Taylor & Francis
Taylor & Francis Group

Journal of Biopharmaceutical Statistics

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

Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data

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

To cite this article: Wei-Chen Chen, Chenguang Wang, Heng Li, Nelson Lu, Ram Tiwari, Yunling Xu & Lilly 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

To link to this article: https://doi.org/10.1080/10543406.2020.1730877

Published online: 06 May 2020.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2020, VOL. 30, NO. 3, 508–520
https://doi.org/10.1080/10543406.2020.1730877

Taylor & Francis
Taylor & Francis Group

Propensity score-integrated composite likelihood approach for augmenting the control arm of a randomized controlled trial by incorporating real-world data

Wei-Chen Chen[a], Chenguang Wang[b], Heng Li[a], Nelson Lu[a], Ram Tiwari[a], 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

ABSTRACT

In this paper, a propensity score-integrated composite likelihood (PSCL) approach is developed for cases in which the control arm of a two-arm randomized controlled trial (RCT) (treated vs control) is augmented with patients from real-world data (RWD) containing both clinical outcomes and covariates at the patient-level. RWD patients who were treated with the same therapy as the control arm of the RCT are considered for the augmentation. The PSCL approach first estimates the propensity score for every patient as the probability of the patient being in the RCT rather than the RWD, and then stratifies all patients into strata based on the estimated propensity scores. Within each propensity score stratum, a composite likelihood function is specified and utilized to down-weight the information contributed by the RWD source. Estimates of the stratum-specific parameters are obtained by maximizing the composite likelihood function. These stratum-specific estimates are then combined to obtain an overall population-level estimate of the parameter of interest. The performance of the proposed approach is evaluated via a simulation study. A hypothetical two-arm RCT and a hypothetical RWD source are used to illustrate the implementation of the proposed approach.

ARTICLE HISTORY
Received 4 September 2019
Accepted 20 December 2019

KEYWORDS
Augmentation; composite likelihood; propensity scores; real-world data; real-world evidence

1. Introduction

In recent years, real-world data (RWD) have been playing an increasingly important role in medical product development as a source of external evidence. Sources of RWD include national and international product and disease registries, electronic health records (EHRs), claims and billing data, and data gathered through personal devices. Based on RWD of acceptable quality and with appropriate analysis, real-world evidence (RWE) can be generated and utilized in making regulatory decisions. Xu et al. (2019) describe a two-stage study design for incorporating external data (including RWD) to augment the control arm of a randomized controlled trial (RCT). In a typical application of this design, one starts with a two-arm RCT and then augments the control arm with the down-weighted external data. With a proper study design, the augmented RCT can reduce the number of patients enrolled into the study, thereby shortening its duration and lowering its cost (Lin et al. 2018a, 2018b), while still providing high-level evidence to support the evaluation of safety and effectiveness for new medical products in regulatory settings.

Two issues need to be considered when using external data (RWD in this paper) to augment the control arm of an RCT (which will also be called the current study): 1) the similarity, in terms

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

of baseline covariates, between patients from the external data source and patients from the RCT, and 2) the amount of information that the external data source contributes to the statistical inference. If the RWD patients and RCT patients are not similar, then it may be difficult to justify incorporating the RWD patients into the control arm of the RCT. Since typically an RWD source contains many more patients than an RCT, we need to figure out how to down-weight the RWD patients so that they do not dominate study results if we decide to leverage the RWD patients. To address these issues, we develop a statistical methodology, henceforth called the propensity score-integrated composite likelihood (PSCL) approach. As suggested by its name, our approach involves integrating propensity score methodology with the technique of composite likelihood. The propensity score methodology is utilized to address the first issue, by forming strata of patients according to their propensity scores and then considering the incorporation of RWD into the control arm of the RCT within each propensity score stratum. The rationale for doing this is that within each propensity score stratum patients from RWD and RCT tend to be more similar in terms of baseline covariates than they are overall, which makes within-stratum leveraging of RWD more justified. The composite likelihood is utilized to address the second issue. As we will see in Section 2, the method of composite likelihood provides a very natural theoretical basis for down-weighting individual observations.

The PSCL approach can be implemented within a "two-stage study design" framework. The two-stage design was first introduced by Yue et al. (2014) for objective causal inference for non-randomized comparative studies and later adapted by Xu et al. (2019), and Tiwari for leveraging RWD to augment an RCT. The two-stage design is a streamlined procedure that maintains the objectivity and integrity of the study by ensuring that outcome data are not in sight during study design, which, in our context, consists of estimating patients' propensity scores, stratifying patients according to their propensity scores, and determining the weights assigned to RWD patients. Key to the procedure proposed by Yue et al. (2014) and Xu et al. (2019) is the requirement that the study protocol details how the statistician (or statistical team) in charge of carrying out those activities are blinded to the outcome data, thereby achieving the so-called "outcome-free design." The two-stage design is becoming a standard practice (Campbell and Yue 2016) for observational studies to support regulatory decisions for medical devices. The fact that the two-stage design framework is applicable to the PSCL approach makes the latter a methodological innovation that can readily be adopted in regulatory settings.

The rest of the paper is organized as follows. In Section 2, we briefly review the propensity score methodology, and then introduce the PSCL approach augmenting the control arm of an RCT with patients from an RWD source. In Section 3, we present the results of simulation studies for the PSCL approach. As an example, included in Section 4 is a hypothetical 2:1 RCT with augmented control arm data to achieve 1:1 ratio in terms of nominal sample size. Section 5 concludes with some discussions.

2. Method

2.1. *Notation*

For the ith patient, let Y_i be the random variable that represents the outcome data and y_i be a realization of Y_i. Let $\mathbf{X_i}$ with dimension $p \times 1$ denote the vector of covariates; let $Z_i = 1$ if patient i is from the current randomized study and 0 if patient i is from external data; let $T_i = 1$ if patient i is from the current randomized study and assigned to the treatment arm, and 0 if patient i is from the current randomized study and assigned to the control arm, or is from the external data. Note that the patients in the external data undergo the same therapy as that of the control arm of the current study. Correspondingly, let $\mathbf{X}$, Z, and T without the subscript be the random variables/vectors representing the values of those quantities for a randomly chosen patient.

We assume that all the patients satisfy the inclusion and exclusion criteria of the *current* randomized study. Furthermore, 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.

2.2. Propensity score method

Formulated by Rosenbaum and Rubin (1983), the propensity score (PS), $e(\mathbf{X})$, for a patient with a vector $\boldsymbol{X}$ 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 $\boldsymbol{X}$:

$$e(\mathbf{X}) = Pr(T = 1|\mathbf{X}).$$

The propensity score $e(\mathbf{X})$ is a balancing score in the sense that among patients with the same propensity score, the distribution of observed covariates is the same between the two treatment groups. 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 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 (Agostino and Rubin 2000; Austin 2011; 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 (sub-classification). 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 particular, they can be implemented in such a way that allows the separation of the study design and the analysis of outcome data, which is fundamental to ensuring the integrity and credibility of study results.

In this paper the propensity score methodology is used for the purpose of leveraging RWD to augment an RCT (the current study) by augmenting its control arm with patients from RWD. The objective is to balance the covariates between the current study and the RWD to make the leveraging of RWD more justified. Accordingly, we define the propensity score as follows:

$$e(\mathbf{X}) = Pr(Z = 1|\mathbf{X})$$

where, as mentioned earlier, $Z = 1$ if the patient is in the current study and $Z = 0$ if the patient is in the RWD. The propensity score so defined has the following balancing property:

$$\begin{aligned} f(\mathbf{X}|e(\mathbf{X}), Z = 1) &= f(\mathbf{X}|e(\mathbf{X}), T = 1, Z = 1) \\ &= f(\mathbf{X}|e(\mathbf{X}), T = 0, Z = 1) \\ &= f(\mathbf{X}|e(\mathbf{X}), T = 0, Z = 0) \end{aligned} \tag{1}$$

where $f(\mathbf{X}|\cdot)$ indicates the conditional density function of $\mathbf{X}$.

To prove Equation (1), note that

$$T \perp \mathbf{X}|Z = 1$$

because of randomization. Since $e(\mathbf{X})$ is a function of $\mathbf{X}$ only, we also have

$$T \perp \mathbf{X}|e(\mathbf{X}), Z = 1.$$

Therefore,

$$f(\mathbf{X}|e(\mathbf{X}), Z = 1) = f(\mathbf{X}|e(\mathbf{X}), T = 1, Z = 1) = f(\mathbf{X}|e(\mathbf{X}), T = 0, Z = 1). \tag{2}$$

Furthermore, we have

$$Z \perp \mathbf{X} \,|\, e(\mathbf{X})$$

by Theorem 1 in Rosenbaum and Rubin (1984). Then,

$$f(\mathbf{X} \,|\, e(\mathbf{X}), Z = 1) = f(\mathbf{X} \,|\, e(\mathbf{X}), Z = 0). \tag{3}$$

Since $Z = 0$ implies $T = 0$ in our setting, we have

$$f(\mathbf{X} \,|\, e(\mathbf{X}), Z = 1) = f(\mathbf{X} \,|\, e(\mathbf{X}), T = 0, Z = 0). \tag{4}$$

Combining Equations (2–4) that completes the proof of the balancing property.

Equation (1) tells us that given the value of the propensity score (PS) $e(\mathbf{X}) = \Pr(Z = 1|\mathbf{X})$, not only are the covariates balanced between patients from the current study and from the RWD, they are also balanced between the following three groups: 1) the control arm of the current study, 2) the treatment arm of the current study, and 3) RWD. It lays the foundation of our PSCL approach to leveraging RWD to augment the control arm of an RCT.

2.3. *Using composite likelihood to discount external data*

In the setting of this paper, where we have the current study and an external data source, it is often desirable to properly down-weight the information from the external data source. The statistical tool that we will use to achieve such discounting is the method of composite likelihood (Lindsay 1988).

Let $f(y; \boldsymbol{\omega})$ denote the density function indexed by the parameter $\boldsymbol{\omega} = (\theta, \phi)$, where θ represents the parameter of interest and ϕ represents the nuisance parameters. In order to conduct statistical inference for θ based on the outcome data from all the n patients in the current study and the external data source (after the exclusion of certain external patients, called *trimming*, as described in Section 2.4), the composite likelihood function is the following weighted product:

$$L(\boldsymbol{\omega}; \mathbf{Y}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\mathrm{y}_{\mathrm{i}}; \boldsymbol{\omega})^{\gamma_{\mathrm{i}}} \tag{5}$$

where γ_i's are nonnegative weights to be chosen. In this paper, those weights are used to discount the external data. We set $\gamma_i = 1$ if patient i is from the current study and $0 < \gamma_i \leq 1$ if patient i is from RWD. Details of the weighting scheme are given in Section 2.4. See Varin et al. (2011) for a thorough review of the composite likelihood methods.

Let $\widehat{\boldsymbol{\omega}}$ be the maximum likelihood estimator of $\boldsymbol{\omega}$ based on Equation (5). In general, $\widehat{\boldsymbol{\omega}}$ can be obtained by solving the equation

$$\Psi(\boldsymbol{\omega}) = \sum_{i=1}^{n} \gamma_i \frac{\partial \log f(y_i; \boldsymbol{\omega})}{\partial \boldsymbol{\omega}} = 0. \tag{6}$$

Note if $\boldsymbol{\omega} = (\omega_1 = \theta, \omega_2 = \phi_1, \ldots, \omega_k = \phi_{k-1})$ is k-dimensional, Equation (6) represents the set of equations

$$\sum_{i=1}^{n} \gamma_i \frac{\partial \log f(y_i; \boldsymbol{\omega})}{\partial \omega_j} = 0 \qquad j = 1, \ldots, k.$$

This method can be generalized to the case where θ is a vector. However, the extensions are beyond the scope of this paper.

Under mild regularity conditions, by the central limit theorem (Varin et al. 2011), we have

$$\sqrt{n}(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}) \xrightarrow{d} \mathrm{MVN}\left(0, G^{-1}(\boldsymbol{\omega})\right)$$

where $G(\boldsymbol{\omega})$ is the Godambe information matrix (Godambe 1960)

$$G(\boldsymbol{\omega}) = \mathbf{E}_{\boldsymbol{\omega}}\left(-\frac{\partial \boldsymbol{\Psi}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right)(\mathrm{Var}_{\boldsymbol{\omega}}\boldsymbol{\Psi}(\boldsymbol{\omega}))^{-1}\mathbf{E}_{\boldsymbol{\omega}}\left(-\frac{\partial \boldsymbol{\Psi}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}}\right). \tag{7}$$

In order to obtain a more robust estimate of the covariance matrix for $\widehat{\boldsymbol{\omega}}$, especially when n is not sufficiently large, methods such as bootstrap or jackknife can be applied in practice. For example, the jackknife estimate of the covariance matrix is

$$\mathrm{Var}\widehat{\boldsymbol{\omega}} = \frac{n-1}{n}\sum_{i=1}^{n}(\widehat{\boldsymbol{\omega}}^{(-i)} - \widehat{\boldsymbol{\omega}})(\widehat{\boldsymbol{\omega}}^{(-i)} - \widehat{\boldsymbol{\omega}})^{\mathrm{T}}$$

where $\widehat{\boldsymbol{\omega}}^{(-i)}$ is the maximum likelihood estimator of $\boldsymbol{\omega}$ when the ith patient is excluded.

2.4. *PS-integrated composite likelihood*

In this section, we will provide a step-by-step description of the PS-integrated composite likelihood (PSCL) approach for leveraging RWD to augment the control arm of the current study. Recall that $Z_i = 1$ if patient i is from the current study and $Z_i = 0$ if patient i is from RWD, and the PS $e(\mathbf{X_i})$ of patient i is defined relative to Z_i.

The first step of the approach, which we call "*trimming*", is to exclude patients in the external study that are not similar to the patients in the current study. Let $\mathbb{E}_1$ denote $\{\widehat{e}(\mathbf{X_i}) : Z_i = 1\}$, the set of the estimated PS for patients in the current study. Note that the hat symbol emphasizes that it is the estimated PS. To conduct *trimming* of the patients in RWD, we exclude the patients in RWD with the estimated PS out of the range of $\mathbb{E}_1$ (i.e. $[\min\mathbb{E}_1, \max\mathbb{E}_1]$). Note that in a typical treatment effect causal inference context, exclusion of patients may distort the patient population of interest and introduce bias. However, this is not a concern here since patients in the current study represent the population of interest which is not altered by *trimming*.

Next, we construct S strata of PS with inner cut points $\{\widehat{q}_s : s = 1, 2, \ldots, S-1\}$ between $\widehat{q}_0 = \min\mathbb{E}_1$ and $\widehat{q}_S = \max\mathbb{E}_1$, such that $\widehat{q}_1 < \widehat{q}_2 < \cdots < \widehat{q}_s < \cdots < \widehat{q}_{S-1}$ and the number of patients in the current study in each stratum $(\widehat{q}_{s-1}, \widehat{q}_s]$ is equal for all strata $s = 1, 2, \ldots, S$. Given the constructed PS strata, we then assign patients to a stratum s according to the estimated PS, e.g., ith patient is assigned to the stratum s if $\widehat{e}(\mathbf{X_i}) \in (\widehat{q}_{s-1}, \widehat{q}_s]$ for some $s \in \{1, 2, \ldots, S\}$. For each stratum s, we denote two index sets based on patients in the current study and from RWD as $\mathbb{W}_{s,1} = \{i : \widehat{e}(\mathbf{X_i}) \in (\widehat{q}_{s-1}, \widehat{q}_s], Z_i = 1\}$ and $\mathbb{W}_{s,0} = \{i : \widehat{e}(\mathbf{X_i}) \in (\widehat{q}_{s-1}, \widehat{q}_s], Z_i = 0\}$, respectively. Figure 1 presents the *trimming* and stratification scheme.

The third step is to specify the (composite) likelihood function. For the stratum s, let $\theta_s^{(1)}$ and $\theta_s^{(0)}$ be the parameters of interest for the treatment and control arms, respectively. We specify the likelihood of $\theta_s^{(1)}$ as

$$L(\theta_s^{(1)}) = \prod_{i \in \mathbb{W}_{s,1}^{(1)}} f(y_i; \theta_s^{(1)}) \tag{8}$$

where $\mathbb{W}_{s,1}^{(1)}$ is a subset of $\mathbb{W}_{s,1}$ that contains only indices of patients who were in the treatment arm of the current study. Similarly, let $\mathbb{W}_{s,1}^{(0)}$ be a subset of $\mathbb{W}_{s,1}$ that contains only indices of patients who were in the control arm of the current study, i.e. $\mathbb{W}_{s,1}^{(1)} \uplus \mathbb{W}_{s,1}^{(0)} = \mathbb{W}_{s,1}$. We then specify the composite likelihood of $\theta_s^{(0)}$ to be of the following form:

$$L(\theta_s^{(0)}) = \prod_{i \in \mathbb{W}_{s,1}^{(0)} \bigcup \mathbb{W}_{s,0}} f(y_i; \theta_s^{(0)})^{\gamma_i} \tag{9}$$

where $\gamma_i = 1$ for all $i \in \mathbb{W}_{s,1}^{(0)}$, $\gamma_i = \frac{\lambda_{s,0}}{n_{s,0}}$ for all $i \in \mathbb{W}_{s,0}$, $0 < \lambda_{s,0} \leq n_{s,0}$ is a discounting parameter to be determined, and $n_{s,0}$ is the number of patients in $\mathbb{W}_{s,0}$. Here, $\lambda_{s,0}$ can be interpreted as a nominal

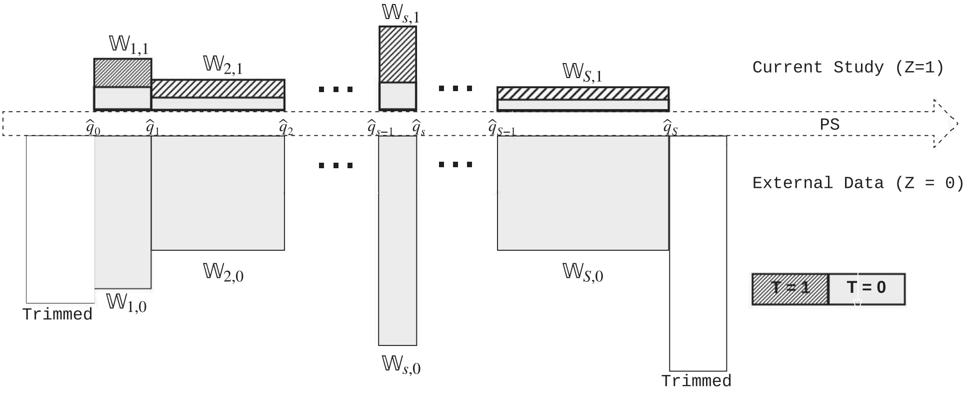

Figure 1. Propensity score trimming and stratification scheme. Area of the bars is proportional to the number of patients in the corresponding group.

"number of patients leveraged" in stratum s, which cannot exceed $n_{s,0}$, the number of RWD patients in stratum s. Note that 1) for simplicity we suppress the dependence on nuisance parameters and 2) we only specify the form of the likelihood function, so no outcome data are needed.

The fourth step is to determine the value of $\lambda_{s,0}$ for all s. To do so, we consider a *similarity weighting* scheme and let the amount of information contributed by the patients in $\mathbb{W}_{s,0}$ (i.e., $\lambda_{s,0}$) be related to the *similarity* between the patients in $\mathbb{W}_{s,0}$ and in $\mathbb{W}_{s,1}$. As a measure of similarity, we use the overlapping coefficient (Inman and Bradley 1989), between the PS distributions in $\mathbb{W}_{s,0}$ and in $\mathbb{W}_{s,1}$, denoted as r_s, i.e.,

$$r_s = \int_0^1 \min[g_{s,0}(e), g_{s,1}(e)]de \tag{10}$$

which is essentially the overlapping area of the density curves $g_{s,0}$ and $g_{s,1}$ of $\hat{e}(\mathbf{X}_i)$ for i in $\mathbb{W}_{s,0}$ and $\mathbb{W}_{s,1}$, respectively. The maximum total number of patients to be leveraged from RWD, denoted by A, is elicited as a fixed constant. This number is usually determined based on clinical and regulatory considerations. Since we want $\lambda_{s,0}$ to be proportional to r_s but not exceed $n_{s,0}$, we let

$$\lambda_{s,0} = \min\left(\frac{Ar_s}{\sum_{s=1}^S r_s}, n_{s,0}\right). \tag{11}$$

2.5. *Inference*

The maximum likelihood estimator of $\theta_s^{(1)}$ and $\theta_s^{(0)}$, denoted as $\widehat{\theta}_s^{(1)}$ and $\widehat{\theta}_s^{(0)}$, can be obtained by maximizing $\log L(\theta_s^{(1)})$ and $\log L(\theta_s^{(0)})$ as in Equations (8) and (9), respectively. Methods such as bootstrap or jackknife can be applied to obtain the variance for $\widehat{\theta}_s^{(1)}$ and $\widehat{\theta}_s^{(0)}$. For example, the jackknife estimate of variance for $\widehat{\theta}_s^{(1)}$ is

$$\mathrm{Var}\widehat{\theta}_s^{(1)} = \frac{n_s - 1}{n_s} \sum_{i \in \mathbb{W}_{s,1} \bigcup \mathbb{W}_{s,0}} \left(\widehat{\theta}_{s,-i}^{(1)} - \widehat{\theta}_s^{(1)}\right)^2$$

where n_s is the number of patients in the stratum s and $\widehat{\theta}^{(1)}_{s,-i}$ is the maximum likelihood estimator of $\theta^{(1)}_s$ when the ith patient is excluded. Similarly, we can obtain the jackknife estimate of variance, $\mathrm{Var}\,\widehat{\theta}^{(0)}_s$, for $\widehat{\theta}^{(0)}_s$.

The treatment effect of interest is μ which can be estimated by a weighted average of $\widehat{\theta}^{(1)}_s - \widehat{\theta}^{(0)}_s$ for all s. That is,

$$\widehat{\mu} = \sum_{s=1}^{S} \frac{n_{s,1}}{n_{\cdot,1}} \left(\widehat{\theta}^{(1)}_s - \widehat{\theta}^{(0)}_s \right) \tag{12}$$

where $n_{s,1}$ is the number of patients in $\mathbb{W}_{s,1}$ and $n_{\cdot,1} = \sum_{s=1}^{S} n_{s,1}$ is the number of patients in the current study. The (approximate) sampling variance of $\widehat{\mu}$ can be obtained as

$$\mathrm{Var}\widehat{\mu} = \sum_{s=1}^{S} \left(\frac{n_{s,1}}{n_{\cdot,1}} \right)^2 \left(\mathrm{Var}\widehat{\theta}^{(1)}_s + \mathrm{Var}\widehat{\theta}^{(0)}_s \right). \tag{13}$$

3. Simulation study

Simulation studies are conducted to evaluate the proposed PSCL approach. The simulation setting and results are reported in this section.

3.1. *Simulation settings*

For $Z = 0$ (external data) and $Z = 1$ (current randomized study), we assume that covariates $\mathbf{X}$ follow a mixture of multivariate normal distribution of K components such that $\mathbf{X}_{\mathbf{p}\times\mathbf{1}} | Z \sim F_Z$ and

$$F_Z = \sum_{k=1}^{K} \psi_{\mathbf{Zk}} MVN(\mu_{\mathbf{Zk}}, \mathbf{\Sigma}_{\mathbf{Z}})$$

with $\sum_{k=1}^{K} \psi_{\mathbf{Zk}} = 1$, the diagonal of $\mathbf{\Sigma}_{\mathbf{Z}}$ being σ^2_Z, and the off-diagonal elements of $\mathbf{\Sigma}_{\mathbf{Z}}$ being $\rho\sigma^2_Z$.

Both continuous and binary outcomes are considered in the simulation studies. For continuous outcomes, we simulate Y_i from model

$$Y_i | \mathbf{X}_i, T_i = \beta_0 + \tau T_i + \beta^{\mathrm{T}} \mathbf{X}_{\mathbf{i}} + \epsilon_{\mathbf{i}}$$

where ϵ_i is the random error. For binary outcomes, we simulate $P(Y_i = 1 | \mathbf{X}_{\mathbf{i}}, T_i)$ from

$$\mathrm{logit}\, P(Y_i = 1 | \mathbf{X}_i, T_i) = \beta_0 + \tau T_i + \beta^{\mathrm{T}} \mathbf{X}_{\mathbf{i}}$$

and simulate Y_i from its binary distribution. In the simulation studies, τ is the treatment effect of interest conditioning on the covariates $\mathbf{X}$.

For all the simulation studies, we set $\rho = 0.1$ and $p = 10$. We convert $X_1, \ldots, X_4$ to binary covariates using cut point 0 in order to evaluate the performance of the proposed approach for binary covariates. For continuous outcomes, we set $\tau = 3$, $\beta = \mathbf{1}_{\mathbf{p}\times\mathbf{1}}$ and $\beta_0 = 0$. For binary outcomes, we set $\beta = \mathbf{1}_{\mathbf{p}\times\mathbf{1}}$ and choose β_0 and τ such that $E(Y|T = 1) = 0.4$ and $E(Y|T = 0) = 0.2$. Note that we assume β_0 and β are identical for $Z = 0$ and 1. When there are no unmeasured confounders, covariate effects on the outcome are constant across data sources. For continuous outcomes, we assume $\epsilon_i \sim N(0, 1)$. 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 a practical issue that patients in the current study may not be a subgroup of the patients in the external data source. In this scenario, we set $K = 1$, $\mu_{\mathbf{01}} = \mathbf{1.2}_{\mathbf{p}\times\mathbf{1}}$, $\mu_{\mathbf{11}} = \mathbf{1}_{\mathbf{p}\times\mathbf{1}}$. Furthermore, we set $\sigma^2_1 = 1$ and $\sigma^2_0 = 1.5$. In Scenario II, we assume the patients in the external data have a mixture distribution. For $Z = 1$, we set $K = 1$, $\mu_{11} = \mathbf{1}_{\mathbf{p}\times\mathbf{1}}$, and $\sigma_1^2 = 1$. For $Z = 0$, we set $K = 2$, $\psi_{01} = \psi_{02} = 0.5$, $\mu_{01} = \mathbf{1}_{\mathbf{p}\times\mathbf{1}}$, $\mu_{02} = \mathbf{1.5}_{\mathbf{p}\times\mathbf{1}}$, and $\sigma_0^2 = 1$. For each scenario, we consider the dimension of $\mathbf{X}$ to be $p = 10$ and $p = 15$, sample sizes of the current study $N_1 = 300$ and 420 with randomization ratio 2:1 (treatment vs. control), and set A to be 50 and 100 for sample size 300, and 70 and 140 for sample size 420. The A caps the information of RWD, so that the ratios of final sample size (treatment vs. control) will be 4:3 and 1:1 after down-weighting from RWD, respectively.

We compare two strategies for the PSCL approach. The first strategy, *No PS Stratification*, sets the number of strata $S = 1$, and the discounting parameter $\lambda_0 = A/N_0$, where A is the intended number of patients to be leveraged from the N_0 patients who have remained in the external data source after *trimming*. The second strategy, *PS Stratification*, sets the number of strata $S = 5$ and chooses the discounting parameters $\lambda_{s,0}$ following Equation (11).

3.2. *Simulation results*

In Table 1, we report for each strategy the estimate of μ, bias, and mean squared error (MSE) for the treatment effect following Equation (12). "No Stra." and "Stra. (rs)" in Table 1 are indicated for the *No PS Stratification* and *PS Stratification* approaches, respectively. The simulation results are based on 1,000 replications.

Table 1. Simulation study results for the PSCL approach. "No Stra." and "Stra. (rs)" are the *No PS Stratification* and *PS Stratification* approaches, respectively.

				Treatment effect		
Scenario	N_1	A	Approach	Mean	Bias × 100	MSE × 100
			Binary outcomes			
I	300	50	No Stra.	0.154	−4.553	0.402
			Stra. (rs)	0.178	−2.180	0.222
I	300	100	No Stra.	0.132	−6.826	0.627
			Stra. (rs)	0.167	−3.251	0.247
I	420	70	No Stra.	0.153	−4.683	0.363
			Stra. (rs)	0.178	−2.227	0.177
I	420	140	No Stra.	0.131	−6.948	0.603
			Stra. (rs)	0.167	−3.254	0.209
II	300	50	No Stra.	0.151	−4.875	0.424
			Stra. (rs)	0.196	−0.352	0.111
II	300	100	No Stra.	0.126	−7.445	0.713
			Stra. (rs)	0.193	−0.677	0.096
II	420	70	No Stra.	0.147	−5.251	0.413
			Stra. (rs)	0.193	−0.672	0.087
II	420	140	No Stra.	0.122	−7.825	0.729
			Stra. (rs)	0.191	−0.949	0.078
			Continuous outcomes			
I	300	50	No Stra.	2.658	−34.186	22.631
			Stra. (rs)	2.952	−4.831	9.463
I	300	100	No Stra.	2.493	−50.719	34.480
			Stra. (rs)	2.929	−7.111	7.404
I	420	70	No Stra.	2.663	−33.720	19.733
			Stra. (rs)	2.953	−4.668	6.978
I	420	140	No Stra.	2.500	−49.975	31.605
			Stra. (rs)	2.935	−6.516	5.491
II	300	50	No Stra.	2.479	−52.052	39.126
			Stra. (rs)	2.956	−4.434	3.834
II	300	100	No Stra.	2.223	−77.703	70.252
			Stra. (rs)	2.935	−6.477	3.238
II	420	70	No Stra.	2.480	−52.032	35.328
			Stra. (rs)	2.955	−4.526	2.449
II	420	140	No Stra.	2.217	−78.253	67.953
			Stra. (rs)	2.933	−6.694	2.185

Overall, the improvement of the *PS Stratification* approach (in terms of the reduction of the bias and MSE) is shown for both binary and continuous outcomes and in different sample size settings. Further, the simulation results also show that the *PS Stratification* approach performs better in both scenarios where the distributions of baseline covariates are different between the current study and the external data. For down-weighting of RWD information, when A is increased, the bias and MSE also increase in *No PS Stratification* approach that is allowing more RWD information may distort the evidence of the current study. However, the magnitude of distortion can be reduced through the proposed PSCL approach. Estimation based on the *PS Stratification* approach greatly reduces the bias and MSE compared with the *No PS Stratification* approach.

Note that the down-weighting of RWD information in the *PS Stratification* approach is determined by $\lambda_{s,0}$ in Equation (11) which is proportional to the r_s in Equation (10) for each stratum. Based on Table 1, we found that the r_s in the *PS Stratification* approach yielded reasonable reductions for the bias and MSE of the treatment effects. However, it appears that the construction of $\lambda_{s,0}$ may also affect the reduction of the bias and MSE. In general, any reasonable measures of the similarity between two distributions can be utilized for determining $\lambda_{s,0}$. For example, similar to Equation (11), one may define $d_s = 1 - r_s$ and also let $\lambda_{s,0} = \min\left(\frac{A\frac{1}{d_s}}{\sum_{s=1}^{S}\frac{1}{d_s}}, n_{s,0}\right)$ provided $d_s > 0$ for all s. The simulation results (not included) show that the value $\lambda_{s,0}$ based on d_s has slightly better performance improvement than that based on r_s.

4. An illustrative example

In this section, we use an example to illustrate the proposed PSCL approach. It involves a randomized clinical trial conducted to demonstrate the safety and effectiveness of a cardiovascular device in patients with congestive heart failure, and utilizing a patient registry that could be appropriately leveraged to augment the control arm of the randomized clinical trial (the current study).

4.1. *The first design stage*

The hypotheses associated with the primary endpoint, a binary clinical outcome variable, were

$$H_0 : \mu = 0 \quad vs. \quad H_a : \mu \neq 0$$

where $\mu = \theta^{(1)} - \theta^{(0)}$ is the treatment effect with respect to one-year adverse event rates $\theta^{(0)}$ (control arm) and $\theta^{(1)}$ (investigational device arm). At the first design stage, a total of 17 baseline covariates that may affect the clinical outcome were identified based on prior knowledge. All these key covariates and the outcome information are also collected in the registry.

In order to determine the sample size, the expected $\theta^{(0)}$ and $\theta^{(1)}$ were assumed to be 0.29 and 0.20, respectively. At the significance level of 0.05, a power of 80% would be obtained with a total sample size of approximately 354 at 1:1 randomization ratio. It was proposed to enroll 267 patients in the current study and randomize those patients at 2:1 ratio to the investigational device arm and the control arm. The information to be borrowed from the registry to augment the control arm will be equivalent to 87 patients. In other words, the value of A is set at 87. Note that the number of patients to be borrowed should be determined based on clinical judgment on a case by case basis in practice.

4.2. *The second design stage*

When all the 267 patients were enrolled, which in this example included 183 patients assigned to the investigational device arm and 84 patients assigned to the control arm, and all their covariate data had been collected, the second design stage was started. Based on the pre-specified inclusion/exclusion criteria of the randomized current study, a total of 1570 patients, treated with the same therapy as the control arm of the current study, from the registry were identified as potential patients to be borrowed.

Based on the 267 patients in the current study and 1570 external patients, propensity scores were estimated using logistic regression with all 17 baseline covariates included in their linear terms. Missing data in the covariates (about 6%) were imputed using the classification and regression trees (CART) method implemented in the R package mice (Buuren and Groothuis-Oudshoorn 2010). Note that we consider a single imputation in this example for simplicity.

After *trimming*, 1192 out of 1570 patients were retained for the study design and outcome analysis. Five propensity score strata were formed for all the patients (267 + 1192) with each stratum containing nearly equal number (i.e., ~54) of patients in the current study. In each stratum, the distributions of each covariate in the current study and the external data are examined and compared. Based on Figure 2, we see that the covariates are well balanced between the current study and external data source for all strata. Then, overlapping coefficients of the propensity score distributions defined in Section 2.4 are calculated for all strata using Equation (11) (Figure 3). The nominal number of external patients leveraged in stratum s, $\lambda_{s,0}$, can then be obtained using Equation (11) Table 2.

Table 2. Number of patients, the overlapping coefficient, and the weights $\lambda_{s,0}$ for all s.

		Stratum					
		$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	Total
Current Study	$n_{s,1}$	54	53	53	53	54	267
Treatment		41	28	39	36	39	183
Control		13	25	14	17	25	84
Registry	$n_{s,0}$	332	270	233	201	156	1192
Overlapping Coefficient	r_s	0.85	0.81	0.82	0.74	0.82	
	$r_s/\sum r_s$	22%	20%	20%	18%	20%	100%
Approx. patients borrowed	$Ar_s/\sum r_s$	19	17	17	16	18	87
	$\lambda_{s,0}/n_{s,0}$	0.06	0.06	0.08	0.08	0.11	

4.3. *Outcome analysis*

The final analysis was conducted after the clinical outcome had been collected from all the 267 patients in the current study.

Based on the analysis with $A = 87$, the estimates of $\theta_s^{(1)}$ and $\theta_s^{(0)}$ and the associated standard errors are reported in Table 3. The estimate of the overall treatment effect $\widehat{\mu}$ is –0.18 with SE equal to 0.04. The p-value from the Wald test is 0.01, which indicates that the adverse event rate of the investigational device is statistically significantly lower than that of the control.

5. Discussion

When an external data source for the control arm of an RCT is available, one can synthesize the external data and the RCT by augmenting its control arm with the external data via proper study design and statistical analyses for making regulatory decisions. It is important to bear in mind that this approach can be applied only if the synthesis of those two types of data is considered reasonable from a clinical and regulatory point of view, and the external data source needs to be of high quality (U.S. Food and Drug Administration 2017).

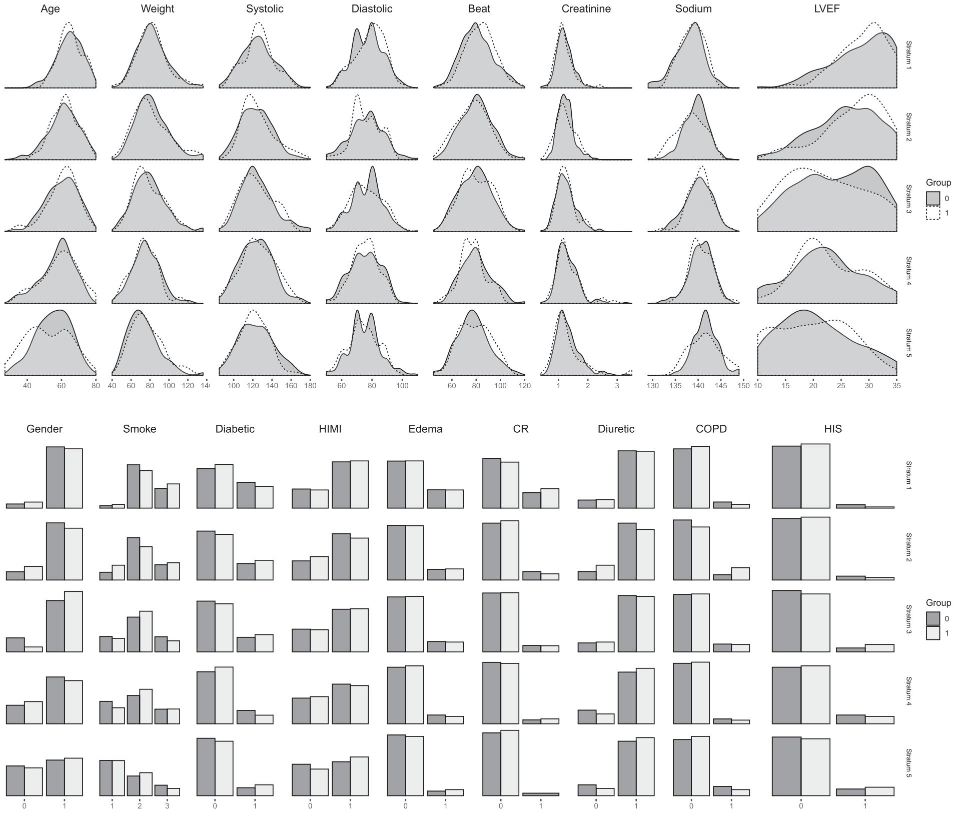

Figure 2. Balance checking of covariates between the registry study (Group = 0) and the current study (Group = 1).

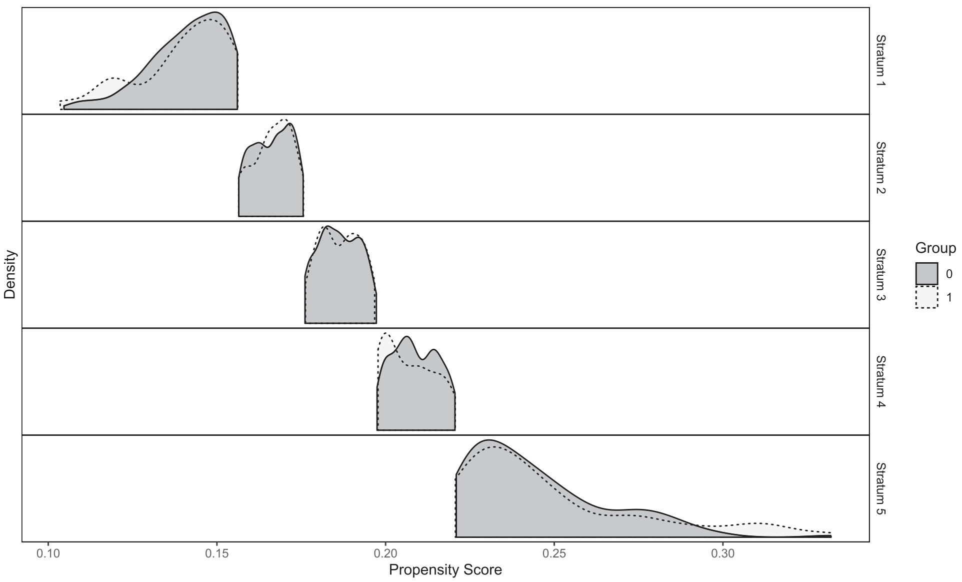

Figure 3. Densities of PS for the registry study (Group = 0) and the current study (Group = 1).

Table 3. Estimate of treatment effect and the associated standard errors (SEs).

	Stratum					
	1	2	3	4	5	Overall
Estimate	−0.16	−0.11	−0.23	−0.30	−0.10	−0.18
SE	0.09	0.09	0.09	0.10	0.09	0.04

The PSCL approach to incorporating RWD into the control arm of an RCT applies the propensity score methodology to design the study and utilizes the composite likelihood technique to analyze outcome data. The propensity score stratification in the proposed approach enables us to leverage information of comparable RWD patients within each stratum. The composite likelihood technique allows us to perform outcome data analysis, incorporating information obtained from RWD. It should be noted the proposed method for augmenting the control arm of an RCT can be applied to augment the treated (or investigational) arm of an RCT by incorporating RWD, for example when indication expansion is sought for an investigational device and its off-label use is captured in a registry or outside the United States.

To study the performance of the PSCL approach, we conducted simulations for various outcome types, sample sizes of RCT, amount of borrowing from RWD, and stratification strategies. The results show that the proposed approach with propensity score stratification reduces the bias in addition to reducing MSE for the estimation of treatment effect.

6. Implementation and software

We have implemented the proposed method for normal and binary outcomes in an R package named psrwd. The package is hosted at github and can be installed in R by install_github("olssol/psrwd").

Acknowledgments

The author (CW) is a consultant for the U.S. Food and Drug Administration on this project.

ORCID

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

References

Agostino, R. B. D., 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.

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.

Buuren, S. V., and K. Groothuis-Oudshoorn. 2010. mice: Multivariate imputation by chained equations inR. *Journal of statistical software* 45 (3):1–67. doi:10.18637/jss.v045.i03.

Campbell, G., and L. Q. Yue. 2016. Statistical innovations in the medical device world sparked by the FDA. *Journal of biopharmaceutical statistics* 26 (1):3–16. doi:10.1080/10543406.2015.1092037.

Godambe, V. P. 1960. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* 31 (4):1208–1211. doi:10.1214/aoms/1177705693.

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.

Lin, J., M. Gamalo-Siebers, and R. Tiwari. 2018a. Propensity-score-based priors for bayesian augmented control design. *Pharmaceutical statistics* 18 (2):223–238.

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.

Lindsay, B. G. 1988. Composite likelihood methods. *Contemporary mathematics* 80 (1):221–239.

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/sim.2739.

Rubin, D. B. 2008. For objective causal inference, design trumps analysis. *The annals of applied statistics* 2:808–840. 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. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices

Varin, C., N. Reid, and D. Firth. 2011. An overview of composite likelihood methods. *Statistica Sinica* 21 (1):5–42.

Xu, Y., N. Lu, L. Q. Yue, and R. Tiwari. 2019. A study design for augmenting the control group in a randomized controlled trial: A quality process for interaction among stakeholders. *Therapeutic Innovation & Regulatory Science* 216847901983038. doi:10.1177/2168479019830385.

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.