Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies

Chenguang Wang, Heng Li, Wei-Chen Chen, Nelson Lu, Ram Tiwari, Yunling Xu & Lilly Q. Yue

To cite this article: Chenguang Wang, Heng Li, Wei-Chen Chen, Nelson Lu, Ram Tiwari, Yunling Xu & Lilly Q. Yue (2019) Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies, Journal of Biopharmaceutical Statistics, 29:5, 731-748, DOI: 10.1080/10543406.2019.1657133

To link to this article: https://doi.org/10.1080/10543406.2019.1657133

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2019, VOL. 29, NO. 5, 731–748
https://doi.org/10.1080/10543406.2019.1657133

Taylor & Francis
Taylor & Francis Group

Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies

Chenguang Wang [a], Heng Li [b], Wei-Chen Chen [b], Nelson Lu [b], Ram Tiwari [b], Yunling Xu [b], and Lilly Q. Yue [b]

[a]Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD, USA; [b]Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, MD, USA

ABSTRACT

We are now at an amazing time for medical product development in drugs, biological products and medical devices. As a result of dramatic recent advances in biomedical science, information technology and engineering, ``big data'' from health care in the real-world have become available. Although big data may not necessarily be attuned to provide the preponderance of evidence to a clinical study, high-quality real-world data can be transformed into scientific evidence for regulatory and healthcare decision-making using proven analytical methods and techniques, such as propensity score methodology and Bayesian inference. In this paper, we extend the Bayesian power prior approach for a single-arm study (the current study) to leverage external real-world data. We use propensity score methodology to pre-select a subset of real-world data containing patients that are similar to those in the current study in terms of covariates, and to stratify the selected patients together with those in the current study into more homogeneous strata. The power prior approach is then applied in each stratum to obtain stratum-specific posterior distributions, which are combined to complete the Bayesian inference for the parameters of interest. We evaluate the performance of the proposed method as compared to that of the ordinary power prior approach by simulation and illustrate its implementation using a hypothetical example, based on our regulatory review experience.

ARTICLE HISTORY
Received 27 August 2018
Accepted 21 December 2018

KEYWORDS
Covariate balance; overlapping coefficient; power prior; propensity score; real-world data; real-world evidence

1. Introduction

In recent years, there is a growing interest in leveraging real-world data (RWD) in medical product development. RWD are data from such sources as electronic health records (EHRs), claims and billing data, product and disease registries, and data gathered through personal devices and health applications. Proper analysis of high-quality RWD can produce scientific evidence, called real-world evidence (RWE), which can then be utilized by stakeholders in public health to inform decision-making. In particular, the 21st Century Cures Act (U.S. House of Representatives 2015) requires an expanded role of RWE in the approval process of medical products. Thus, there is an emerging literature on analytical methods that can transform RWD into RWE and is readily implementable in regulatory settings. An example is the application of propensity score (PS) methodology in the pre-market confirmatory non-randomized studies for medical devices: Yue et al. (2016), (2014) and Li et al. (2016) have implemented a two-stage design in which appropriate blinding to outcome data (outcome-free design) during PS modeling is ensured to meet the regulatory requirements. Various methods of Bayesian inference can be applied in the regulatory settings as well (Bonangelino et al. 2011; Gamalo et al. 2014; Lin et al. 2016; Zhang et al. 2019; Zhao et al. 2014).

CONTACT Lilly Q. Yue lilly.yue@fda.hhs.gov Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, New Hampshire Avenue, Silver Spring, MD 10903, USA

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lbps.

In this paper, we develop an analytical method for leveraging RWD that combines Bayesian inference and PS methodology. To illustrate how our method works, consider a prospective study (to be referred to as the current study) that will provide evidence to support a new indication for an approved medical device. The device has already been used off-label for this indication in the real world, and data from the off-label use have been captured in a patient registry, forming a source of RWD (to be referred to as external data). It is determined that these external data are relevant to the current study, and it is appropriate for the current study to borrow from them. One way to apply Bayesian inference in this scenario is to construct an informative prior based on the external data using the power prior approach (Chen and Ibrahim 2000), in which the prior is constructed by raising the likelihood function of the external data to the power α, where $0 \leq \alpha \leq 1$. The power parameter α, sometimes referred to as discounting parameter, controls the amount of prior information contributed by the external data, with $\alpha = 0$ corresponding to non-informative prior (borrowing no external data) and $\alpha = 1$ corresponding to direct pooling of the external data with the current study data (borrowing all the external data). In practice, it is often desirable to fix in advance the maximum amount of information that the external data contributes, or the maximum number of patients to be borrowed, so that the prior does not dominate the posterior. For "big" external data (RWD tend to be "big"), this means choosing a very small α. Such a strategy, however, may be too simplistic. The external data may contain patients that are not very relevant to the current study, and it may not be desirable to include those patients in the power prior. The analytical method that we propose addresses this issue by taking the following preliminary steps to process the data before the construction of power prior. First, a propensity score model is built using the indicator of the current study versus the external data as dependent variable and baseline covariates as independent variables, and a propensity score is estimated for each patient. Then, we look at the range of propensity scores of the current study and exclude the patients in the external data whose propensity score is not in this range. We refer to this operation as ``trimming,'' which depicts the process of retaining only the patients in the external data that are more relevant to the current study. Finally, to make the patients in the power prior to be even more similar to those in the current study, data on the patients in the current study together with the external data after trimming are stratified into S propensity score strata. To conduct Bayesian inference, a power parameter is specified for each stratum that is derived from a similarity measure between PS distributions of the patients in the current study and the external data. To elicit the power parameters we can either treat them as constants or as random variables (and assign a prior distribution to them). In this paper, we look into both options. Note that the propensity score modeling, propensity score estimation, trimming, stratification, and the specification of the power parameters (or their prior distribution) constitutes the study design and should be completed with no outcome data in sight. After the design is complete and all the outcome data are available, a power prior is constructed in each stratum based on the pre-specified power parameter and the external data in that specific stratum. Stratum-specific posteriors are then combined to obtain the posterior distribution for the parameters of interest. We call the analytical method developed in this paper the PS-integrated power prior approach. As will be demonstrated, the PS-integrated power prior methodology can be implemented in a regulatory setting by applying the two-stage outcome-free design proposed by Yue et al. (2014), which is critical for ensuring the validity and integrity of the clinical study.

The rest of the paper is organized as follows. In Section 2, we review the PS and power prior methods and describe the details of the PS-integrated power prior approach. Simulation studies are conducted in Section 3 to evaluate the performance of the proposed method in different scenarios. As an illustration, the two-stage design of a single-group clinical study of a hypothetical cardiovascular medical product is described in Section 4, followed by an analysis of the outcome data. Finally, Section 5 is devoted to discussions.

2. Method

2.1. *Notation*

Let Y_i be a random variable that represents the outcome data from the ith patient, and y_i be the realization of Y_i. Let $\boldsymbol{X}_i$ with dimension $p \times 1$ denote the vector of covariates collected from the

same patient. Let $Z_i = 1$ if patient i is from the current study and 0 if patient i comes from external data. For simplicity, we assume that there is only one external data source and the covariates in $\boldsymbol{X}$ are collected in both the current study and the external study.

Let D_1 denote all the data, $\{(y_i, \boldsymbol{X}_i)\}$, collected from the current study and D_0 denote the data observed from patients in the external data that satisfy the inclusion and exclusion criteria of the current study. Let N_1 and N_0 denote the number of patients in D_1 and D_0, respectively. In the context of this article, where the external data are from RWD sources such as registries, it is expected that N_0 is sufficiently large for the proposed method in Section 2.4. Lastly, denote the parameters of interest as θ.

2.2. *Propensity score*

Formulated by Rosenbaum and Rubin (1983), the propensity score $e(\boldsymbol{X})$ for a patient with a vector $\boldsymbol{X}$ of observed baseline covariates is the conditional probability of receiving one treatment ($Z = 1$) rather than the other ($Z = 0$) given $\boldsymbol{X}$:

$$e(\boldsymbol{X}) = Pr(\mathrm{Z} = 1|\boldsymbol{X}).$$

The propensity score $e(\boldsymbol{X})$ is a balancing score in the sense that for patients with the same propensity score, the distribution of observed covariates is the same between the two treatment groups. If treatment assignment is strongly ignorable, that is, the treatment assignment Z and the potential outcome Y are conditionally independent given the observed covariates, $\boldsymbol{X}$, i.e., $\Pr(Z|\boldsymbol{X}, Y) = \Pr(Z|\boldsymbol{X})$, and if $0 < e(\boldsymbol{X}) < 1$ for all $\boldsymbol{X}$, then the average treatment effect on outcome at each value of the propensity score is an unbiased estimate of the true treatment effect at that value of propensity score (Rosenbaum and Rubin 1983). The above assumption requires that all covariates relevant to both treatment assignment and outcome are measured and captured in $\boldsymbol{X}$ (i.e., there is no hidden bias). In practice, the propensity score is estimated by modeling the probability of treatment group membership as a function of the observed covariates, typically via logistic regression. There are other flexible methods available for the propensity score estimation such as machine learning algorithms (Lee et al. 2010).

The propensity score methodology refers to a collection of versatile statistical tools based on the concept of propensity score for causal inference in observational studies, and can improve treatment comparison by adjusting for a relatively large number of potentially confounding covariates (Austin 2011; D'Agostino and Rubin 2000; Lunceford and Davidian 2004; Rosenbaum and Rubin 1983, 1984; Rubin 1997, 2001, 2007, 2008; Stuart 2010). Commonly used propensity score methods include propensity score matching, stratification (sub-classification) on the propensity score, inverse probability of treatment weighting using the propensity score, covariate adjustment using the propensity score, and some combinations of these approaches. The propensity score methods could be used to design and analyze an observational study, mimicking some of the characteristics of a randomized controlled trial (Rubin 2001, 2007, 2008). While the first three of the methods enable the separation of the design and the analysis, which is an appealing feature in clinical studies conducted in the regulatory settings, covariate adjustment using the propensity score does not enable such separation. We focus on the technique of stratification in this paper, which consists of grouping patients with similar propensity scores into strata.

2.3. *Power prior approach*

A power prior is an informative prior that takes the form:

$$\pi(\theta|D_0, \alpha) \propto [L(\theta|D_0)]^{\alpha}\pi_0(\theta) \tag{1}$$

where $L(\theta|D_0)$ is the likelihood of the external data, $\pi_0(\theta)$ is the initial prior distribution for θ, and $0 \leq \alpha \leq 1$ is the *power* parameter.

When $\alpha = 0$, the power prior reduces to the initial prior $\pi_0(\theta)$ and the external data D_0 does not contribute to the prior distribution of θ. On the other hand, when $\alpha = 1$, the power prior reduces to $\pi(\theta|D_0)$, the posterior distribution of $\theta|D_0$. In the latter case, D_0 contributes to the posterior distribution of θ in the same strength as the current data D_1, which is D_0's maximum strength. The simplicity of using a single parameter to control the impact of D_0 on the current study has made the power prior approach a widely applied Bayesian methodology for incorporating external information in various settings (Duan et al. 2005; Gamalo et al. 2014; Lin et al. 2016; Murray et al. 2014; Neelon and O'Malley 2010; Zhang et al. 2019; Zhao et al. 2014).

The power parameter α can be elicited as a fixed constant or viewed as a random variable with its hyperprior distribution (Chen and Ibrahim 2000; Duan et al. 2005; Hobbs et al. 2011).

2.4. PS-integrated power prior

In this section, we will introduce the PS-integrated power prior approach, or PS-power prior for short, for leveraging real-world data.

The rationale of the proposal is very simple. In Bayesian inference it is usually not desirable for the prior to be too strong, as we do not want the prior to dominate posterior, thereby risking too much bias. For this reason, in the power prior approach, if D_0 is much larger than D_1, which is often the case when D_0 is real-world data to be leveraged, one has the incentive to choose a very small power parameter α. However, such deep discounting may not solve the problem. The posterior would be sensitive to α because of the size of D_0, especially if the data generation mechanism for D_0 is different from that of D_1. This motivates us to consider an alternative approach, namely using propensity score to select a subset of D_0 more relevant to D_1 and to stratify this subset together with D_1 into propensity score strata, and then applying the power prior approach within each stratum.

The first step of the PS-integrated power prior approach is to use propensity score to select patients in D_0 (e.g., from a registry) that are more relevant to the patients in the current study D_1. We call this selection procedure *trimming*. The selected patients are then stratified on PS.

To apply the propensity score methodology described in Section 2.2, think of the patients constituting the external data and patients from the current study as two treatment groups, so that the propensity score can still be defined as the probability of receiving the investigational device in the current study for each patient in those studies. Recall that we let $Z_i = 1$ if patient i is in the current study and $Z_i = 0$ if patient i comes from the external data.

Let $\mathbb{E}_1$ denote $\{\widehat{e}(\boldsymbol{X}_i) : Z_i = 1\}$, the set of the estimated PSs for all the patients in the current study. Let

$$\dot{\mathbb{Q}} = \{\dot{q}_s : s = 0, \ldots, S, S \geq 1, \dot{q}_0 = 0, \dot{q}_S = 1, \dot{q}_s < \dot{q}_{s'} \forall s < s'\}$$

denote a set of distinct, ordered cut points in $[0, 1]$. Let the function $q(\dot{q}_s, \mathbb{E}_1)$ be the $100 \times \dot{q}_s$-th percentile of $\mathbb{E}_1$. For notational convenience, let $\widehat{q}_s$ denote $q(\dot{q}_s, \mathbb{E}_1)$, $\widehat{q}_0 = \min \mathbb{E}_1$, and $\widehat{q}_S = \max \mathbb{E}_1$. The hat symbol emphasizes that it is the quantiles of estimated propensity scores that are being denoted.

Trimming of the patients in D_0 is realized by excluding the patients with PSs out of the range of $[\widehat{q}_0, \widehat{q}_S]$. Note that patients in D_1 represents the population of interest which is not altered by *trimming*.

Next, we conduct a PS stratification based on $\{\widehat{q}_s\}$ as follows. Let $X_s = \{\boldsymbol{X} : \widehat{e}(\boldsymbol{X}) \in (\widehat{q}_{s-1}, \widehat{q}_s]\}$, $D_{s,z} = \{(y_i, \boldsymbol{X}_i) : \boldsymbol{X}_i \in X_s, Z_i = z\}$, and $\omega_{s,z} = \{i : \boldsymbol{X}_i \in X_s, Z_i = z\}$ denote the subset of patients in the current study ($z = 1$) or the external study ($z = 0$) with PS in $(\widehat{q}_{s-1}, \widehat{q}_s]$. Correspondingly, we formulate S strata with the s-th stratum, ω_s, constituted by the patients in $\{\omega_{s,0}, \omega_{s,1}\}$. Denote $\widetilde{D}_0 = \cup_{s=1}^{S} D_{s,0}$. Note that $\widetilde{D}_0 \subseteq D_0$ because of trimming. Figure 1 presents the *trimming* and stratification scheme.

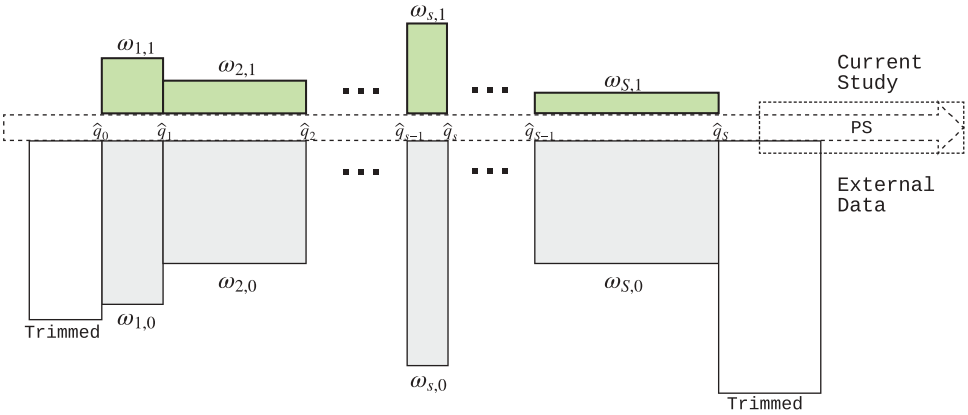

Figure 1. Propensity score trimming and stratification scheme. Area of the bars is proportional to the number of patients in the corresponding group.

Let $n_{s,z} = |\omega_{s,z}|$ be the number of patients in $\omega_{s,z}$. By definition, $n_{s,1} = N_1(\dot{q}_s - \dot{q}_{s-1})$. Note that it is assumed that $n_{s,0}$ is large enough for eliciting information of interest for all s. This assumption is considered plausible in the setting we are considering since size of RWD is typically large.

The second step of the approach is to conduct the statistical inference for the stratum-specific parameter of interest, θ_s. We propose to formulate a power prior for θ_s as follows:

$$\pi(\theta_s) \propto \left[L(\theta_s|D_{s,0})\right]^{\alpha_s}\pi_0(\theta_s) \tag{2}$$

where $0 \leq \alpha_s \leq 1$ for all s.

In the degenerate case of a single stratum, prior (2) becomes prior (1). That is,

$$\pi(\theta|\widetilde{D}_0, \alpha) \propto [L(\theta|\widetilde{D}_0)]^{\alpha}\pi_0(\theta). \tag{3}$$

Furthermore, just as the power parameter in (1), $\alpha_s = 0$ for all s corresponds to non-informative prior with no patients borrowed and $\alpha_s = 1$ for all s corresponds to direct pooling, respectively, within each stratum.

In the proposed PS stratification setting, since the analysis needs to simultaneously take into account the S-dimensional vector $\{\alpha_s\}$, it is more complex to set the values of $\{\alpha_s\}$. Hence, it is preferable to reduce the dimension of the power parameters $\{\alpha_s\}$. The following is one way to achieve this. We specify α_s as

$$\alpha_s = \min\left(\frac{Av_s}{n_{s,0}}, 1\right), \tag{4}$$

where A and v_s are design parameters. The parameter A, with a range of $[0, N_0]$, can be interpreted as the target number of patients the current study intends to borrow from the external data. To see this, note that the division by $n_{s,0}$ roughly standardizes the likelihood $L(\theta_s|D_{s,0})$ into the individual contribution to the likelihood. Thus, the value of A can be viewed as the overall number of patients to be borrowed from the external data, with Av_s being the target number of patients to be borrowed in the s-th stratum. The choice of A can be elicited as a fixed constant. For v_s, we put on the constraint $v_s \geq 0$ for all s and $\sum_{s=1}^{S} v_s = 1$. Furthermore, we let v_s, and consequently the amount of information contributed by $D_{s,0}$ to θ_s, be proportional to the similarity between the patients in $\omega_{s,0}$ and in $\omega_{s,1}$ with respect to the baseline covariates. As a measure of this similarity, we consider the overlapping coefficient (Inman and Bradley 1989) between the PS distributions in $\omega_{s,0}$ and $\omega_{s,1}$. Specifically, let $F_{s,z}$ and $f_{s,z}$ denote the distribution and the density function, respectively, of the

observed PS for patients in $\omega_{s,z}$. The similarity between the patients in $\omega_{s,0}$ and in $\omega_{s,1}$ is then measured by the overlapping coefficient between $f_{s,0}$ and $f_{s,1}$, defined as

$$r_s = \int_0^1 \min[f_{s,0}(e), f_{s,1}(e)]de. \tag{5}$$

The overlapping coefficient r_s can be interpreted as the overlapping area of the two density curves $f_{s,0}$ and $f_{s,1}$. Note that r_s can be equivalently defined as

$$r_s = \int_0^1 \min[1, \frac{d}{du} F_{s,0}(F_{s,1}^{-1}(u))]du.$$

Thus, it can also be interpreted as integration of the slope of the percentile-percentile curve between $F_{s,0}$ and $F_{s,1}$. In practice, we can estimate r_s based on a kernel estimation of $f_{s,0}$ and $f_{s,1}$, or a kernel estimation of $F_{s,0}(F_{s,1}^{-1}(u))$ with its corresponding derivative when such derivative exists. Note that there are alternative ways for measuring the similarity between the PS distributions in $\omega_{s,0}$ and $\omega_{s,1}$ (e.g., the Kolmogorov-Smirnov distance). In general, r_s could be any reasonable measures of the similarity of two distributions.

Two strategies can be used to specify the parameters v_s based on r_s, the *fixed-proportion* strategy and the *fully Bayesian* strategy. With the *fixed-proportion* strategy, we let $v_s = r_s / \sum_{s=1}^{S} r_s$. That is, A is split among the s strata in proportion to r_s, the similarity between the patients in $\omega_{s,0}$ and in $\omega_{s,1}$. With the *fully Bayesian* strategy, a prior distribution is assigned to $v_1, \ldots, v_S$. For example, we may assign $v_1, \ldots, v_S$ the Dirichlet prior

$$\pi_0(v_1, \ldots, v_S) = \text{Dir}(\frac{r_1}{R}, \ldots, \frac{r_s}{R}, \ldots, \frac{r_S}{R}), \tag{6}$$

through which the expected number of patients to be borrowed in stratum s is proportional *a priori* to r_s. The constant R controls the prior variances for v_s. In fact, the *fixed-proportion* strategy can be viewed as a special case of the Dirichlet prior with $R \to 0$.

With the *fixed-proportion* strategy, the resulting power prior is

$$\pi(\theta_1, \ldots, \theta_S) \propto \prod_{s=1}^{S} \left[L(\theta_s|D_{s,0})\right]^{\alpha_s} \pi_0(\theta_1, \ldots, \theta_S). \tag{7}$$

With the *fully Bayesian* strategy, we modify the prior (2) to address the likelihood principle concern following Duan et al. (2005). The resulting power prior is thus

$$\pi(\theta_1, \ldots, \theta_S, v_1, \ldots, v_S) = \frac{\prod_{s=1}^{S} \left[L(\theta_s|D_{s,0})\right]^{\alpha_s}}{\int \prod_{s=1}^{S} \left[L(\theta_s|D_{s,0})\right]^{\alpha_s} d\theta_1 \ldots d\theta_S} \pi_0(\theta_1, \ldots, \theta_S)\pi_0(v_1, \ldots, v_S). \tag{8}$$

To specify $\pi_0(\theta_1, \ldots, \theta_S)$, one may consider $\theta_1, \ldots, \theta_S$ *a priori* independent such that $\pi_0(\theta_1, \ldots, \theta_S) = \prod_{s=1}^{S} \pi(\theta_s)$. In the Appendix, we present the derivation of this proposed prior for normal and binary outcomes.

2.5. *Inference*

With the *fixed-proportion* strategy, given the prior (7), the posterior distribution of $\{\theta_s\}$ is as follows:

$$\pi(\theta_1, \ldots, \theta_S|D_1) \propto \prod_{s=1}^{S} [L(\theta_s|D_{s,1})]\pi(\theta_1, \ldots, \theta_S).$$

With the *fully Bayesian* strategy, given the prior (8), the posterior distribution of $\{\theta_s\}$ is as follows:

$$\pi(\theta_1,\ldots,\theta_S,\nu_1,\ldots,\nu_S|D_1)\propto\prod_{s=1}^{S}[L(\theta_s|D_{s,1})]\pi(\theta_1,\ldots,\theta_S,\nu_1,\ldots,\nu_S).$$

For both strategies, the posterior means can be obtained from the posterior distribution of θ_s and used in inference.

The parameter of interest in this paper, θ, is expressed as

$$\sum_{s=1}^{S}(\dot{q}_s-\dot{q}_{s-1})\theta_s. \tag{9}$$

For example, if the cut points in $\dot{\mathbb{Q}}$ are equally spaced with $\dot{q}_s=s/S$, the weighted average becomes $\frac{1}{S}\sum_{s=1}^{S}\theta_s$. Based on (9), the posterior distribution of θ can be readily derived from $\pi(\theta_1,\ldots,\theta_S,\nu_1,\ldots,\nu_S|D_1)$.

3. Simulation study

In this section, we conduct simulation studies to evaluate the proposed PS-power prior approach.

3.1. *Simulation settings*

In our simulation, the covariates $\boldsymbol{X}$ follow a mixture of multivariate normal distribution. That is, $\boldsymbol{X}_{p\times1}|Z=z\sim F_z$ and

$$F_z=\sum_{k=1}^{K}\psi_{zk}MVN(\mu_{zk},\Sigma_z)$$

with $\sum_{k=1}^{K}\psi_{zk}=1$, the diagonal of Σ_z being σ_z^2, and the off-diagonal elements of Σ_z being $\rho\sigma_z^2$.

We consider both continuous and binary outcomes. For continuous outcomes, we simulate Y_i from model

$$Y_i|\boldsymbol{X}_i,Z_i=\beta_0+\beta^T\boldsymbol{X}_i+\epsilon_i$$

where ϵ_i is the random error. For binary outcomes, we simulate $P(Y_i=1|\boldsymbol{X}_i,Z_i)$ from

$$\text{logit }P(Y_i=1|\boldsymbol{X}_i,Z_i)=\beta_0+\beta^T\boldsymbol{X}_i$$

and simulate Y_i from its binary distribution. Note that the covariate effects on the outcome (i.e., β) are assumed to be independent of Z.

For all the simulation studies, we set $\rho=0.1$. We convert $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_4$ to binary covariates using cut point 0. For continuous outcomes, we set $\beta=\mathbf{1}_{p\times1}$ and $\beta_0=0$. For binary outcomes, we set $\beta=\mathbf{1}_{p\times1}$ and choose β_0 such that $E(Y|Z=1)=0.4$. Note that we assume β_0 and β are identical for $z=0$ and 1. The rationale is that covariate effects on the outcome are constant across studies. For continuous outcomes, we assume $\epsilon_i\sim N(0,1)$. Lastly, we set the number of patients in the external data to be 3000.

Two major simulation scenarios are contemplated. In Scenario I, we assume the covariates in the current study and the external data have different distributions, which reflects the situation where patients in the current study may not be a subgroup of the patients in the external data. In this scenario, we set $K=1$, $\mu_{01}=\mathbf{1.2}_{p\times1}$, $\mu_{11}=\mathbf{1}_{p\times1}$. Furthermore, we set $\sigma_1^2=1$ and $\sigma_0^2=1.5$. In Scenario II, we assume the patients in the external data have a mixture distribution. For $z=1$, we set $K=1$, $\mu_{11}=\mathbf{1}_{p\times1}$, and $\sigma_1^2=1$. For $z=0$, we set $K=2$, $\psi_{01}=\psi_{02}=0.5$, $\mu_{01}=\mathbf{1}_{p\times1}$, $\mu_{02}=\mathbf{1.5}_{p\times1}$, and $\sigma_0^2=1$. For each scenario, we consider the dimension of $\boldsymbol{X}$ to be $p=10$ and $p=15$, sample sizes to be $N_1=100$ and 200, and set A to be 20 and 40 for $N_1=200$, and 40 and 80 for $N_1=400$.

The parameter of interest is θ given by $\frac{1}{S}\sum_{s=1}^{S}\theta_s$. In the simulation studies, we compare three strategies for PS-power prior. The first strategy is a degenerate case where the number of strata

$S = 1$, and the prior $\pi(\theta|\widetilde{D}_0, \alpha)$ is constructed following (3). Note $\widetilde{D}_0$ is the subset of D_0 after PS trimming. We choose $\alpha = A/N_0$. The second and third strategies are the *fixed-proportion* and *fully Bayesian* PS-power prior approach, respectively, as has been described in Section 2.4. For these two PS-power prior strategies, we choose $S = 5$ (Cochran 1968; Yue 2007) and stratify the patients into 5 PS strata with equal number of patients from $\omega_{s,1}$ for all s. Moreover, we assume that variances are stratum-specific and independent of each other *a priori* for continuous outcomes.

3.2. Results

In order to evaluate the bias-variance trade-off with the different strategies, we report for each strategy the estimate of θ, bias, mean squared error (MSE), the width and the coverage rate of the 95% credible interval. The simulation results are based on 10,000 replications.

In Table 1, we report the simulation study results for continuous outcomes. Several observations can be made. First, the *fully Bayesian* strategy (Strategy 3) and the *fixed-proportion* strategy (Strategy 2) outperform Strategy 1 ($S = 1$, i.e. no stratification) in terms of bias in all the cases. In Scenario I, the Strategies 2 and 3 are comparable to Strategy 1 in terms of MSE, whereas Strategies 2 and 3 generate smaller MSEs in Scenario II. These results clearly show the advantage of stratification compared to no stratification. Between Strategies 2 and 3, the *fully Bayesian* strategy tends to result in smaller biases with slightly larger MSEs than the *fixed-proportion* strategy. Furthermore, as the dimension of $\boldsymbol{X}$ increases from 10 to 15, we observe that all three strategies have increased biases and MSEs. On the other hand, as the sample size of the current study N_1 increases from 200 to 400, all three approaches tend to have smaller biases and MSEs. However, this bias and MSE reduction is much less for Strategy 1 (without stratification) compared to Strategies 2 and 3 (with stratification), which seems to suggest that Strategies 2 and 3 are more sensitive to the increase of the sample size in the current study. Moreover, as A increases, we observe larger biases and MSEs for all three strategies. Lastly, we observe that biases and MSEs in Scenario II, where we assume patients in the external data were from a mixture distribution, are larger than Scenario I in general. This seems to suggest that the mixture distribution setting is more challenging in terms of identifying comparable patients for the current study. It is worth noting that the variances (not reported) in Strategies 2 and 3 are in general larger than Strategy 1. The implication is that stratifying patients into different strata does introduce additional variation into the estimates. However, this is less of an issue when N_1 is large.

In Table 2, we report the simulation study results for binary outcomes. What we observe in the binary outcome case is almost the same as in the continuous outcome case. The results also show an advantage, with respect to both bias and MSE, of applying stratification compared to no stratification. The *fully Bayesian* strategy also results in smaller biases with slightly larger MSEs compared to the *fixed-proportion* strategy. Furthermore, all three strategies have larger bias and MSE the larger the dimension of $\boldsymbol{X}$.

4. An illustrative example

This straw man example is a simple illustration of how the PS-integrated power prior approach could be implemented when leveraging RWD/RWE.

4.1. The first-stage design

A single-arm clinical study was proposed to demonstrate the safety and effectiveness of a cardiovascular device, with a plan to leverage real-world data obtained from a patient registry.

The hypothesis test associated with the binary clinical outcome variable was

$$H_0 : \theta \geq 36\% \quad vs. \quad H_a : \theta < 36\%$$

Table 1. Simulation study results for continuous outcomes. No Stratification corresponds to the power prior without stratification strategy. Fully Bayesian and Fixed-Proportion correspond to the *fully Bayesian* and the *fixed-proportion* strategies, respectively. N_0^*: the average number of patients in the external data after trimming.

		True						Bias	MSE	95% CI	
Scenario	p	θ	N_1	N_0^*	A	Strategy	Mean	(× 100)	(× 100)	Width	Coverage
I	10	9.37	200	2893	20	No Stratification	9.41	4.44	5.46	0.93	0.95
						Fully Bayesian	9.37	0.40	5.67	0.88	0.93
						Fixed-Proportion	9.37	0.82	5.62	0.87	0.93
					40	No Stratification	9.45	8.84	5.64	0.91	0.94
						Fully Bayesian	9.37	0.66	5.71	0.86	0.92
						Fixed-Proportion	9.38	1.90	5.56	0.84	0.92
I	10	9.37	400	2925	40	No Stratification	9.41	4.49	3.00	0.66	0.94
						Fully Bayesian	9.36	0.02	3.06	0.62	0.92
						Fixed-Proportion	9.37	0.67	3.00	0.61	0.92
					80	No Stratification	9.45	8.81	3.36	0.64	0.92
						Fully Bayesian	9.36	−0.24	3.11	0.61	0.91
						Fixed-Proportion	9.38	1.50	2.97	0.60	0.91
I	15	14.37	200	2874	20	No Stratification	14.45	8.49	12.52	1.37	0.95
						Fully Bayesian	14.38	0.86	12.72	1.27	0.92
						Fixed-Proportion	14.38	1.59	12.64	1.26	0.92
					40	No Stratification	14.53	16.64	13.70	1.34	0.93
						Fully Bayesian	14.38	1.21	12.78	1.25	0.92
						Fixed-Proportion	14.40	3.35	12.59	1.23	0.91
I	15	14.37	400	2918	40	No Stratification	14.45	8.38	6.59	0.97	0.94
						Fully Bayesian	14.37	−0.17	6.40	0.90	0.92
						Fixed-Proportion	14.38	1.02	6.32	0.89	0.92
					80	No Stratification	14.53	16.48	8.17	0.95	0.90
						Fully Bayesian	14.36	−0.75	6.49	0.89	0.92
						Fixed-Proportion	14.39	2.36	6.31	0.86	0.91
II	10	9.37	200	2926	20	No Stratification	9.50	13.11	6.62	0.91	0.92
						Fully Bayesian	9.38	1.26	5.66	0.56	0.76
						Fixed-Proportion	9.38	1.39	5.64	0.55	0.76
					40	No Stratification	9.60	24.04	10.03	0.88	0.83
						Fully Bayesian	9.38	1.94	5.59	0.54	0.75
						Fixed-Proportion	9.39	2.32	5.56	0.53	0.74
II	10	9.37	400	2951	40	No Stratification	9.49	12.67	4.12	0.65	0.89
						Fully Bayesian	9.37	0.65	2.89	0.35	0.70
						Fixed-Proportion	9.37	0.82	2.88	0.35	0.69
					80	No Stratification	9.60	23.70	7.79	0.62	0.69
						Fully Bayesian	9.38	1.23	2.86	0.34	0.68
						Fixed-Proportion	9.38	1.74	2.84	0.33	0.67
II	15	14.37	200	2906	20	No Stratification	14.57	20.31	15.62	1.35	0.91
						Fully Bayesian	14.38	1.17	13.13	0.83	0.74
						Fixed-Proportion	14.38	1.42	13.10	0.83	0.74
					40	No Stratification	14.75	38.04	24.56	1.31	0.80
						Fully Bayesian	14.39	2.14	13.02	0.81	0.73
						Fixed-Proportion	14.40	2.91	12.98	0.79	0.72
II	15	14.37	400	2936	40	No Stratification	14.58	21.51	10.15	0.96	0.87
						Fully Bayesian	14.39	2.05	6.37	0.51	0.69
						Fixed-Proportion	14.39	2.44	6.36	0.51	0.69
					80	No Stratification	14.76	39.48	20.43	0.93	0.62
						Fully Bayesian	14.39	2.79	6.36	0.50	0.68
						Fixed-Proportion	14.40	3.95	6.37	0.48	0.67

where θ is the one-year adverse event rate. At the first design stage, a total of 17 baseline covariates that may affect the clinical outcome were identified based on prior knowledge: age, weight, gender, stroke, left ventricular ejection fraction (LVEF), smoking, creatinine, sodium, diabetes, myocardial infarction, diuretic, chronic obstructive pulmonary disease (COPD), dependent edema, cardiothoracic ratio, systolic and diastolic blood pressure, and heart rate. An appropriate registry was selected, and all these key covariates and the outcome information had already been planned to be collected in the registry. The study was planned to be based on Bayesian inference. The study success criterion is that the posterior probability of θ being less than 0.36 is greater than 0.95.

Table 2. Simulation study results for binary outcomes. No Stratification corresponds to the power prior without stratification strategy. Fully Bayesian and Fixed-Proportion correspond to the *fully Bayesian* and the *fixed-proportion* strategies, respectively. Note that the average number of patients in the external data after trimming is the same as in the continuous case reported in Table 1.

						Bias	MSE	95% CI	
Scenario	p	N_1	A	Strategy	Mean	(× 100)	(× 100)	Width	Coverage
I	10	200	20	No Stratification	0.41	0.99	0.11	0.13	0.95
				Fully Bayesian	0.41	0.63	0.11	0.12	0.93
				Fixed-Proportion	0.41	0.72	0.11	0.12	0.93
			40	No Stratification	0.42	1.76	0.12	0.12	0.93
				Fully Bayesian	0.41	0.74	0.10	0.12	0.93
				Fixed-Proportion	0.41	0.95	0.10	0.11	0.92
I	10	400	40	No Stratification	0.41	0.95	0.06	0.09	0.94
				Fully Bayesian	0.40	0.37	0.05	0.09	0.93
				Fixed-Proportion	0.40	0.49	0.05	0.08	0.93
			80	No Stratification	0.42	1.73	0.07	0.09	0.90
				Fully Bayesian	0.40	0.44	0.05	0.08	0.93
				Fixed-Proportion	0.41	0.72	0.05	0.08	0.92
I	15	200	20	No Stratification	0.41	1.20	0.12	0.13	0.94
				Fully Bayesian	0.41	0.62	0.11	0.12	0.93
				Fixed-Proportion	0.41	0.70	0.11	0.12	0.92
			40	No Stratification	0.42	2.18	0.13	0.12	0.91
				Fully Bayesian	0.41	0.74	0.11	0.11	0.92
				Fixed-Proportion	0.41	0.94	0.10	0.11	0.91
I	15	400	40	No Stratification	0.41	1.23	0.06	0.09	0.93
				Fully Bayesian	0.40	0.42	0.05	0.08	0.93
				Fixed-Proportion	0.41	0.53	0.05	0.08	0.93
			80	No Stratification	0.42	2.21	0.09	0.09	0.86
				Fully Bayesian	0.40	0.50	0.05	0.08	0.93
				Fixed-Proportion	0.41	0.76	0.05	0.08	0.92
II	10	200	20	No Stratification	0.42	1.61	0.12	0.13	0.93
				Fully Bayesian	0.41	0.53	0.10	0.10	0.88
				Fixed-Proportion	0.41	0.54	0.10	0.10	0.88
			40	No Stratification	0.43	2.84	0.16	0.12	0.87
				Fully Bayesian	0.41	0.54	0.10	0.10	0.87
				Fixed-Proportion	0.41	0.57	0.10	0.09	0.87
II	10	400	40	No Stratification	0.42	1.53	0.07	0.09	0.91
				Fully Bayesian	0.40	0.24	0.05	0.07	0.86
				Fixed-Proportion	0.40	0.25	0.05	0.07	0.86
			80	No Stratification	0.43	2.80	0.12	0.09	0.78
				Fully Bayesian	0.40	0.27	0.05	0.07	0.85
				Fixed-Proportion	0.40	0.28	0.05	0.06	0.85
II	15	200	20	No Stratification	0.42	1.78	0.13	0.13	0.93
				Fully Bayesian	0.40	0.46	0.10	0.10	0.87
				Fixed-Proportion	0.40	0.48	0.10	0.10	0.87
			40	No Stratification	0.43	3.22	0.19	0.12	0.85
				Fully Bayesian	0.40	0.48	0.10	0.09	0.87
				Fixed-Proportion	0.41	0.51	0.10	0.09	0.86
II	15	400	40	No Stratification	0.42	1.81	0.08	0.09	0.90
				Fully Bayesian	0.40	0.26	0.05	0.06	0.84
				Fixed-Proportion	0.40	0.27	0.05	0.06	0.84
			80	No Stratification	0.43	3.29	0.15	0.09	0.71
				Fully Bayesian	0.40	0.29	0.05	0.06	0.83
				Fixed-Proportion	0.40	0.31	0.05	0.06	0.83

In the sample size determination, assuming $\theta = 0.30$ and a significance level of 0.05, a power of 80% would require a sample size of approximately 380. It was proposed to enroll 290 patients in the current investigational study and borrow 90 patients from the registry. Thus, the value of A was set at 90. In practice, the number of patients to be borrowed is determined mainly based on clinical judgment. It should be noted that this number will affect key design parameters. However, what is acceptable is decided on a case by case basis with clinical input.

4.2. *The second-stage design*

The second-stage design was started when the enrollment of all patients was completed, and all covariate data had been collected, cleaned, and locked. Based on the pre-specified inclusion/exclusion criteria of the current investigational study, a total of 987 patients from the registry were identified as potential patients to be borrowed.

Based on the 290 patients in the current study and 987 external patients, propensity scores were estimated using logistic regression with all 17 baseline covariates included in their linear terms. Missing data in the covariates were imputed using the classification and regression trees (CART) method implemented in the R package mice. The mice package creates multiple imputations for multivariate missing data. Each incomplete variable is imputed by a separate model conditioning on all the other variables. With the CART option, a classification or regression tree is fit to the data for each incomplete variable at each iteration. Then, each missing value is located in a terminal node according to the fitted tree. A random draw among the members in the node is made and the observed value from that draw is considered as the imputation of the missing value. More details can be found in Buuren and Groothuis-Oudshoorn (2011). Note that we consider a single imputation in this example for simplicity. In practice, however, global sensitivity analysis should be conducted to address this thorny missing data issue (National Research Council 2010).

After trimming, 941 out of 987 patients were retained for the study design and outcome analysis. Five propensity score strata were formed for all the patients (290 + 941) in such a way that each stratum contains equal number of patients in the current study. Overlapping coefficient of the propensity score distributions was then calculated for each propensity score stratum (Figure 2). The balance of covariate distributions between the current study and the external data source in each PS stratum was examined for each covariate (Figure 3). As we can see, those distributions are well balanced.

Note that at the end of the second stage, the following features are determined: the PS stratum each patient would belong to, how many external patients will be ``borrowed'' for each stratum, as well as how much information each external patient would contribute (i.e. α_s) (Table 3).

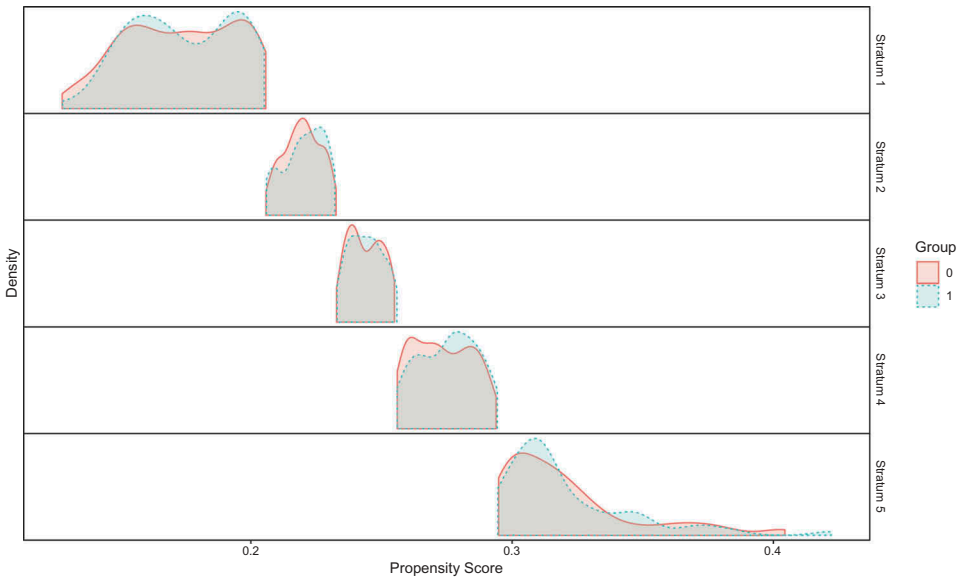

Figure 2. Densities of PS for the registry study (Group = 0) and the current study (Group = 1).

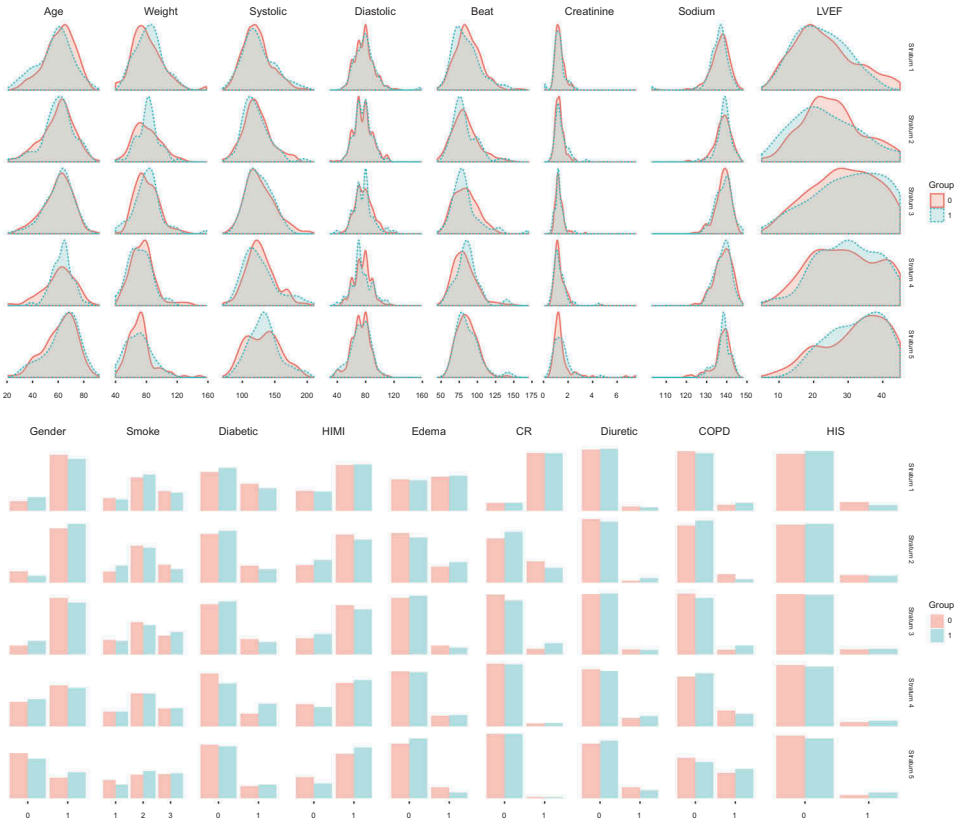

Figure 3. Balance checking of covariates between the registry study (Group = 0) and the current study (Group = 1).

Table 3. Number of patients, the overlapping coefficient, and the fixed power parameters a_s in each PS stratum.

		Stratum					
		$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	Total
Current Study	$n_{s,1}$	58	58	58	58	58	290
Registry	$n_{s,0}$	281	210	154	187	109	941
Overlapping Coefficient (OC)	r_s	0.87	0.78	0.86	0.84	0.77	
Standardized OC	$r_s / \sum r_s$	21%	19%	21%	19%	19%	100%
Approx. patients borrowed	$Ar_s / \sum r_s$	19	17	19	18	17	90
Power parameter	a_s	0.07	0.08	0.12	0.10	0.15	

4.3. *Outcome analysis*

After the clinical outcome was observed from all the patients, the final analysis was conducted.

Figure 4 presents the point estimates for the parameter of interest, θ, on the current study and the external data source for each stratum. Overall there were no large discrepancies between those two point estimates in most strata.

The posterior distribution of θ and θ_s for all s based on the analysis using $A = 90$ is presented in Figure 5. The corresponding posterior means and 95% credible intervals are reported in Table 4. In Figure 6, we compare the posterior distribution of θ based on the analysis using $A = 90$ and the posterior distribution of θ based on the current study only, with the fitting Jeffreys prior

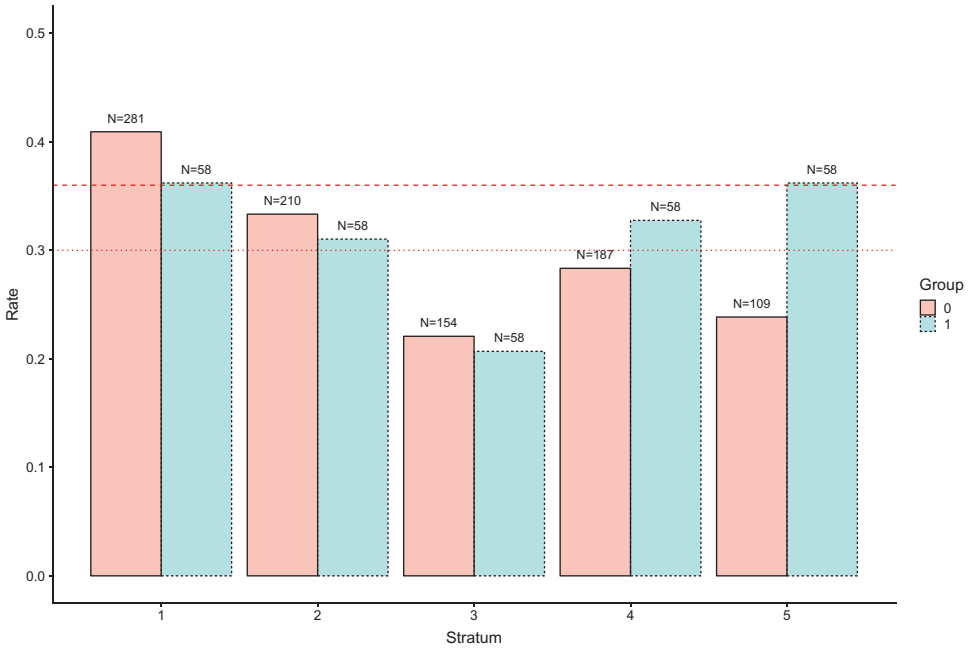

Figure 4. Stratum-Specific observed clinical outcome in the current study and the registry study. The reference horizontal lines correspond to 0.30 and 0.36, respectively.

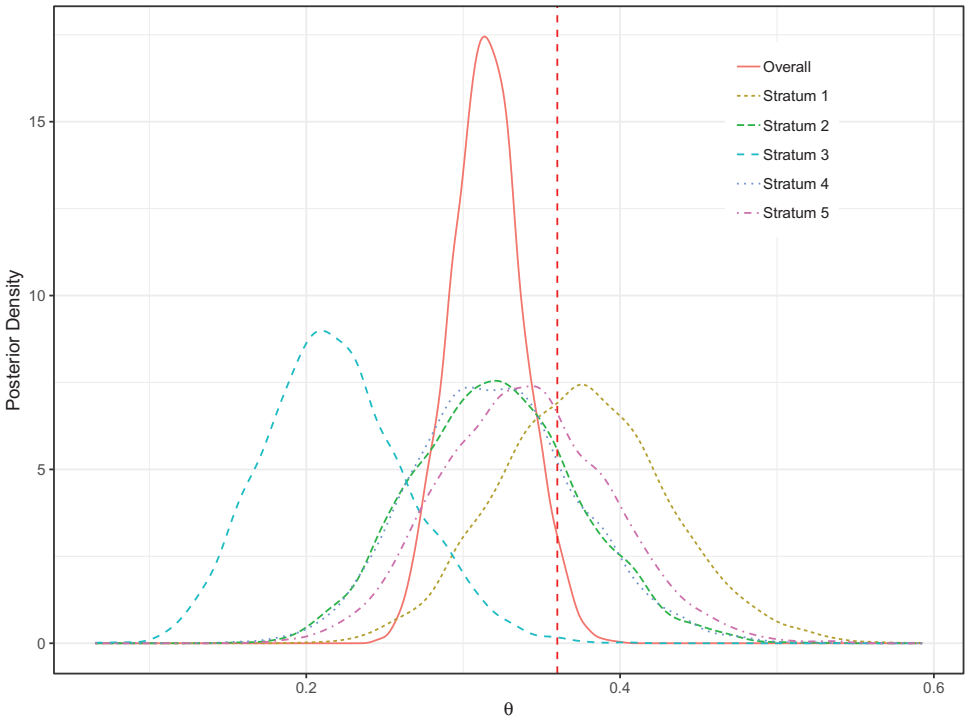

Figure 5. Posterior density of θ (Overall) and θ_s (Stratum s) for all strata. The vertical reference line corresponds to 0.36.

Table 4. Posterior mean and 95% credible interval for θ and θ_s for all s.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ
Posterior mean	37.6%	32.1%	21.8%	32.1%	34.0%	31.5%
95% credible interval (lower bound)	27.3%	22.3%	13.5%	22.5%	24.1%	27.1%
95% credible interval (upper bound)	48.7%	42.7%	31.2%	43.0%	45.0%	36.2%

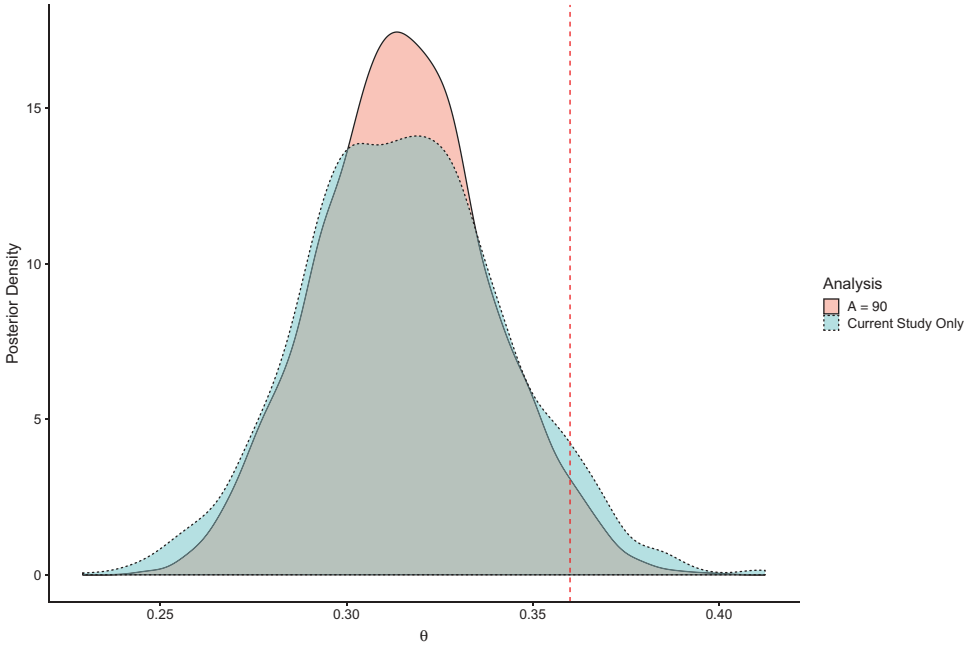

Figure 6. Posterior density of θ based on the analysis with $A = 90$ and the current study only. The vertical reference line corresponds to 0.36.

$Beta(0.5, 0.5)$. It is observed that there is a reduction in the posterior variance of θ by borrowing information via the *fixed-proportion* PS-power prior approach. Based on the analysis, the posterior mean of θ is 31.5% with 95% quantile 35.4%. The posterior probability of $\theta < 36\%$ is 96.9%, which meets the study success criterion.

Based on the results from the analysis, we conclude that the hypothetical medical device meets the study success criterion.

5. Discussion

We developed an analytical method integrating the propensity score methodology with the Bayesian methodology of power prior to designing and analyzing single-arm clinical studies. The goal of this PS-integrated power prior approach is to allow the regulatory decision-making to leverage the real-world evidence derived from a real-world data source. The real-world data source, or external data, considered in this paper is a well-designed and well-executed registry.

Our proposed method extends the basic idea of power prior by first using the propensity score (PS) methodology to pre-select relevant patients in the registry to be used in the construction of power prior for the current study. We then stratify patients using PS and specify a power prior within each stratum. We proposed to choose each stratum-specific power parameter based on the similarity of the registry patients and the current study patients measured by the overlap of PS

distributions of those two groups of patients for that stratum. Two strategies for specifying the power parameters were investigated: the *fixed-proportion* strategy and the *fully Bayesian* strategy. Simulations are conducted for both the *fixed-proportion* strategy and the *fully Bayesian* strategy, and compared to no stratification. The simulation results clearly show an advantage of stratification in terms of bias and MSE. As to the *fully Bayesian* and the *fixed-proportion* strategies, the former tends to result in smaller biases with slightly larger MSEs. It should be noted that with the *fixed-proportion* strategy the amount of borrowing from the external data is known *a priori*, whereas with the *fully Bayesian* strategy the amount of borrowing depends on data in the current study. In regulatory settings it is often desirable to fix how much information is borrowed from the external data *a priori*, therefore the *fixed-proportion* approach may be preferable.

Note that the purpose of this paper is to propose the PS-integrated power prior approach and compare its performance to that of the ordinary power prior approach. The simulations show that the PS-integrated power prior approach clearly has an advantage. The effect of the amount borrowed is an interesting issue in its own right, meriting an extensive and in-depth discussion. But it is a separate issue that is beyond the scope of this paper and would be better addressed in a separate paper.

It is important to point out that, in a regulatory setting, using PS to bolster the leveraging of external information does not need any access to the outcome data at the design stage; thereby allowing the design to maintain the objectivity. The *outcome-free design principle* is a fundamental principle for any pre-market studies seeking regulatory approval (Rubin 2008; Yue et al. 2016, 2014).

We adopt the power prior approach as the Bayesian method for the current study to borrow information from RWD. It is worth mentioning that Bayesian hierarchical modeling is another, maybe even more popular, approach to leveraging information from external sources. When there is only one prior study, however, Bayesian hierarchical models are known to have difficulty in estimating study-to-study variation (Gelman 2006), which motivates us to consider the alternative power prior approach.

There are several important extensions we are considering for the next steps. First, the current approach considers only a single source of real-world data. In practice, there may be cases where there exists prior information from multiple registries or even sources other than registries. Second, the proposed method considers single-arm clinical studies in which the parameter of interest is tested against a fixed number. Extension is required for the method to be applied when the current study is a randomized clinical trial. Third, in this paper, we use PS to pre-select relevant patients from a real-world data source, to stratify patients, and to measure similarity between patients in the external data and current study within each propensity score stratum. It is worthwhile to explore the performance of other methods that also do the above in the same settings (Diamond and Sekhon 2013; Imai and Ratkovic 2014; Li et al. 2018). Finally, although this paper deals with Bayesian inference, the machinery developed herein can be integrated with frequentist approaches to leveraging external real-world data as well.

6. Implementation and software

We have implemented the proposed PS-power prior to normal and binary outcomes in STAN (Stan Development Team 2018). Source code of the STAN models is available on request from the authors. The authors are developing an R package to fully implement the proposed Bayesian approach and a user-friendly software to facilitate designing and analyzing medical device single-group clinical studies by leveraging RWE.

Acknowledgments

The first author (CW) was a consultant for the Food and Drug Administration on this project and was compensated for his consultation services. The authors are grateful to the referees for their great suggestions and comments which helped to improve the article.

ORCID

Chenguang Wang http://orcid.org/0000-0002-7085-3303

References

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46 (3):399–424. doi:10.1080/00273171.2011.568786.

Bonangelino, P., T. Irony, S. Liang, X. Li, V. Mukhi, S. Ruan, Y. Xu, X. Yang, and C. Wang. 2011. Bayesian approaches in medical device clinical trials: A discussion with examples in the regulatory setting. *Journal of Biopharmaceutical Statistics* 21 (5):938–953. doi:10.1080/10543406.2011.589650.

Buuren, S., and K. Groothuis-Oudshoorn. 2011. Mice: Multivariate imputation by chained equations in rR. Journal of Statistical Groothuis-Oudshoorn 1–6845 (3): 1–67. 45 doi:10.18637/jss.v045.i03

Chen, M.-H., and J. G. Ibrahim. 2000. Power prior distributions for regression models. *Statistical Science* 15 (1):46–60. doi:10.1214/ss/1009212673.

Cochran, W. G. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 295–313. doi:10.2307/2528036.

D'Agostino, R. B., Jr, and D. B. Rubin. 2000. Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* 95 (451):749–759. doi:10.1080/01621459.2000.10474263.

Diamond, A., and J. S. Sekhon. 2013. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 95 (3):932–945. doi:10.1162/REST_a_00318.

Duan, Y., K. Ye, and E. P. Smith. 2005. Evaluating water quality using power priors to incorporate historical information. *Environmetrics* 17 (1):95–106. doi:10.1002/env.752.

Gamalo, M. A., R. C. Tiwari, and L. M. LaVange. 2014. Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. *Pharmaceutical Statistics* 13 (1):25–40. doi:10.1002/pst.1588.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis* 1(3):515–534. doi:10.1214/06-BA117A.

Hobbs, B. P., B. P. Carlin, S. J. Mandrekar, and D. J. Sargent. 2011. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* 67 (3):1047–1056. doi:10.1111/biom.2011.67.issue-3.

Imai, K., and M. Ratkovic. 2014. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (statistical Methodology)* 76 (1):243–263. doi:10.1111/rssb.12027.

Inman, H. F., and E. L. Bradley Jr. 1989. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-Theory and Methods* 18 (10):3851–3874. doi:10.1080/03610928908830127.

Lee, B. K., J. Lessler, and E. A. Stuart. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine* 29 (3):337–346.

Li, F., K. L. Morgan, and A. M. Zaslavsky. 2018. Balancing covariates via *Balancing covariates via propensity score weighting*. Journal of the American Statistical Association 1–11113 (521): 390–400. 113 1-400 doi:10.1080/01621459.2016.1260466

Li, H., V. Mukhi, N. Lu, Y.-L. Xu, and L. Q. Yue. 2016. A note on good practice of objective propensity score design for premarket nonrandomized medical device studies with an example. *Statistics in Biopharmaceutical Research* 8 (3):282–286. doi:10.1080/19466315.2016.1148071.

Lin, J., M. Gamalo-Siebers, and R. Tiwari. 2016. Non-inferiority and networks: Inferring efficacy from a web of data. *Pharmaceutical Statistics* 15 (1):54–67. doi:10.1002/pst.v15.1.

Lunceford, J. K., and M. Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23 (19):2937–2960. doi:10.1002/sim.v23:19.

Murray, T. A., B. P. Hobbs, T. C. Lystig, and B. P. Carlin. 2014. Semiparametric bayesian commensurate survival model for post-market medical device surveillance with non- exchangeable historical data. *Biometrics* 70 (1):185–191. doi:10.1111/biom.v70.1.

National Research Council. 2010. *The prevention and treatment of missing data in clinical trials*. Washington, DC: The National Academies Press. doi:10.17226/12955.

Neelon, B., and A. J. O'Malley. 2010. Bayesian analysis using power priors with application to pediatric quality of care. *Journal of Biometrics and Biostatistics* 1 (1):1–9. doi:10.4172/2155-6180.1000103.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1):41–55. doi:10.1093/biomet/70.1.41.

Rosenbaum, P. R., and D. B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79 (387):516–524. doi:10.1080/01621459.1984.10478078.

Rubin, D. B. 1997. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 127 (8_Part_2):757–763. doi:10.7326/0003-4819-127-3-199708010-00022.

Rubin, D. B. 2001. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2 (3–4):169–188. doi:10.1023/A:1020363010465.

Rubin, D. B. 2007. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* 26 (1):20–36. doi:10.1002/(ISSN)1097-0258.

Rubin, D. B. 2008. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* 808–840. doi:10.1214/08-AOAS187.

Stan Development Team. 2018. Stan modeling language users guide and reference manual, Version 2.18.0. http://mc-stan.org.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics* 25 (1):1. doi:10.1214/09-STS313.

U.S. House of Representatives (2015). 21st century cures act. http://docs.house.gov/meetings/IF/IF00/20150519/103516/BILLS-1146ih.pdf.

Yue, L. Q. 2007. Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. *Journal of Biopharmaceutical Statistics* 17 (1):1–13. doi:10.1080/10543400601044691.

Yue, L. Q., G. Campbell, N. Lu, Y. Xu, and B. Zuckerman. 2016. Utilizing national and international registries to enhance pre-market medical device regulatory evaluation. *Journal of Biopharmaceutical Statistics* 26 (6):1136–1145. doi:10.1080/10543406.2016.1226336.

Yue, L. Q., N. Lu, and Y. Xu. 2014. Designing premarket observational comparative studies using existing data as controls: Challenges and opportunities. *Journal of Biopharmaceutical Statistics* 24 (5):994–1010. doi:10.1080/10543406.2014.926367.

Zhang, J., C.-W. Ko, L. Nie, Y. Chen, and R. Tiwari. 2019. Bayesian hierarchical methods for meta-analysis combining randomized-controlled and single-arm studies. Statistical Methods in Medical Ko 28 (5):1293–1310. 0962280218754928. doi:10.1177/0962280218754928

Zhao, Y., J. Zalkikar, R. C. Tiwari, and L. M. LaVange. 2014. A bayesian approach for benefit-risk assessment. *Statistics in Biopharmaceutical Research* 6 (4):326–337. doi:10.1080/19466315.2014.965845.

Appendix A. Appendix

A.1. Derivation of the PS-power prior for normal and binary cases

Let Y_i denote the outcome of patient i ($i = 1, \ldots, N_0$) in an existing dataset.

First, consider $Y_i \sim N(\theta_s, \sigma_s^2)$ in stratum s with σ_s^2 known. In practice, we set σ_s to be the standard deviation of Y_i in stratum s.

The likelihood in stratum s is given by

$$L(\theta_s|D_{s,0}) = \left(\frac{1}{\sqrt{2\pi\sigma_s^2}}\right)^{n_{s,0}} \exp\left[-\frac{\sum_{i=1}^{n_{s,0}}(Y_i-\theta_s)^2}{2\sigma_s^2}\right].$$

Note that

$$\begin{aligned}\int [L(\theta_s|D_{s,0})]^{\alpha_s} d\theta_s &= \left(\frac{1}{\sqrt{2\pi\sigma_s^2}}\right)^{\alpha_s n_{s,0}} \exp\left[-\frac{\sum_{i=1}^{n_{s,0}}(Y_i-\bar{Y}_s)^2}{2\frac{\sigma_s^2}{\alpha_s}}\right] \int \exp\left[-\frac{(\theta_s-\bar{Y}_s)^2}{2\frac{\sigma_s^2}{\alpha_s n_{s,0}}}\right] d\theta_s \\ &= \left(\frac{1}{\sqrt{2\pi\sigma_s^2}}\right)^{\alpha_s n_{s,0}} \sqrt{\frac{2\pi\sigma_s^2}{\alpha_s n_{s,0}}} \exp\left[-\frac{\sum_{i=1}^{n_{s,0}}(Y_i-\bar{Y}_s)^2}{2\frac{\sigma_s^2}{\alpha_s}}\right].\end{aligned}$$

Thus,

$$\frac{[L(\theta_s|D_{s,0})]^{\alpha_s}}{\int [L(\theta_s|D_{s,0})]^{\alpha_s} d\theta_s} = \frac{1}{\sqrt{2\pi\frac{\sigma_s^2}{\alpha_s n_{s,0}}}} \exp\left[-\frac{(\theta_s-\bar{Y}_s)^2}{2\frac{\sigma_s^2}{\alpha_s n_{s,0}}}\right] = \phi\left(\frac{\theta_s-\bar{Y}_s}{\frac{\sigma_s}{\sqrt{\alpha_s n_{s,0}}}}\right).$$

Assign $\pi(\theta_s) \propto 1$. Then,

$$\pi(\theta_1, \ldots, \theta_S, \nu_1, \ldots, \nu_S) \propto \prod_s \phi\left(\frac{\theta_s-\bar{Y}_s}{\frac{\sigma_s}{\sqrt{\alpha_s n_{s,0}}}}\right) \nu_s^{\frac{r_s}{R}}.$$

Next, consider $Y_i \sim \text{Bern}(\theta_s)$ in stratum s. The likelihood in stratum s is given by

$$L(\theta_s|D_{s,0}) = \theta_s^{n_{s,0}\bar{Y}_s}(1-\theta_s)^{n_{s,0}-n_{s,0}\bar{Y}_s}.$$

Assign $\pi(\theta_s) = \text{Beta}(1,1)$. Then,

$$\int L(\theta_s|D_{s,0})^{\alpha_s} d\theta_s = \int \theta_s^{\alpha_s n_{s,0}\bar{Y}_s}(1-\theta_s)^{\alpha_s n_{s,0}-\alpha_s n_{s,0}\bar{Y}_s} d\theta_s = B(\alpha_s n_{s,0}\bar{Y}_s + 1, \alpha_s n_{s,0} - \alpha_s n_{s,0}\bar{Y}_s + 1)$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Thus,

$$\pi(\theta_1, \ldots, \theta_S, \nu_1, \ldots, \nu_S) \propto \prod_s \text{Beta}(\alpha_s n_{s,0}\bar{Y}_s + 1, \alpha_s n_{s,0} - \alpha_s n_{s,0}\bar{Y}_s + 1)\nu_s^{\frac{r_s}{R}}.$$