Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

# Accounting for matching structure in post-matching analysis of observational studies

Yuyang Zhang & Bo Lu

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
2022, VOL. 51, NO. 6, 3081–3099
https://doi.org/10.1080/03610918.2019.1708928

Taylor & Francis
Taylor & Francis Group

# Accounting for matching structure in post-matching analysis of observational studies

Yuyang Zhang and Bo Lu

Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, USA

**ABSTRACT**

Matching design is commonly used in social science and health research with observational data, as it is robust to outcome model misspecification and has the intuitive interpretation similar to blocked randomization design. Estimate the population average treatment effect with propensity score adjustment is very popular. From a practical perspective, however, it is not clear whether the post-matching analysis should adjust for the matching structure. Analytical strategies with and without accounting for matching design have appeared in literature. For continuous outcomes, the implication is more on the variance estimation. But for binary outcomes, the non-collapsibility problem for the odds ratio adds another layer of complexity in choosing between estimation strategies. We have conducted extensive simulation studies to compare several matching estimators and the propensity score weighting estimator for both continuous and binary outcomes. Especially, we consider three measures for binary outcomes, risk difference, relative risk and odds ratio. Our simulation results suggest that statistical methods accounting for matching structure are more advantageous and among binary effect measures, odds ratio tends to have higher power than other measures. We also apply different estimation strategies to a U.S. trauma care database to examine mortality difference between trauma centers and non-trauma centers.

**ARTICLE HISTORY**
Received 26 September 2018
Accepted 18 December 2019

**KEYWORDS**
Propensity score matching; Weighting; Causal inference; Conditional inference

## 1. Introduction

In social science and health research, observational studies provide a rich source of data for evaluating the impact of interventions or programs. Samples in observational studies are usually more representative of the real population that the intervention is intended to be applied. Matching is a commonly used design to balance the covariate distribution, hence to remove the confounding bias. Comparing to parametric modeling, matching is robust to outcome model misspecification and has the intuitive interpretation similar to blocked randomization designs. When there are only several covariates to be balanced, direct matching on individual variables may be used. When there is a large number of confounding variables, which is likely to be the case in practice,

**CONTACT** Bo Lu lu.232@osu.edu Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, USA

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

Supplemental data for this article can be accessed here.

matching needs to be done based on some composite summary of the data, such as Mahalanobis distance. To estimate the average causal effect, Rosenbaum and Rubin (1983) showed that propensity score matching is an effective way of removing bias. Propensity score can also be used in the form of stratification, weighting or covariance adjustment (Lunceford and Davidian 2004).

To make valid causal inference, it requires more assumptions than merely estimating the associatinal relationship. First, stable unit treatment value assumption (SUTVA) dictates that the treatment applied to one subject does not affect the outcomes for other subjects and there is only one version of the treatment. Second, no unmeasured confounding assumption implies that adjusting for observed pretreatment covariates is sufficient to remove all confounding bias. Third, common support assumption assures that the treated and untreated groups are comparable in terms of observed covariates. The first assumption is important for both randomized experiments and observational studies. The latter two assumptions may be referred as strongly ignorable treatment assignment assumption that is more relevant for observational studies. The common support assumption is more relevant for matching design. If the covariates in two groups do not overlap well, it is very difficult to create high quality matched pairs, which subsequently may bias the analysis.

Unlike modeling based strategies, matching has some unique design features (Stuart 2010). First, it needs a measure of closeness. When the dimension of covariates is high, it is impossible to match exactly on every single one. A composite distance metric is often used, i.e., Mahalanobis distance or propensity score based distance, etc. To further improve the matching quality, caliper matching may be used, where the caliper refers to additional constraints on the within-pair discrepancy of certain variables or the propensity score. Second, it needs some special algorithm to implement matching. The popular choices are nearest neighbor matching algorithm with caliper or optimal matching algorithm. Third, there are different options for constructing the matched sets. Pair matching is a special case of the $1:k$ matching design, where it matches one treated to $k$ untreated subjects ($k \geq 1$). Full matching is known to minimize the discrepancy between treated and untreated groups, but Rosenbaum (1991) mentioned that the post-matching analysis is more complicated as not all matched sets have the same size. Fourth, just like residual analysis for regression models, it is important to conduct post-matching balance diagnostics. The purpose of matching is to create matched groups that are comparable in terms of covariate distributions. In the sense, it recreates a randomization-like scenario, which can be used to infer causal relationship without much modeling. If the two groups of data, collected via an observational manner, cannot be matched well. It is a warning signal that any effort to infer causal effect might be futile, because the two groups of subjects are different in some important ways. This is a distinctive feature from the modeling based strategy, where one can always run models regardless of the data overlap. But the resulting treatment effect estimates are likely just speculations based on the model.

With a well-matched dataset, the post-matching inference can be straightforward. One can just contrast the difference in outcomes between two groups without further adjustment, as covariates are balanced by matching as a priori. Post-matching modeling may be needed either as a means to reduce residual confounding in matched pairs or to improve the efficiency of the estimate by taking advantage of predictive information from important covariates. Based on Rubin (1973)'s work, combining matching with regression modeling is regarded as a more robust way of estimating causal effects. However, it is less clear to the practitioners what is the most appropriate model to run with matched data. To stick with the matching design, it seems natural to account for the pairing structure in the analysis. This is referred as conditional modeling, which requires more sophisticated statistical tools. If one views the matching as a pre-processing step to create balanced covariate distributions at group level, a marginal modeling strategy without adjusting for the pairing structure may be used. For continuous outcomes, the implication is more on the variance estimation. But for binary outcomes, the popular measure of odds ratio is known to have the non-collapsibility problem, which adds another layer of complexity in choosing between marginal and conditional models.

In this paper, we try to fill the gap by providing practical advice on post-matching analytical strategies, based on extensive simulation studies. In Sec. 2, we review both conditional and marginal modeling methodologies for continuous and binary outcomes. Secs. 3 and 4 present simulation setup and results for continuous and binary outcomes, respectively. In Sec. 5, we apply various estimating strategies to a U.S. trauma care database to examine mortality difference between trauma centers and non-trauma centers. Section 6 discusses the findings and provides practical advices.

## 2. Post-matching inference strategies

Matching design in observational studies has received a lot of discussions, from the matching algorithms, distance metrics to balance assessment in matched data. But, relatively little attention has been paid to post-matching analysis in terms of whether the matching structure should be accounted for. In practice, after matched sets are generated satisfactorily based on a certain criterion, researchers tend to pick a convenient analytical strategy without too much thoughts on the matching structure. The analytical strategy that accounts for the matching structure is referred as conditional modeling and the strategy ignoring the matching structure is referred as marginal modeling. In this section, we will review the two strategies and discuss their impacts on treatment effect estimation in details when they are applied to continuous and binary outcomes. To keep notations simple, we focus on pair matching design for the rest of the paper.

### 2.1. Two views on post-matching analysis

The conditional modeling strategy accounts for the pairing structure in the analysis by using statistical methods such as the paired t-test, the McNemar's test or the conditional logistic regression, etc. It follows the matching design closely, which mimics a paired randomization design. Under the assumption that matching has successfully removed all observed confounding, Rosenbaum (2002) discussed a class of nonparametric test statistics for matched pair data, namely sign-score statistics. It includes the commonly used Wilcoxon's signed rank statistic and McNemar's statistic as special cases. It takes advantage of the randomization distribution under the mull hypothesis of no treatment effect. It does not rely on parametric modeling assumption on the outcomes, hence it is robust to potential model misspecifications. If good information regarding the predictive relationship between the outcome and covariates is available, we may expand the framework by including important predictors in outcome models. For continuous outcomes, a natural choice is to use the linear mixed model to account for the pairing structure. Another option is to model the pairwise difference using a linear regression model as suggested by Imbens and Rubin (2015) for pairwise randomized experiments. It becomes more complex for binary outcomes, as there are multiple ways of measuring the treatment effect. Three commonly used measures are risk difference (RD), relative risk (RR) and odds ratio (OR). Conditional logistic regression models are popular for estimating conditional OR. Generalized linear mixed models with identity or logarithm link function may be used for RD or RR, respectively. The generalized estimating equation (GEE) method is popular for correlated binary data, which can also be used to estimate the marginal effect while accounting for correlation in pairs.

A different view of matching regards it as a nonparametric pre-processing step (Ho et al. 2007). Matching is a simple way to eliminate or reduce the relationship between observed covariates and the treatment indicator. With successfully matched data, covariates and the treatment assignment are uncorrelated, which mimics a completely randomized experiment. Such process may also eliminate an important source of model dependence in the post-matching parametric analysis stemming from the functional form misspecification. Ho et al. (2007) argued that matching does not require pairing observations, as long as the resulting covariate distributions are close in treated and untreated groups. In this sense, post-matching analysis does not necessarily need to account for the matching structure. There could be the situation where the distributions of an important covariate are very similar between different treatment groups, but within some pairs, there are big difference in covariate values. For example, in matching with multiple covariates, it is likely to have matched data with similar proportions of males in both treated and untreated groups, but some pairs matching a treated male to an untreated female and other pairs matching a treated female to an untreated male. Without the need to account for the pairing structure, post-matching analysis would be easier to implement.

Examples of real data analysis following either view point can be found in literature (Austin and Mamdani 2006; Saposnik et al. 2012). For continuous outcomes, when estimating the average treatment effect, the difference is in variance estimation. Because the mean differences are the same with and without considering the paired structure, the point estimates of treatment effects are the same using either conditional or marginal models. For binary outcomes, it is more complicated as some popular measures, such as odds ratio, suffer from the collapsibility problem, which dictates that the conditional model and the marginal model will produce different results. Austin (2011a) compared paired and non-paired inference methods for estimating risk difference in matched data and recommended that statistical methods accounting for the paired structure should be used. We will expand the investigation by including RR and OR, and evaluate their statistical performance in terms of estimation bias, variance and power. The next two subsections review the estimating strategies for continuous and binary outcomes, respectively, with the following notations:

$Y$: observed outcome

$T$: treatment indicator (1 for treated and 0 for untreated)

$\mathbf{X}$: a vector of observed covariates

$\pi$: propensity score, $\pi = P(T = 1|\mathbf{X})$

$m$: index of matched pairs, $m = 1, ..., M$

$\{mi\}$: index of subject $i$ in matched pair $m$

## 2.2. Inference with continuous outcomes

It is easy to see that the post-matching point estimate of treatment effect is not affected by the matching structure,

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}[Y_{m1} - Y_{m2}] &= \frac{1}{M}\sum_{m=1}^{M}Y_{m1} - \frac{1}{M}\sum_{m=1}^{M}Y_{m2}\\ &= \frac{1}{M}\sum_{i\in\{matched, treated\}} Y_i - \frac{1}{M}\sum_{j\in\{matched, untreated\}} Y_j\end{aligned}$$

where $\{m1\}$ indexes the treated subject and $\{m2\}$ indexes the untreated subject in pair $m$.

Assuming $Y$ follows a normal distribution, we consider four estimating strategies: conditional model with no covariate adjustment (CON), conditional model with covariate adjustment (CON_ADJ), marginal model with no covariate adjustment (MAR), and marginal model with covariate adjustment (MAR_ADJ). CON and CON_ADJ use linear mixed models to account for the pairing structure and the only difference is whether the model adjusts for covariates. MAR and MAR_ADJ use regular linear regression models and the difference is whether covariates are adjusted for. Since the GEE method provides a marginal interpretation of the treatment effect while adjusting for the correlation within pairs, we include it in the comparison under MAR and MAR_ADJ.

We focus on the variance estimation as we do not expect a difference in the point estimates. Since there is no covariate adjustment, the variance of CON is smaller than that of MAR if there is positive correlation within pairs, which is similar to comparing a paired t-test to an independent t-test. With covariate adjustment, the difference between the two methods becomes smaller since the inclusion of covariates accounts for part of the variation observed in outcomes. This also depends on whether the outcome model is correctly specified. In our simulation studies, we explore both situations when the correct model is used and when an incorrect model is used. With misspecified models, statistical methods have been derived to adjust the variance estimate to make robust inference such as Gail, Tan, and Piantadosi (1988) and Lin and Wei (1989)'s work. Suppose the true parameter is $\theta$ and the maximum likelihood estimator under the misspecified model is $\widehat{\theta}$. Under some mild conditions, $\widehat{\theta}$ converges to $\theta^*$ with the asymptotic distribution:

$$\widehat{\theta} \sim N(\theta^*, \widehat{A}(\widehat{\theta})^{-1}\widehat{B}(\widehat{\theta})\widehat{A}(\widehat{\theta})^{-1})$$

where $\widehat{A}(\widehat{\theta})^{-1}\widehat{B}(\widehat{\theta})\widehat{A}(\widehat{\theta})^{-1}$ is the so-called sandwich estimator. For many misspecified parametric models, the bias between $\theta^*$ and $\theta$ becomes zero asymptotically (Gail, Wieand, and Piantadosi 1984). So using the sandwich estimator provides a more robust inference strategy for the variance estimation.

### 2.3. Inference with binary outcomes

With binary outcomes, the marginal inference and the conditional inference may yield quite different results. Among the three measures that we consider, OR is arguably the most popular one in medical and health research and is known to have the noncollapsibility problem, where the marginal OR does not equal to the conditional OR. Pang, Kaufman, and Platt (2016) discussed extensively the noncollapsibility issue for OR in presence of confounding bias. We are more interested in evaluating the estimation of the three measures and comparing their performance in an observation study setup. In this subsection, we review the inference under unadjusted marginal and conditional models for each of three measures to provide some theoretical insights.

With $Y$ being dichotomous, denote $Y=1$ for a success and $Y=0$ for a failure, where success/failure are broadly defined for any event, i.e., the onset of a disease, etc. Suppose the data are successfully matched with a paired design (one treated subject matched to one untreated subject) and presented as Table 1 (there are $n$ pairs):

It is easy to re-organize it as an independent table by ignoring the matching structure, shown in Table 2:

Denote $p_1 = P(Y = 1|T = 1)$ be the success probability in the treated group and $p_0 = P(Y = 1|T = 0)$ be the success probability in the untreated group.

For RD, the marginal inference is straightforward as a two-sample independent proportion comparison with following point and variance estimates (the subscript ".m" for marginal):

$$\widehat{\Delta}_{RD.m} = \widehat{p}_1 - \widehat{p}_0 = \frac{a+c}{n} - \frac{a+b}{n} = \frac{c-b}{n}$$

$$\widehat{V}(\widehat{\Delta}_{RD.m}) = \frac{(a+c)(b+d)}{n^3} + \frac{(a+b)(c+d)}{n^3}$$

The conditional inference for RD is more complicated. Based on Chen (1996)'s work, the maximum likelihood estimation yields the following results (the subscript ".c" for conditional):

**Table 1.** Paired design table.

| | T = 1 | | |
|---|---|---|---|
| T = 0 | Y = 1 | Y = 0 | Total |
| Y = 1 | a | b | a + b |
| Y = 0 | c | d | c + d |
| Total | a + c | b + d | n |

**Table 2.** Independent design table.

| | Y = 1 | Y = 0 | Total |
|---|---|---|---|
| T = 1 | a + c | b + d | n |
| T = 0 | a + b | c + d | n |
| Total | 2a + b + c | b + c + 2d | 2n |

$$\widehat{\Delta}_{RD.c} = \frac{c-b}{n}$$
$$\widehat{V}(\widehat{\Delta}_{RD.c}) = \frac{b+c-(b-c)^2/n}{n^2}$$

Since RD does not suffer from the noncollapsibility, both methods yield the same point estimate. With the paired design, Agresti (2002) and Agresti and Min (2004) showed that the covariance between $p_1$ and $p_0$ could be estimated by $(ad-bc)/n^3$. Assuming a positive correlation, i.e., $ad-bc>0,$ we have

$$\widehat{V}(\widehat{\Delta}_{RD.m}) - \widehat{V}(\widehat{\Delta}_{RD.c}) = \frac{2(ad-bc)}{n} > 0$$

where it implies that the conditional model is more efficient than the marginal model.

For RR, we are interested in estimating, $p_1/p_0$ or $\log(p_1/p_0),$ where the latter is better approximated by a Normal distribution. The marginal inference yields (Zou 2004):

$$\log(\widehat{RR.m}) = \log(\widehat{p}_1) - \log(\widehat{p}_0) = \log\frac{(a+c)}{(a+b)}$$
$$\widehat{V}(\log(\widehat{RR.m})) = \frac{1}{a+c} - \frac{1}{n} + \frac{1}{a+b} - \frac{1}{n}$$

Chen (1996) presented that the maximum likelihood estimation for RR yields the following results:

$$\log(\widehat{RR.c}) = \log\frac{(a+c)}{(a+b)}$$
$$\widehat{V}(\log(\widehat{RR.c})) = \frac{(b+c)}{(a+b)(a+c)}$$

Since RR does not suffer from the noncollapsibility, both methods yield the same point estimates. Assuming a positive correlation, i.e., $ad-bc>0,$ we have

$$\widehat{V}(\log(\widehat{RR.m})) - \widehat{V}(\log(\widehat{RR.c})) = 2\frac{ad-bc}{(a+b)(a+c)n} > 0$$

where it implies that the conditional model is more efficient than the marginal model.

Because OR suffers from the noncollapsibility, the marginal and conditional models do not estimate the same thing. We use the logarithm transformation of OR, as it is better approximated by a Normal distribution. Based on a multinomial distribution, Agresti and Min (2004) showed that the marginal inference yields the following results:

$$\log(\widehat{OR.m}) = \log\frac{(a+c)(c+d)}{(b+d)(a+b)}$$
$$\widehat{V}(\log(\widehat{OR.m})) = \frac{1}{a+b} + \frac{1}{a+c} + \frac{1}{c+d} + \frac{1}{b+d}$$

Based on a conditional likelihood approach, Agresti and Min (2004) derived the estimates for conditional OR as below:

$$\log(\widehat{OR.c}) = \log\frac{c}{b}$$
$$\widehat{V}(\log(\widehat{OR.c})) = \frac{1}{b} + \frac{1}{c}$$

There is no consensus which measure should be used to capture the treatment effect of a dichotomous outcome. OR is often seen in medical and health research because the logistic regression model is convenient with easy interpretation. But the downside is that it suffers the noncollapsibility even without confounding (Pang, Kaufman, and Platt 2016). Different designs or different covariate adjustment strategies would produce estimates that are not directly comparable. Though RD and RR are collapsible, the outcome modeling strategies are not as straightforward as the OR case. RD often involves an identity link in generalized linear models, which is not that appropriate as it assumes a Normal distribution for the probability that should be always between 0 and 1. RR involves an logarithm link in generalized linear models, which is also restrictive as the range of the predictive component needs to be positive. In later sections, we conduct an extensive simulation study to evaluate the performance of three measures in terms of estimation bias, accuracy, and power. By summarizing our results, we try to provide some practical guidelines for estimating treatment effects of binary outcomes with a paired design.

## 3. Simulation study: continuous outcomes

We simulate continuous outcomes from Normal distributions and the focus is on comparing the estimation efficiency. To assess the quality of matching on the inference, we consider two distribution overlap scenarios – large overlap and moderate overlap, where the latter is supposed to result in less well matched pairs. Along the same line, we also consider three matching algorithms, optimal matching (OPTM), nearest neighbor matching (NNM) with a caliper of 0.5 standard deviation of the estimated propensity scores (NNM_0.5), NNM with a caliper of 0.2 standard deviation of the estimated propensity scores (NNM_0.2). OPTM and NNM are two popular algorithms used in matching design. Unlike NNM, OPTM minimizes the the sum of pairwise distances among all possible pairings. It is optimal in the global sense that results in a matched set with smaller total distance than NNM (Hansen and Klopfer 2006). The resulting number of pairs is different for the three algorithms. OPTM uses all the treated subjects and an equal number of untreated ones. NNM_0.5 may discard treated subjects if they cannot be matched with untreated ones within the caliper. NNM_0.2 tend to discard more treated subjects to satisfy the more stringent caliper. To investigate the impact due to misspecified outcome models, we consider two data generating models— simple and complex, where the simple model only has linear covariate terms and the complex model has exponential and quadratic terms. In the moderate overlap scenario, we consider three covariates with the following distributions:

$$X_1 \sim N(2,1), X_2 \sim N(-3,1.5^2), X_3 \sim Bern(0.4).$$

The treatment assignment model is:

$$\text{logit}[P(T=1)] = 1.5 + \log(0.5)x_1 + \log(1.3)x_2.$$

The true outcome generating models are $Y \sim N(\mu, 3^2)$, where

- Simple model:

$$\mu = -2 + 2T + \log(10)x_1 - \log(3)x_2 + \log(0.01)x_3,$$

- Complex model:

$$\mu = -2 + 2T - \log(1.5)e^{x_1} + \log(1.75)x_2^2 + \log(0.15)x_3.$$

Various analytical strategies are applied to the simulated data. For the propensity score matched data, the following models are considered:

- IND: A linear regression model ignoring the pairing structure, including T as the only predictor.
- IND_ADJ: A linear regression model ignoring the pairing structure, including T and all covariates X's.
- PAR: A paired t-test analysis, without adjusting for any covariate.
- RBT: IND_ADJ using robust variance estimation for potentially misspecified models.
- CON: A linear mixed model preserving the pairing structure, including T as the only predictor.
- CON_ADJ: A linear mixed model preserving the pairing structure, including T and all covariates X's.
- MAR: A GEE model (to account for the pair level correlation), including T as the only predictor.
- MAR_ADJ: A GEE model (to account for the pair level correlation), including T and all covariates X's.

To compare with other methods commonly seen in literature, we apply the following models to the full dataset:

- NAI: A linear model, including T as the only predictor.
- NAI_ADJ: A linear model, including T and all covariates X's.
- IPTW: A linear model weighted by the inverse of the estimated propensity scores, including T as the only predictor.
- IPTW_ADJ: A linear model weighted by the inverse of the estimated propensity scores, including T and all covariates X's.
- RAN: A linear model applied to the dataset with a randomly assigned treatment, including T as the only predictor.
- RAN_ADJ: A linear model applied to the dataset with a randomly assigned treatment, including T and all covariates X's.

In addition to the unadjusted analysis, we also run adjusted analysis that includes all three covariates in each of the above outcome models.

In practice, the true propensity scores are not known and need to be estimated. It is not a simple task to guess the correct functional form of the propensity score model. So we also consider two scenarios of propensity score modeling – a correct model and a misspecified model. For the moderate overlap case, the misspecified model only includes $X_1$. Each dataset contains 1000 simulated observations and the simulation is repeated for 2000 times for all scenarios.

Figure 1 depicts the results for unadjusted and adjusted analyses under the moderate overlap scenario. The two plots on the left panel present the percentage of bias for each

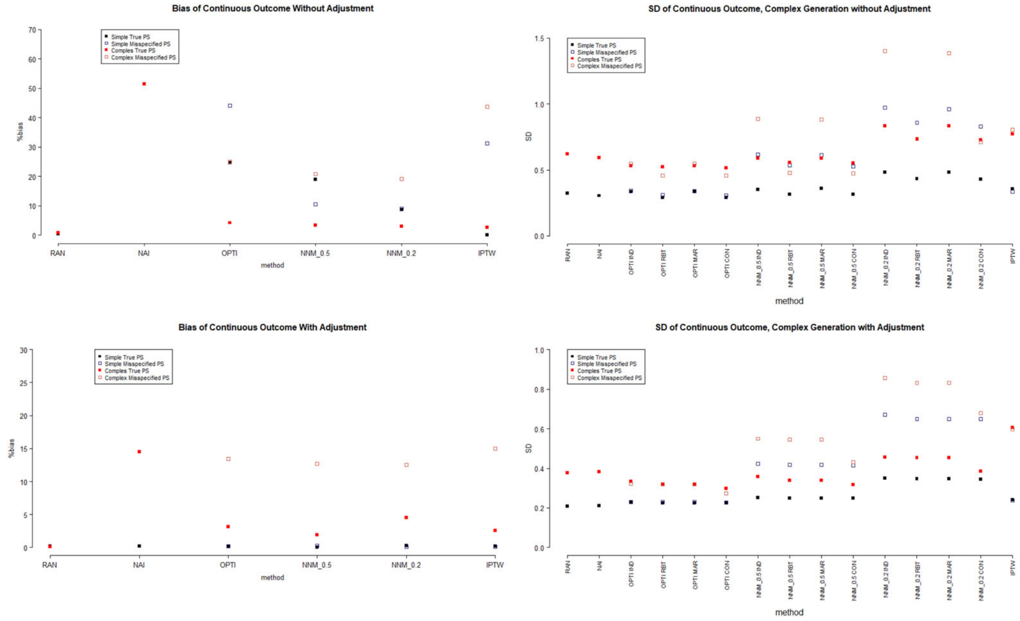

**Figure 1.** Results for continuous outcome.

estimator, i.e., $(\widehat{\beta} - 2)/2 \times 100\%$. The two plots on the right panel present the average standard deviation (SD) estimates under different methods. Results under the simple data generation are shown with the dark color and results under the complex data generation are shown with the red color. The results for unadjusted analyses are plotted in the top panel. In the upper left plot, the first two columns of points show the results for simple two-group comparison when the treatment is randomly assigned (RAN) and when the treatment is not randomly assigned (NAI). It is easy to see that the two-group comparison under random assignment is unbiased and the naive analysis introduces substantial biases as the assignment is not random. The remaining four columns represent different estimation strategies for non-randomized data when the correct propensity score model is used (solid point) and when the misspecified propensity score model is used (hollow point). Three different types of matching and the weighting method are compared. Consistent with Austin (2011b)'s findings, NNM_0.2 tends to yield the least bias because it obtains the best matching quality with a smaller caliper. In the upper right plot, with the correct propensity score model, the conditional methods (PAR and CON) produce smaller variance estimates than the marginal methods (IND and MAR). The main reason that matching with caliper methods produce large variance is because they use a much smaller subset of observations than the optimal matching (see the footnote). IPTW yields unbiased point estimates with moderate to large variance, because the correct propensity score model is used. With the misspecified propensity score model, all methods present biased point estimates, which is expected due to the model misspecification. The biases are big for the IPTW method as the doubly robust property does not hold when there is no model adjustment and incorrect propensity score model is used. The variance estimate pattern is similar to matching scenarios with conditional methods having smaller variance than marginal methods.

The results for adjusted analyses are plotted in the bottom panel. For model adjusted analyses, we consider robust variance estimation method (RBT) for potential model misspecification. There is no bias for randomized studies. For the non-randomized study, NAI shows no bias with model adjustment under the simple data generation because the true outcome model is used for adjustment. It presents some moderate bias under the complex data generation because the outcome model is misspecified. When the correct propensity score model is used, all matching methods present no bias under the simple data generation. The biases are slightly larger under the complex generation, but comparable to IPTW results. With model adjustment, the variance estimates are close with the linear mixed model having the smallest values. The variance estimates under robust estimation and GEE model are almost identical because both methods use sandwich type of estimator. Using the misspecified propensity score model, the model adjustment also improve the estimates. Under the simple data generation, all methods are unbiased. Under the complex data generation, all methods show more bias than the simple case and matching methods tend to perform slightly better than IPTW.

Due to the space limit, we only present results for the moderate overlap scenario as it is more practically relevant. The simulation results for the large overlap scenario are included in the online supplementary document. The findings are similar to the moderate overlap scenario except that the IPTW method tends to perform a bit better than matching methods overall. This is likely due to the fact that, with large covariate distribution overlap, the estimated propensity score values are more regular (further away from 0's or 1's). Therefore, there are no extreme weights and IPTW tends to do better.

## 4. Simulation study: binary outcomes

We simulate binary outcomes using the approach described in Austin (2010)'s study, where the data are simulated for a pre-specified risk difference through an iterative process. Similar to the simulation study of continuous outcomes, we consider two covariate overlap scenarios (moderate and large), three matching algorithms (OPTM, NNM_0.5, NNM_0.2), two outcome generating models (simple and complex), two propensity score scenarios (correctly and incorrectly specified propensity score models). Results for large overlap and moderate overlap with simple generation are included in the online supplementary document for space consideration.

In the moderate overlap scenario, we consider three covariates with the same distributions and the same assignment probability model as the continuous case. The true outcome generating models are:

- Simple model:

$$\operatorname{logit}[P(Y=1)] = 1 + \beta_t \times T + \log(3)x_1 + \log(1.5)x_2 + \log(1.5)x_3,$$

where $\beta_t$ is determined iteratively for a pre-specified risk difference by evaluating marginal risk differences using Monte Carlo integration. For example, $\mathrm{RD} = 0.1$ corresponds to $\mathrm{RR} = 1.798$ and a marginal $\mathrm{OR} = 2.030$.

- Complex model:

$$\text{logit}[P(Y=1)] = -7 + \beta_t \times T + \log(0.5) \times e^{x_1} + \log(1.5) \times x_2^2 + \log(1.5) \times x_3.$$

where $\beta_t$ is determined in the same way as above. For example, RD $= 0.1$ corresponds to RR $= 1.422$ and a marginal OR $= 1.636$.

Various analytical strategies can be applied to binary data. Because RD and RR are collapsible, both marginal and conditional modeling strategies yield the same point estimate. For marginal models, we consider generalized linear models (GLM) with appropriate links ignoring the matching structure. We also use GEE models accounting for the correlation due to matching. For conditional models, generalized linear mixed models (GLMM) are often used to account for the matching structure. However, Greenland, Robins, and Pearl (1999) discussed OR is non-collasible, which implies that the marginal and conditional models would yield different results. To compare all three measures on the same ground, we choose to only use marginal models, which carry the interpretation of the average effect of the study population. Specifically, the following models are consider for the propensity score matched data:

- GLM: A generalized linear regression model ignoring the pairing structure, including T as the only predictor. An identify link is used for RD estimation, a log link is used for RR estimation and a logit link is used for OR estimation (same for the models below).
- GLM_ADJ: A generalized linear regression model with appropriate link functions ignoring the pairing structure, including T and all covariates X's.
- GEE: A GEE model (to account for the pair level correlation) with appropriate link functions, including T as the only predictor.
- GEE_ADJ: A GEE model with appropriate link functions (to account for the pair level correlation), including T and all covariates X's.

To compare with other commonly used methods, we also apply the following models to the full dataset:

- NAI: A generalized linear model with appropriate link functions, including T as the only predictor.
- NAI_ADJ: A generalized linear model with appropriate link functions, including T and all covariates X's.
- IPTW: A generalized linear model with appropriate link functions, weighted by the inverse of the estimated propensity scores, including T as the only predictor.
- IPTW_ADJ: A generalized linear model with appropriate link functions, weighted by the inverse of the estimated propensity scores, including T and all covariates X's.
- RAN: A generalized linear model with appropriate link functions applied to the dataset with a randomly assigned treatment, including T as the only predictor.
- RAN_ADJ: A generalized linear model with appropriate link functions applied to the dataset with a randomly assigned treatment, including T and all covariates X's.

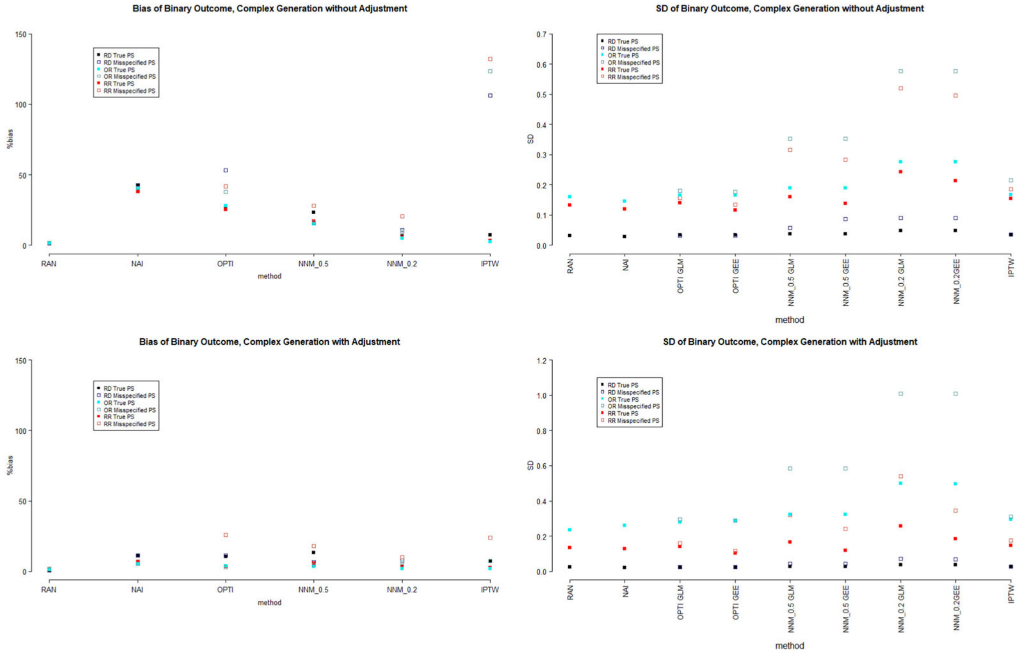

**Figure 2.** Results for binary outcome.

In addition to the unadjusted analysis, we also run adjusted analysis that includes all three covariates in each of the above outcome models.

Figure 2 depicts the results for adjusted and unadjusted analyses under the moderate overlap and complex generation scenario. The plots are arranged in a similar way as in Figure 1. For the unadjusted analyses (top panel), with correctly specified propensity scores, the estimation bias of matching methods are similar for all three measures and IPTW yields the least biased estimates. With misspecified propensity scores, the estimation bias of matching methods are similar for all three measures and IPTW yields the most biased results. Among different matching algorithms, NNM_0.2 tends to produce less biased estimates and this is likely due to the more stringent matching criteria. GLM and GEE variance estimates are close except for RR, where GEE estimates are much smaller. With covariate adjustment (bottom panel), the estimation biases are generally smaller than those without adjustment. With the true propensity score model (solid point), IPTW has the best overall result, thanks to the doubly robust property. But when the propensity score model is misspecified (hollow point), IPTW produces the worst results. Among the three effect measures, OR estimates tend to be less biased than RD and RR estimates. GLM and GEE variance estimates are close except for RR, where GEE estimates are much smaller. It is interesting that the naive regression analysis performs better than IPTW when the incorrect propensity score model is used. This is consistent with the findings in Kang and Schafer (2007)'s study that IPTW may perform poorly if both the propensity score and the outcome models are misspecified.

To gain more insights on the statistical performance of using each measure, we further examine the powers of detecting pre-specified risk differences. For a series of risk

Table 3. Power for binary outcomes with moderate overlap and complex generation.

| | RD = 0 | RD = 0.01 | RD = 0.03 | RD = 0.05 | RD = 0.07 | RD = 0.09 | RD = 0.10 |
|---|---|---|---|---|---|---|---|
| | OPTM | | | | | | |
| rd.glm.no* | 0.045 | 0.084 | 0.319 | 0.600 | 0.871 | 0.970 | 0.987 |
| rd.gee.no | 0.045 | 0.090 | 0.329 | 0.612 | 0.876 | 0.972 | 0.988 |
| rd.glm.yes† | 0.069 | 0.101 | 0.372 | 0.685 | 0.926 | 0.986 | 0.999 |
| rd.gee.yes | 0.070 | 0.106 | 0.379 | 0.696 | 0.930 | 0.988 | 0.999 |
| rr.glm.no | 0.019 | 0.043 | 0.176 | 0.428 | 0.740 | 0.919 | 0.967 |
| rr.gee.no | 0.044 | 0.089 | 0.321 | 0.605 | 0.874 | 0.969 | 0.987 |
| rr.glm.yes | 0.019 | 0.026 | 0.112 | 0.277 | 0.552 | 0.782 | 0.865 |
| rr.gee.yes | 0.067 | 0.091 | 0.280 | 0.543 | 0.796 | 0.935 | 0.975 |
| or.glm.no | 0.045 | 0.083 | 0.318 | 0.599 | 0.871 | 0.970 | 0.987 |
| or.gee.no | 0.044 | 0.090 | 0.325 | 0.608 | 0.875 | 0.972 | 0.988 |
| or.glm.yes | 0.057 | 0.101 | 0.427 | 0.785 | 0.967 | 0.998 | 1.000 |
| or.gee.yes | 0.061 | 0.108 | 0.442 | 0.799 | 0.969 | 0.998 | 1.000 |
| | IPTW | | | | | | |
| rd.no | 0.017 | 0.028 | 0.134 | 0.395 | 0.693 | 0.905 | 0.957 |
| rd.yes | 0.052 | 0.080 | 0.274 | 0.620 | 0.882 | 0.980 | 0.991 |
| rr.no | 0.016 | 0.027 | 0.128 | 0.385 | 0.675 | 0.898 | 0.952 |
| rr.yes | 0.065 | 0.078 | 0.216 | 0.465 | 0.728 | 0.905 | 0.953 |
| or.no | 0.016 | 0.027 | 0.130 | 0.388 | 0.684 | 0.901 | 0.953 |
| or.yes | 0.049 | 0.080 | 0.325 | 0.709 | 0.932 | 0.992 | 0.999 |

*: ".no" means no covariate adjustment; † : ".yes" means with covariate adjustment.

difference values (RD = 0, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1), we simulate 2000 datasets with 1000 sample in each dataset. Then we calculate the power of detecting a significant effect for RD, RR and OR, with and without covariate adjustment. Results are reported in Table 3. The column of "RD = 0" literally shows the type-I error estimates. The unadjusted analyses have type-I errors less than 5%, while the covariate adjusted analyses tend to inflate type-I errors to slightly larger than 5%. Overall, OR has the best power results, while RR has the lowest power. The matching method tends to produce better power than the corresponding weighting method.

## 5. Real data analysis: trauma care evaluation

Injury is the leading cause of death among young Americans aged 1-44 (source: CDC https://www.cdc.gov/injury/wisqars/LeadingCauses.html). Trauma centers provide specialized medical services and resources to patients suffering from traumatic injuries. Hospitals in the US are categorized as Trauma Centers (TC) and Non-Trauma Centers (NTC), according to resources and expertise. Admitting patients to TC or NTC is not a random process, which depends on various factors, including severity of the injury, geographical locations or other patient characteristics. The observational nature of the data presents challenges in evaluating care quality at different trauma centers. Using 2006–2010 NEDS data (Agency for Healthcare Research and Quality 2016), we plan to assess the performance of two levels of trauma care (NTC vs. TC) with respect to a key outcome, emergency department (ED) mortality. We consider trauma patients, aged 18-64, characterized by a severe trauma (injury severity score greater or equal to 25). Shi et al. (2016) described the detail of the data.

The binary exposure variable of interest is the admission to NTC and the binary outcome is ED mortality. The primary research question is "would the mortality of patients treated at a NTC be different if these patients had been treated at a TC?" A matching

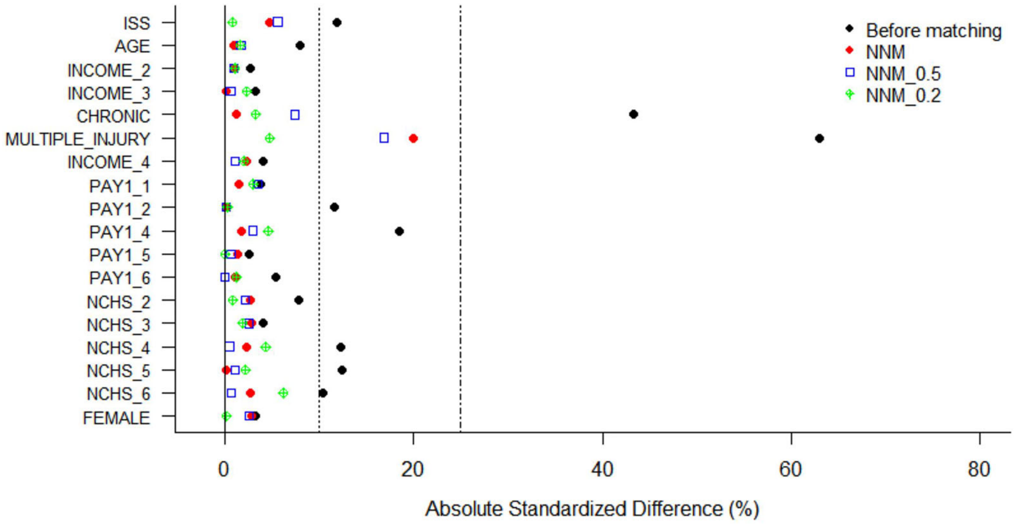

**Figure 3.** Covariate balance before and after matching.

design is appropriate to evaluate this causal effect, which corresponds to the exposure effect on the exposed. Our analytical dataset consists of 21,855 patients, of whom 5,314 (24.3%) and 16,541 (75.7%) patients were admitted to NTC and TC, respectively.

The propensity score of being admitted to NTC is estimated with a logistic regression model including 16 important covariates identified by content experts. They include age, sex, injury severity score, comorbidity of chronic conditions, multiple injuries, median household income by zip code, expected primary insurance payer, and urban rural designation for patient's county of residence. Matching are conducted with three algorithms (NNM without caliper, NNM_0.5, NNM_0.2) based on the estimated propensity scores. We do not use the optimal matching algorithm because the sample size of trauma data exceeds the limit of the R package used in our simulation study. Figure 3 presents the absolute standardized differences (ASD) for each covariate before and after matching as a means of checking balance. Usually, ASD $< 10\%$ is considered a good balance. Matching tends to obtain well matched pairs on all covariates, except the multiple injury indicator. The distributions of the multiple injury variable are very imbalanced in the original dataset, where the prevalence is 8% in NTC group and 33% in TC group. To achieve the best balance, we have to discard more observations. So NNM with 0.2 SD caliper produces the ideal result, which results in 3436 pairs. In contrast, NNM without caliper has 5314 pairs and NNM with 0.5 SD caliper has 4542 pairs. In addition to checking ASD, we also examine the variance ratios after matching as suggested by Rubin (2001) All post-matching variance ratios are less than 1.5 with NNM with 0.2 SD caliper showing the best balance (results not shown). Therefore, we believe that matching methods have produced well matched pairs and we can proceed with the outcome analysis.

We apply both unadjusted and adjusted analyses to the matched data. For comparison purpose, IPTW approach using full data is included. The top panel of Table 4 presents the unadjusted analysis. The treatment effect estimates are significant for all three measures, indicating a beneficial effect of being treated at trauma centers. For RD,

**Table 4.** Trauma data analysis without covariate adjustment.

| | | RD | | | RR | | | OR | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | n | Point Est. | SD (GLM) | SD (GEE) | Point Est. | SD (GLM) | SD (GEE) | Point Est. | SD (GLM) | SD (GEE) |
| Without covariate adjustment | | | | | | | | | | |
| NNM | 10628 | 0.066 | 0.006 | 0.006 | 1.867 | 0.061 | 0.057 | 2.012 | 0.065 | 0.065 |
| NNM_0.5 | 9084 | 0.061 | 0.006 | 0.006 | 2.018 | 0.074 | 0.071 | 2.159 | 0.077 | 0.077 |
| NNM_0.2 | 6872 | 0.062 | 0.006 | 0.006 | 2.760 | 0.106 | 0.103 | 2.950 | 0.109 | 0.109 |
| IPTW | 21855 | 0.046 | 0.007 | NA | 1.474 | 0.063 | NA | 1.553 | 0.070 | NA |
| With covariate adjustment | | | | | | | | | | |
| NNM | 10628 | 0.053 | 0.006 | 0.006 | 1.634 | 0.062 | 0.057 | 1.825 | 0.069 | 0.068 |
| NNM_0.5 | 9084 | 0.051 | 0.006 | 0.006 | 1.853 | 0.075 | 0.068 | 2.056 | 0.081 | 0.079 |
| NNM_0.2 | 6872 | 0.058 | 0.006 | 0.006 | 2.613 | 0.106 | 0.101 | 2.893 | 0.111 | 0.110 |
| IPTW | 21855 | 0.044 | 0.007 | NA | 1.453 | 0.059 | NA | 1.600 | 0.072 | NA |

all three matching designs yield very similar results, which show a bigger effect than the IPTW approach. For RR, estimates from NNM and NNM with 0.5 SD caliper are close, indicating that being treated at the non-trauma center may double the mortality risk. NMM with 0.2 SD caliper yields a bigger effect, while IPTW yields a smaller effect. The same pattern is observed for OR. In terms of efficiency, GEE models tend to provide smaller variance estimates.

The bottom panel of Table 4 presents the adjusted analysis where we include all linear terms of covariates in the regression models. After covariate adjustment, treatment effect sizes of matching methods become a bit smaller, while IPTW estimates do not change much. Also, covariate adjustment has a small impact on reducing the variance estimates.

We observe some discrepancies on treatment effect estimates between matching and weighting methods. This is likely due to the fact that the two methods are applied to different patient populations and the treatment effects are heterogeneous in subpopulations. Generally, matching methods work better when the two treatment groups have quite different sample sizes, then a well-matched sample can be selected. So the matching estimators are usually used for estimating the average treatment effect on the treated population (ATT). On the other hand, the weighting estimators are used for the average treatment effect on the entire population (ATE). In our simulation studies, we assume a constant effect. So there is no difference between ATT and ATE estimates. In practice, like our trauma center evaluation study, treatment effects are likely to be different in subpopulations. It seems the treatment effect in the entire trauma patient population is smaller than that in the NTC patient population.

There are also some small discrepancies on treatment effect estimates between matching algorithms. NNM with 0.2 SD caliper yields larger effect estimates than the other two algorithms. This is likely due to the matched sample differences. NNM with 0.2 caliper has the smallest sample size with the prevalence difference of multiple injury being 1.8%. The prevalence differences of multiple injury in other matched samples are 8% and 6% respectively. It seems that impact due to such imbalance can not be mitigated by using simple model adjustment as shown in the bottom panel of Table 4. Using OR as the effect measure, we observe an increase of 189% in odds of death if the patient had been treated at NTC in the well matched data. In the less well matched data, the increase in odds of death becomes 105% (NNM with 0.5 SD caliper).

## 6. Summary

We have compared various estimating strategies for matched observational data. In general, matching with covariate adjustment modeling provides better results than weighting methods when you are not sure about the propensity score model and the outcome model. For continuous outcomes, statistical approaches accounting for matching structure provide smaller variance estimates, which are about 5% less than other methods. For binary data, RD and RR are collapsible. Our simulation study shows that their performance are close under different matching algorithms. GEE models produce smaller variance than GLMs for RR, but the variance estimates are comparable for RD. Conditional models, such as GLMM, can also be run for RD and RR. For RD, the variance estimates from GLMM are close to both GLM and GEE models. But we encountered non-convergences issue with R function "glmer" (in the package "lme4") when we ran covariate-adjusted GLMMs to estimate RR. So we do not report any GLMM results. OR suffers the non-collapsibility, so we only consider marginal models in our simulations. Overall, it has slightly less biased estimates than RD and RR. The variance estimates of OR are similar under GLM and GEE models. In practice, researchers may choose the appropriate measure suitable for their study. As far as power is concerned, OR with matching and adjusted modeling yields better results over RD and RR.

For post-matching variance estimation, statistical methods accounting for the matching structure are recommended. The gains are more noticeable for continuous outcome and RR, while minimal for RD and OR. In comparison to IPTW method, the decision needs more thoughts. If the researchers are very confident about correctly specifying either the propensity score model or the outcome model, IPTW has an advantage. For complex observational studies, this is less likely to be the case. Matching seems to be a more robust method as long as good covariate balance is achieved. Another advantage of matching is that, as a design based method, the propensity score values do not go into the treatment effect estimation process, aside from matching. In this sense, we may ignore the variability in estimating propensity scores as long as we use it as a design tool to create well-matched pairs. On the contrary, the estimated propensity scores are part of the estimating equations for the IPTW. So the variability in estimating propensity scores should be accounted for, which adds another layer of complexity for practitioners.

A final note on matching is that there are more sophisticated matching algorithms to obtain better balance on covariates. As shown in our simulation, better covariate balance may improve the estimation accuracy. If the researchers prefer a high level balance on some important variables, they could consider exact matching on certain variables in addition to propensity score matching on other variables. Alternatively, fine balance matching (Rosenbaum 2002) is a way that forces a nominal variable to be balanced, and "rcbalance" is a R package for fine or near-fine balance (Pimentel et al. 2015).

## Acknowledgments

The authors thank Junxin Shi and Henry Xiang for the dataset preparation and insightful discussion on comparative trauma care research.

## Funding

This work was partially supported by grant 1R01 HS024263-01 from the Agency of Healthcare Research and Quality of the U.S. Department of Health and Human Services.

## References

Agency for Healthcare Research and Quality. 2016. Overview of the nationwide emergency department sample (NEDS). www.hcup-us.ahrq.gov/nedsoverview.jsp.

Agresti, A. 2002. *Categorical data analysis*. Hoboken, New Jersey: John Wiley & Sons.

Agresti, A., and Y. Min. 2004. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Statistics in Medicine* 23 (1):65–75. doi:10.1002/sim.1589.

Austin, P. C. 2010. A data-generation process for data with specified risk differences or numbers needed to treat. *Communications in Statistics - Simulation and Computation* 39 (3):563–77. doi:10.1080/03610910903528301.

Austin, P. C. 2011a. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Statistics in Medicine* 30 (11):1292–301. doi:10.1002/sim.4200.

Austin, P. C. 2011b. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* 10 (2):150–61. doi:10.1002/pst.433.

Austin, P. C., and M. M. Mamdani. 2006. A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 25 (12): 2084–106. doi:10.1002/sim.2328.

Chen, J. 1996. On the conditional and mixture model approaches for matched pairs. *Journal of Statistical Planning and Inference* 55 (3):319–29. doi:10.1016/S0378-3758(96)00080-8.

Gail, M., W.-Y. Tan, and S. Piantadosi. 1988. Tests for no treatment effect in randomized clinical trials. *Biometrika* 75 (1):57–64. doi:10.1093/biomet/75.1.57.

Gail, M. H., S. Wieand, and S. Piantadosi. 1984. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71 (3):431–44. doi:10.1093/biomet/71.3.431.

Greenland, S., J. M. Robins, and J. Pearl. 1999. Confounding and collapsibility in causal inference. *Statistical Science* 14 (1): 29–46. doi:10.1214/ss/1009211805.

Hansen, B. B., and S. O. Klopfer. 2006. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* 15 (3):609–27. doi:10.1198/106186006X137047.

Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15 (3):199–236. doi:10.1093/pan/mpl013.

Imbens, G. W., and D. B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press.

Kang, J. D., and J. L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22 (4): 523–39. doi:10.1214/07-STS227.

Lin, D. Y., and L.-J. Wei. 1989. The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association* 84 (408):1074–8. doi:10.1080/01621459.1989.10478874.

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): 2937–60. doi:10.1002/sim.1903.

Pang, M., J. S. Kaufman, and R. W. Platt. 2016. Studying noncollapsibility of the odds ratio with marginal structural and logistic regression models. *Statistical Methods in Medical Research* 25 (5):1925–37. doi:10.1177/0962280213505804.

Pimentel, S. D., R. R. Kelz, J. H. Silber, and P. R. Rosenbaum. 2015. Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association* 110 (510):515–27. doi:10.1080/01621459.2014.997879.

Rosenbaum, P. R. 1991. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society: Series B (Methodological)* 53 (3):597–610. doi:10.1111/j.2517-6161.1991.tb01848.x.

Rosenbaum, P. R. 2002. *Observational studies*. New York: Springer.

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.

Rubin, D. B. 1973. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 29 (1):185–203. doi:10.2307/2529685.

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):169–88.

Saposnik, G., M. K. Kapral, R. Cote, P. A. Rochon, J. Wang, S. Raptis, M. Mamdani, and S. E. Black. 2012. Is pre-existing dementia an independent predictor of outcome after stroke? A propensity score-matched analysis. *Journal of Neurology* 259 (11):2366–75. doi:10.1007/s00415-012-6508-4.

Shi, J., B. Lu, K. K. Wheeler, and H. Xiang. 2016. Unmeasured confounding in observational studies with multiple treatment arms: Comparing emergency department mortality of severe trauma patients by trauma center level. *Epidemiology* 27 (5):624–32. doi:10.1097/EDE.0000000000000515.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25 (1):1–21. doi:10.1214/09-STS313.

Zou, G. 2004. A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* 159 (7):702–6. doi:10.1093/aje/kwh090.