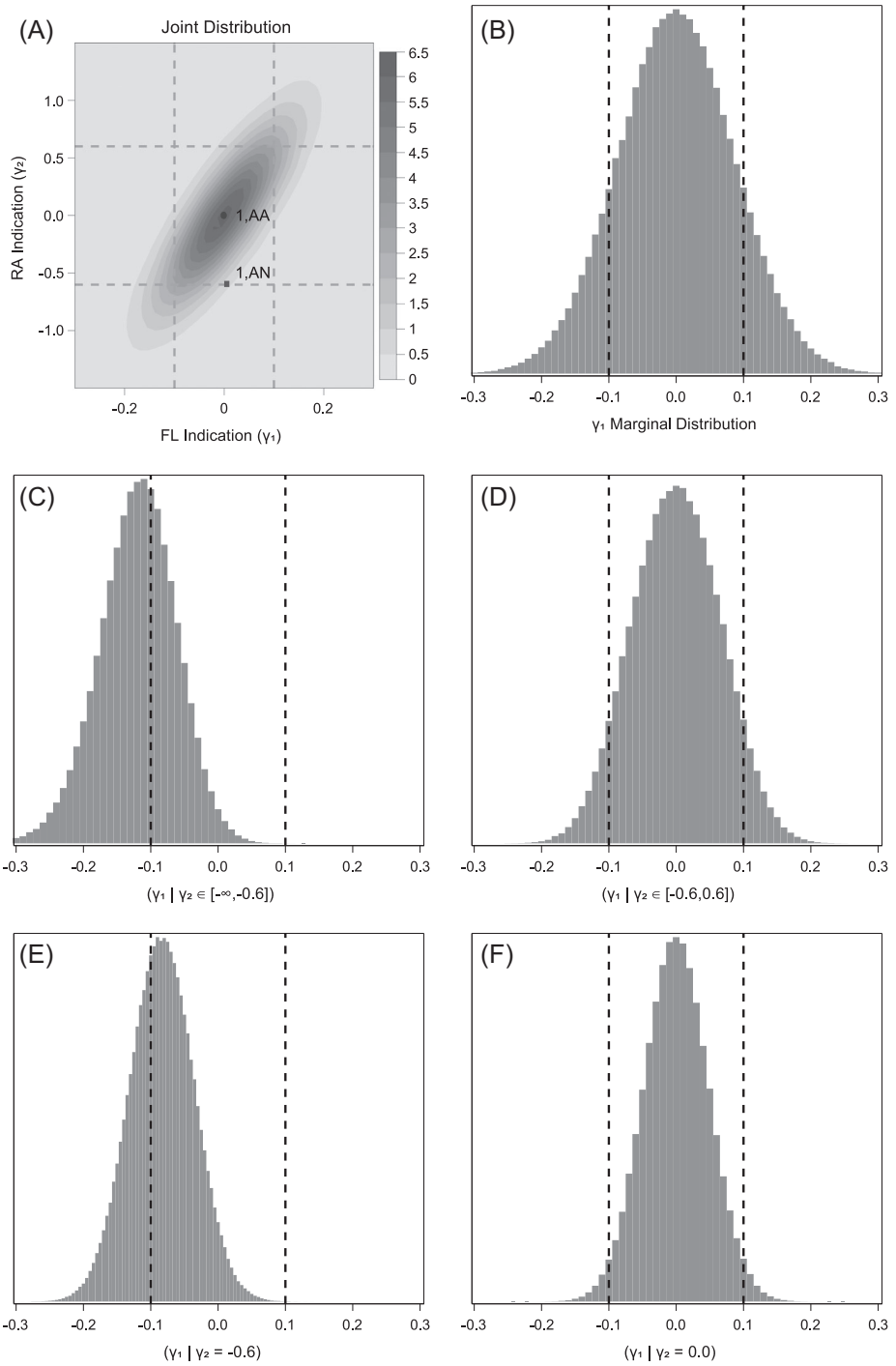

FIGURE 1 A, Bivariate prior $\pi^{(D)}(\boldsymbol{\gamma})$ for FL indication treatment effect γ_1 (difference in proportions, *x*-axis) and RA indication treatment effect γ_2 (difference in means, *y*-axis). B, Marginal sampling prior for FL effect. C, Sampling prior for FL effect conditional on RA inferiority (ie, $\gamma_2 \in (-\infty, -0.6]$). D, Sampling prior for FL effect conditional on RA equivalence (ie, $\gamma_2 \in [-0.6, 0.6]$). E, Sampling prior for FL effect conditional on RA inferiority boundary value (ie, $\gamma_2 = -0.6$). F, Sampling prior to FL effect conditional on RA perfect equivalence (ie, $\gamma_2 = 0.0$). FL, follicular lymphoma; RA, rheumatoid arthritis

approval of the proposed biologic as a biosimilar for the reference biologic in both indications and potentially others. For the FL indication, an asymptotic sample size calculation based on the score test identifies that 328 subjects per group are needed (total $N_1 = 756$) to have power equal to 0.90 ($\alpha = .05$) to demonstrate equivalent efficacy between the proposed and reference biologics when both groups have a response probability of .81 and the equivalence margin for the difference in response probabilities is .10. For the RA indication, an exact sample

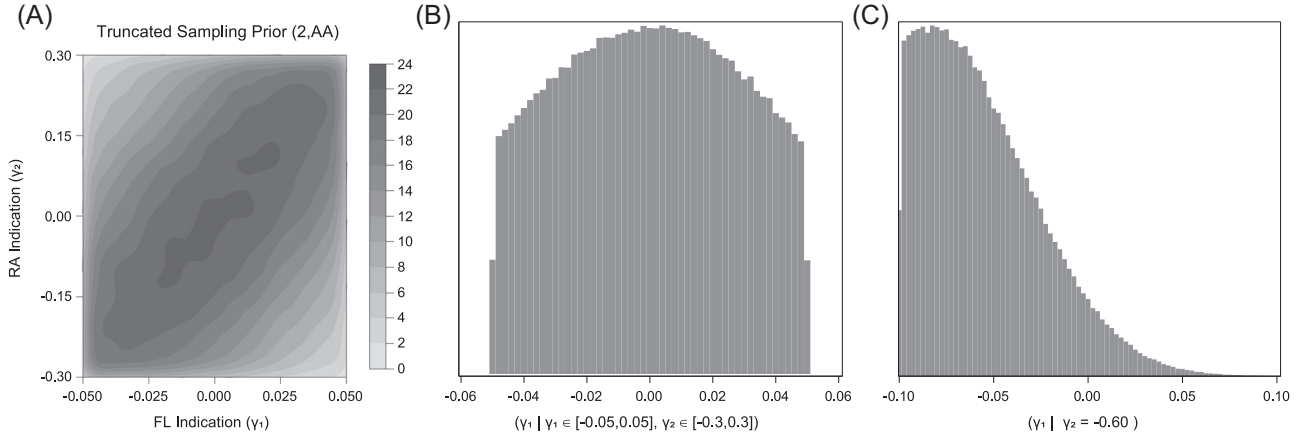

FIGURE 2 A, Truncated bivariate alternative sampling prior $\pi^{(s)}_{2,AA}(\gamma_1, \gamma_2)$ based on $c = 0.5$. B, Marginal alternative sampling prior for FL effect based on $c = 0.5$. C, Marginal alternative sampling prior for FL effect conditional on RA inferiority boundary value (ie, $\gamma_2 = -0.6$). FL, follicular lymphoma; RA, rheumatoid arthritis

size calculation identifies that 119 subjects per group are needed (total $N_2 = 238$) to have power equal to 0.90 ($\alpha = .05$) to demonstrate equivalent efficacy based on mean DAS28-CRP change from baseline when the standard deviation for the change is 1.4 for both groups. Thus, if two independent trials were conducted using these sample sizes (when the assumptions are met), the power to demonstrate global equivalence would be approximately 0.81 (ie, $0.9^2 = 0.81$).

A key benefit of using an informative CPP is that doing so permits sample size reduction relative to that required for two independent equivalence trials without sacrificing statistical power (provided one is willing to relax traditional frequentist type I error control requirements). We evaluate the performance of CPP-based designs based on two strategies for sample size reduction. For the first approach (presented in this section), we consider proportional sample size reduction for both the FL and RA indications relative to the sample sizes that would be needed for two independent trials (ie, N_1 and N_2 above). For the second approach, we keep the total sample size for the RA group equal to $N_2 = 238$ and consider a reduction in the number of FL subjects required. The motivation and results for this example are described in Appendix A of the Supplementary Materials. In Appendix B of the Supplementary Materials, we consider an additional example with $J = 3$ indications to demonstrate that the method can be easily applied with more than two indications.

7.2 | Design with proportional sample size modification in both indications

Table 2 presents estimated null hypothesis rejection rates for the FL and RA indications as well as the global null hypothesis rejection rate for the sampling priors described in Section 6.3. These quantities are denoted as $r^{(s)}_{\text{FL}}$, $r^{(s)}_{\text{RA}}$, and $r^{(s)}$, respectively.

Estimates for each combination of sampling prior and CPP used for analysis are based on 20 000 simulated data sets. For each CPP, the operating characteristics presented are from the design having the smallest sample size in each indication such that the Bayesian power to demonstrate equivalent efficacy using the $\pi^{(s)}_{1,AA}$ sampling prior is $\geqslant$0.90 for each indication and $\geqslant$0.80 globally with a Bayesian type I error rate of no greater than 0.10 based on the $\pi^{(s)}_{2,AN}$ and $\pi^{(s)}_{2,NA}$ sampling priors. The sample sizes considered ranged from 30% to 100% of the standard sample sizes identified above for the FL and RA indications, using 5% increments.

The CPP prior defined by $\pi_0 = \pi_1 = 0.33$ is noninformative and results in no information borrowing. The prior affords no sample size reduction due to the power requirements imposed on the design but provides excellent type I error control for indication-specific and global null hypotheses regardless of whether the more liberal null sampling priors (eg, $\pi^{(s)}_{2,AN}$) or conservative null sampling priors (eg, $\pi^{(s)}_{1,AN}$) are used to define the Bayesian type I error rate. Hence, the prior is a useful reference for evaluating the efficiency gains (eg, increased power) and tradeoffs (eg, elevated type I error rates) associated with the use of more informative CPPs.

Under the imposed type I error and power constraints (based on the $\pi^{(s)}_{1,AA}$ sampling prior), trials may be performed using up to a 40% reduced sample size when the CPP based on $\pi_0 = 0.75$ and $\Delta_{\text{RU}} = 50$ ($\pi_1 = 0.875$) is used for analysis. However, for the less optimistic $\pi^{(s)}_{2,AA}$ sampling prior (and for the same analysis prior), the power to demonstrate global equivalence drops to 0.68 with indication-specific hypotheses having power slightly more than 0.8. Nonetheless, for *any* CPP the power to

TABLE 2 Design operating characteristics based on proportional sample size modification for each indication

	$\Delta_{RU} = 0$						$\Delta_{RU} = 25$					$\Delta_{RU} = 50$				
Sampling prior	π_0	π_1	%SS	$r_{FL}^{(s)}$	$r_{RA}^{(s)}$	$r^{(s)}$	π_1	%SS	$r_{FL}^{(s)}$	$r_{RA}^{(s)}$	$r^{(s)}$	π_1	%SS	$r_{FL}^{(s)}$	$r_{RA}^{(s)}$	$r^{(s)}$
$\pi_{1,AA}^{(s)}$	0.33	0.33	1.00	0.91	0.91	0.81	0.50	1.00	0.91	0.92	0.81	0.67	0.95	0.91	0.92	0.82
	0.50	0.50	1.00	0.92	0.92	0.82	0.63	0.90	0.91	0.91	0.80	0.75	0.80	0.91	0.91	0.81
	0.67	0.67	0.90	0.91	0.91	0.80	0.75	0.85	0.92	0.92	0.83	0.84	0.70	0.92	0.92	0.82
	0.75	0.75	0.85	0.91	0.92	0.80	0.81	0.75	0.91	0.91	0.80	0.88	0.60	0.91	0.91	0.81
$\pi_{1,AN}^{(s)}$	0.33	0.33	1.00	0.91	0.06	0.04	0.50	1.00	0.91	0.06	0.04	0.67	0.95	0.91	0.07	0.05
	0.50	0.50	1.00	0.92	0.06	0.05	0.63	0.90	0.90	0.08	0.06	0.75	0.80	0.87	0.11	0.09
	0.67	0.67	0.90	0.91	0.08	0.06	0.75	0.85	0.91	0.10	0.08	0.84	0.70	0.84	0.18	0.14
	0.75	0.75	0.85	0.91	0.09	0.07	0.81	0.75	0.89	0.12	0.10	0.88	0.60	0.80	0.24	0.20
$\pi_{1,NA}^{(s)}$	0.33	0.33	1.00	0.06	0.91	0.05	0.50	1.00	0.06	0.91	0.05	0.67	0.95	0.07	0.91	0.06
	0.50	0.50	1.00	0.06	0.92	0.05	0.63	0.90	0.08	0.90	0.06	0.75	0.80	0.13	0.88	0.10
	0.67	0.67	0.90	0.08	0.92	0.06	0.75	0.85	0.11	0.91	0.09	0.84	0.70	0.19	0.85	0.16
	0.75	0.75	0.85	0.10	0.92	0.08	0.81	0.75	0.14	0.89	0.11	0.88	0.60	0.25	0.81	0.21
$\pi_{1,NN}^{(s)}$	0.33	0.33	1.00	0.06	0.06	0.00	0.50	1.00	0.06	0.05	0.00	0.67	0.95	0.05	0.05	0.00
	0.50	0.50	1.00	0.06	0.06	0.00	0.63	0.90	0.05	0.05	0.00	0.75	0.80	0.05	0.05	0.00
	0.67	0.67	0.90	0.08	0.08	0.00	0.75	0.85	0.06	0.06	0.01	0.84	0.70	0.05	0.05	0.01
	0.75	0.75	0.85	0.10	0.09	0.01	0.81	0.75	0.07	0.06	0.00	0.88	0.60	0.06	0.07	0.01
$\pi_{2,AA}^{(s)}$	0.33	0.33	1.00	0.79	0.79	0.58	0.50	1.00	0.79	0.80	0.59	0.67	0.95	0.80	0.80	0.60
	0.50	0.50	1.00	0.80	0.80	0.61	0.63	0.90	0.79	0.79	0.59	0.75	0.80	0.81	0.81	0.63
	0.67	0.67	0.90	0.80	0.81	0.61	0.75	0.85	0.81	0.82	0.64	0.84	0.70	0.83	0.83	0.67
	0.75	0.75	0.85	0.81	0.82	0.62	0.81	0.75	0.81	0.82	0.63	0.88	0.60	0.83	0.82	0.68
$\pi_{2,AN}^{(s)}$	0.33	0.33	1.00	0.42	0.06	0.02	0.50	1.00	0.42	0.06	0.02	0.67	0.95	0.39	0.06	0.02
	0.50	0.50	1.00	0.44	0.06	0.02	0.63	0.90	0.39	0.06	0.02	0.75	0.80	0.35	0.07	0.03
	0.67	0.67	0.90	0.45	0.08	0.02	0.75	0.85	0.40	0.08	0.03	0.84	0.70	0.34	0.10	0.05
	0.75	0.75	0.85	0.47	0.09	0.03	0.81	0.75	0.39	0.09	0.04	0.88	0.60	0.33	0.12	0.07
$\pi_{2,NA}^{(s)}$	0.33	0.33	1.00	0.06	0.43	0.02	0.50	1.00	0.06	0.43	0.02	0.67	0.95	0.06	0.42	0.02
	0.50	0.50	1.00	0.06	0.44	0.02	0.63	0.90	0.06	0.41	0.02	0.75	0.80	0.07	0.37	0.04
	0.67	0.67	0.90	0.08	0.46	0.03	0.75	0.85	0.08	0.42	0.04	0.84	0.70	0.10	0.37	0.06
	0.75	0.75	0.85	0.10	0.48	0.03	0.81	0.75	0.10	0.41	0.05	0.88	0.60	0.13	0.36	0.08

Abbreviations: %SS, fraction of standard sample size; AA, both alternative; AN, indication 2 null; NA, indication 1 null; NN, both null.

demonstrate global equivalence is ⩾0.58, which may be acceptable to stakeholders given the relative conservativeness of that sampling prior.

For the AN and NA scenarios, the proposed biologic does not meet the criteria of having equivalent efficacy for both indications and so the global null hypothesis rejection rate $r^{(s)}$ should be viewed as a type I error rate. The point-mass sampling priors for the AN and NA scenarios (ie, $\pi_{1,AN}^{(s)}$ or $\pi_{1,NA}^{(s)}$) help to characterize a worst-case global Bayesian type I error rate for this set of hypotheses. When there is no information borrowing ($\pi_0 = \pi_1$), the Bayesian type I error rate for the global equivalence hypothesis is estimated to be no greater than 0.08 even when $\pi_0 = 0.75$. Bayesian type I error rates for indication-specific hypotheses reach as high as 0.10 for the same analysis prior. In contrast, for a CPP inducing substantial information borrowing (ie, $\Delta_{RU} = 50$) coupled with relatively informative priors marginally, the Bayesian type I error rate for the global equivalence hypothesis is estimated to be as high as 0.21 (FL indication null) with indication-specific type I error rates as high as 0.25.

Fundamentally, the type I error rates above based on the $\pi_{1,AN}^{(s)}$ or $\pi_{1,NA}^{(s)}$ sampling priors should not be viewed as the only pertinent type I error rates with which to evaluate a design. Bayesian type I error rates based on the $\pi_{2,AN}^{(s)}$ and $\pi_{2,NA}^{(s)}$ sampling priors (as defined in Section 6.3) are designed to characterize type I error rates in more plausible null scenarios that reflect the relatedness of the treatment effects in the FL and RA indications. These priors take into account the fact that, if the proposed biologic has a null (eg, inferior) treatment effect in one indication, then its effect is probably worse than the reference biologic in the other, even

if the equivalence criterion is met for the second indication. This property is illustrated by the marginal alternative sampling prior for the FL indication shown in panel C of Figure 2. Not surprisingly, Bayesian type I error rates defined based on the $\pi^{(s)}_{2,AN}$ and $\pi^{(s)}_{2,NA}$ sampling priors are significantly lower and can be controlled at say, level 0.05 while still permitting meaningful sample size reductions of 25% to 30%.

The part of Table 2 corresponding to the $\pi^{(s)}_{1,NN}$ sampling prior characterizes the Bayesian type I error rates when the treatment effects for the proposed biologic are uniformly inferior to those for the reference biologic. One can see the global type I error rate is conservatively controlled regardless of the informativeness of the CPP chosen for analysis.

7.3 | Comparison of CPP to BHM

In this section, we compare CPP-based design performance to a design based on the BHM. For the BHM design, the same constraints were applied with regards to power and type I error control and the same range of possible sample sizes were considered. For the BHM, we implemented the naïve hierarchical prior $\gamma_j \sim \mathrm{N}(\gamma_j|\mu, SD = \sigma)$ for $j = 1, 2$, $\mu \sim \mathrm{N}(\mu|0, SD = 100)$, and $\sigma \sim \mathrm{Gamma}(\sigma|0.01, \mathrm{scale}=100)$. For the intercept parameters, we assumed $\alpha_j \sim \mathrm{N}(\alpha_j|0, SD = 100)$ for $j = 1, 2$. The chosen BHM prior results in a design with the operating characteristics as shown in Table 3, which are presented alongside selected CPP-based design results associated with comparable sample size reductions. Under the Bayesian type I error constraints, the BHM prior permitted 25% sample size reduction, had greater power than the CPP priors for the AA scenario, and worse type I error control and indication-specific power for the AN and NA scenarios. All designs perform well for the NN scenario.

Increasing power in the AA scenario necessarily comes at the price of more inflation of type I error rates in the AN and NA scenarios. However, the reduced indication-specific power for the BHM in the AN and NA scenarios relative to the CPP is a manifestation of how strongly the BHM encourages borrowing unless the hyperpriors are carefully calibrated.

8 | DISCUSSION

In this paper, we develop a strategy for a biosimilars clinical program that leverages an informative CPP to increase the efficiency of the trials conducted (ie, lowers their required sample sizes). The CPP is developed based on an elicited prior probability of treatment efficacy equivalence and a conditional probability of treatment efficacy equivalence for the set of indications to be investigated. As described in Appendix C of the Supplementary Materials, this elicitation strategy generates a covariance matrix where each treatment effect's prior standard deviation is a common multiple of its corresponding equivalence margin and results in a common prior correlation between all treatment effects. In the biosimilars setting, the hyperparameters in the CPP having appealing interpretations. Several obvious generalizations to the CPP are possible. If desired, one could elicit a distinct prior probability of treatment

TABLE 3 Operating characteristics for select CPP and BHM designs

						SP = 1			SP = 2		
True hypothesis	**Method**	π_0	π_1	Δ_{RU}	**%SS**	$\hat{r}^{(s)}_{FL}$	$\hat{r}^{(s)}_{RA}$	$\hat{r}^{(s)}$	$\hat{r}^{(s)}_{FL}$	$\hat{r}^{(s)}_{RA}$	$\hat{r}^{(s)}$
AA	CPP	0.67	0.84	50	0.70	0.92	0.92	0.82	0.83	0.83	0.67
	CPP	0.75	0.81	25	0.75	0.91	0.91	0.80	0.81	0.82	0.63
	BHM				0.75	0.95	0.93	0.88	0.86	0.83	0.74
AN	CPP	0.67	0.84	50	0.70	0.84	0.18	0.14	0.34	0.10	0.05
	CPP	0.75	0.81	25	0.75	0.89	0.12	0.10	0.39	0.09	0.04
	BHM				0.75	0.72	0.18	0.16	0.27	0.11	0.08
NA	CPP	0.67	0.84	50	0.70	0.19	0.85	0.16	0.10	0.37	0.06
	CPP	0.75	0.81	25	0.75	0.14	0.89	0.11	0.10	0.41	0.05
	BHM				0.75	0.20	0.73	0.18	0.11	0.31	0.09
NN	CPP	0.67	0.84	50	0.70	0.05	0.05	0.01			
	CPP	0.75	0.81	25	0.75	0.07	0.06	0.00			
	BHM				0.75	0.04	0.05	0.02			

Abbreviations: 1, point-mass; 2, nondegenerate; %SS, fraction of standard sample size; AA, both alternative; AN, indication 2 null; BHM, Bayesian hierarchical model; CPP, correlated parameter prior; NA, indication 1 null; NN, both null; SP, sampling prior.

efficacy equivalence for each of the indications (ie, π_{0j}, $j = 1, ..., J$). For this generalization of the CPP, there is no simple analog to the conditional probability π_1 and so direct elicitation of the correlation parameter (or matrix) for the J treatment effects would be required. An alternative generalization of the CPP would be to model the correlation parameter directly and give the parameter a prior in an attempt to let the observed data determine (or at least influence) how much information should be borrowed. Such a generalization could be termed a hierarchical CPP. However, we would advocate against using such a prior. Our opinion is based in part on the view that the hyperparameter π_1 is highly interpretable and therefore it may be reasonable to elicit by way of opinion from nonstatistician stakeholders. Perhaps more importantly, our numerical investigations suggest that, regardless of whether J is small or large, a hierarchical CPP will result in *increased* borrowing even for observed data where one would desire the opposite behavior (eg, when half the data suggest inferiority and the other half suggest equivalence). We discuss this more in Appendix E of the Supplementary Materials. In that appendix, we illustrate that this challenge is equally applicable to the CPP with a random correlation parameter and the BHM.

It is clear from Table 3 that the BHM tends to shrink estimates toward one another more strongly than the CPP for the chosen BHM hyperpriors. This phenomenon explains the comparative behavior of the designs in the AA and NN scenarios where both treatment effects are equal, and the AN and NA scenarios where they are not. For the design comparisons in Section 7.3, we chose a single BHM to evaluate against the CPP. Different hyperpriors would lead to different properties for a BHM-based design, some perhaps more desirable and comparable to properties obtained using the CPP (eg, using a BHM hyperprior for the standard deviation that suggests comparatively larger values would result in less borrowing leading to performance closer to that obtained by the CPP). Given that hierarchical priors have difficulty identifying how much information *should* be borrowed (see Appendix E of the Supplementary Materials) and the superior interpretability of the CPP hyperparameters (due to being direct statements about the hypotheses being tested), we argue for use of the CPP over the BHM or other meta-analytic priors.

If the scales of the treatment effects differ substantially (eg, $\delta_2 = 6$ for the normal endpoint in our example instead of $\delta_2 = 0.6$), the BHM's performance will degrade substantially unless some type of modification is applied (eg, rescaling the data or modifying the hyperprior). However, the CPP can be applied without any such modification. The problem of borrowing information when data are on different scales is more apparent in our example application based on $J = 3$ indications presented in Appendix B of the Supplementary Materials.

Characterizing type I error rates is particularly challenging when one evaluates multiple related hypotheses and wishes to make a global claim about treatment equivalence (ie, that H_1 is true). In situations where a subset of indications are null (should they occur), even when one takes the standard approach of assuming a boundary null effect for the null indications (ie, $\gamma_j = -\delta_j$), the type I error rate will be a function of the treatment effect in the other indications. Care must be taken when evaluating the performance of any prior to determine whether its use results in adequate type I error control in this complex setting. Constructing a reasonable sampling prior to distribution for nonnull treatment effects is critical. Our procedure for generating a sampling prior for the nonnull treatment effects provides a rational framework for doing this. Although considering a worst-case sampling prior for type I error evaluation (eg, the $\pi_{1,AN}^{(s)}$ sampling prior) is helpful for developing a comprehensive understanding of the benefits and risks of information borrowing, to fully benefit from the Bayesian approach, stakeholders must be willing to accept type I error control based on more likely null scenarios given prior knowledge about the similarity of the proposed and reference products.

ACKNOWLEDGMENTS

The authors wish to thank the associate editor and reviewers for helpful comments and suggestions, which have led to improvements of this article. This research was partially supported by NIH (grant no. GM 70335 and P01CA142538).

ORCID

Matthew A. Psioda http://orcid.org/0000-0002-4450-6981

REFERENCES

Barry, W.T., Perou, C.M., Marcom, P.K., Carey, L.A. and Ibrahim, J.G. (2015) The use of Bayesian hierarchical models for adaptive randomization in biomarker-driven phase II studies. *Journal of Biopharmaceutical Statistics*, 25(1), 66–88. https://doi.org/10.1080/10543406.2014.919933

Berry, S.M., Reitsma, D.J., Combest, A.J., Bryan, J.K., Adair, J.W., Healey, B.T. *et al.* (2011) A novel approach to rituximab biosimilar drug development. *Journal of Clinical Oncology*, 29, e13064.

Chen, M.-H., Ibrahim, J.G., Lam, P., Yu, A. and Zhang, Y. (2011) Bayesian design of non-inferiority trials for medical devices using historical data. *Biometrics*, 67, 1163–1170.

Chen, M.-H., Ibrahim, J.G., Zeng, D., Hu, K. and Jia, C. (2014) Bayesian design of superiority clinical trials for recurrent events data with applications to bleeding and transfusion events in myelodyplastic syndrome. *Biometrics*, 70, 1003–1013.

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

Coiffier, B., Sancho, J.-M., Jurczak, W., Kim, J.S., Nagarkar, R.V., Zhavrid, E. *et al.* (2016) Pharmacokinetic and safety of CT-P10, a biosimilar candidate to the rituximab reference product, in patients with newly diagnosed advanced stage follicular lymphoma (AFL). *Blood*, 128, 1807–1807.

Deeks, E.D. (2017) CT-P10 (truxima): a rituximab biosimilar. *BioDrugs*, 31, 275–278.

Haitao, P., Ying, Y. and Jielai, X. (2017) A calibrated power prior approach to borrow information from historical data with application to biosimilar clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66, 979–996.

Ibrahim, J.G., Chen, M.-H., Xia, H.A. and Liu, T. (2012) Bayesian meta-experimental design: evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. *Biometrics*, 68, 578–586.

Kim, W.S., Jurczak, W., Sancho, J.-M., Javrid, E., Kim, J.S., HernandezRivas, J.A. *et al.* (2017) Double-blind, randomized phase 3 study to compare efficacy and safety of the biosimilar CT-P10 to rituximab combined with CVP therapy in patients with previously untreated advanced-stage follicular lymphoma. *Journal of Clinical Oncology*, 35, 7532–7532.

Liu, R., Liu, Z., Ghadessi, M. and Vonk, R. (2017) Increasing the efficiency of oncology basket trials using a Bayesian approach. *Contemporary Clinical Trials*, 63, 67–72.

O'Hagan, A. and Stevens, J.W. (2001) Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21, 219–230.

Psioda, M.A. and Ibrahim, J.G. (2018) Bayesian design of a survival trial with a cured fraction using historical data. *Statistics in Medicine*, 37, 3814–3831.

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

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

Thall, P.F., Kyle, W.J., Nebiyou, B.B., 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.

U.S. Food and Drug Administration. (2015) Scientific considerations in demonstrating biosimilarity to a reference product: guidance for industry. U.S. Food and Drug Administration.

Wang, F. and Gelfand, A.E. (2002) A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17, 193–208.

Yoo, D.H., Suh, C.-H., Shim, S.C., Jeka, S., Cons-Molina, F.F., Hrycaj, P. *et al.* (2017) A multicentre randomised controlled trial to compare the pharmacokinetics, efficacy and safety of CT-P10 and innovator rituximab in patients with rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 76, 566–570.

SUPPORTING INFORMATION

Web Appendices referenced in Sections 2, 5, 6, 7, and 8 are available with this paper at the Biometrics website on Wiley Online Library. A GitHub repository contains the programs and other resources needed to reproduce the analyses presented in this paper (https://github.com/psioda/Bayes-Biosimilars). The software provided was written using R (R Core Team, 2016) version 3.3.1.

How to cite this article: Psioda MA, Hu K, Zhang Y, Pan J, Ibrahim JG. Bayesian design of biosimilars clinical programs involving multiple therapeutic indications. *Biometrics*. 2020;76:630–642. https://doi.org/10.1111/biom.13163