



## Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies

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

To cite this article: Chenguang Wang, Nelson Lu, Wei-Chen Chen, Heng Li, Ram Tiwari, Yunling Xu & Lilly Q. Yue (2019): Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies, Journal of Biopharmaceutical Statistics, DOI: [10.1080/10543406.2019.1684309](https://doi.org/10.1080/10543406.2019.1684309)

To link to this article: <https://doi.org/10.1080/10543406.2019.1684309>



Published online: 10 Nov 2019.



Submit your article to this journal [↗](#)



Article views: 2




View related articles [↗](#)



View Crossmark data [↗](#)



# Propensity score-integrated composite likelihood approach for incorporating real-world evidence in single-arm clinical studies

Chenguang Wang <sup>a</sup>, Nelson Lu<sup>b</sup>, Wei-Chen Chen<sup>b</sup>, Heng Li<sup>b</sup>, Ram Tiwari<sup>b</sup>, Yunling Xu<sup>b</sup>, and Lilly Q. Yue<sup>b</sup>

<sup>a</sup>Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland, USA; <sup>b</sup>Division of Biostatistics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

## ABSTRACT

In medical product development, there has been an increased interest in utilizing real-world data which have become abundant with recent advances in biomedical science, information technology, and engineering. High-quality real-world data may be analyzed to generate real-world evidence that can be utilized in the regulatory and healthcare decision-making. In this paper, we consider the case in which a single-arm clinical study, viewed as the primary data source, is supplemented with patients from a real-world data source containing both clinical outcome and covariate data at the patient-level. Propensity score methodology is used to identify real-world data patients that are similar to those in the single-arm study in terms of the baseline characteristics, and to stratify these patients into strata based on the proximity of the propensity scores. In each stratum, a composite likelihood function of a parameter of interest is constructed by down-weighting the information from the real-world data source, and an estimate of the stratum-specific parameter is 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 example based on our experience is provided to illustrate the implementation of the proposed approach.

## ARTICLE HISTORY

Received 6 March 2019


Accepted 13 October 2019

## KEYWORDS

Covariate balance; overlapping coefficient; composite likelihood; propensity score; PSCL; 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 sources such 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 be 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.

**CONTACT** Lilly Q. Yue  [Lilly.Yue@fda.hhs.gov](mailto: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, Maryland 10903, USA

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/lbps](http://www.tandfonline.com/lbps).

© 2019 Taylor & Francis Group, LLC

In this paper, we develop a statistical methodology for leveraging RWD to augment a prospective, single-arm clinical study to save the sample size of the latter (to be called the current study). Such a methodology has a wide range of applications. For example, it may be applicable when a prospective, single-arm clinical study is to be conducted to support an indication extension for an approved medical device, where the device has been used off-label for the extended indication in clinical practice, the data of such off-label use are collected in a registry database (the source of RWD), and the registry is relevant to the questions under consideration and is sufficiently reliable. When leveraging RWD to augment the current study, two critical questions arise: the first concerns the similarity between the patient population constituting the RWD and that of the current study (intuitively, leveraging is less justifiable if those two populations are dissimilar), and the second concerns the amount of information the RWD are to contribute to the statistical inference. To address the first question, let us first explain what we mean by similarity between two patient populations. Two individual patients are similar if they have similar characteristics in terms of demographics, health status, genetic background, etc. Likewise, we consider two patient populations to be similar if the distributions of those characteristics in one population are close to that in the other. Of course, if a patient characteristic is unmeasured, we do not know its distribution. Therefore, only the distribution of measured patient characteristics, or baseline covariates, is used to define similarity. As part of the proposed methodology, we use propensity score stratification to construct strata within which the group of patients from the current study is more similar to the group of patients constituting RWD than the two groups are overall. This is achieved in the same manner in which treatment and control groups are rendered more similar in terms of measured baseline covariates after propensity score stratification in the causal inference setting. Data leveraging is then implemented within each propensity score stratum. Finally, we combine stratum-specific statistical inference.

In practice, clinical judgment provides an answer to the second question, i.e., how much do RWD contribute to the statistical inference for the parameter of interest. Typically, it is appropriate to “down-weight” or “discount” the RWD. Here we propose to use composite likelihood to realize the down-weighting (discounting). Given its theoretical foundation, the proposed method will be called the propensity score-integrated composite likelihood approach (or PS-integrated composite likelihood, in short, PSCL).

The rest of the paper is organized as follows. In [Section 2](#), we briefly review the propensity score methodology and the composite likelihood technique, and introduce the PSCL approach. In [Section 3](#), we present the results of simulation studies. Included in [Section 4](#) is a hypothetical example to illustrate our proposed approach. [Section 5](#) concludes with some discussion points.

## 2. Method

### 2.1. Notation

For patient  $i$ , let  $Y_i$  be the random variable that represents the outcome and  $y_i$  be the realization of  $Y_i$ . Let  $\mathbf{X}_i$  denote the vector with dimension  $p \times 1$  of covariates for patient  $i$ . Let  $Z_i = 1$  if patient  $i$  is from the current study and 0 if patient  $i$  is from RWD, or more generally external data (i.e. data external to the current study), and let  $n_1$  be the number of patients in the current study. In this paper, we assume that there is only one external data source and covariates collected in the current study are contained in the external data. Furthermore, we assume that all patients with  $Z_i = 0$  satisfy the inclusion and exclusion criteria of the current study. Since the external data considered in this paper are from RWD sources such as registries, we assume that the number of patients in the external data is sufficiently large for the proposed method in [Section 2.4](#).

### 2.2. Propensity score methodology

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

$$e(X) = \Pr(Z = 1|X).$$

The propensity score  $e(\mathbf{X})$  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. 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; Lin et al. 2018a; Lin et al. 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 Jr 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 (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 this paper, the propensity score methodology is used for the purpose of leveraging RWD to augment the current study instead of causal inference. We focus on the technique of stratification, which consists of grouping patients with similar propensity scores into strata (e.g., five propensity score quintiles). Here patients from the current study are labeled  $Z = 1$  and patients from the RWD source are labeled  $Z = 0$ . Due to the balancing property of propensity score, within each stratum patients from the RWD source are expected to be more similar to those in the current study in terms of baseline covariates than they are overall, which makes leveraging RWD within strata more justifiable. In this context, the propensity score methodology also allows us to separate study design and analysis of outcome data, so it can be applied in a regulatory setting.

### 2.3. Using composite likelihood to discount external data

When combining information from multiple data sources for statistical inference, simply pooling all the data without addressing the heterogeneity between data sources is in general not a valid strategy. In the setting of this paper, where we have the current study and an external data source, it is often desirable to “discount” 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 parameters  $\boldsymbol{\omega} = (\theta, \boldsymbol{\phi})$ , where  $\theta$  represents the parameter of interest and  $\boldsymbol{\phi}$  represents the nuisance parameters. In order to conduct statistical inference for  $\theta$  based on the outcome data from 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:

$$L(\boldsymbol{\omega}; \mathbf{Y}) = \prod_{i=1}^n f(y_i; \boldsymbol{\omega})^{\Gamma_i}. \quad (1)$$

where  $\Gamma_i$ ’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  $\hat{\boldsymbol{\omega}}$  be the maximum likelihood estimator of  $\boldsymbol{\omega}$  based on (1). In general,  $\hat{\boldsymbol{\omega}}$  can be obtained by solving the equation

$$\Psi_n(\boldsymbol{\omega}) = \sum_{i=1}^n \Gamma_i \frac{\partial \log f(y_i; \boldsymbol{\omega})}{\partial \boldsymbol{\omega}} = 0. \quad (2)$$

Note if  $\omega = (\omega_1 = \theta, \omega_2 = \phi_1, \dots, \omega_k = \phi_{k-1})$  is  $k$ -dimensional, (2) represents the set of equations

$$\sum_{i=1}^n \Gamma_i \frac{\partial \log f(y_i; \omega)}{\partial \omega_j} = 0 \quad j = 1, \dots, k.$$

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

$$\sqrt{n}(\hat{\omega} - \omega) \xrightarrow{d} \text{MVN}(0, G^{-1}(\omega))$$

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

$$G(\omega) = E_{\omega} \left( -\frac{\partial \Psi_n(\omega)}{\partial \omega} \right) (\text{Var}_{\omega} \Psi_n(\omega))^{-1} E_{\omega} \left( -\frac{\partial \Psi_n(\omega)}{\partial \omega} \right).$$

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

$$\text{Var } \hat{\omega} = \frac{n-1}{n} \sum_{i=1}^n (\hat{\omega}^{(-i)} - \hat{\omega})(\hat{\omega}^{(-i)} - \hat{\omega})^T \quad (3)$$

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

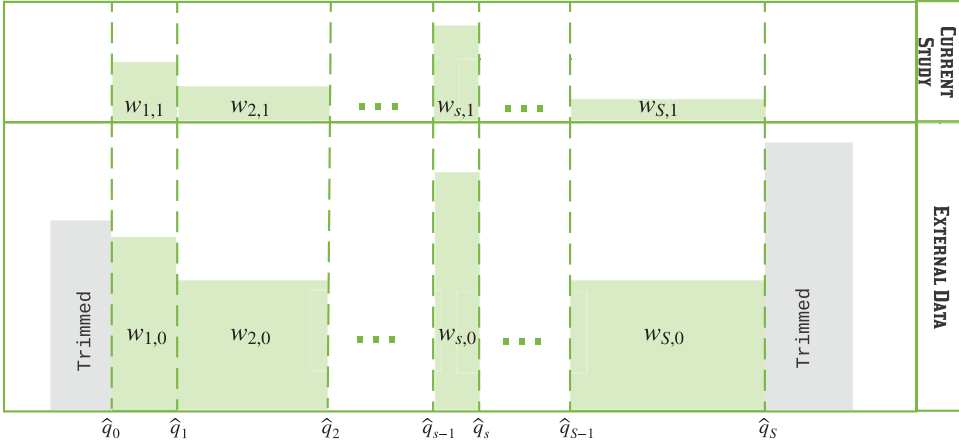
## 2.4. The PS-integrated composite likelihood approach

In this section, we will introduce PS-integrated composite likelihood (PSCL) approach for leveraging RWD. This approach is implemented in two parts, the study design part and the outcome analysis part. The study design part is performed in two stages (Yue et al. 2014). Below is a description of the activities in the first design stage and steps constituting the second design stage. It should be pointed out that no outcome data are needed, nor should they be accessed, during study design.

The first design stage includes the following activities: 1) formulating the statistical hypotheses, 2) specifying a comprehensive list of pertinent covariates to be measured, 3) choosing an appropriate RWD source to be used to augment the current study, 4) estimating the sample size, and 5) identifying an independent statistician to carry out the second design stage.

The first step of the second design stage is to identify the patients in the external data that are similar to the patients in the current study using PS (Figure 1). To apply the PS 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 be defined as the probability of a patient coming from the current study as opposed to the external data. 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  $\mathbb{E}_1$  be the set of the estimated PSs of the patients in the current study. We identify the patients in the external data that are similar to the patients in the current study by excluding patients in the external data source with PS outside of the range of  $\mathbb{E}_1$ , which is referred to as *trimming*. In general, excluding patients may distort the patient population of interest and introduce bias. However, this is not a concern here since the interest is in the patient population represented by the current study, which is not altered by *trimming*.

The second step of the second design stage is propensity score stratification. Let  $\hat{q}_0 < \hat{q}_1 < \hat{q}_2 < \dots < \hat{q}_{S-1} < \hat{q}_S$  be a set of cut points based on which  $S$  propensity score strata are formed,  $\hat{e}_i$  be the estimated propensity score for patient  $i$ ,  $w_{s,0} = \{i : \hat{e}_i \in (\hat{q}_{s-1}, \hat{q}_s], Z_i = 0\}$  and  $w_{s,1} = \{i : \hat{e}_i \in (\hat{q}_{s-1}, \hat{q}_s], Z_i = 1\}$  (i.e.,  $w_{s,0}$  is the index set for patients from external data in the  $s$ th stratum, and  $w_{s,1}$  is the index set for the patients from the current study). In this paper, the inner cut points  $\hat{q}_1, \hat{q}_2, \dots, \hat{q}_{S-1}$  are chosen such that the number of patients from the current study in each stratum are equal (the size of  $w_{s,1}$  are the same for all  $s$ ).



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

Let  $\theta_s$  be the parameter of interest in stratum  $s$ . The third step of the second design stage is to specify the form of the composite likelihood function for  $\theta_s$  for all  $s = 1, \dots, S$ . We specify the composite likelihood of  $\theta_s$  to be of the following form:

$$L(\theta_s) = \prod_{i \in w_{s,0} \cup w_{s,1}} f(y_i; \theta_s)^{\Gamma_i} = \prod_{i \in w_{s,1}} f(y_i; \theta_s) \prod_{i \in w_{s,0}} f(y_i; \theta_s)^{\lambda_{s,0}/n_{s,0}}, \quad (4)$$

where  $n_{s,0}$  is the number of patients in  $w_{s,0}$  and  $\lambda_{s,0}$  is the design parameter given to external data in stratum  $s$ . We allow nuisance parameters to be different between the current study and RWD, but, for simplicity, we suppress the dependence on nuisance parameters in (4). Note that here we are only specifying the form of the likelihood function, so no outcome data are needed.

The fourth step of the second design stage is to determine the value of  $\lambda_{s,0}$  for all  $s$ . We propose that the amount of information contributed by the patients in  $w_{s,0}$  be related to the similarity between the patients in  $w_{s,0}$  and in  $w_{s,1}$  with respect to the baseline covariates. We use the overlapping coefficient (Inman and Bradley Jr, 1989)  $r_s$  as the similarity measure. The overlapping coefficient

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

is the overlapping area of the density curves  $g_{s,0}$  and  $g_{s,1}$  of the  $\hat{e}_i$  for  $i$  in  $w_{s,0}$  and  $w_{s,1}$ , respectively. The overall number of patients to be borrowed from the external data, denoted by  $A$ , is elicited as a fixed constant. Since we want  $\lambda_{s,0}$  to be proportional to  $r_s$  but not exceed  $n_{s,0}$ , we let

$$\lambda_{s,0} = \min \left( \left\lceil \frac{A r_s}{\sum_{s=1}^S r_s} \right\rceil, n_{s,0} \right) \quad (6)$$

where  $\lceil u \rceil$  is the integer closest to  $u$ .

## 2.5. Inference

The parameter of interest,  $\theta$ , can be estimated as a weighted average with weights proportional to  $n_{s,1}$ , the size of  $w_{s,1}$ . Since in this paper the size of  $w_{s,1}$  is the same for all  $s$ , we have

$$\hat{\theta} = \frac{1}{S} \sum_{s=1}^S \hat{\theta}_s \quad (7)$$

where  $\hat{\theta}_s$  is the maximum likelihood estimator based on maximizing (4) with  $\lambda_{s,0}$  given by (6). Furthermore, the variance of  $\hat{\theta}$  can be computed as

$$\text{Var } \hat{\theta} = \frac{1}{S^2} \sum_{s=1}^S \text{Var } \hat{\theta}_s \quad (8)$$

where  $\text{Var } \hat{\theta}_s$  can be obtained following (3). With (7) and (8), a Wald test can be constructed to infer the true value of  $\theta$ .

### 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

We consider the following simulation study settings. Covariates  $\mathbf{X}$  are assumed to follow a mixture of multivariate normal distribution of  $K$  components such that  $\mathbf{X}_{p \times 1} | Z \sim F_Z$  and

$$F_Z = \sum_{k=1}^K \psi_{Zk} \text{MVN}(\boldsymbol{\mu}_{Zk}, \boldsymbol{\Sigma}_Z)$$

with  $\sum_{k=1}^K \psi_{Zk} = 1$ , the diagonal of  $\boldsymbol{\Sigma}_Z$  being  $\sigma_Z^2$ , and the off-diagonal elements of  $\boldsymbol{\Sigma}_Z$  being  $\rho\sigma_Z^2$ .

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

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

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

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

and simulate  $Y_i$  from its binary distribution.

For all the simulation studies, we set  $\rho = 0.1$ . We convert  $X_1, \dots, X_4$  to binary covariates using cut point 0. For continuous outcomes, we set  $\boldsymbol{\beta} = \mathbf{1}_{p \times 1}$  and  $\beta_0 = 0$ . For binary outcomes, we set  $\boldsymbol{\beta} = \mathbf{1}_{p \times 1}$  and choose  $\beta_0$  such that  $E(Y | Z = 1) = 0.4$ . Note that we assume  $\beta_0$  and  $\boldsymbol{\beta}$  are identical for  $Z = 0$  and 1. 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. In this scenario, we set  $K = 1$ ,  $\boldsymbol{\mu}_{01} = \mathbf{1.2}_{p \times 1}$ ,  $\boldsymbol{\mu}_{11} = \mathbf{1}_{p \times 1}$ . 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  $K = 1$ ,  $\boldsymbol{\mu}_{11} = \mathbf{1}_{p \times 1}$ , and  $\sigma_1^2 = 1$ . For  $Z = 0$ , we set  $K = 2$ ,  $\psi_{01} = \psi_{02} = 0.5$ ,  $\boldsymbol{\mu}_{01} = \mathbf{1}_{p \times 1}$ ,  $\boldsymbol{\mu}_{02} = \mathbf{1.5}_{p \times 1}$ , and  $\sigma_0^2 = 1$ . For each scenario, we consider the dimension of  $\mathbf{X}$  to be  $p = 10$ , sample sizes of the current study to be  $n_1 = 200$  and 400, and set  $A$  to be 20 and 40 for  $n_1 = 200$ , and 40 and 80 for  $n_1 = 400$ .

We compare two strategies for the PSCL approach. The first strategy, *No PS Stra.*, sets the number of strata  $S = 1$ , and the design parameter  $\lambda_{1,0} = \min(A, n_{1,0})$  where  $A$  is the intended number of patients to be borrowed from the external data source after trimming. The second strategy, *PS Stra.*, sets the number of strata  $S = 5$  and chooses the design parameters  $\lambda_{s,0}$  following (6).

3.2. Results

In Table 1, we report for each strategy the true  $\theta$ , the estimate of  $\theta$ , bias, and mean squared error (MSE) in Scenarios I and II. The *true* value of  $\theta$  is set as the mean of the outcome in the current study (i.e. 0.40 and 9.36 for binary and continuous outcomes, respectively). We also report  $\bar{r}_s = \frac{1}{5} \sum_{s=1}^5 r_s$  to assess the PS overlapping in each scenario. The simulation results are based on 1,000 replications. Based on the simulation configurations, the estimates of  $\theta$  by the current study or the external data only are reported in Table 2.

Several observations can be made. First, the *PS Stra.* strategy outperforms the *No PS Stra.* strategy in terms of bias in all the cases. For binary outcomes, in Scenario I, the *PS Stra.* strategy has smaller MSE compared to the *No PS Stra.* strategy. As  $A$  increases, the reduction in MSE by the *PS Stra.* strategy is larger compared to the *No PS Stra.* strategy. In Scenario II, we assume patients in the external data were from a mixture distribution, so that the baseline covariates are considered to be more heterogeneous comparing to Scenario I in general. The difference between the two strategies are more obvious in Scenario II compared to Scenario I, favoring the *PS Stra.* strategy. In the case of continuous outcomes, what we observe are mainly the same as in the case of binary outcomes: the *PS Stra.* strategy gives smaller bias and MSE compared to the *No PS Stra.* strategy. In summary, the simulation results clearly show the advantage of stratification compared to no stratification.

Table 1. Simulation study results.  $\bar{r}_s = \frac{1}{5} \sum_{s=1}^5 r_s$ .

Scenario	$\theta$	$n_1$	$\bar{r}_s$	A	Model	$\hat{\theta}$	Bias ( $\times 100$ )	MSE ( $\times 100$ )
Binary Outcomes								
I	0.4	200	0.799	20	No PS Stra.	0.408	0.792	0.103
					PS Stra.	0.400	0.043	0.112
				40	No PS Stra.	0.416	1.582	0.107
					PS Stra.	0.402	0.235	0.109
		400	0.840	40	No PS Stra.	0.409	0.899	0.060
					PS Stra.	0.402	0.159	0.060
				80	No PS Stra.	0.417	1.678	0.072
					PS Stra.	0.403	0.346	0.059
II	0.4	200	0.798	20	No PS Stra.	0.417	1.713	0.128
					PS Stra.	0.405	0.545	0.118
				40	No PS Stra.	0.429	2.946	0.170
					PS Stra.	0.408	0.844	0.118
		400	0.844	40	No PS Stra.	0.415	1.517	0.079
					PS Stra.	0.403	0.305	0.066
				80	No PS Stra.	0.428	2.792	0.125
					PS Stra.	0.406	0.611	0.066
Continous Outcomes								
I	9.36	200	0.799	20	No PS Stra.	9.464	9.624	5.780
					PS Stra.	9.393	2.545	5.536
				40	No PS Stra.	9.539	17.096	7.186
					PS Stra.	9.411	4.361	5.505
		400	0.840	40	No PS Stra.	9.451	8.722	3.270
					PS Stra.	9.380	1.642	2.885
				80	No PS Stra.	9.526	16.167	4.799
					PS Stra.	9.398	3.427	2.893
II	9.36	200	0.799	20	No PS Stra.	9.501	13.650	7.106
					PS Stra.	9.392	2.685	6.048
				40	No PS Stra.	9.617	25.231	10.921
					PS Stra.	9.420	5.506	6.098
		400	0.844	40	No PS Stra.	9.504	14.135	4.326
					PS Stra.	9.390	2.745	2.731
				80	No PS Stra.	9.624	26.107	8.840
					PS Stra.	9.419	5.610	2.889



**Table 2.** Estimate of  $\theta$  by current study only or external data only.

Scenario	Outcome Type	$\hat{\theta}$	
		External Data Only	Current Study Only
I	Binary	0.50	0.40
II	Binary	0.57	0.40
I	Continuous	10.34	9.36
II	Continuous	10.94	9.36

**4. An illustrative example**

In this section, we use a straw man example to illustrate how the PSCL approach could be implemented when leveraging RWD. It involves a single-arm clinical study conducted to demonstrate the safety and effectiveness of a cardiovascular device in patients with congestive heart failure, and a device registry that could be appropriately leveraged to augment the clinical study.

**4.1. The first design stage**

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

$$H_0 : \theta \geq 35\% \quad \text{vs.} \quad H_a : \theta < 35\%$$

where  $\theta$  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. All these key covariates and the outcome information are also collected in the registry.

In order to determine the sample size,  $\theta$  was assumed to be 0.29. At the significance level of 0.05, a power of 80% would require a sample size of approximately 380. It was proposed to enroll 300 patients in the current study and borrow 80 patients from the registry. That is, the value of  $A$  is set at 80. Note that the number of patients to be borrowed should be determined mainly based on clinical judgment on a case by case basis in practice.

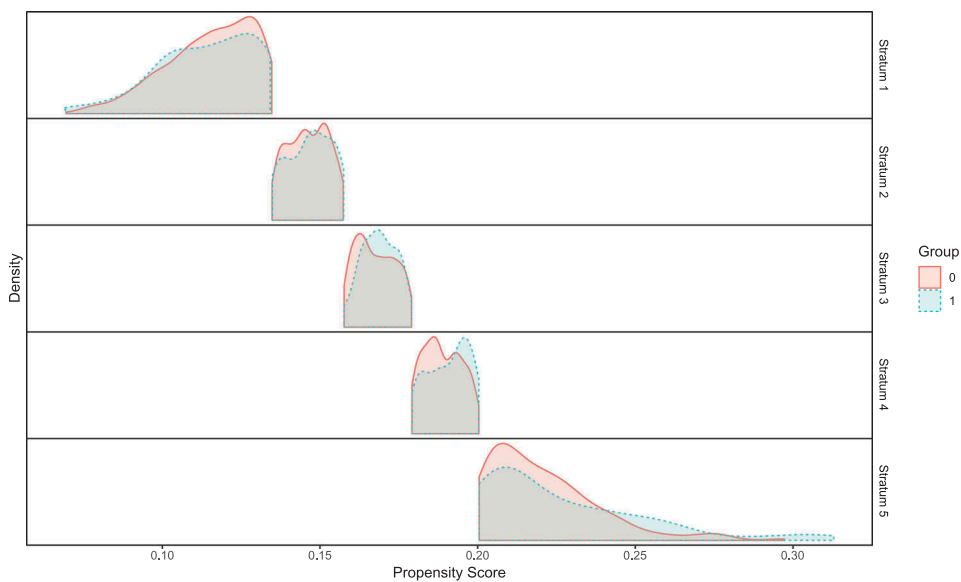
**4.2. The second design stage**

When all the 300 patients were enrolled, and all their covariate data had been collected, the second design stage was started. Based on the pre-specified inclusion/exclusion criteria of the current investigational study, a total of 1,787 patients from the registry were identified as potential patients to be borrowed.

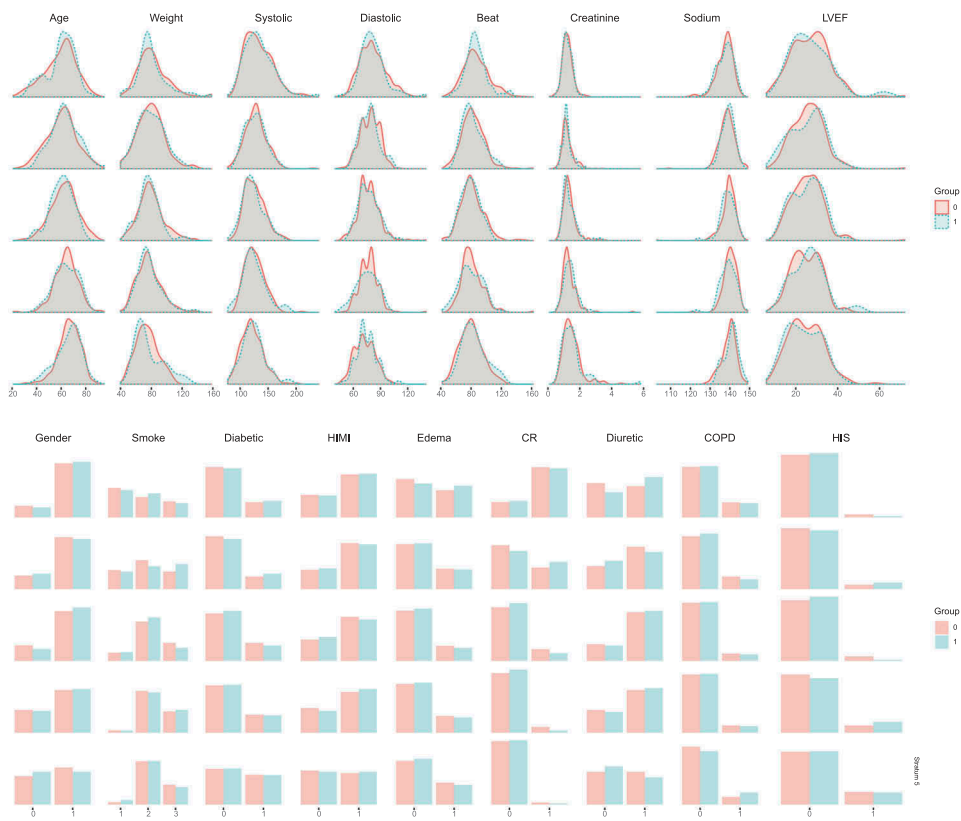
Based on the 300 patients in the current study and 1,787 external patients, propensity scores were estimated using logistic regression with all 17 baseline covariates included in their linear terms. Missing data in the covariates (about 6%) were imputed using the classification and regression trees (CART) method implemented in the R package mice. 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, 1,575 out of 1,787 patients were retained for the study design and outcome analysis. Five propensity score strata were formed for all the patients (300 + 1575) with each stratum containing equal number of patients in the current study. Then, overlapping coefficients of the propensity score distributions defined in Section 2.4 are calculated for all strata (Figure 2). The balance of each covariate distribution in each stratum was examined visually between the current study and the external data (Figure 3). Note that the balance can also be assessed quantitatively by metrics such as the standardized mean difference (Franklin et al., 2014).

In summary, the second stage determined the following features: the PS strata, how many external patients were “borrowed” for each stratum, as well as how much information each external patient contributed (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 design parameters  $\lambda_{s,0}$  for all  $s$ .

		Stratum					Total
		$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	
Current Study	$n_{s,1}$	60	60	60	60	60	300
Registry	$n_{s,0}$	434	344	369	228	198	1575
Overlapping Coefficient	$r_s$	0.85	0.83	0.81	0.75	0.78	
	$r_s / \sum r_s$	21%	21%	20%	19%	19%	100%
Approx. Patients Borrowed	$Ar_s / \sum r_s$	17	16	16	15	16	80
	$\lambda_{s,0} / n_{s,0}$	0.04	0.05	0.04	0.07	0.08	

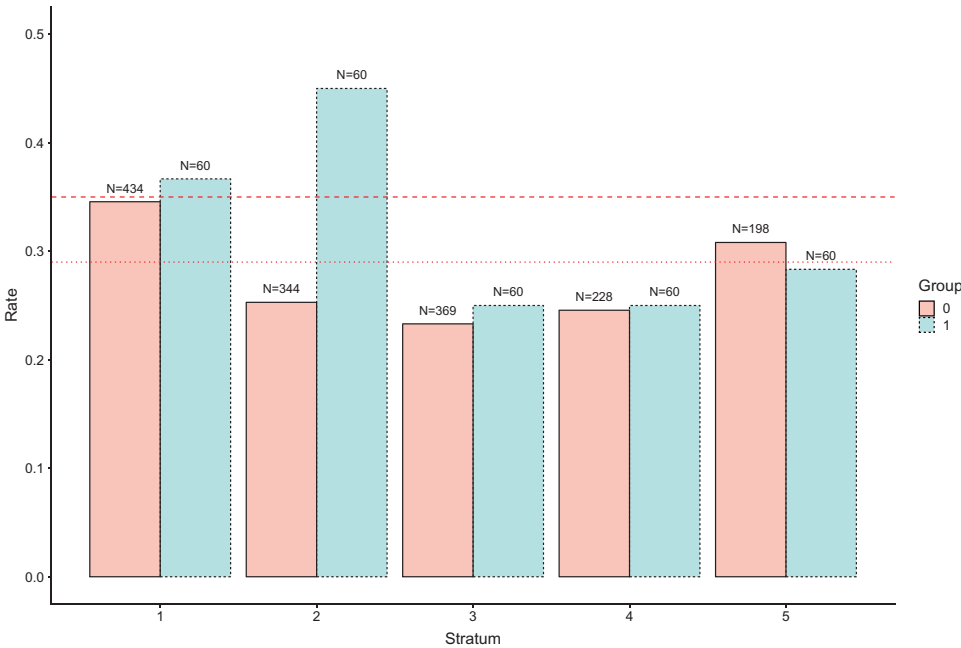
### 4.3. Outcome analysis

The final analysis was conducted after the clinical outcome had been collected from all the 300 patients in the current study. The averages of outcomes  $Y_i$  are 0.32 and 0.28 based on the current study and the external data only, respectively. Figure 4 presents the point estimates for  $\theta_s$ , on the current study and the external data source for each stratum. Overall there are no large discrepancies between those two point estimates in most strata except for Stratum 2.

Based on the analysis with  $A = 80$ , the estimates of  $\theta_s$  and  $\theta$  and the associated standard errors are reported in Table 4. The  $p$ -value from the Wald test is 0.033, which is significant. Based on the results, we conclude that the hypothetical medical device meets the study success criterion.

## 5. Discussion

We developed an analytical approach to incorporating RWD into a single-arm clinical study (the current study) by utilizing the composite likelihood technique to analyze outcome data while applying propensity score methodology to design the study. The RWD source should be of acceptable quality and considered a similar data source to the current study. As the amount of missing data is an important component of



**Figure 4.** Stratum-specific observed clinical outcome in the current study and the registry study. The reference horizontal lines correspond to 0.29 and 0.35, respectively.

**Table 4.** Estimates and associated standard errors (SEs).

Method		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}$
PSCL	Estimate	0.36	0.41	0.25	0.25	0.29	0.31
	SE	0.05	0.05	0.04	0.05	0.05	0.02

data quality, it should be taken into consideration when determining whether an RWD data source is suitable for being leveraged to augment the current study. In reality, some missing data are inevitable. For some advice on handling missing data in the implementation of propensity score methodology see D’Agostino Jr and Rubin (2000) and Mitra and Reiter (2016).

In our proposed method, we first apply the propensity score methodology to identify the RWD patients relevant to the current study. We then group patients into strata based on their propensity scores. Within each stratum, RWD patients are down-weighted based on the similarity of the two data sources, measured by the overlap of propensity score distributions of the two groups based on patients in that stratum. A composite likelihood is then constructed to analyze outcome data.

In evaluating data in the scenario that we consider, an analysis based on direct pooling of current study and RWD without addressing heterogeneity between the current study and the real-world data source may not be an appropriate strategy. Bias may be introduced due to differences in the patient baseline characteristics and other aspects such as time, medical practice, adjunct therapy, measurements in endpoint, data collection mechanisms, etc. Bias due to imbalance of measured patient characteristics would be mitigated via the application of propensity score methodology, while the impact of other factors requires clinical assessment. Such assessment determines the maximum amount of borrowed information from the RWD. A smaller amount of borrowing is warranted if there is less confidence that the two data sources are comparable regarding various factors from the clinical perspective.

When there is a large number of patients from the RWD source who meet the inclusion/exclusion criteria pre-specified in the current investigational study, which is usually the case, one will likely be able to borrow the pre-specified number of patients from the RWD source so that the study power is maintained. In the unlikely event that one is not able to select enough patients from the RWD source to augment the current study (e.g. due to the lack of overlap between propensity score distributions), one may increase the enrollment into the current study. Given that typically it is relatively expensive to enroll patients into a medical device clinical study, incorporating RWD may result in saving a substantial amount of money. Perhaps more importantly, when enrollment is not very fast, the incorporation of RWD could shorten the duration of the study by months if not longer, which would be a considerable amount of time.

The objective of this paper was to provide a frequentist approach when it is desired to leverage the information from an appropriate RWD source, and the amount of borrowing is determined by similarity of patient covariates in combination with clinical judgment. Our simulation studies suggest that stratification on propensity score tend to improve the performance in terms of both bias and MSE. It should be noted that after propensity score stratification we may observe variability in the outcome across strata. Outcome heterogeneity is an important phenomenon in clinical studies in general. In practice, when heterogeneity is observed, it is advisable to investigate whether it could be due to chance and to investigate its source.

It cannot be emphasized enough that an outcome-free design paradigm needs to be implemented when applying our approach under a regulatory framework. There should be no knowledge about outcome measures of RWD and the primary data source (the current study) during the propensity score stratification process. Without such a practice, PS strata could be formed in such a way that a favorable outcome analysis result is selected, and the objectivity and validity of the study are undermined.

Our proposed approach may be extended to other scenarios. For example, currently we only consider a single RWD source to be leveraged. In practice, it is possible that multiple data sources can be leveraged. As another example, an extension is needed for a case in which the current study is a randomized clinical trial, and RWD are to be leveraged to supplement one or both treatment groups.

## 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 first author (CW) was a consultant for the Food and Drug Administration on this project and was compensated for his consultation services.

## 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.
- 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.
- Franklin, J. M., J. A. Rassen, D. Ackermann, D. B. Bartels, and S. Schneeweiss. 2014. Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine* 33 (10):1685–1699. doi:10.1002/sim.6058.
- 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. doi:10.1002/sim.3782.
- 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.
- 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.
- Mitra, R., and J. P. Reiter. 2016. A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research* 25 (1):188–204. doi:10.1177/0962280212445945.
- National Research Council. 2010. *The prevention and treatment of missing data in clinical trials*. Washington (DC): The National Academies Press.
- 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/(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.
- 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>.
- Varin, C., N. Reid, and D. Firth. 2011. An overview of composite likelihood methods. *Statistica Sinica* 21 (1):5–42.
- 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](https://doi.org/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](https://doi.org/10.1080/10543406.2014.926367).