JOURNAL OF BIOPHARMACEUTICAL STATISTICS
https://doi.org/10.1080/10543406.2021.1998098

Propensity score-integrated power prior approach for augmenting the control arm of a randomized controlled trial by incorporating multiple external data sources

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

[a]Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, USA; [b]Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland, USA; [c]18426 Polynesian Lane, Boyds, Maryland 20841

ABSTRACT

In this paper, a propensity score-integrated power prior approach is developed to augment the control arm of a two-arm randomized controlled trial (RCT) with subjects from multiple external data sources such as real-world data (RWD) and historical clinical studies containing subject-level outcomes and covariates. The propensity scores for the subjects in the external data sources versus the subjects in the RCT are first estimated, and then subjects are placed in different strata based on their estimated propensity scores. Within each propensity score stratum, a power prior is formulated with the information contributed by the external data sources, and Bayesian inference on the treatment effect is obtained. The proposed approach is implemented under the two-stage study design framework utilizing the outcome-free principle to ensure the integrity of a study. An illustrative example is provided to demonstrate the implementation of the proposed approach.

ARTICLE HISTORY
Received 21 April 2021
Accepted 5 October 2021

KEYWORDS
Real-world data; real-world evidence; propensity score; power prior; multiple external data sources; outcome-free design

1. Introduction

There has been great interest in leveraging external data, which include real-world data (RWD) and historical clinical studies, in medical product development. RWD refer to data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources such as electronic health records, claims and billing data, and product and disease registries (U.S. Food and Drug Administration 2017). Real-world evidence (RWE) is the clinical evidence regarding the usage and potential benefits or risks of a medical product derived from analysis of RWD. Oftentimes, similar to RWD, data from clinical studies conducted in the past or in other regions may be leveraged to generate evidence. In practice, external data could be utilized in various ways in regulatory applications.

Some statistical methods and procedures that are readily implementable in a regulatory setting for leveraging RWD can be found in the literature. For Yue et al. (2014, 2016), Li et al. (2016), and Lu et al. (2020) discuss some considerations and good practice in a regulatory setting when using propensity score methodology. Particularly, they stress the importance of the utilization of a two-stage design to ensure that the outcome data are blinded during the design stage so that the validity of the study can be maintained. Levenson et al. (2021) provide some statistical considerations when using RWD and RWE for regulatory purposes. Bayesian analytical methods can be naturally and conceptually adapted when

CONTACT Lilly Q. Yue Lilly.Yue@fda.hhs.gov Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, Maryland, 20993, USA

#Chenguang Wang was a consultant for the Food and Drug Administration on this project and was compensated for his consultation services.

information is planned to be borrowed from external data. As pointed out by Gottlieb (2018), innovative statistical approaches such as propensity scores and Bayesian methods can combine information from different sources and potentially reduce the size and duration of clinical trials. Indeed, more innovative methods and procedures are needed and may be developed by utilizing statistical approaches such as the ones mentioned above to handle various emerging regulatory application scenarios in which RWD are leveraged.

The statistical procedure described in this paper is for the scenario that external data such as RWD are leveraged to augment the control arm of a randomized controlled trial (RCT). This scenario was previously discussed in papers such as Xu et al. (2020), Lu et al. (2019), Lin et al. (2018a, 2018b), and Chen et al. (2020). The need to augment the control arm of an RCT is often due to limited number of available patients and low enrollment rate. In a 1:1 RCT that aims to compare the performance of an investigational medical device with an active control, the enrollment rate may be unacceptably low if clinical equipoise is questionable, i.e., patients and physicians believe that the investigational device is superior to the control. Sometimes the enrollment may be more difficult when several trials of similar products are simultaneously conducted resulting the trials to compete for potential subjects, which is not an uncommon occurrence in medical device studies. To speed up the enrollment, increasing the randomization ratio in favor of the investigational device may enhance potential subjects' willingness to participate in a trial. To compensate the sample size in the control group and/or to utilize available information on the control, external data sources may be considered to augment the control arm. This approach helps streamline clinical evidence generation and speeds up medical product development and regulatory decision-making so that new medical technologies can be available to patients faster.

It is possible that the control arm can be augmented from multiple external data sources, including historical clinical studies and registries. The sample size of a historical clinical study tends to be smaller, and the data quality is typically better, than those of a registry. Furthermore, the historical clinical study data may be years, or decades old. In comparison, a registry database often consists of a huge number of more contemporary patients whose data are collected under a real-world practice setting. Regardless of the type of an external data source, it is possible to leverage data from a data source when this data source is relevant to the questions under consideration and sufficiently reliable.

In order to leverage external data in our scenario, two important issues that need to be considered are described below. The first consideration is the similarity of the characteristics of patients between the current RCT and external data sources. The clinical outcomes are often associated with patient characteristics, so, when the distributions of patient characteristics among data sources are dissimilar, direct pooling of all data sources (i.e. ignoring differences in patient characteristics between them) to make a statistical inference is inappropriate. One strategy to address this issue is to use the propensity scores methodology. For example, in Wang et al. (2019), Wang et al. (2020) and Chen et al. (2020), the propensity score strata are constructed such that the inference is made within each propensity score stratum in which the distribution of baseline characteristics for the patients from the external data sources is similar to the distribution for those from the current study. The same strategy is applied in the method proposed in this paper.

The second important issue is the determination of amount of information that each external data source is to contribute to the statistical inference for the parameter of interest. Such a decision is largely based on clinical and regulatory considerations, depending on the data quality, relevance, and reliability (U.S. Food and Drug Administration 2017), data quantity, and time of data collection. In instances where one wants to avoid that the external data source(s) dominates the inference, it may be desirable to "down-weight" or "discount" the external data sources so that they do not dominate final study results, based on clinical and regulatory considerations. In Chen et al. (2020), data from a single external data source are discounted by applying the composite likelihood method under the frequentist framework. In this paper, the Bayesian power prior approach is adopted to down-weight data from multiple external data sources.

Our proposed propensity score-integrated power prior approach can be implemented under the "two-stage study design" framework first introduced by Yue et al. (2014). As the two-stage design has become a standard practice (Yue et al. 2016) when observational studies are used to support regulatory decisions for medical devices, our proposed approach can readily be implemented in the regulatory setting.

The rest of the paper is organized as follows. In Section 2, a brief review of the propensity score methodology and Bayesian power prior is provided, then the propensity score-incorporated power prior approach are described. An illustrative example is presented in Section 3. Some discussion points are provided in Section 4.

2. Method

The method presented in this section can be adopted in an application when leveraging external data is fit for purpose in addressing the specific objectives from a regulatory perspective and when the data quality of each external data source is considered to be adequate based on clinical and regulatory judgments. The number of nominal subjects to be borrowed from each external data source needs to be determined from clinical evaluations and regulatory considerations.

2.1. Notation

Suppose that there are J external data sources. Let n be the total number of subjects from all data sources. For subject i $(i = 1, \ldots n)$, let y_i be the observed response; the vector X_i be the covariates; Z_i be the data source that the subject is from; T_i be the treatment that the subject receives; R_i be the indicator whether the subject is in the current RCT. Specifically, $Z_i = 0$ if subject i is in the current study, and $Z_i = j$ $(j = 1, \ldots, J)$ if subject i is in the j^{th} external data source. $T_i = 0$ if subject i receives the control, and $T_i = 1$ if subject i receives the investigational device. $R_i = 0$ if subject i is in the current RCT (i.e. $Z_i = 0$), and $R_i = 1$ if subject i is not in the current RCT (i.e. $Z_i = j, j = 1, \ldots, J$).

For $j = 0, 1, \ldots, J$, let $S_j = \{i : Z_i = j\}$ denote the set of subjects who are from the j^{th} data source. Set S_0 is further divided into two sets based on the treatment type: $S_0^{(0)} = \{i : Z_i = 0, T_i = 0\}$, denoting the set of RCT subjects who receive the control, and $S_0^{(1)} = \{i : Z_i = 0, T_i = 1\}$, denoting the set of RCT subjects who receive the investigational device. The nominal number of subjects intended to be borrowed from the j external data source is denoted as A_j.

Let $D_{\text{cur}} = \{(y_i, X_i, T_i, Z_i) : i \in \{S_0\}\}$ denote the collection of data from the current RCT, and $D_{\text{ext}} = \{(y_i, X_i, T_i, Z_i) : i \in \{S_1, \ldots, S_J\}\}$ denote the set of subjects who are from the external data sources. Lastly, let the parameter of interest be denoted as θ.

2.2. Propensity score methodology

Formulated by Rosenbaum and Rubin (1983), the propensity score (PS), $e(X)$, for a subject with a vector X of observed baseline covariates in a comparative study is the conditional probability of being in one treatment group $(T = 1)$ rather than the other $(T = 0)$ given X:

$$e(X) = \Pr(T = 1|X)$$

Under the strongly ignorable treatment assignment assumption, the propensity score $e(X)$ is a balancing score in the sense that, for subjects with the same propensity score, the distribution of observed covariates is the same between the two treatment groups. PS can be viewed as a scalar function of X to summarize the information required to balance the distribution of X. Conditioning on PS, the estimate of average treatment effects is unbiased. In practice, propensity scores are often 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; Lin et al. 2018a, 2018b).

Originally developed for causal inference in observational studies to 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), the propensity score methodology refers to a collection of versatile statistical tools based on the concept of propensity score, which include propensity score matching and stratification (subclassification). Those propensity score methods could be used to design and analyze an observational study so that it assumes some of the characteristics of a randomized controlled trial (Rubin 2001, 2007, 2008).

In this paper, the propensity score methodology is utilized for the purpose of leveraging external data sources to augment the control arm of the current RCT. The objective is to balance the covariates between the current RCT and external data sources to make the leveraging more justified. Therefore, the propensity score is defined as the conditional probability of being in the current RCT ($R = 0$) rather than the external data sources given X:

$$e(X) = \Pr(R = 0|X)$$

Technique of stratification, which consists of grouping subjects with similar propensity scores into strata, is adopted in this paper.

Due to the balancing property of propensity score, within each stratum subjects from the external data sources are expected to be more similar to those in the current study in terms of baseline covariates than they are overall, which makes leveraging external data within strata more justifiable. In this context, the propensity score methodology also allows us to separate study design from the analysis of outcome data.

2.3. Using power prior to discount external data

From the clinical and regulatory perspectives, it is often desirable to downweight the information from the external data sources. The statistical tool used in this paper to achieve such

discounting is the method of power prior, first introduced by Ibrahim and Chen (2000).

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

$$\pi(\theta|D_{\text{ext}}, \alpha) = [L(\theta|D_{\text{ext}})]^{\alpha}\pi_0(\theta)$$

where $L(\theta|D_{\text{ext}})$ is the likelihood function of the external data D_{ext}, $\pi_0(\theta)$ is the initial prior distribution for θ, and $0 \le \alpha \le 1$ is the power parameter. When $\alpha = 0$, the power prior reduces to the initial prior $\pi_0(\theta)$, and the external data D_{ext} does not contribute to the prior distribution of θ. On the other hand, when $\alpha = 1$, the power prior reduces to $L(\theta|D_{\text{ext}})$, the posterior distribution of $\theta|D_{\text{ext}}$. In the latter case, each individual in D_{ext} contributes to the posterior distribution of θ in the same strength as each individual in the current data D_{cur}. The simplicity of using a single parameter, α, to control the impact of D_{ext} on the current study has made the power prior approach a widely applied Bayesian methodology for incorporating external information in various settings.

2.4. PS-integrated power prior

In this section, a description is provided regarding the PS-integrated power prior approach for leveraging multiple external data sources to augment the control arm of the current RCT. In this approach, propensity score methodology serves for the design purpose, placing subjects from all data sources into several PS strata such that the subjects within each stratum are more similar. The stratum-specific power prior is then formulated to discount the external data, and then outcome analysis is conducted. The approach can be implemented using two-stage design paradigm proposed by Yue et al. (2014), which follows the principle that study design and data analysis are separated. In the design stage, no outcome data are accessed.

Once all baseline covariates of all data sources are available, the propensity scores can be estimated. The propensity score is defined as the probability of a subject coming from the current study as opposed to external data sources:

$$e_i = \Pr(R_i = 0|X_i) \tag{1}$$

With current RCT being in one group and the pooled external data another group, a logistic regression model, or any other method such as one referred to in Section 2.2, may be used to obtain the propensity score estimates, $\hat{e}_i$.

A step, referred to as *trimming*, is taken to for the purpose of selecting external subjects that are more similar to the RCT subjects. External subjects whose $\hat{e}_i$, where $i \in \{S_1, \ldots, S_J\}$, fall outside the range of all $\hat{e}_i$ of subjects in the RCT (i.e. $i \in S_0$) are excluded from the analysis set, as they are not similar to the RCT subjects. The rationale for trimming is based on the notion that the distribution of the subject characteristics in the target population is best represented by the RCT subjects.

Next, propensity score strata can be formed so that patients of the external data sources and the current RCT in each stratum are comparable with respect to their baseline covariate distributions. Let $0 < \hat{q}_0 < \hat{q}_1 < \ldots < \hat{q}_K < 1$ be a set of cut points based on which K propensity score strata are formed for S_0. In this paper, these cut points are selected such that the number of RCT patients in each stratum are roughly equal. Let $w_{k,j} = \left\{i : \hat{e}_i \in \left(\hat{q}_{k-1}, \hat{q}_k\right], Z_i = j, j = 0, \ldots, J\right\}$, the index set for patients from j^{th} data source in the k stratum. In the current study, let $w_{k,0}^{(0)} = \left\{i : \hat{e}_i \in \left(\hat{q}_{k-1}, \hat{q}_k\right], Z_i = 0, T_i = 0\right\}$, the index set for current RCT patients receiving the control in the k stratum, and $w_{k,0}^{(1)} = \left\{i : \hat{e}_i \in \left(\hat{q}_{k-1}, \hat{q}_k\right], Z_i = 0, T_i = 1\right\}$, the index set for current RCT patients receiving the investigational device in the k^{th} stratum. The number of patients within each index set is denoted as $N_{k,j} = \left|w_{k,j}\right|$, $N_{k,0}^{(0)} = \left|w_{k,0}^{(0)}\right|$, and $N_{k,0}^{(1)} = \left|w_{k,0}^{(1)}\right|$. Consequently, $N_{.,j} \equiv \sum_k N_{k,j}$ and $N_{.,.} \equiv \sum_j N_{.,j}$. Note that the covariate distributions in each stratum between the RCT and pooled external data need to be demonstrated to be balanced, otherwise the propensity scores may need to be reestimated. Figure 1 presents the trimming and stratification strategy.

Let $\theta_k^{(0)}$ and $\theta_k^{(1)}$ denote the parameters of interest in stratum k for patients receiving the control and the investigational device, respectively. The stratum-specific power prior of $\theta_k^{(0)}$ is expressed as

$$\pi\left(\theta_k^{(0)}|D_{\text{ext}}\right) = \prod_{j=1}^{J}\left[\prod_{i \in w_{k,j}} f\left(y_i; \theta_k^{(0)}\right)\right]^{\alpha_{k,j}} \pi_0\left(\theta_k^{(0)}\right) \tag{2}$$

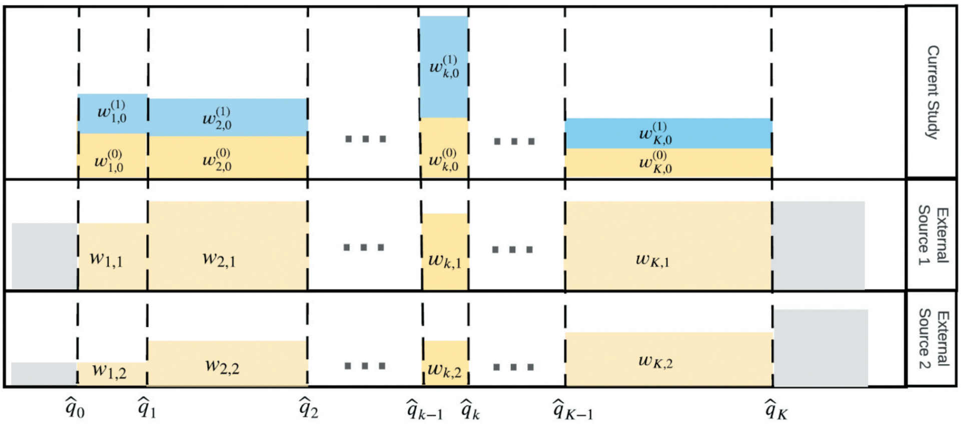

Figure 1. Trimming and stratification based on propensity scores with two external data sources as an example.

where $0 \leq \alpha_{k,j} \leq 1$, for all $j = 1, \ldots, J$. The power parameters are proposed to be

$$\alpha_{k,j} = \lambda_{k,j} / N_{k,j}$$

where $\lambda_{k,j}$ can be interpreted as the nominal number of patients to be borrowed in the k^{th} stratum from the j^{th} external data source. Specifically, $\lambda_{k,j}$ is proposed to be

$$\lambda_{k,j} = \min\left(\frac{A_j r_{k,j}}{\sum_{k'=1}^{K} r_{k',j}}, N_{k,j} \right) \tag{3}$$

Note that $\lambda_{k,j}$ is proportional to $r_{k,j}$ but is not exceeding $N_{k,j}$. The $r_{k,j}$'s are measures of similarity between patients in $w_{k,j}$ and $w_{k,0}$ with respect to their baseline covariates. Recall that A_j represents the nominal number of patients intended to be borrowed from the j^{th} external data source. One possibility for specifying $r_{k,j}$ is to use the overlapping area of PS density curves $g_{k,0}$ in $w_{k,0}$ and $g_{k,j}$ in $w_{k,j}$ $(j = 1, \ldots, J)$ (Inman and Bradley 1989):

$$r_{k,j} = \int_{0}^{1} \min\left[g_{k,0}(u), g_{k,j}(u) \right] \mathrm{d}u \tag{4}$$

The overall actual nominal number of patients to be borrowed from the j^{th} external data source is denoted as $\Lambda_j = \sum\limits_{k=1}^{K} \lambda_{k,j}$. Note that $\Lambda_j = \mathrm{A}_j$ when $N_{k,j} \geq A_j r_{k,j} / \sum\limits_{k'=1}^{K} r_{k',j}$ for all k.

2.5. Inference

Within stratum k, the posterior distribution $\theta_k^{(1)}$ is

$$\pi\left(\theta_k^{(1)}\right) = \prod_{i \in w_{k,0}^{(1)}} f\left(y_i; \theta_k^{(1)}\right) \pi_0\left(\theta_k^{(1)}\right)$$

where $\pi_0\left(\theta_k^{(1)}\right)$ is the prior which is usually considered to be non-informative; and the posterior distribution $\theta_k^{(0)}$ is

$$\pi\left(\theta_k^{(0)} | D_{\text{cur}}\right) = \prod_{i \in w_{k,0}^{(0)}} f\left(y_i; \theta_k^{(0)}\right) \pi\left(\theta_k^{(0)} | D_{\text{ext}}\right)$$

where $\pi\left(\theta_k^{(0)} | D_{\text{ext}}\right)$ is the power prior as specified in Equation (2).
The treatment effect of interest is denoted as μ, which is the weighted average of $\theta_k^{(1)} - \theta_k^{(0)}$:

$$\mu = \frac{1}{K} \sum_{k=1}^{K} \left(\theta_k^{(1)} - \theta_k^{(0)} \right). \tag{5}$$

The posterior distribution of μ can be readily derived from the posterior distribution of $\left(\theta_k^{(1)}, \theta_k^{(0)}\right)$'s.

3. An illustrative example

In this section, an example is presented to illustrate the proposed approach, utilizing two-stage design (Yue et al. 2014). An investigational device was newly developed to treat a coronary artery disease (CAD), which had been commonly treated with an approved medical device, considered as an active control. The clinical data of the active control had been captured in some databases. To demonstrate that the clinical performance of the investigational device was superior to the active control, the plan was to conduct a randomized clinical trial in which the control arm was to be augmented by two identified external data sources: a registry database and a historical clinical study. There were a total of 1053 registry patients and 350 historical clinical study patients who met the inclusion/exclusion criteria set for the current study. These two data sources, containing patient-level clinical outcomes and baseline covariates data, were thought to have good data quality. For the historical clinical study, while the data quality was considered very good, the size of 350 was thought to be relatively small, and the time lag between this study and the planned RCT was relatively large. In contrast, the registry contained data from a larger number of patients who were more contemporary. From the clinical and regulatory perspectives, the nominal number of patients to be leveraged from external data could be increased in this case when incorporating both external data sources as opposed to only one of them.

3.1. The first design stage

The primary endpoint was the occurrence of any specified adverse events at one year. The null and alternative hypotheses were

$$H_0: \mu \geq 0 \text{ vs. } H_a: \mu < 0$$

where μ, the difference in adverse event rate between the investigational device and the control, is expressed as $\theta^{(1)} - \theta^{(0)}$, where $\theta^{(1)}$ and $\theta^{(0)}$ are the one-year adverse event rates for patients treated with the investigational device and the control, respectively. The null hypothesis was to be rejected if the posterior probability of $\mu < 0$ is greater than 0.975. In order to determine the sample size, the true $\theta^{(0)}$ and $\theta^{(1)}$ were assumed to be 0.192 and 0.12, respectively. A sample size of approximately 400 per group was required to achieve a power of 80% with 1:1 randomization ratio. Based on clinical and regulatory considerations, it was proposed to borrow the information equivalent to 200 control subjects from the two external data sources, the registry and the historical clinical study. Therefore, the current randomized controlled trial would enroll 600 subjects that were to be 2:1 randomized to the investigational device arm and the active control arm. Furthermore, it was planned that (1) 130 nominal number of subjects were to be leveraged from the registry and (2) 70 nominal number of subjects were to be leveraged from the historical clinical study. By indexing j as C, R, and H for current RCT, registry, and historical clinical study, respectively, $A_R = 130$ and $A_H = 70$.

A total of 10 baseline covariates that likely affect the clinical outcome were identified based on prior clinical knowledge and are to be collected in the investigational study. All these key covariates and the outcome information were collected in the registry and the historical clinical study. An independent statistician, who was going to perform the PS design in the second design stage, was identified. This statistician was to be blinded to the outcomes throughout the design stages.

3.2. The second design stage

The second design stage started once complete covariate information was available for all enrolled 600 RCT subjects. Per the pre-specified inclusion/exclusion criteria of the current study, a total of 1053 registry subjects and 350 historical clinical study subjects were identified as potential subjects to be borrowed.

Based on the 600 RCT subjects and 1403 (= 1053 + 350) external subjects, the propensity scores, defined in Expression (1), were estimated using a logistic regression model in which the dependent variable was data source (RCT vs. external) and the independent variables include all the 10 baseline covariates in their first-order terms. Missing data in the covariates were imputed using the classification and regression trees (CART) method implemented in the R package mice. Note that a single imputation is considered in this example for simplicity.

Table 1. Number of subjects, the overlapping coefficient, and the design parameters for all k.

		Stratum					
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	Total
Current study ($j=C$)	$N_{k,C}$	120	120	121	119	120	600
Investigational device	$N_{k,C}^{(1)}$	79	83	87	71	84	404
Active control	$N_{k,C}^{(0)}$	41	37	34	48	36	196
Historical clinical study ($j=H$)	$N_{k,H}$	70	74	63	48	94	349
Overlapping coefficient	$r_{k,H}$	0.78	0.85	0.70	0.69	0.87	
	$r_{k,H}/\sum r_{k,H}$	0.20	0.22	0.18	0.18	0.22	
$\lambda_{k,H}$	$A_H r_{k,H}/\sum r_{k,H}$	13.96	15.33	12.68	12.38	15.65	70
$\alpha_{k,H}$	$\lambda_{k,H}/N_{k,H}$	0.20	0.21	0.20	0.26	0.17	
Registry ($j=R$)	$N_{k,R}$	376	278	174	97	117	1042
Overlapping coefficient	$r_{k,R}$	0.82	0.78	0.79	0.74	0.79	
	$r_{k,R}/\sum r_{k,R}$	0.21	0.20	0.20	0.19	0.20	
$\lambda_{k,R}$	$A_R r_{k,R}/\sum r_{k,R}$	27.08	25.74	26.38	24.71	26.10	130
$\alpha_{k,R}$	$\lambda_{k,R}/N_{k,R}$	0.07	0.09	0.15	0.25	0.22	

After the trimming step, 1042 out of 1053 registry subjects and 349 out of 350 historical clinical study subjects were retained. Five propensity score strata were formed from all the subjects (1042 + 349 + 600 = 1991) with each stratum containing ~120 subjects from the current RCT. The distribution of subjects across strata from the three data sources for the two treatment arms is presented in Table 1.

In every stratum, the distributions of each covariate in the current study and each external data source (the registry or the historical clinical study) are examined and compared. Based on Figure 2 and Figure 3, it can be observed that the covariates are reasonably balanced within every stratum. The

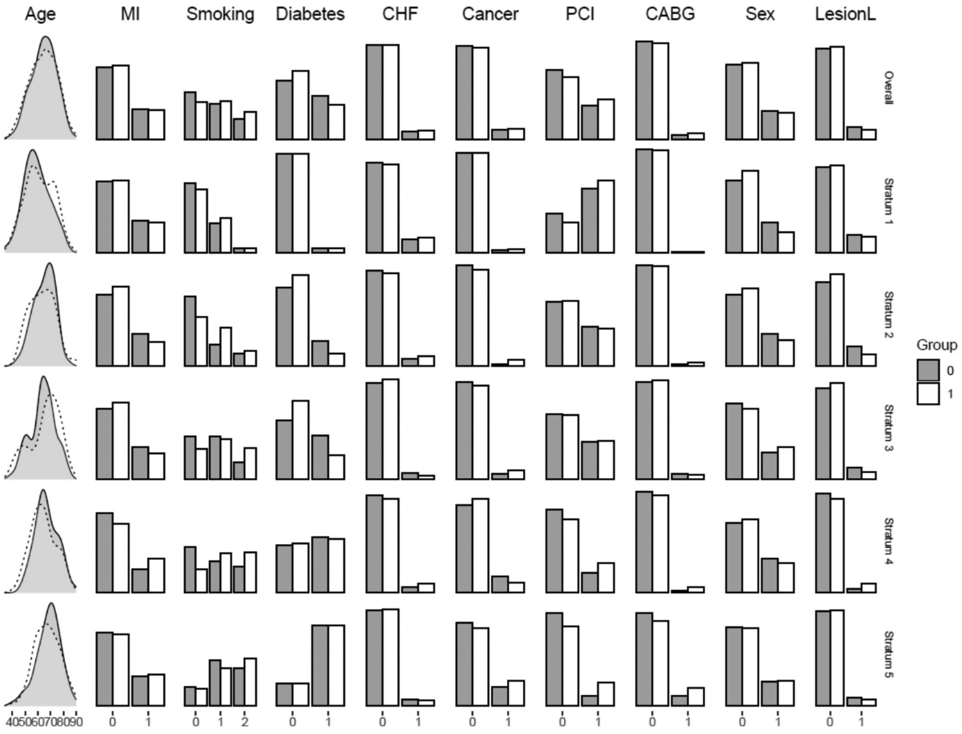

Figure 2. Balance checking of covariates between the current study (Group = 0) and historical clinical study (Group = 1).

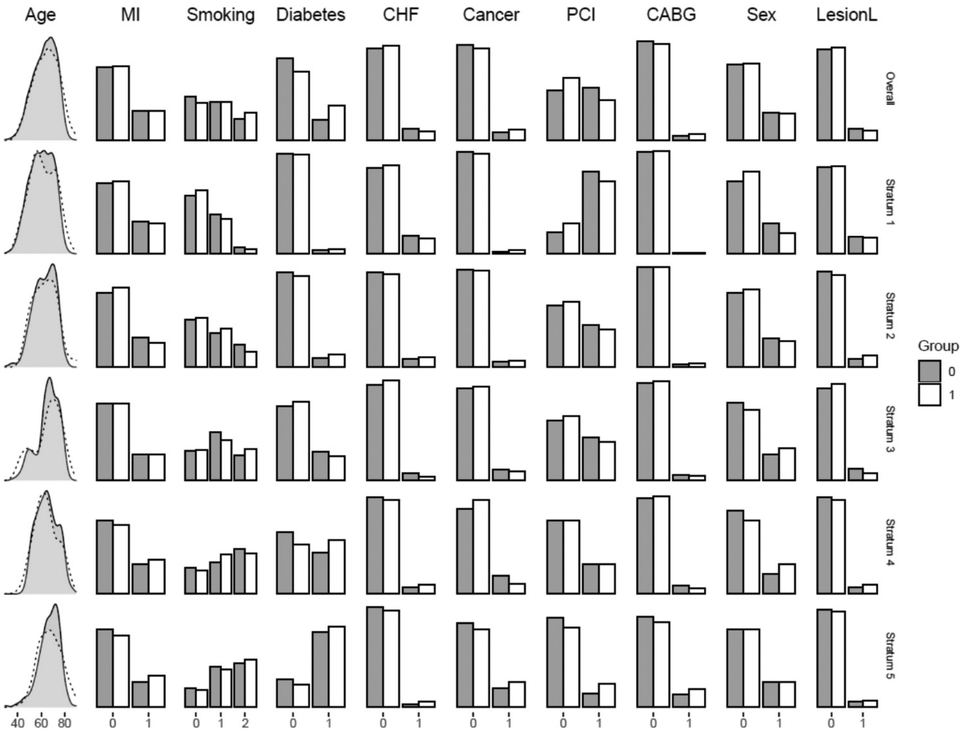

Figure 3. Balance checking of covariates between the current study (Group = 0) and the registry (Group = 1).

balance can also be assessed quantitatively by metrics such as the standardized mean difference by strata (Li et al. 2016; Yue et al. 2016). Then, overlapping coefficients of the propensity score distributions defined in Section 2.4 are calculated for all strata using Equation (4). The nominal number of subjects leveraged in stratum k from external data source $j(j = R, H)$, $\lambda_{k,j}$, can then be obtained using expressions in (3).

Table 1 presents number of subjects in each data source, the overlapping coefficient, and the design parameters $\lambda_{k,j}$ for each stratum k.

3.3. Outcome analyses

The final analysis was conducted after the clinical outcome data had been collected from all the 600 subjects in the current study and external data sources. The means and 95% credible intervals of the posterior distributions for $\theta_k^{(0)}$, $\theta_k^{(1)}$ and μ are displayed in Table 2. Based on the analysis, the posterior probability of $\mu < 0$ is 99.8%, which meets the study success criterion.

4. Discussion

An analytical approach is described when data from multiple external sources are leveraged to augment the control arm of an RCT, if the external data are deemed to be relevant and reliable as outlined in the FDA RWE guidance document (U.S. Food and Drug Administration 2017). The proposed approach is based on the notion that subjects in the RCT best represent the population of interest. Subjects in the RCT are divided into K strata with roughly equal sample size based on the estimated propensity scores, and each stratum may be viewed as a quasi-RCT. The weighting scheme across strata for the inference of the treatment effect of interest (μ) is devised accordingly: The weights

Table 2. Outcome analysis.

	Stratum					Overall
	1	2	3	4	5	
	$\theta_1^{(1)}$	$\theta_2^{(1)}$	$\theta_3^{(1)}$	$\theta_4^{(1)}$	$\theta_5^{(1)}$	
Posterior mean	0.156	0.101	0.119	0.049	0.159	
Lower 95% credible interval	0.084	0.047	0.062	0.012	0.090	
Upper 95% credible interval	0.241	0.175	0.194	0.108	0.242	
	$\theta_1^{(0)}$	$\theta_2^{(0)}$	$\theta_3^{(0)}$	$\theta_4^{(0)}$	$\theta_5^{(0)}$	
Posterior mean	0.187	0.160	0.153	0.193	0.242	
Lower 95% credible interval	0.113	0.089	0.082	0.117	0.155	
Upper 95% credible interval	0.277	0.248	0.243	0.280	0.341	
	$\theta_1^{(1)} - \theta_1^{(0)}$	$\theta_2^{(1)} - \theta_2^{(0)}$	$\theta_3^{(1)} - \theta_3^{(0)}$	$\theta_4^{(1)} - \theta_4^{(0)}$	$\theta_5^{(1)} - \theta_5^{(0)}$	μ
Posterior mean	−0.031	−0.059	−0.033	−0.144	−0.084	−0.070
Lower 95% credible interval	−0.147	−0.166	−0.141	−0.242	−0.204	−0.119
Upper 95% credible interval	0.082	0.043	0.070	−0.050	0.036	−0.023

are specified proportional to the total sample size per stratum, i.e. weights = $1/K$ in Equation (5). Control subjects in the external data sources are mainly utilized to improve the estimation of $\theta_k^{(0)}$ (the parameter associated with the control in the k^{th} stratum) for all k's. Note that the external controls are not viewed as a sample that well represents the target population of interest, thus the sample sizes in external data sources are not incorporated into the weights. To assess the performance of the proposed method, simulation studies were conducted. The description and the results are included in the supplementary material.

A key step in our proposed approach is to group subjects into K strata based on the selected $\hat{q}_0$, $\hat{q}_1$, . . ., $\hat{q}_K$ such that, within each stratum, subjects from two data sources are roughly similar in terms of the distribution of baseline characteristics. Although K is set at 5 in the example presented in Section 3, other choices for K are also plausible. In general, the selection of K may depend on size of external data sources, sample size of the current RCT, and whether covariates are reasonably balanced within each stratum based on the PS design. Different $\hat{q}_0, \hat{q}_1, \ldots, \hat{q}_K$ may be selected, and the approach described in Sections 2.4 and 2.5 can be generally followed, except that $\lambda_{k,j}$ may need to be modified, and the weights in the Equation (5) need to be adjusted using the following formula:

$$\mu = \sum_{k=1}^{K} \frac{N_{k,0}}{N_{\cdot,0}} \left(\theta_k^{(1)} - \theta_k^{(0)}\right).$$

In the first design stage, the sample size of the current study and nominal number of subjects to be borrowed from all external data sources are specified to obtain a certain level of power. However, it is possible that, due to relatively severe incomparability of patient baseline characteristics between an external data source and the current RCT, the actual nominal number of subjects to be leveraged is less than the originally planned nominal number of subjects to be leveraged (i.e. $\Lambda_j < \mathrm{A}_j$). This is due to $\lambda_{k,j} < A_j r_{k,j} / \sum_{k'=1}^{K} r_{k',j}$ for some k by observing Equation (3). As a result, the power may be reduced. One possible remedy is to reallocate $\left(A_j r_{k,j} / \sum_{k'=1}^{K} r_{k',j} - \lambda_{k,j}\right)$ to other strata and/or other external data source(s). It may be reallocated in a way such that the resulting updated Λ_j does not exceed A_j, and the updated actual nominal number of subjects to be borrowed within stratum k^{th} from all external data sources may not greatly exceed $A\left(N_{k,0}/N_{\cdot,0}\right)$. Another possibility is to use different number of strata, K, or different set of cut points $\hat{q}_0, \hat{q}_1, \ldots, \hat{q}_K$ (used to create K), as discussed in the previous paragraph.

If the power is still unsatisfactory after these attempts, more current RCT subjects may need to be enrolled. Such an action is only feasible if it is pre-planned. Note that all these remedies are only plausible if the outcome data are blinded.

It cannot be emphasized enough the importance of the principle that the design of a study and outcome analysis need to be separated. That is, during the design stage, outcome data are not supposed to be accessed due to any reasons to ensure the integrity of study. Under a regulatory framework, this can be accomplished by implementing the two-stage design discussed above. Note that, when the outcome-free design is violated, the objectivity and validity of the study may be compromised.

Acknowledgments

The authors are grateful to the referees for their great suggestions and comments that helped improve the article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this 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.

Chen, W.-C., C. Wang, H. Li, N. Lu, R. Tiwari, Y. Xu, and L. Q. Yue. 2020. Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data. *Journal of Biopharmaceutical Statistics* 30 (3):508–520. doi:10.1080/10543406.2020.1730877.

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.

Gottlieb, S. 2018. FDA budget matters: a cross-cutting data enterprise for real world evidence. https://www.fda.gov/news-events/fda-voices/fda-budget-matters-cross-cutting-data-enterprise-real-world-evidenc

Ibrahim, J. G., and M. H. Chen. 2000. Power prior distributions for regression models. *Statistical Science* 15:46–60.

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. doi:10.1002/sim.3782.

Levenson, M., W. He, J. Chen, Y. Fang, D. Faries, B. A. Goldstein, M. Ho, K. Lee, P. Mishra-Kalyani, F. Rockhold, et al. 2021. Biostatistical considerations when using RWD and RWE in clinical studies for regulatory purposes: A landscape assessment. *Statistics in Biopharmaceutical Research* 1–11. doi:10.1080/19466315.2021.1883473.

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. 2018a. Propensity-score-based priors for Bayesian augmented control design. *Pharmaceutical Statistics* 18 (2):223–238. doi:10.1002/pst.1918.

Lin, J., M. Gamalo-Siebers, and R. Tiwari. 2018b. Propensity score matched augmented controls in randomized clinical trials: A case study. *Pharmaceutical Statistics* 17 (5):629–647.

Lu, N., Y. Xu, and L. Yue. 2019. Good statistical practice in utilizing real world data in a comparative study for premarket evaluation of medical devices. *Journal of Biopharmaceutical Statistics* 29 (4):580–591. doi:10.1080/10543406.2019.1632880.

Lu, N., Y. Xu, and L. Yue. 2020. Some considerations on design and analysis plan on a nonrandomized comparative study utilizing propensity score methodology for medical device premarket evaluation. *Statistics in Biopharmaceutical Research* 12 (2):155–163. doi:10.1080/19466315.2019.1647873.

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.

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:10.7326/0003-4819-127-8_Part_2-199710151-00064.

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/()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.

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. Food and Drug Administration. 2017. Use of real-world evidence to support regulatory decision-making for medical devices - guidance for industry and food and drug administration staff. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices

Wang, C., H. Li, W.-C. Chen, N. Lu, R. Tiwari, Y. Xu, and L. 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.

Wang, C., N. Lu, W.-C. Chen, H. Li, R. Tiwari, Y. Xu, and L. Q. Yue. 2020. Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies. *Journal of Biopharmaceutical Statistics* 30 (3):495–507. doi:10.1080/10543406.2019.1684309.

Xu, Y., N. Lu, L. Yue, and R. Tiwari. 2020. A study design for augmenting the control group in a randomized controlled trial: A quality process. *Therapeutic Innovation & Regulatory Science* 54 (2):269–274. doi:10.1177/2168479019830385.

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.