Received October 2017; | Revised ; | Accepted
DOI: xxx/xxxx

RESEARCH ARTICLE

Safety Surveillance and the Estimation of Risk in Select Populations: Flexible Methods to Control for Confounding while Targeting Marginal Comparisons via Standardization

Xu Shi[1] | **Robert Wellman**[3] | **Patrick J. Heagerty**[2] | **Jennifer C. Nelson**[2,3] | **Andrea J. Cook**[2,3]

[1]Department of Biostatistics, University of Michigan, MI, USA
[2]Department of Biostatistics, University of Washington, WA, USA
[3]Biostatistics Unit, Kaiser Permanente Washington Health Research Institute, WA, USA

Correspondence
Xu Shi, Email: shixu@umich.edu

Present Address
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Summary

We consider the critical problem of pharmacosurveillance for adverse events once a drug or medical product is incorporated into routine clinical care. When making inference on comparative safety using large-scale electronic health records, we often encounter an extremely rare binary adverse outcome with a large number of potential confounders. In this context, it is challenging to offer flexible methods to adjust for high-dimensional confounders, whereas use of the propensity score can help address this challenge by providing both confounding control and dimension reduction.

Among propensity score methods, regression adjustment using the propensity score as a covariate in an outcome model has been incompletely studied and potentially misused. Previous studies have suggested that simple linear adjustment may not provide sufficient control of confounding. Moreover, no formal representation of the statistical procedure and associated inference has been detailed. In this paper, we characterize a three-step procedure which performs flexible regression adjustment of the estimated propensity score followed by standardization to estimate the causal effect in a select population. We also propose a simple variance estimation method for performing inference. Through a realistic simulation mimicking data from the FDA Sentinel Initiative comparing the effect of angiotensin-converting enzyme inhibitors and beta-blockers on incidence of angioedema, we show that flexible regression on the propensity score resulted in less bias without loss of efficiency, and can outperform other methods when the propensity score model is correctly specified. In addition, the direct variance estimation method is a computationally fast and reliable approach for inference.

KEYWORDS:
causal inference, electronic health records, pharmacosurveillance, propensity score, rare adverse event

1 | INTRODUCTION

The increasing availability of electronic health record (EHR) and claims data has created the potential for population scale observational biomedical research. One transformative national effort is the Food and Drug Administration's (FDA) Sentinel Initiative that aims to monitor and evaluate the safety of all regulated medical products [1]. For example, a recent observational cohort study from the FDA Sentinel Initiative compared the effect of Angiotensin-Converting Enzyme Inhibitors (ACEI) and Beta Blockers (BB), two blood pressure control medications, on incidence of angioedema in the first 30 days after starting treatment. The FDA Sentinel system utilizes a distributed data network that provides access to electronic healthcare data for approximately 193 million patients. Data of this quantity is powerful for safety studies because it represents a broader population than typically enrolled in clinical trials, and it enables early detection of safety signals for less common adverse outcomes.

While the use of such large-scale healthcare data presents numerous opportunities for postmarketing safety research, there are also many inferential challenges. One key challenge is the need to control for a large number of potential confounders in EHR data, which is further complicated by the fact that an adverse event is often extremely rare. When the outcome is rare, using regression adjustment with a large number of covariates can result in model fitting issues such as non-convergence or extreme coefficients. Flexible nonparametric regression is even more challenging due to the well-known "curse of dimensionality" [2]. In contrast, if there is sufficient uptake of both the new medical product of interest and the control, usually a comparator medical product, then fitting a propensity score (PS) model with a large number of covariates may be more feasible.

Because use of propensity scores can provide both control of confounding and dimension reduction, it is attractive to consider propensity score methods in the postmarket surveillance setting [3]. One such propensity score method is the direct regression adjustment of the propensity score as a covariate in an outcome regression model. To estimate marginal, population-level contrasts that are the central focus of causal inference, the use of a regression model is often considered as the intermediate summary that is then used in a final standardization step, which takes the empirical average of the pair of predicted risks over the entire target population under hypothetical exposure and control conditions, also referred to as G-computation formula for point exposures [4, 5, 6].

Propensity score regression adjustment coupled with standardization has not been well-studied and is potentially underused. [7] and [8] have shown in simulation studies that regression adjustment on the propensity score can result in a biased effect estimate. However, in these studies, the propensity score was adjusted as a linear term, which may not fully capture the relationship between the outcome and the propensity score. Therefore, previously observed bias could be due to model misspecification or residual confounding, rather than the validity of the propensity score regression adjustment method [9, 10, 11, 12]. Because the propensity score and the outcome may not have a linear relationship, efforts have been made to relax model assumptions by flexible adjustment of the propensity score. In the context of missing data, [13] proposed the Propensity Penalized Spline Prediction (PSPP) method which adjusts for the propensity score using a penalized spline model, and in addition adjusts for covariates under parametric assumptions. Similarly, [14] viewed causal inference as a missing data problem and proposed multiple imputation with two subclassification splines (MITSS) method, which uses Bayesian modeling with a spline function of the propensity score and a simple linear adjustment for the covariates to impute the pair of potential outcomes. The rationale for both [13] and [14] is that the spline function provides flexibility in the propensity score adjustment which achieves balance and the covariate adjustment further improves efficiency. In the spirit of flexible adjustment of propensity score, [15] suggested to include quantiles of the propensity score as additional dummy variables, which is equivalent to adjusting for the propensity score in a step function, while [16] considered fitting a generalized additive model (GAM) of the treatment and the propensity score. The aforementioned methods either rely on fairly limited scenarios such as linear or log-linear models without treatment heterogeneity, which are generally collapsible [13, 16], or require adjustment for both covariates and the propensity score, which may not perform well in a rare outcome setting [13, 14, 15]. Moreover, when the number of adverse events are extremely rare, statistical inference using the bootstrap procedure can be challenging due to potential bootstrap samples with nearly zero event. However, to our knowledge there is no existing closed form standard error estimator for the propensity score regression adjustment method.

Despite the benefits and popularity of propensity score methods in postmarket surveillance, there have been few studies comparing the performance of propensity score methods in the setting of rare outcomes with many confounders. [17] compared propensity score-based estimators of the marginal relative risk mimicking confounders from EHR studies and assumed a relatively high event rate of 5% yielding 250 events among 5,000 patients. However, in postmarket surveillance, we often encounter less events with a much smaller event rate and a larger population, such as 50 events among 100,000 patients at an event rate of 0.05%. In addition, there was no statistical inference or variance estimation considered in their study.

In this paper, we focus on developing a propensity-based estimator for analyzing data with rare binary adverse events. In particular, we characterize a simple three-step procedure that provides flexible nonlinear adjustment of the estimated propensity score in an outcome regression model, followed by standardization to estimate a marginal causal effect in a select population. In particular, using this methodology, our inferential procedure targets common population-level contrasts that have a simple and direct causal interpretation, such as the causal risk difference, risk ratio, or odds ratio. In addition, we characterize variance estimation through adoption of influence functions that fully accounts for the uncertainty from propensity score estimation, outcome modeling, and standardization [18, 19]. We also propose a direct and simple variance estimator that is

particularly attractive under the rare outcome setting. We conduct a realistic simulation study by mimicking real data from the FDA Sentinel Initiative comparing the effect of ACEIs and BBs on incidence of angioedema in the first 30 days. We look at both the relative performance of different methods and the validity of our variance estimator.

Our paper is organized as follows. In Section 2 we detail flexible propensity score adjustment and provide an empirical variance estimator of the causal effect of interest. Section 3 presents an overview of existing methods for causal inference using the propensity score which estimate the exposure effect in a specific population. In Section 4, we conduct a simulation study to compare flexible regression adjustment of the propensity score with existing methods. In Section 5, we apply the various methods to the FDA Sentinel investigation comparing the effect of ACEIs and BBs on incidence of angioedema. We close with a discussion in Section 6. R code for implementing the methods is available at https://github.com/shixu0830/PSregress.

2 | FLEXIBLE REGRESSION ON THE PROPENSITY SCORE

2.1 | Notation and background

Causation is inferred by the difference in outcomes when all circumstances are the same except for one factor whose condition was changed. Accordingly, for each subject in a target population, there is a pair of variables $\big(Y(1),\ Y(0)\big)$ that characterizes the hypothetical outcomes that would have been observed under exposure and control, respectively. The main goal of causal inference is to provide a comparison of the population-level averages between the two potential (counterfactual) outcomes. For example, a common causal effect is the average treatment effect, defined as $E[Y(1)] - E[Y(0)]$.

In this paper, we focus on statistical inference for the average treatment effect (ATE). Let A denote the binary exposure, taking on value 1 (exposed) or 0 (unexposed). We denote the observed outcome as Y, with $Y = Y(1)$ if $A = 1$, and $Y = Y(0)$ if $A = 0$ under the consistency assumption. Hereafter we take $Y(1)$ as an example. Because only one outcome per subject can be observed at a time, $Y(1)$ is unobserved in the control group and $Y(0)$ is unobserved in the exposure group. In observational studies, the exposed and unexposed may have systematic differences in their characteristics, which are potentially associated with the outcome. These patient characteristics are referred to as confounding variables, denoted by X. Consequently, the distribution of the observed $Y(1)$'s in the exposure group may not represent the distribution of the unobserved $Y(1)$'s in the control group or the $Y(1)$'s in the entire population.

To mitigate this issue, [3] proposed the strongly ignorable treatment assignment assumption $\big(Y(1),\ Y(0)\big) \perp A \mid X$, which implies that the common causes of the outcome and the treatment are fully observed and thus the treatment assignment is uninformative of the potential outcomes given X. This key assumption, combined with the stable unit treatment value assumption and the positivity assumption [3], allows one to estimate the mean of potential outcomes $Y(1)$ using only the observed portion of $Y \mid A = 1$ by restricting to a stratum of X, i.e.,

$$E[Y(1) \mid X] = E[Y \mid A = 1, X].$$

We immediately have that $E[Y(1)] = E_X\{E[Y \mid A = 1, X]\}$. In addition, [3] defined the propensity score as the probability of being exposed given the subject's characteristics, i.e., $S = P(A = 1 \mid X)$. One can show that $E[Y(1)] = E_X\{E[Y \mid A = 1, X]\} = E[\frac{YA}{P(A=1|X)}]$. This dual representation of the mean potential outcomes has led to a widely used method which is the inverse probability of treatment weighting.

The propensity score is a one-dimensional balancing score: within a stratum of S, the covariates are similar between both exposure and control groups. Therefore, adjustment for the scalar S is sufficient to remove bias due to X. Consequently, we have an alternative estimation of the mean of potential outcomes through the fact that

$$E[Y(1) \mid S] = E[Y \mid A = 1, S]. \tag{1}$$

Therefore, besides the dual representation of the mean potential outcomes, we in fact have a third representation which is

$$E[Y(1)] = E\{E[Y \mid A = 1, S]\}. \tag{2}$$

The idea of estimating a sufficient statistic S and then substituting the set of confounders X with a one-dimensional S in subsequent analysis has led to numerous methods. These propensity score methods could potentially gain efficiency in finite sample and are particularly useful in pharmacosurveillance with a rare outcome, a common exposure, and many confounders.

2.2 | Flexible regression on the estimated propensity score

In this section, we detail a three-step procedure that flexibly adjusts for confounding using the estimated propensity score, then standardizes to a target population for causal comparison. Compared to direct covariate adjustment, which will be discussed in Section 3, our approach replaces the

set of covariates with a function of the propensity score in the model to reduce the dimensionality of the covariates while attempting to minimize model assumptions. We further derive direct variance estimates that are computationally efficient and more feasible for a rare outcome setting.

In the first step, we estimate the propensity score based on a parametric model such as a logistic regression. Recall that we take A as the binary exposure to a drug or medical product. Thus, the propensity score, i.e., the probability of being exposed can be predicted as

$$\hat{S} = \hat{P}[A = 1 \mid X; \hat{\gamma}] = \left[1 + \exp(-X^T \hat{\gamma})\right]^{-1}, \tag{3}$$

where $\hat{\gamma}$ is a set of estimated coefficients obtained from fitting a logistic regression.

In the second step, we fit a nonparametric or semiparametric model of the outcome taking the estimated propensity score as the only covariate. We consider the varying coefficient model proposed by [20]:

$$g\left(E[Y \mid A, S = \hat{S}]\right) = \alpha(\hat{S}) + \beta(\hat{S})A, \tag{4}$$

where $\alpha(\cdot)$ and $\beta(\cdot)$ are unknown and potentially nonlinear functions, and $g(\cdot)$ is a known link function. This is equivalent to fitting two separate outcome curves in each of the exposure and control arms. Note that when S balances the two arms, conditioning on S is sufficient to control for confounding. For binary adverse outcomes, $g(\cdot)$ is often the logit link function. When there is little evidence of treatment effect heterogeneity conditional on the propensity score, we could potentially gain substantial efficiency by assuming that $\beta(\cdot) = \beta$. Then the varying coefficient model reduces to a partially linear model $g\left(E[Y \mid A, S = \hat{S}]\right) = \alpha(\hat{S}) + \beta A$, which tends to stabilize the outcome model when the number of events is low [21]. Combining (1) and (4), we can see that the varying coefficient model allows us to estimate two treatment-specific outcome curves that predict the means of potential outcomes given patient's propensity:

$$\begin{aligned} E[Y(1) \mid \hat{S}] = E[Y \mid A = 1, \hat{S}] &= g^{-1}\left[\alpha(\hat{S}) + \beta(\hat{S})\right], \\ E[Y(0) \mid \hat{S}] = E[Y \mid A = 0, \hat{S}] &= g^{-1}\left[\alpha(\hat{S})\right]. \end{aligned}$$

To estimate the nonlinear functions $\alpha(\cdot)$ and $\beta(\cdot)$, several methods of nonparametric regression on an one-dimensional covariate could be adopted. Here we apply the spline regression, which is a special case of sieve estimation [22]. A spline is a piece-wise polynomial function that is smooth at the joint of each piece, referred to as the knots. Any spline function on a given set of knots can be expressed as a linear combination of B-splines [23, 24]. Denote a set of B-spline basis functions with evenly spaced knots as $\mathbf{B}(S) = [b_1(S), \ldots, b_K(S)]$. The dimension $K = K(n)$ grows to infinity with sample size n, which can be selected by cross-validation procedure in practice [22, 25]. Finally, we fit the outcome on the basis functions and the exposure indicator. The estimated risk given the patient's propensity score is

$$\hat{E}[Y(a) \mid \hat{S}] = \hat{E}[Y \mid A = a, \hat{S}] = g^{-1}\left[\mathbf{B}(\hat{S})\hat{\boldsymbol{\alpha}} + a \cdot \mathbf{B}(\hat{S})\hat{\boldsymbol{\beta}}\right],\ a = 0, 1$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are the coefficients of the B-spline basis functions. When assuming a partially linear model with constant treatment effect conditional on the propensity score, i.e., $\beta(S) = \beta$, we instead predict the patient's risk as $\hat{E}[Y(a) \mid \hat{S}] = \hat{E}[Y \mid A = a, \hat{S}] = g^{-1}\left[\mathbf{B}(\hat{S})\hat{\boldsymbol{\alpha}} + \hat{\beta}a\right]$, $a = 0, 1$.

As discussed in Section 2.1, causal inference is a comparison of the population-level averages. Thus, the outcome regression model should often be considered as the intermediate summary that is then used in a final standardization step to yield marginal, population-level contrasts. One popular approach is to take the empirical averages of the predicted risks resulting from creating a pair of predictions for each patient as if they were exposed to each of the two different drugs regardless of their actual exposure condition. Such a procedure has been called model-based standardization in epidemiology, which we will refer to generally as standardization [26, 27, 6]. Therefore, at the third step, we take the empirical average of the estimated potential outcomes over the target population which gives

$$\begin{aligned} \hat{E}[Y(1)] &= \frac{1}{n}\sum_{i=1}^{n} \hat{E}[Y_i(1) \mid \hat{S}_i] = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left[\mathbf{B}(\hat{S}_i)\hat{\boldsymbol{\alpha}} + \mathbf{B}(\hat{S}_i)\hat{\boldsymbol{\beta}}\right], \\ \hat{E}[Y(0)] &= \frac{1}{n}\sum_{i=1}^{n} \hat{E}[Y_i(0) \mid \hat{S}_i] = \frac{1}{n}\sum_{i=1}^{n} g^{-1}\left[\mathbf{B}(\hat{S}_i)\hat{\boldsymbol{\alpha}}\right] \end{aligned}$$

if the target population is the entire study population. One can also average over other target populations such as the exposure arm to estimate $E[Y(a)|A = 1], a = 0, 1$, which yields the average treatment effect among the treated (ATT).

With the pair of population-level averages of potential outcomes, we can now make simple comparisons that have explicit causal interpretations. For example, for a binary adverse event outcome like angioedema, we have a pair of mean risks denoted by $\hat{p}_1 = \hat{E}[Y(1)]$ and $\hat{p}_0 = \hat{E}[Y(0)]$, which are the risks of angioedema among the full population in need of high blood pressure control medications (a combination of the ACEI and BB groups), had they taken ACEI ($\hat{p}_1$) or BB ($\hat{p}_0$). We plug in the estimated mean risks to estimate the parameter of interest such as the risk difference $\hat{\text{RD}} = \hat{p}_1 - \hat{p}_0$, the relative risk $\hat{\text{RR}} = \hat{p}_1/\hat{p}_0$, or the odds ratio $\hat{\text{OR}} = \left[\hat{p}_1/(1 - \hat{p}_1)\right]/\left[\hat{p}_0/(1 - \hat{p}_0)\right]$.

2.3 | Variance estimation

The key challenge in studying the variance of the proposed estimator is the need to incorporate the variability due to estimation of the propensity score, which impacts both the estimation of the nonlinear functions $\hat{\alpha}(\cdot)$ and $\hat{\beta}(\cdot)$, and the evaluation of the functions $\hat{\alpha}(\hat{S})$ and $\hat{\beta}(\hat{S})$ when plugging in $\hat{S}$. [19] studied the asymptotic distribution of a class of estimators that employ covariates estimated from a preliminary regression, such as the use of the estimated propensity score. Motivated by [19], we now consider variance estimation that takes in to account the distinctive contributions of propensity score estimation, nonparametric regression on the propensity score, and standardization, while remaining a feasible inferential procedure for the rare outcome setting.

To this end, we introduce the notion of an influence function [18]. Under certain regularity conditions, many estimators of $E[Y(a)]$ are asymptotically equivalent to the sample average of an object referred to as the influence function. The influence function is a function of the observed data with mean zero and finite variance, which contains all information about an estimator's asymptotic behavior. For example, we say the estimator $\hat{p}_1$ defined in Section 2.2 has influence function IF_1 if $\hat{p}_1 - p_1 = n^{-1}\sum_{i=1}^{n} \text{IF}_1(O_i) + o_p(n^{-1/2})$, where O_i denotes the i-th observed data point. The influence function is particularly useful because its variance is the variance of the asymptotic distribution of the estimator [28]. For example, the variance of $\hat{p}_1$ is equal to the sample variance of the estimated IF_1 divided by the sample size n. Therefore, the influence function is a key component for variance estimation and efficiency comparisons, and the one with the smallest variance is referred to as the efficient influence function. In parametric models the influence function is a scaled version of the score function. For semiparametric and nonparametric models, derivation of the influence function varies by the parameter of interest and the model assumptions. We refer the interested readers to the broad literature on semiparametric theory such as [29], [30], [31], and [32].

The work of [19] and [33] studied the influence function of the three-step estimator based on nonparametric regression on an estimated propensity score $\hat{S}$ followed by standardization. They considered scenarios when the propensity score is generated under either a parametric or a nonparametric model in a prior step and provided two versions of the influence function. In particular, for binary outcomes with $p_1 = E[Y(1)]$ and $p_0 = E[Y(0)]$, when the propensity score is estimated under a nonparametric model, the influence functions for p_1 and p_0 are

$$\text{IF}_1 = E[Y|A=1,S] - p_1 + \frac{A}{S}(Y - E[Y|A=1,S]) - \frac{E[Y|A=1,X] - E[Y|A=1,S]}{S}(A-S)$$

and

$$\text{IF}_0 = E[Y|A=0,S] - p_0 + \frac{1-A}{1-S}(Y - E[Y|A=0,S]) - \frac{E[Y|A=0,X] - E[Y|A=1,S]}{1-S}(A-S),$$

respectively. When the propensity score is estimated in a parametric model, such as a logistic regression with $S = \varphi(X;\gamma) = [1+\exp(-X\gamma)]^{-1}$, the influence functions are

$$\begin{aligned}\text{IF}_1 =& E[Y \mid A=1,S] - p_1 + \frac{A}{S}(Y - E[Y \mid A=1,S]) - \\ & E\left[\frac{E[Y \mid A=1,X] - E[Y \mid A=1,S]}{S}\frac{\partial\varphi(X;\gamma)}{\partial\gamma}\right]\phi(\gamma)\end{aligned}$$

and

$$\begin{aligned}\text{IF}_0 =& E[Y \mid A=0,S] - p_0 + \frac{1-A}{1-S}(Y - E[Y \mid A=0,S]) - \\ & E\left[\frac{E[Y \mid A=0,X] - E[Y \mid A=0,S]}{1-S}\frac{\partial\varphi(X;\gamma)}{\partial\gamma}\right]\phi(\gamma),\end{aligned}$$

where γ is the coefficients of the propensity score model, $\phi(\gamma) = -E[\frac{\partial\dot{\ell}(\gamma)}{\partial\gamma}]^{-1}\dot{\ell}(\gamma)$ is the influence function for γ, and $\dot{\ell}(\gamma)$ is the score function for γ. The variance can then be estimated using the sample variances of one version of the IFs according to how the propensity score is estimated. The above influence functions are derived in detail in [33]. Regularity conditions for root-n consistency and asymptotic normality have been detailed in Theorem 3.1 of [34]. In particular, the parameter of interest in [34] is a known functional of the two-step sieve estimates, where the second step involves sieve estimation of unknown functions that may use the nonparametric estimates from the first step as inputs. In our paper, the corresponding known functional of the two-step sieve estimates is $E\{\hat{E}[Y \mid A=a,\hat{S}]\}$, which is further estimated in the third step via an empirical average. By the same stochastic equicontinuity argument as Lemma 1.4 of [34], the empirical average of the sieve estimates has the same asymptotic distribution as the expectation of the sieve estimates. Therefore our estimator is root-n consistent and asymptotically normal under the regularity conditions stated in [34].

Note that evaluation of the above influence functions requires estimation of the outcome models $E[Y \mid A, X]$. For our proposed regression on the propensity score method, estimating the influence functions introduces an extra step of estimating the mean outcome conditional on all covariates, which may be unstable when the outcome is rare. Therefore, in a rare outcome setting we propose to substitute $E[Y \mid A, X]$ with $E[Y \mid A, S]$, which is readily computed in our estimation procedure. Specifically, we estimate the influence functions as

$$\widetilde{\text{IF}}_1 = \hat{E}[Y \mid A=1,\hat{S}] - \hat{p}_1 + \frac{A}{\hat{S}}(Y - \hat{E}[Y \mid A=1,\hat{S}])$$

and

$$\widetilde{\mathrm{IF}}_0 = \hat{E}[Y \mid A=0, \hat{S}] - \hat{p}_0 + \frac{1-A}{1-\hat{S}}(Y - \hat{E}[Y \mid A=0, \hat{S}]),$$

where $\hat{S}$ is estimated parametrically in the first step, $\hat{E}[Y(\cdot) \mid \hat{S}]$ is estimated by a nonparametric regression on the estimated propensity score, and $\hat{p}_a, a=0,1$ is a population-level sample average. It turns out that $\widetilde{\mathrm{IF}}_1$ and $\widetilde{\mathrm{IF}}_0$ are the influence functions assuming propensity score is known with

$$\mathrm{Var}(\widetilde{\mathrm{IF}}) \geq \mathrm{Var}(\mathrm{IF}) \tag{5}$$

under a nonparametric model [33]. This corresponds to the well-known phenomenon that ignoring knowledge about the propensity score and estimating it using the observed data instead may improve efficiency for certain estimators. Therefore, using $\widetilde{\mathrm{IF}}$ will generally yield a conservative standard error. Moreover, using $\widetilde{\mathrm{IF}}$ instead of IF is equivalent to setting the last term of IF to zero, which is a product of two residuals. Under a rare outcome scenario, this term often has negligible contribution and thus variance is well approximated using $\widetilde{\mathrm{IF}}$. However, when the outcome is non-rare, the precise influence functions IF_1 and IF_0 should be computed for valid inference.

By an application of the delta method to functions of p_1 and p_0, one can show that the influence function for the risk difference (RD) is $\widetilde{\mathrm{IF}}_{\mathsf{RD}} = \widetilde{\mathrm{IF}}_1 - \widetilde{\mathrm{IF}}_0$; the influence function for log of the risk ratio (RR) is $\widetilde{\mathrm{IF}}_{\mathsf{logRR}} = \frac{\widetilde{\mathrm{IF}}_1}{\hat{p}_1} - \frac{\widetilde{\mathrm{IF}}_0}{\hat{p}_0}$; and the influence function for log of the odds ratio (OR) is $\widetilde{\mathrm{IF}}_{\mathsf{logOR}} = \frac{\widetilde{\mathrm{IF}}_1}{\hat{p}_1(1-\hat{p}_1)} - \frac{\widetilde{\mathrm{IF}}_0}{\hat{p}_0(1-\hat{p}_0)}$. Finally, the variance of an estimator, e.g. the log of odds ratio ($\log$OR), is estimated by

$$\frac{1}{n}\hat{\sigma}^2 = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}\widetilde{\mathrm{IF}}^2_{\log \mathsf{OR},\, i}\right] = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\widetilde{\mathrm{IF}}_{1i}}{\hat{p}_1(1-\hat{p}_1)} - \frac{\widetilde{\mathrm{IF}}_{0i}}{\hat{p}_0(1-\hat{p}_0)}\right)^2\right]. \tag{6}$$

Accordingly, a 95% confidence interval can be constructed using $\widehat{\log \mathsf{OR}} \pm c \cdot \hat{\sigma}/\sqrt{n}$, where c is the 97.5-th percentile of the normal distribution.

We applied this variance estimation to all approaches that utilize standardization following regression adjustment in our simulation and application studies, such as regression on covariates (detailed in Section 3.1) using IF, and linear adjustment of the propensity score using $\widetilde{\mathrm{IF}}$.

3 | COMMON PROPENSITY SCORE METHODS

In this section, we briefly review common propensity-based methods for estimating a pair of population-level average risks for binary outcomes, which will then be plugged in to estimate a population-level summary such as risk difference, risk ratio, or odds ratio that has a causal interpretation. We focus on methods that utilize the propensity score. In Section 4, we will compare these strategies to our proposed flexible regression on the propensity score with standardization method.

3.1 | Regression on covariates followed by standardization

As outlined in Section 2, standardization following flexible propensity score adjustment is a viable approach to estimate causal effects. A common alternative is to use standardization following direct adjustment for confounders in the outcome regression model. Specifically, we build an outcome regression model with both the exposure and all confounders as covariates in the model $g\left(E[Y \mid A, X]\right) = \beta A + X\boldsymbol{\alpha}$, then standardize to the target population as outlined in Section 2.2 to estimate the marginal, population-level mean risks $E[Y(a)], a=0,1$. The estimated causal effect can then be obtained by plugging in the population risks into risk difference, risk ratio, or odds ratio. Compared to Section 2.2, there is no estimation of the propensity score, and all confounders are directly adjusted for in the outcome regression model. Therefore, application of this method may be unstable, or have model fitting issues, when there are few outcome events and many potential confounders.

3.2 | Inverse probability of treatment weighting

As mentioned in Section 2.1, the dual representation of the causal estimand has led to a widely used method which is the inverse probability of treatment weighting (IPTW). The IPTW method aims to achieve balance by reweighting every subject to create a pseudo-population in which every exposed/unexposed pseudo-subject has equal possibility of being exposed/unexposed. Such a pseudo-population is representative of one from a randomized study [35]. A commonly used weight is the inverse of the propensity score, that is, to use $\frac{1}{S}$ if subject is exposed and $\frac{1}{1-S}$ if subject is unexposed. A well-known challenge with IPTW is the potential instability from inverting the estimated propensity score. To address this issue, stabilized weights have been proposed [35]. Truncation of the propensity score using either a pre-specified threshold or a quantile is also widely used in practice [36, 37], and will be implemented in our simulation study in Section 4.

3.3 | Augmented inverse probability of treatment weighting

Simple IPTW requires that the propensity score must be correctly specified. To relax this assumption, the Augmented IPTW (AIPTW) approach was proposed, which is a combination of the propensity score model and the outcome regression model [38, 39]. It has also been referred to as the doubly robust estimator because it consistently estimates the truth when either the propensity score model or the outcome regression model is correctly specified. The AIPTW estimates the distribution of the observed data under a parametric (or semiparametric) model and then evaluates a particular estimating equation under such working model. The estimating equation is defined by the efficient influence function for the mean potential outcome under a nonparametric model. Specifically, we solve for the mean potential outcome $E[Y(a)]$ in

$$\mathbb{P}_n \left\{ \frac{\mathbb{1}(A=a)}{P(A=a|X)} Y + \frac{P(A=a|X) - \mathbb{1}(A=a)}{P(A=a|X)} E[Y|A=a,X] - E[Y(a)] \right\} = 0, a = 0, 1 \quad (7)$$

where $\mathbb{P}_n$ is the empirical average operator, i.e., $\mathbb{P}_n(V) = \frac{1}{n}\sum_{i=1}^{n} V_i$. From the above estimating equation, we can see that AIPTW can be viewed as IPTW with a bias correction term that involves the predicted risks by the outcome regression model. However, modeling of the outcome is likely to be unstable when there is insufficient number of outcome events and many confounders. In addition, when the risk of adverse events is extremely low, the magnitude of bias correction from the predicted risks may be small.

3.4 | Targeted maximum likelihood estimation with data-adaptive estimation of the outcome and exposure models

Similar to the AIPTW approach, the targeted maximum likelihood estimation (TMLE) is a methodology that incorporates both outcome and exposure mechanisms to exhibit doubly robust finite sample performance [40]. This is achieved by a targeting step that solves the efficient influence function via maximum likelihood estimation in order to improve the initial outcome regression model. It is more robust to outliers than the AIPTW in the sense that for binary outcomes, the estimated mean potential outcome is always between 0 and 1, whereas the AIPTW estimates can fall outside of this range.

Another advantage of TMLE is that one can incorporate a collection of algorithms to estimate the outcome and exposure mechanisms, which may further reduce bias. This is particularly of interest when there are a large number of confounders and model misspecification is likely to occur. An ensemble procedure called Super Learner was developed which implements a library of data-adaptive algorithms and finds an optimal combination of the predictions from the algorithms using cross-validated weights. The weighted combination has equal or better performance than the best-fitting algorithm in the library [41]. Therefore, including a rich set of algorithms in the library provides better chance of predicting the outcome and exposure mechanisms well. However, the computation time grows significantly as the number of algorithms considered increases.

4 | REALISTIC SIMULATION

In this section, we perform extensive simulation studies to investigate the performance of our proposed flexible regression on the propensity score method (Section 2) and the existing methods outlined in Section 3. We consider estimating a marginal OR for observational surveillance within the rare outcomes setting since it is the most common estimand of interest in observational cohort studies for binary outcomes. Moreover, in the rare event setting, the marginal OR is approximately the relative risk. Our simulation study will mimic real data from the FDA Sentinel Initiative study comparing the effect of angiotensin-converting enzyme Inhibitors (ACEI) and beta blockers (BB) on incidence of angioedema in the first 30 days [42]. Further details of the study are outlined in the simulation setting in Section 4.1 and the real data application in Section 5.

In the context of rare binary outcomes, [17] evaluated propensity-based estimators of the marginal relative risk using a data simulation framework referred to as Plasmode [43] mimicking two cohort studies constructed from healthcare claims data with 70 to 178 confounders and 250 number of events in most scenarios. They found that nonlinear adjustment for the propensity score provides lower bias and mean squared error regardless of the propensity score estimation method. Our study will further evaluate the performance of propensity score methods targeting marginal odds ratios to provide additional guidance on selecting methods that control for confounding. In particular, we use a new realistic simulation method which does not require sharing of individual-level data with similar performance as the Plasmode method [44].

4.1 | Simulation setting

We generate a realistic population of 10,000 subjects mimicking data from the ACEI and BB example. Specifically, there are nine binary clinically relevant covariates (NSAIDs (Nonsteroidal anti-inflammatory drugs), aspirin, ORAL-CS (optimizing recovery after laparoscopic colon surgery), allergic reaction, diabetes, heart disease, Ischemic HD (heart disease), inpatient hospitalization, and gender) and one categorical variable which is age category with four levels, corresponding to three dummy variables (binary indicators). See Table 1 for the prevalence (prev) of each confounder.

To simulate the ACEI and BB dataset, we used the following procedure and generated:

1. Binary and categorical covariates X that have the same mean and pairwise covariance as the real data, yielding correlated confounders.
2. A binary exposure A (ACEI = 1 and BB = 0) generated based on a logistic regression on the covariates (the propensity score model, see Table 1), using coefficients observed from fitting the real data.
3. A pair of binary potential outcomes $(Y(1), Y(0))$ (angioedema within 30 days under exposure and control for the same subject) based on a logistic regression on the exposure and covariates (the outcome regression model, see Table 1), using the coefficients observed from fitting the real data.

We specify the intercepts in the exposure and outcome models such that the exposure prevalence is the same as the real data and the number of events is similar to the real data. In particular, the exposure rate in the real data is 69%, which is uncommon in postmarketing surveillance because there are more exposed (ACEI, 69%) than controls (BB, 31%). Therefore we also simulated another scenario when the exposure rate is 20%, which is more commonly seen for new medications. In addition, we hold the event rate in the control group such that the total number of events in the simulated data is around 50 across all scenarios. Because there are 12 confounders including nine binary variables and three indicators for age categories, we have less than five events per covariate on average. This indicates a relatively small amount of information in the data, which may lead to model fitting issues such as imprecise estimation and model misspecification. In addition, we increased the strength of confounding by scaling up the coefficient of covariates in the propensity score model (multiply coefficients on the $\log$OR scale by 1.5), while still holding the exposure prevalence and the baseline event rate the same (see Table 1 stronger propensity). In a rare outcome setting, sufficient power to detect

TABLE 1 Prevalence (%) of each confounder, relationship (adjusted odds ratio) between exposure (ACEI and BB) and confounders (propensity score model) for different simulation scenarios, and relationship (adjusted odds ratio) between outcome and the exposure and confounders (outcome regression model).

Confounders	Prevalence %	Propensity Score Model		Outcome Regression Model	
		Observed Relationship	Stronger Relationship	Observed Relationship	Treatment Heterogeneity
Heart Disease	2.0	0.5	0.4	0.7	0.7
Aspirin	4.4	1.2	1.3	0.9	0.9
Ischemic HD	5.3	0.3	0.2	1.4	1.4
OptRec Colon Surg	5.6	1.0	1.0	1.4	1.4
Inpatient Hosp.	7.9	0.3	0.2	1.9	1.9
Allergic Reaction	8.3	0.9	0.8	0.6	0.6
NSAIDS	11.8	1.0	1.0	5.4	5.4
Diabetes	15.6	4.5	9.6	2.4	2.4
Female	51.3	0.6	0.4	1.5	1.5
Age (Ref: 18-44)					
45-54	26.6	2.0	2.7	0.8	0.8
55-64	29.7	2.1	2.9	0.5	0.5
65-99	22.3	1.7	2.2	0.5	0.5
Exposure					
ACEI	69.1			5.4	$\exp(d)$*

* d = 0.7$\mathbb{1}$(Heart Disease) - 0.1$\mathbb{1}$(Aspirin) + 0.3$\mathbb{1}$(Ischemic HD) + 0.3$\mathbb{1}$(OptRec Colon Surg) + 0.7$\mathbb{1}$(Inpatient Hosp.) - 0.6$\mathbb{1}$(Allergic Reaction) + 1.7$\mathbb{1}$(NSAIDS) + 0.9$\mathbb{1}$(Diabetes) + 0.4$\mathbb{1}$(Female) - 0.3$\mathbb{1}$(Age 45-54) - 0.8$\mathbb{1}$(Age 55-64) - 0.7$\mathbb{1}$(Age 65-99)

a safety signal often requires a moderate to strong exposure effect. We generated the pair of potential outcomes under both the null model (marginal OR = 1, i.e., no elevated risk of angioedema due to treatment with ACEI), and the alternative where the marginal OR = 3. Note that, the marginal OR defined as $\frac{p_1/(1-p_1)}{p_0/(1-p_0)}$ is not equal to the conditional OR typically obtained from the treatment coefficient in a logistic regression due to non-collapsibility. We use the algorithm described in [8] to find a conditional treatment effect that would yield a marginal OR of 3. In addition, we considered heterogeneous treatment effect in the sense that the log odds ratio is a linear combination of all covariates (see Table 1 treatment heterogeneity). We call the third scenario strong treatment heterogeneity scenario. Note that unlike the main effects presented in Table 1 , the

coefficients for treatment heterogeneity are pre-specified since treatment heterogeneity was not present in the real data. We use the simulated pairs of potential outcomes to compute the sample average treatment effects for bias comparison, which are marginal OR = 1 (log(OR) = 0) under the null, marginal OR = 3 (log(OR) = 1.1) under the alternative, and marginal OR = 1.6 (log(OR) = 0.5) under treatment heterogeneity on average.

In each scenario assessed we used 5000 simulated datasets. For simplicity, we use the abbreviation PS to refer to the propensity score hereafter.

4.2 | Methods under comparison

We consider the following marginal OR estimators:

(i) Crude estimate from regression on the exposure without confounding adjustment;

(ii) Regression on covariates without interaction terms between the exposure and covariates;

(iii) Regression on main term of the PS (linear adjustment);

(iv) Flexible regression of the PS using B-spline basis functions with data-adaptive degrees of freedom selected by cross-validation to fit a nonlinear function $\alpha(S)$, while imposing a marginal structure that the conditional treatment effect is a constant, i.e., $\beta(S) = \beta$. Referred to as the PS one-spline method;

(v) Flexible regression of the PS using B-spline basis functions with data-adaptive degrees of freedom selected by cross-validation to fit both $\alpha(S)$ and $\beta(S)$. Referred to as the PS two-spline method;

(vi) Regression on indicators of five strata of PS that fits a nonlinear function $\alpha(S)$ by step functions, while imposing a constant conditional treatment effect $\beta(S) = \beta$;

(vii) IPTW with stabilized PS (truncated at the (0.025, 0.975) percentile);

(viii) AIPTW with parametric models for exposure and outcome both adjusting for main terms of all covariates, and stabilized PS (truncated at the (0.025, 0.975) percentile);

(ix) Targeted maximum likelihood estimation with parametric PS (truncated at the (0.025, 0.975) percentile) and outcome models;

(x) Targeted maximum likelihood estimation with data-adaptive estimation of the PS (truncated at the (0.025, 0.975) percentile) and outcome models using Super Learner.

All model-based methods are followed by standardization to estimate the population average of potential outcomes and the marginal OR. The PS used in (iii)-(ix) is estimated from the same parametric model adjusting for all covariates. Due to the extremely long computation time of Super Learner (method (x)), we considered a limited library of algorithms including the generalized linear model, step-wise regression, and penalized regression. Method (iii) has been shown to be biased in other simulation studies [7, 8] and we were therefore interested in assessing it in our simulations for verification, as well as to investigate the potential for bias correction via flexible semiparametric or nonparametric modeling. To this end, we considered method (iv) that fits a function $\alpha(S)$ which is then shifted by a constant treatment effect β on the logit scale under the assumption of homogeneous treatment effect conditional on the PS, and method (v) which fits two exposure-specific curves nonparametrically. For simplicity, we refer to method (iv) as PS one-spline, and method (v) as PS two-spline. In addition, we considered method (vi) which essentially fits a piecewise constant function or step function. These methods entail a more flexible nonlinear function of the PS, and may reduce the residual confounding from regression on a simple linear term in method (iii). Method performance was assessed in terms of bias and variance of the estimates on the log OR scale.

To investigate the performance of our proposed variance estimation for statistical inference, we estimated the standard error of regression on PS estimators (iii)-(v) using equation (6) as detailed in Section 2.3. Performance was then assessed in terms of type I error and power. To provide a benchmark for comparison, we approximated the variance of the other estimators as follows. First of all, the variance of TMLE estimates is available in the R package `tmle`. The variance estimator is computed using the sample variance of the estimating equation that is the efficient influence function evaluated at each data point. In the simulation study, we adopt such variance estimator for TMLE with and without super learning. Likewise, because both IPTW and AIPTW methods are asymptotically linear estimators solving an estimating equation evaluated under a specified model, we computed the variance using their corresponding estimating equations as an approximation of the variance of $\hat{p}_1$ and $\hat{p}_0$, which did not take into account the variability of the propensity score estimation. For simplicity, we use the same variance approximation as AIPTW for the outcome regression method, which will provide a conservative variance estimator.

We considered three model misspecification scenarios which are (a) the PS model is misspecified, (b) the outcome model is misspecified, and (c) both the PS and the outcome models are misspecified. All model misspecification is due to omitting age, which is strongly associated with both the outcome and the exposure, and has sufficient number of observations in all categories according to Table 1 . Note that under the strong treatment heterogeneity scenario, the outcome model is specified without the corresponding interaction terms between the exposure and covariates, therefore is misspecified even without omitting age categories. In this scenario, only TMLE with super learner may be able to pick up such interaction terms in the outcome model.

4.3 | Results

Table 2 and 3 shows the bias and variance of the various methods on the log(OR) scale, the type I error and power, as well as the mean estimated variance, when both the PS model and outcome regression model are correctly specified. We consider two scenarios: an exposure rate of 69% (observed in the real data) and 20%. Each scenario is further tabulated by null and alternative exposure effect, as well as strong treatment heterogeneity as described in Section 4.1. Table 2 focuses on the original strength of confounding observed in the real data. In Table 3 we provide simulation results under stronger relationship between the confounders and the exposure in the PS model which generates stronger confounding.

First of all, simple linear adjustment of the PS had a notable increase in bias under the alternative with either homogeneous or heterogeneous effect, which agrees with the findings of [7] and [8]. PS one-spline method outperformed the simple linear adjustment and removed the residual confounding. It had equivalent or smaller variance relative to all unbiased methods across all scenarios, indicating an efficiency gain due to assuming constant conditional exposure effect which greatly reduces the dimension. PS two-spline method tended to have larger variance under homogeneous effect due to overfitting assuming a heterogeneous effect. In contrast, under effect heterogeneity, PS one-spline had increased bias whereas PS two-spline had the smallest bias among all methods since it correctly captured the heterogeneity through the exposure-specific curves. Regression on PS strata also had relatively good performance, with the smallest bias under homogeneous effect when the exposure rate was 69%, although it had larger bias when the exposure rate was 20%. In addition, it had small variance under homogeneous effect because the degree of freedom was the smallest across all regression methods. Although regression on covariates is an oracle estimator under homogeneous effect which essentially fits the true data generating model, it frequently encountered the "perfect fit" phenomenon with predicted probabilities that are exactly zero or one, which is expected since the outcome is rare and the number of covariates is large. There were also several non-convergence cases among the replications. IPTW had large variance and bias potentially due to the variability from inverting the propensity scores which may include extreme values in finite sample. In terms of the three doubly robust estimators, i.e., AIPTW and TMLE with and without super learning, the bias of TMLE with super learning was generally the smallest among all three. In addition, their variances tended to be larger than methods that only depend on either the outcome or propensity score models. This is expected because the doubly robust methods essentially trade efficiency for robustness by operating in a larger model. The performance of AIPTW was very similar to that of the IPTW estimator, which may be due to the fact that when the event rate is extremely small, the bias correction which involves the predicted outcome risk can be small. In Table 3 under stronger confounding effect, all methods had larger biases and variances in general. We observed similar relative performance as Table 2 .

In terms of inference, our direct variance estimator had generally good performance. In particular, under correctly specified models in Table 2 and 3 , all regression on PS methods followed by standardization had similar and valid type I error, except that the PS two-spline method had inflated type I error which may indicate insufficient characterization of the variation using our simplifying approximation. Nonetheless, the PS two-spline method had similar power as the other regression on PS methods. The valid type I error showed that our proposed direct variance estimation is a reliable approach for inference. Particularly, it does not require computationally intensive methods such as the bootstrap. Covariate adjustment and TMLE methods also had valid type I error and similar power. In addition, IPTW and AIPTW had inflated type I error and slightly less power. We also observed that under the treatment heterogeneity, the power of the confounding adjusted methods was less than 0.4 and the power of the crude estimate was at most 0.123. This is because when the number of events is extremely small, the power is highly dependent on the strength of treatment effect, which is not strong enough to yield a large power in this scenario.

Table 4 and 5 present method performance under model misspecification due to omitting the age categories. All methods that depend on only the outcome regression model or the PS model had increased bias when their corresponding model is misspecified. In particular, regression on covariates had a notable increase of bias, although the variance was smaller due to omitting the three age category indicators which reduced the dimension but introduced misspecification. Among regression on PS methods, PS one-spline and PS strata methods had relative smaller bias under homogeneous effect, whereas PS two-spline method had the smallest bias under heterogeneous effect.

In contrast, doubly robust methods were less sensitive to model misspecification. In particular, under a homogeneous null or alternative effect, all doubly robust methods were insensitive to misspecification of the outcome models, and had relative smaller bias when the PS model was misspecified. When both the outcome and PS models were misspecified, AIPTW had relatively larger bias than TMLE methods. Moreover, TMLE with super learning generally had smaller bias than TMLE without super learning, and can outperform all other methods under the alternative with either a homogeneous or heterogeneous effect. Therefore, using ensemble learning with data adaptive algorithms can reduce bias from model

misspecification. Moreover, under effect heterogeneity, only TMLE with super learning may be able to pick up the interaction term between covariates and exposure to reduce bias.

Our simulation indicated that the bias of regression on the PS observed in previous studies [7, 8] could be due to residual confounding from simple linear adjustment. Based on our consistent observation that flexible adjustment of PS reduces bias from insufficient linear adjustment and potentially outperforms traditional methods, it is promising in safety surveillance with rare outcomes to estimate a propensity score that sufficiently controls for confounding and reduces dimension to allow for flexible outcome modeling. Moreover, the valid type I error showed that our proposed direct estimation of variance is a fast and valid approach for inference.

TABLE 2 Median bias, variance, type I error, power, and mean estimated variance in estimating the marginal odds ratio (on the log scale) using correctly specified propensity score model and outcome regression model, under the null, the alternative, and strong treatment heterogeneity, with original confounding effect observed in the real data.

Methods	Null: OR = 1			Alternative: OR = 3			Heterogeneity: OR = 1.6		
	Bias (Var)	Type I Error	Mean Est Var	Bias (Var)	Power	Mean Est Var	Bias (Var)	Power	Mean Est Var
Exposure rate = 69%									
(i) Crude	-0.178 (0.149)	0.079	0.152	-0.180 (0.116)	0.841	0.110	-0.238 (0.129)	0.123	0.133
(ii) Covariate adj	0.010 (0.162)	0.048	0.167	0.006 (0.123)	0.912	0.131	0.059 (0.145)	0.312	0.165
(iii) PS adj	0.008 (0.169)	0.051	0.168	0.029 (0.128)	0.916	0.134	0.084 (0.157)	0.333	0.169
(iv) PS one-spline	0.008 (0.163)	0.048	0.167	0.006 (0.124)	0.912	0.131	0.058 (0.145)	0.308	0.164
(v) PS two-spline	0.004 (0.168)	0.058	0.154	0.001 (0.129)	0.909	0.119	-0.003 (0.143)	0.304	0.134
(vi) PS strata	-0.000 (0.159)	0.046	0.167	-0.000 (0.122)	0.911	0.130	0.049 (0.142)	0.297	0.164
(vii) IPTW	0.024 (0.182)	0.067	0.167	0.022 (0.147)	0.871	0.131	0.021 (0.162)	0.307	0.165
(viii) AIPTW	0.025 (0.182)	0.066	0.167	0.023 (0.147)	0.872	0.131	0.021 (0.163)	0.307	0.165
(ix) TMLE w/o SL	0.025 (0.183)	0.056	0.164	0.024 (0.147)	0.910	0.128	0.023 (0.163)	0.325	0.144
(x) TMLE w/ SL	0.022 (0.182)	0.055	0.163	0.021 (0.146)	0.913	0.128	0.020 (0.163)	0.325	0.143
No. events	34			78			48		
Exposure rate = 20%									
(i) Crude	-0.211 (0.243)	0.035	0.372	-0.201 (0.094)	0.676	0.150	-0.354 (0.167)	0.037	0.341
(ii) Covariate adj	-0.008 (0.258)	0.055	0.277	-0.003 (0.105)	0.869	0.119	-0.106 (0.188)	0.176	0.240
(iii) PS adj	-0.012 (0.254)	0.054	0.277	-0.020 (0.102)	0.858	0.122	-0.109 (0.185)	0.172	0.242
(iv) PS one-spline	-0.012 (0.260)	0.057	0.277	-0.001 (0.106)	0.867	0.120	-0.103 (0.190)	0.180	0.240
(v) PS two-spline	0.058 (0.287)	0.085	0.223	-0.014 (0.121)	0.887	0.108	-0.013 (0.192)	0.282	0.174
(vi) PS strata	-0.036 (0.261)	0.053	0.288	-0.021 (0.106)	0.853	0.122	-0.139 (0.191)	0.157	0.252
(vii) IPTW	-0.065 (0.353)	0.095	0.277	-0.025 (0.126)	0.866	0.119	-0.063 (0.236)	0.191	0.240
(viii) AIPTW	-0.064 (0.354)	0.095	0.277	-0.022 (0.126)	0.866	0.119	-0.058 (0.236)	0.191	0.240
(ix) TMLE w/o SL	-0.065 (0.349)	0.060	0.269	-0.016 (0.122)	0.872	0.112	-0.057 (0.229)	0.243	0.189
(x) TMLE w/ SL	-0.069 (0.350)	0.061	0.267	-0.013 (0.123)	0.876	0.110	-0.057 (0.233)	0.252	0.187
No. events	36			48			39		

TABLE 3 Median bias, variance, type I error, power, and mean estimated variance in estimating the marginal odds ratio (on the log scale) using correctly specified propensity score model and outcome regression model, under the null, the alternative, and strong treatment heterogeneity, with stronger confounding effect.

Methods	Null: OR = 1			Alternative: OR = 3			Heterogeneity: OR = 1.6		
	Bias (Var)	Type I Error	Mean Est Var	Bias (Var)	Power	Mean Est Var	Bias (Var)	Power	Mean Est Var
Exposure rate = 69%									
(i) Crude	-0.235 (0.149)	0.099	0.182	-0.239 (0.113)	0.748	0.128	-0.314 (0.127)	0.087	0.167
(ii) Covariate adj	0.012 (0.177)	0.059	0.202	0.004 (0.127)	0.852	0.158	0.082 (0.155)	0.306	0.212
(iii) PS adj	0.009 (0.191)	0.068	0.202	0.044 (0.137)	0.856	0.164	0.122 (0.180)	0.343	0.220
(iv) PS one-spline	0.008 (0.179)	0.063	0.202	0.004 (0.130)	0.849	0.158	0.076 (0.156)	0.303	0.211
(v) PS two-spline	0.015 (0.213)	0.077	0.160	0.024 (0.144)	0.905	0.123	0.024 (0.169)	0.311	0.141
(vi) PS strata	0.005 (0.174)	0.059	0.202	0.002 (0.126)	0.852	0.159	0.064 (0.152)	0.292	0.212
(vii) IPTW	0.040 (0.226)	0.079	0.202	0.050 (0.183)	0.841	0.158	0.044 (0.204)	0.308	0.212
(viii) AIPTW	0.041 (0.227)	0.082	0.202	0.051 (0.183)	0.841	0.158	0.047 (0.204)	0.311	0.212
(ix) TMLE w/o SL	0.039 (0.263)	0.072	0.183	0.050 (0.241)	0.864	0.142	0.042 (0.221)	0.324	0.162
(x) TMLE w/ SL	0.036 (0.279)	0.072	0.181	0.052 (0.210)	0.865	0.142	0.041 (0.229)	0.327	0.161
No. events	34			78			48		
Exposure rate = 20%									
(i) Crude	-0.293 (0.252)	0.061	0.507	-0.278 (0.098)	0.493	0.225	-0.494 (0.192)	0.017	0.556
(ii) Covariate adj	-0.020 (0.285)	0.079	0.334	-0.003 (0.123)	0.787	0.161	-0.168 (0.235)	0.139	0.338
(iii) PS adj	-0.020 (0.275)	0.079	0.329	-0.043 (0.115)	0.761	0.168	-0.171 (0.223)	0.128	0.338
(iv) PS one-spline	-0.018 (0.293)	0.085	0.335	-0.003 (0.125)	0.787	0.160	-0.155 (0.238)	0.145	0.333
(v) PS two-spline	0.067 (0.484)	0.132	0.239	-0.021 (0.168)	0.831	0.128	-0.004 (0.255)	0.262	0.205
(vi) PS strata	-0.061 (0.292)	0.079	0.351	-0.033 (0.126)	0.765	0.166	-0.213 (0.242)	0.115	0.361
(vii) IPTW	-0.137 (0.486)	0.148	0.334	-0.055 (0.180)	0.758	0.161	-0.119 (0.355)	0.146	0.338
(viii) AIPTW	-0.130 (0.509)	0.152	0.334	-0.054 (0.186)	0.753	0.161	-0.115 (0.365)	0.152	0.338
(ix) TMLE w/o SL	-0.131 (0.481)	0.097	0.321	-0.023 (0.165)	0.798	0.132	-0.095 (0.350)	0.225	0.238
(x) TMLE w/ SL	-0.137 (0.485)	0.102	0.318	-0.013 (0.167)	0.809	0.129	-0.094 (0.363)	0.243	0.233
No. events	36			48			39		

TABLE 4 Median bias, variance, type I error, power, and mean estimated variance in estimating the marginal odds ratio (on the log scale) when propensity score model is misspecified, or outcome regression model is misspecified, or both are misspecified, under original confounding effect.

Methods	Null: OR = 1			Alternative: OR = 3			Heterogeneity: OR = 1.6		
	Bias (Var)	Type I Error	Mean Est Var	Bias (Var)	Power	Mean Est Var	Bias (Var)	Power	Mean Est Var
Exposure rate = 69%									
Outcome regression model misspecified									
(ii) Covariate adj	-0.040 (0.160)	0.053	0.161	-0.044 (0.123)	0.899	0.125	-0.063 (0.141)	0.222	0.149
(viii) AIPTW	0.027 (0.182)	0.071	0.161	0.022 (0.147)	0.880	0.125	0.024 (0.163)	0.338	0.149
(ix) TMLE w/o SL	0.027 (0.182)	0.057	0.164	0.023 (0.148)	0.911	0.128	0.026 (0.163)	0.326	0.144
(x) TMLE w/ SL	0.023 (0.181)	0.056	0.163	0.022 (0.147)	0.913	0.128	0.022 (0.163)	0.325	0.143
Propensity score model misspecified									
(iii) PS adj	-0.044 (0.166)	0.056	0.164	-0.027 (0.126)	0.897	0.131	-0.049 (0.149)	0.239	0.151
(iv) PS one-spline	-0.041 (0.161)	0.051	0.164	-0.042 (0.123)	0.895	0.128	-0.062 (0.142)	0.225	0.149
(iv) PS two-spline	-0.044 (0.190)	0.058	0.152	-0.047 (0.125)	0.887	0.118	-0.082 (0.138)	0.230	0.133
(vi) PS strata	-0.051 (0.157)	0.053	0.163	-0.050 (0.121)	0.892	0.128	-0.067 (0.139)	0.217	0.149
(vii) IPTW	-0.025 (0.178)	0.062	0.166	-0.028 (0.143)	0.842	0.131	-0.058 (0.159)	0.234	0.164
(viii) AIPTW	0.025 (0.180)	0.063	0.166	0.023 (0.145)	0.867	0.131	0.064 (0.167)	0.353	0.164
(ix) TMLE w/o SL	-0.014 (0.176)	0.054	0.162	-0.005 (0.141)	0.901	0.127	-0.019 (0.157)	0.284	0.143
(x) TMLE w/ SL	0.022 (0.182)	0.056	0.163	0.021 (0.146)	0.913	0.128	0.023 (0.163)	0.328	0.143
Both outcome regression and propensity score model misspecified									
(viii) AIPTW	-0.025 (0.178)	0.066	0.162	-0.028 (0.144)	0.853	0.125	-0.059 (0.158)	0.259	0.149
(ix) TMLE w/o SL	-0.024 (0.178)	0.056	0.161	-0.028 (0.144)	0.890	0.126	-0.059 (0.158)	0.245	0.141
(x) TMLE w/ SL	0.022 (0.182)	0.056	0.163	0.021 (0.146)	0.913	0.128	0.022 (0.162)	0.325	0.143
No. events	33			75			46		
Exposure rate = 20%									
Outcome regression model misspecified									
(ii) Covariate adj	-0.056 (0.256)	0.047	0.297	-0.052 (0.103)	0.833	0.126	-0.182 (0.182)	0.121	0.266
(viii) AIPTW	-0.063 (0.354)	0.086	0.297	-0.024 (0.126)	0.854	0.126	-0.056 (0.237)	0.169	0.266
(ix) TMLE w/o SL	-0.065 (0.350)	0.061	0.270	-0.016 (0.122)	0.873	0.112	-0.052 (0.231)	0.247	0.189
(x) TMLE w/ SL	-0.069 (0.350)	0.061	0.268	-0.014 (0.122)	0.874	0.111	-0.054 (0.234)	0.255	0.187
Propensity score model misspecified									
(iii) PS adj	-0.059 (0.254)	0.051	0.269	-0.066 (0.101)	0.841	0.117	-0.187 (0.181)	0.156	0.208
(iv) PS one-spline	-0.060 (0.258)	0.052	0.272	-0.053 (0.104)	0.850	0.116	-0.182 (0.185)	0.161	0.207
(iv) PS two-spline	0.010 (0.404)	0.105	0.235	-0.067 (0.115)	0.870	0.106	-0.140 (0.212)	0.201	0.168
(vi) PS strata	-0.083 (0.257)	0.050	0.278	-0.071 (0.104)	0.837	0.118	-0.204 (0.183)	0.144	0.214
(vii) IPTW	-0.111 (0.338)	0.094	0.277	-0.074 (0.122)	0.838	0.119	-0.188 (0.227)	0.122	0.240
(viii) AIPTW	-0.058 (0.335)	0.091	0.277	-0.021 (0.122)	0.865	0.119	-0.110 (0.229)	0.167	0.240
(ix) TMLE w/o SL	-0.100 (0.342)	0.057	0.259	-0.029 (0.121)	0.878	0.104	-0.151 (0.235)	0.206	0.177
(x) TMLE w/ SL	-0.069 (0.350)	0.061	0.268	-0.015 (0.122)	0.873	0.111	-0.054 (0.232)	0.254	0.187
Both outcome regression and propensity score model misspecified									
(viii) AIPTW	-0.113 (0.341)	0.087	0.297	-0.073 (0.122)	0.825	0.126	-0.188 (0.227)	0.106	0.266
(ix) TMLE w/o SL	-0.111 (0.337)	0.052	0.264	-0.068 (0.118)	0.854	0.109	-0.184 (0.221)	0.165	0.185
(x) TMLE w/ SL	-0.069 (0.353)	0.061	0.268	-0.014 (0.122)	0.874	0.111	-0.053 (0.232)	0.254	0.186
No. events	36			47			38		

TABLE 5 Median bias, variance, type I error, power, and mean estimated variance in estimating the marginal odds ratio (on the log scale) when propensity score model is misspecified, or outcome regression model is misspecified, or both are misspecified, under stronger confounding effect.

Methods	Null: OR = 1			Alternative: OR = 3			Heterogeneity: OR = 1.6		
	Bias (Var)	Type I Error	Mean Est Var	Bias (Var)	Power	Mean Est Var	Bias (Var)	Power	Mean Est Var
Exposure rate = 69%									
Outcome regression model misspecified									
(ii) Covariate adj	-0.059 (0.173)	0.064	0.193	-0.066 (0.126)	0.830	0.148	-0.101 (0.148)	0.191	0.185
(viii) AIPTW	0.040 (0.228)	0.087	0.193	0.051 (0.184)	0.848	0.148	0.046 (0.206)	0.350	0.185
(ix) TMLE w/o SL	0.039 (0.269)	0.073	0.182	0.049 (0.237)	0.864	0.142	0.042 (0.241)	0.327	0.162
(x) TMLE w/ SL	0.037 (0.277)	0.073	0.181	0.052 (0.212)	0.865	0.142	0.041 (0.231)	0.328	0.161
Propensity score model misspecified									
(iii) PS adj	-0.060 (0.183)	0.068	0.191	-0.036 (0.134)	0.843	0.153	-0.086 (0.163)	0.215	0.179
(iv) PS one-spline	-0.059 (0.174)	0.061	0.191	-0.068 (0.128)	0.836	0.148	-0.102 (0.151)	0.200	0.175
(iv) PS two-spline	-0.036 (0.201)	0.079	0.155	-0.042 (0.129)	0.891	0.117	-0.096 (0.159)	0.218	0.134
(vi) PS strata	-0.064 (0.167)	0.061	0.191	-0.066 (0.123)	0.837	0.149	-0.105 (0.145)	0.189	0.176
(vii) IPTW	-0.032 (0.213)	0.070	0.200	-0.024 (0.171)	0.822	0.157	-0.086 (0.192)	0.202	0.209
(viii) AIPTW	0.046 (0.222)	0.079	0.200	0.052 (0.177)	0.851	0.157	0.104 (0.209)	0.366	0.209
(ix) TMLE w/o SL	-0.017 (0.681)	0.082	0.175	0.005 (0.669)	0.877	0.135	-0.035 (0.633)	0.293	0.154
(x) TMLE w/ SL	0.040 (0.281)	0.072	0.181	0.053 (0.206)	0.868	0.142	0.043 (0.232)	0.331	0.161
Both outcome regression and propensity score model misspecified									
(viii) AIPTW	-0.032 (0.214)	0.076	0.193	-0.023 (0.171)	0.834	0.147	-0.083 (0.190)	0.237	0.184
(ix) TMLE w/o SL	-0.035 (0.758)	0.088	0.173	-0.025 (0.716)	0.870	0.134	-0.087 (0.649)	0.249	0.152
(x) TMLE w/ SL	0.038 (0.280)	0.072	0.181	0.054 (0.206)	0.869	0.142	0.043 (0.224)	0.329	0.160
No. events	33			75			46		
Exposure rate = 20%									
Outcome regression model misspecified									
(ii) Covariate adj	-0.091 (0.281)	0.068	0.371	-0.078 (0.119)	0.728	0.175	-0.271 (0.224)	0.070	0.398
(viii) AIPTW	-0.133 (0.507)	0.140	0.371	-0.056 (0.190)	0.731	0.175	-0.112 (0.365)	0.117	0.398
(ix) TMLE w/o SL	-0.130 (0.485)	0.097	0.321	-0.020 (0.166)	0.799	0.132	-0.086 (0.358)	0.235	0.236
(x) TMLE w/ SL	-0.136 (0.487)	0.102	0.318	-0.014 (0.167)	0.808	0.129	-0.086 (0.366)	0.245	0.232
Propensity score model misspecified									
(iii) PS adj	-0.089 (0.268)	0.069	0.307	-0.110 (0.113)	0.748	0.152	-0.275 (0.218)	0.109	0.261
(iv) PS one-spline	-0.092 (0.280)	0.072	0.309	-0.082 (0.121)	0.769	0.147	-0.269 (0.229)	0.120	0.259
(iv) PS two-spline	0.016 (0.359)	0.102	0.233	-0.100 (0.140)	0.814	0.120	-0.200 (0.218)	0.151	0.194
(vi) PS strata	-0.130 (0.281)	0.067	0.326	-0.108 (0.121)	0.748	0.152	-0.305 (0.226)	0.104	0.271
(vii) IPTW	-0.184 (0.431)	0.146	0.335	-0.125 (0.165)	0.711	0.161	-0.318 (0.323)	0.089	0.339
(viii) AIPTW	-0.096 (0.424)	0.129	0.335	-0.043 (0.167)	0.763	0.161	-0.200 (0.328)	0.126	0.339
(ix) TMLE w/o SL	-0.168 (0.439)	0.085	0.301	-0.040 (0.158)	0.818	0.117	-0.262 (0.344)	0.174	0.216
(x) TMLE w/ SL	-0.136 (0.475)	0.100	0.317	-0.014 (0.167)	0.808	0.129	-0.089 (0.362)	0.243	0.232
Both outcome regression and propensity score model misspecified									
(viii) AIPTW	-0.177 (0.447)	0.133	0.375	-0.123 (0.167)	0.690	0.175	-0.311 (0.331)	0.073	0.396
(ix) TMLE w/o SL	-0.183 (0.433)	0.078	0.309	-0.096 (0.153)	0.773	0.125	-0.302 (0.319)	0.125	0.229
(x) TMLE w/ SL	-0.135 (0.481)	0.101	0.318	-0.014 (0.167)	0.808	0.129	-0.086 (0.362)	0.244	0.232
No. events	36			47			38		

5 | APPLICATION TO THE ACEI AND ANGIOEDEMA STUDY

In this section, we analyze a subset of data obtained from an observational cohort study using EHR data from 2008-2012 at Kaiser Permanente Washington, a managed healthcare system in Washington State that is part of the FDA's Sentinel network [42]. The goal of this evaluation is to compare the effect of Angiotensin-Converting Enzyme Inhibitors (ACEI) and Beta Blockers (BB), which are two medications used to control high blood pressure, on incidence of angioedema in the first 30 days after starting either medication. There is a known elevated risk among those who take ACEIs relative to BBs for incidence of angioedema especially early after initial drug exposure [45].

Our cohort includes 31,269 prescribed to ACEI and 15,025 to BB. Among those prescribed to ACEIs, 49 subjects had an angioedema event (0.157%), and 5 had an angioedema event (0.033%) among BB prescribers yielding an unadjusted OR of 4.715. We reanalyze this data set with all of the methods described in the previous sections. For the analysis, we include all of the following potential confounders: NSAIDs (Nonsteroidal anti-inflammatory drugs), aspirin, ORAL-CS (optimizing recovery after laparoscopic colon surgery), allergic reaction, diabetes, heart disease, Ischemic HD (heart disease), inpatient hospitalization, and gender and one categorical variable which is age (categories: 18-44, 45-54, 55-64, and 65-99). This example was also mimicked for the simulation study presented in the Section 4.1. Details of the confounders including the prevalence and relationship to the exposure and the outcome were shown previously in Table 1 .

TABLE 6 Estimation and inference for a marginal odds ratio (ATE) comparing ACEI and BB on angioedema.

Methods	OR	Std Err log(OR) scale	Risk of angioedema ACEI	Risk of angioedema BB	P value	95% CI
(i) Crude	4.715	0.382	0.157%	0.033%	<0.001	(2.230, 9.967)
(ii) Covariate adj	5.677	0.421	0.167%	0.029%	<0.001	(2.490, 12.946)
(iii) PS adj	5.670	0.420	0.167%	0.029%	<0.001	(2.488, 12.923)
(iv) PS one-spline	5.485	0.413	0.165%	0.030%	<0.001	(2.441, 12.326)
(v) PS two-spline	5.632	0.419	0.167%	0.030%	<0.001	(2.478, 12.800)
(vi) PS strata	5.612	0.418	0.166%	0.030%	<0.001	(2.473, 12.735)
(vii) IPTW	6.363	0.421	0.239%	0.077%	<0.001	(2.790, 14.509)
(viii) AIPTW	6.477	0.421	0.160%	0.025%	<0.001	(2.841, 14.770)
(ix) TMLE w/o SL	6.494	0.491	0.161%	0.025%	<0.001	(2.481, 16.997)
(x) TMLE w/ SL	6.492	0.491	0.161%	0.025%	<0.001	(2.480, 16.994)

Std Err: standard error on the log(OR) scale.

We present in Table 6 results of applying the various methods estimating the marginal OR to the ACEI and BB cohorts. The methods considered are the same as the ones evaluated in our simulation studies described in Section 4.1. All of the methods found a statistically significant association in increased risk for angioedema when the entire population is treated with ACEI, compared to when the entire population is treated with BB. The estimated marginal OR is approximately 6 across all methods. In the study population, the average adjusted risk of angioedema in 30 days under ACEI treatment is around 0.17% (53 events out of 31, 269) whereas the average adjusted risk of angioedema in 30 days under BB is around 0.03% (5 events out of 15, 025). The PS one-spline method had the smallest estimated standard error. All three doubly robust estimation procedures had similar results. In particular, super learning with a library of data-adaptive algorithms produced similar estimation and inference as the ones from TMLE using parametric models.

6 | DISCUSSION

In the era of big data, use of electronic health record and claims data increasingly enables detection of safety signals for rare adverse outcomes in a realistic population once a drug or medical product has been approved and incorporated into routine clinical care. When making inference on comparative safety using routinely collected healthcare data, propensity score methods are particularly advantageous due to the insufficient number of outcome events and the potentially large number of confounders. In this paper, we have shown that there is a great potential in using flexible regression adjustment of the propensity score coupled with standardization to estimate a marginal, causal effect in a select population for rare binary outcomes. We illustrated that simply adjusting for the propensity score as a covariate in the outcome regression model can result

in residual confounding bias [7, 8]. In contrast, fitting a nonlinear function of the propensity score in the outcome regression model is a simple correction of such bias. In addition, we proposed direct estimation of the variance which is a fast and reliable approach for performing inference compared to computationally extensive approaches such as the bootstrap procedure.

Our simulation studies have shown that the PS one-spline method resulted in less bias without loss of efficiency. Moreover, when the propensity score model is correctly specified, it performed equivalently and often better than the existing methods under homogeneous treatment effect. However, under treatment heterogeneity, PS one-spline had notable increase in bias. In contrast, PS two-spline method had the least bias among all competing methods in this scenario, although it had larger variance than other regression on PS methods. The relatively larger variance of the PS two-spline approach under homogeneous effect is expected because it relies on a larger model without incorporating any knowledge about the outcome model, whereas the PS one-spline method gains efficiency by using the extra knowledge that treatment is a constant conditional on the PS, at the price of an increased bias when such assumption is incorrect. IPTW with stabilized weight had similar to slightly larger variance but much larger bias than the PS two-spline method. We note that they rely on the same model in the sense that both assume the same PS model and make no assumption about the outcome model. Therefore, the similar efficiency is expected and the smaller bias of PS two-spline could be attributed to the smoothing procedure of the nonparametric regression which stabilizes the estimate in finite sample. Although regression on covariates is an oracle estimator since it fits the true model under homogeneous effect, it frequently encountered the “perfect fit" phenomenon with predicted probabilities that are exactly zero or one in addition to non-convergence, with a large increase in bias under model misspecification.

The three doubly robust estimators, i.e., AIPTW and TMLE with and without super learning were the least sensitive to model misspecification, and had generally larger variance than regression based methods that depend on correctly specified corresponding models, due to the fact that the doubly robust estimators operate on a larger model under which the estimator is consistent as long as one of the outcome or PS model is correctly specified without necessarily knowing which one is correct. The small bias of TMLE with super learning under model misspecification or effect heterogeneity suggested that using ensemble learning with data adaptive algorithms may be a promising alternative to reduce bias.

Our results confirm prior findings of [17], which evaluated propensity-based estimators of the marginal relative risk in a rare outcome setting and found that PS one-spline method performed the best under the null. They also generate a strong non-null treatment heterogeneity scenario and found that PS one-spline may not work well in this scenario whereas PS two-spline was less biased but with large standard error. They considered evaluation of both ATE and ATT estimation and compared estimators with and without trimming the PS, which are not included in our study. In contrast, they did not consider a homogeneous non-null effect or doubly robust estimators which are studied in our simulation.

Our proposed variance estimation also provided valid type I error and high power, although under a rare outcome setting with small number of events, the power was highly dependent on the strength of the exposure effect. Applying all of the aforementioned methods to a real world application, we see that the estimated effects were similar across different methods. PS one-spline method had the smallest estimated standard error. Moreover, super learning with a library of data-adaptive algorithms produced similar estimation and inference as TMLE using parametric models.

Findings of our simulation study are limited in several aspects. First, we considered only one real-world study, thus the interpretation of our results may not be generalizable to other studies.In addition, the ACEI and BB study consists of only a moderate number of design-based confounders, which may not sufficiently reflect the high-dimensional covariate setting when working with healthcare databases. Nonetheless, the average number of events per covariate is in fact less than five, which parallels the setting of insufficient information for flexible outcome regression adjusting for all covariates in previous studies [17]. In addition, the observed exposure rate was uncommon in postmarketing surveillance. Further study is needed to perform realistic simulation comparing multiple real-world studies with a much larger number of covariates and ranging numbers of events. Moreover, our simulation study is also limited to the particular data generating distribution specified. Future study should consider settings such as a richer set of homogeneous and heterogeneous treatment effects, exposure prevalences and outcome event rates, near violations of the positivity assumption, and model misspecifications. A practical challenge is the modeling of high dimensional covariates to estimate the propensity score which is subject to potential misspecification. Recently, the covariate balancing propensity score has been proposed which mitigates the effect of the potential misspecification of a parametric propensity score model by selecting parameter values that maximize the resulting covariate balance [46]. It has been successfully incorporated in matching and weighting methods but not yet fully studied in regression adjustment methods. Therefore, a potential future research is to investigate the performance of regression on a data-driven propensity score learned targeting covariate balance.

In summary, in postmarketing surveillance with a rare outcome but a common exposure, we suggest the following: first, focus on fitting the propensity score model to balance covariates and reduce dimensionality; second, apply flexible regression adjustment of propensity score to control for confounding, and then standardize to a prespecified target population for marginal causal comparison; lastly, use the proposed variance estimation method which is particularly attractive in the postmarketing surveillance setting using large healthcare databases due to its simple and fast calculation and good approximation of the variance under a rare outcome setting. If there is evidence of treatment heterogeneity, PS two-spline or TMLE with super learning may be preferred. If the knowledge about the propensity score model is insufficient but event size is sufficient, we suggest considering alternative methods such as the TMLE with super learning which had better performance when the propensity score model

was likely to be misspecified. We also suggest regression on the propensity score strata with standardization as a sensitivity analysis as it fits a step function of the propensity score with relatively smaller degrees of freedom, which may work particularly well with discrete covariates.

ACKNOWLEDGEMENTS

This work was supported by the Sentinel System, a project sponsored by the U.S. Food and Drug Administration (FDA), United States to support monitoring the safety of FDA-regulated medical products. The Sentinel System is one piece of the Sentinel Initiative, a multifaceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) contract number HHSF223201400030I. This work was also supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number UL1 TR002319. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Behrman Rachel E, Benner Joshua S, Brown Jeffrey S, McClellan Mark, Woodcock Janet, Platt Richard. Developing the Sentinel System—a national resource for evidence development. *New England Journal of Medicine.* 2011;364(6):498–499.

[2] Härdle Wolfgang. *Applied nonparametric regression*. Cambridge university press; 1990.

[3] Rosenbaum Paul R, Rubin Donald B. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41–55.

[4] Robins James. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling.* 1986;7(9-12):1393–1512.

[5] Snowden J.M., Rose S., Mortimer K.M.. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology.* 2011;173(7):731–738.

[6] Vansteelandt Stijn, Keiding Niels. Invited commentary: G-computation–lost in translation?. *American journal of epidemiology.* 2011;173(7):739–742.

[7] Austin Peter C, Grootendorst Paul, Normand Sharon-Lise T, Anderson Geoffrey M. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in medicine.* 2007;26(4):754–768.

[8] Austin Peter C. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in medicine.* 2007;26(16):3078–3094.

[9] Robins James M, Rotnitzky Andrea. Comment on “Inference for semiparametric models: Some questions and an answer" by P.J. Bickel and J. Kwon. *Statistica Sinica.* 2001;11:920–936.

[10] Hade Erinn M, Lu Bo. Bias associated with using the estimated propensity score as a regression covariate. *Statistics in medicine.* 2014;33(1):74–87.

[11] Vansteelandt Stijn, Daniel Rhian M. On regression adjustment for the propensity score. *Statistics in medicine.* 2014;33(23):4053–4072.

[12] Wan Fei, Mitra Nandita. An evaluation of bias in propensity score-adjusted non-linear regression models. *Statistical Methods in Medical Research.* 2016;27(3):846 - 862.

[13] Little Roderick, An Hyonggin. Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica.* 2004;:949–968.

[14] Gutman Roee, Rubin Donald B. Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in medicine.* 2015;34(26):3381–3398.

[15] Schafer Joseph L, Kang Joseph. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods.* 2008;13(4):279.

[16] Myers Jessica A, Louis Thomas A. Comparing treatments via the propensity score: stratification or modeling?. *Health Services and Outcomes Research Methodology.* 2012;12(1):29–43.

[17] Franklin Jessica M, Eddings Wesley, Austin Peter C, Stuart Elizabeth A, Schneeweiss Sebastian. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Statistics in medicine.* 2017;36(12):1946–1963.

[18] Hampel Frank R. The influence curve and its role in robust estimation. *Journal of the american statistical association.* 1974;69(346):383–393.

[19] Hahn J., Ridder G.. Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica.* 2013;81(1):315–340.

[20] Hastie Trevor, Tibshirani Robert. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological).* 1993;:757–796.

[21] Härdle Wolfgang, Liang Hua, Gao Jiti. *Partially linear models*. Springer Science & Business Media; 2012.

[22] Hansen Bruce E. Nonparametric sieve regression: Least squares, averaging least squares, and cross-validation. In: Oxford University Press 2014. In J. S. Racine, L. Su & A. Ullah (eds.).

[23] De Boor C.. *A practical guide to splines*. Applied Mathematical Sciences, New York: Springer; 1978.

[24] De Boor C.. On calculating with B-splines. *Journal of Approximation Theory.* 1972;6:50–62.

[25] Hansen Bruce E. The integrated mean squared error of series regression and a Rosenthal Hilbert-space inequality. *Econometric Theory.* 2015;31(2):337–361.

[26] Localio A Russell, Margolis David J, Berlin Jesse A. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of clinical epidemiology.* 2007;60(9):874–882.

[27] Greenland Sander. *Introduction to regression modelling. Chapter 21. In: Rothman KJ, Greenland S, Lash TL (eds).* Lippincott Williams & Wilkins; 2008.

[28] Gruber Susan, van der Laan Mark J. TMLE: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software.* 2012;51(13).

[29] Vaart Aad W. On differentiable functionals. *The Annals of Statistics.* 1991;:178–204.

[30] Newey Whitney K. The Asymptotic Variance of Semiparametric Estimators. *Econometrica.* 1994;62(6):1349-1382.

[31] Vaart Aad W. *Asymptotic statistics*. Cambridge university press; 1998.

[32] Tsiatis Anastasios. *Semiparametric theory and missing data*. Springer Science & Business Media; 2007.

[33] Hahn Jinyong, Ridder Geert. *The asymptotic variance of semi-parametric estimators with generated regressors.* : Centre for Microdata Methods and Practice, Institute for Fiscal Studies; 2010.

[34] Hahn Jinyong, Liao Zhipeng, Ridder Geert. Nonparametric two-step sieve M estimation and inference. *Econometric Theory.* 2018;34(6):1281–1324.

[35] Robins James M, Hernan Miguel Angel, Brumback Babette. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.

[36] Potter Frank J. A study of procedures to identify and trim extreme sampling weights. *In Proceedings of the Section on Survey Research Methods, American Statistical Association.* 1990;:225-230.

[37] Potter Frank J. The effect of weight trimming on nonlinear survey estimates. *In Proceedings of the American Statistical Association, Section on Survey Research Methods.* 1993;758–763.

[38] Robins James M, Rotnitzky Andrea, Zhao Lue Ping. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association.* 1994;89(427):846–866.

[39] Bang Heejung, Robins James M. Doubly robust estimation in missing data and causal inference models. *Biometrics.* 2005;61(4):962–973.

[40] van der Laan M.J., Rubin D.. Targeted maximum likelihood learning. *The International Journal of Biostatistics.* 2006;2(1).

[41] van der laan Mark J, Polley Eric C, Hubbard Alan E. Super learner. *Statistical applications in genetics and molecular biology.* 2007;6(1).

[42] Nelson Jennifer C., Boudreau Denise, Wellman Robert, Yu Onchee, Cook Andrea J., et. al.. Improving Sequential Safety Surveillance Planning Methods for Routine Assessments that use Regression Adjustment or Weighting to Control Confounding https://www.sentinelsystem.org/sentinel/methods/routine-prospective-safety-surveillance-new-drugs-vaccines-and-other-biologic/Mini-Sentinel Methods Report; 2016.

[43] Franklin Jessica M, Schneeweiss Sebastian, Polinski Jennifer M, Rassen Jeremy A. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational statistics & data analysis.* 2014;72:219–226.

[44] Cook Andrea J., Wellman Robert., et. al.. Safety signalling methods for survival outcomes to control for confounding in the Mini-sentinel distributed database https://www.sentinelinitiative.org/sites/default/files/Methods/Mini-Sentinel_Methods_Survival_Outcomes_II_Final_Report.pdfFDA's Sentinel Initiative: Project Report; 2018.

[45] Roujeau Jean Claude, Stern Robert S. Severe adverse cutaneous reactions to drugs. *New England Journal of Medicine.* 1994;331(19):1272–1285.

[46] Imai Kosuke, Ratkovic Marc. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2014;76(1):243–263.