Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v12n2p491

Examination of Entropy balancing technique for estimating some standard measures of treatment effects: a simulation study

By Amusa, Zewotir, North

Published: 14 October 2019

This work is copyrighted by Università del Salento, and is licensed under a Creative Commons Attribuzione - Non commerciale - Non opere derivate 3.0 Italia License.
For more information see:
http://creativecommons.org/licenses/by-nc-nd/3.0/it/

Electronic Journal of Applied Statistical Analysis
Vol. 12, Issue 02, October 2019, 491-507
DOI: 10.1285/i20705948v12n2p491

Examination of Entropy balancing technique for estimating some standard measures of treatment effects: a simulation study

Lateef Amusa*, Temesgen Zewotir, and Delia North

Department of Statistics, University of Kwazulu-Natal, Westville campus, South Africa

Published: 14 October 2019

In observational studies, propensity score weighting methods are regarded as the conventional standard for estimating the effects of treatments on outcomes. We consider entropy balancing, which despite its excellent conceptual properties, has been under-utilized in the applied studies. Using an extensive series of Monte Carlo simulations, we evaluated the performance of entropy balancing, in estimating difference in means, marginal odds ratios, rate ratios, and hazard ratios. The performance of entropy balancing was relatively compared with that of inverse probability of treatment weighting using the propensity score. We found that entropy balancing outperformed the IPW method in estimating difference in means, marginal odds ratios, and hazard ratios, but when estimating marginal rate ratios, IPW performed better. Entropy balancing produced more biased estimates in many cases. However, the entropy balancing algorithm is capable of controlling bias by loosening the tightening of the pre-specified tolerance on covariate balance. We report findings as to when one technique is better than the other with no proclamation on whether one method is in every case superior to the other. Entropy balancing merits more widespread adoption in applied studies.

keywords: Entropy balancing, Monte Carlo simulation, Observational studies, Propensity score weighting, Treatment effect, odds ratios, hazard ratios, rate ratios.

*Corresponding author: amusasuxes@gmail.com

©Università del Salento
ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

1 Introduction

The evaluation of a treatment or intervention is particularly straightforward in experiments but very complicated in observational studies where treatment assignment is not random. In observational studies, treatment selection is usually related to the background covariates and can confound estimated treatment effects.

Estimation of treatment effects in observational studies has conventionally been done using propensity score (PS) methods (Austin, 2014; Dehejia and Wahba, 2002; Guo et al., 2006; Guo and Fraser, 2010; Hirshberg and Zubizarreta, 2017). Among the PS methods, PS weighting (Hirano and Imbens, 2001) received more attention. In particular, the inverse probability of treatment weighting (IPW) is the most commonly used weighting method by applied researchers and practitioners, especially in the medical and health sciences (Austin and Stuart, 2015).

Much recently, entropy balancing – an optimization-based method, has gained the attention of applied researchers (Adhikary et al., 2016; Brettschneider et al., 2017; Grupp et al., 2017; Mattke et al., 2015; Pearson et al., 2014). Entropy balancing (Hainmueller, 2012) performs excellently in achieving covariate balance and efficient estimation of treatment effects. Additionally, entropy balancing (EB) is straightforward to implement. EB calibrates weights using the control group's distribution moments as constraints while optimizing the covariate balance apriori. Consequently, EB obviates the need for continually specifying the PS model until the desired covariate balance is achieved.

There is an increasing interest in using entropy balancing to estimate marginal or average treatment effects on outcomes of different types (Adhikary et al., 2016; Brettschneider et al., 2017; Grupp et al., 2017; Mattke et al., 2015; Parish et al., 2018; Pearson et al., 2014). Accordingly, we investigate the performance of entropy balancing in estimating treatment effects on continuous, binary, count, and time-to-event outcomes.

Using the IPW method as a benchmark, the current study used Monte Carlo simulations to examine the performance of entropy balancing in estimating some measures of treatment effects. We considered the estimation of difference in means, odds ratios, rate ratios, and hazard ratios for the continuous, binary, count and time-to-event outcomes, respectively. We also utilized the average treatment effect among the treated (ATT) as our estimand of interest.

This paper is structured as follows: In the next Section, we describe briefly the methodology of the entropy balancing, as well as the inverse probability of treatment weighting. In Section 3, we describe the Monte Carlo simulation scheme that were used to examine the performance of the two considered techniques. In particular, we report on bias, mean squared error (MSE), model-based standard errors, and 95% confidence interval coverage. The simulation results are presented in Section 4. Finally, in Section 5, we summarize our findings and gave final remarks.

2 Methods

We briefly describe the weighting methods that were included in the simulation study. We consider units or subjects indexed i $(i = 1, ..., n)$. We assume that there is a binary treatment variable, T_i, and the size of the treated and control group units, respectively, n_1, n_0, are known, while $\mathbf{X}_k$ denote a K-dimensional column vector of the observed background covariates.

2.1 Entropy balancing

Entropy balancing is a preprocessing method that can guarantee covariates balance, via a reweighting scheme that assigns a scalar weight to each sample unit such that the reweighted groups satisfy a set of balance constraints that are imposed on the sample moments of the covariate distributions (Hainmueller, 2012). The reweighting scheme belongs to the family of Maximum Entropy methods, which has roots in information theory and applied statistics (Kullback, 1959; Golan, 2018; Aria et al., 2018; Ciavolino and Carpita, 2015; Carpita and Ciavolino, 2017). The weights w_i are selected to minimize the relative entropy:

$$\min_{w_i} H(w) = \min_{w_i} \sum_{i|T=0} w_i \; log(w_i/q_i) \tag{1}$$

subject to the constraints:

$$\sum_{i|T=0} w_i \; c_{ri} \; (X_k) = m_r, for \; r = 1, \ldots, R \tag{2}$$

$$\sum_{i|T=0} w_i = 1 \tag{3}$$

$$w_i \geq 0, \forall i, \tag{4}$$

where $q_i = \frac{1}{n_0}$ is a vector of the base weights, and m_r describes a set of R balance constraints imposed on the covariate moments of the reweighted control group. m_r is the formulation containing the rth order moment of a given variable X_k from the treated group, while the moment functions are specified for the control group as $c_{ri} \; (X_k) = X_k^r$ or $c_{ri} \; (X_k) = (X_k - \mu_k)^r$, with mean μ_k. Equation (2) is the balance constraint specified in terms of the rth moment to be achieved on all covariates; (3) is the normalization constraint, while (4) is the non-negativity constraint.

The minimization problem described above is computed from an unconstrained dual problem and reduced to a system of non-linear equations with R Lagrange multipliers (Hainmueller, 2012) of the form:

$$\min_z L^d = log(q' \; e^{-C'z}) + m'z \tag{5}$$

where $\mathbf{z} = (\lambda_1, \ldots, \lambda_R)'$ is a vector $(\mathbf{z}^*)$ of Lagrange multipliers for the balance constraints, rewritten in matrix form as $\mathbf{CW = M}$, with the $(R \times n_0)$ constraint matrix,

$\mathbf{C} = (c_1(\mathbf{X_k}), \ldots, c_R(\mathbf{X_k}))'$, and the vector of moments $\mathbf{m} = (m_1, \ldots, m_R)'$. The corresponding solution of (5) is:

$$W^* = \frac{Q \ . \ exp(-C'Z)}{Q' \ exp(-C'Z)} \tag{6}$$

An iterative Levenberg-Marquardt algorithm exploits the 2nd order information to solve the dual problem:

$$z^{new} = z^{old} - l \ \nabla_z^2 \ L^{d-1} \ \nabla_z \ L^d \tag{7}$$

Where l is the step length, ∇_z and ∇_z^2 is the gradient and Hessian, respectively. The optimal step length is selected for each iteration.

We utilized the ATT weights (Parish et al., 2018), which are defined for the entropy balancing as fixing treated units' weight at unity and reweighting the control group units using the algorithm described above.

2.2 Inverse Probability of Treatment Weighting

The propensity score, defined by $e(x) = P(T = 1|X), 0 < e < 1$, is the probability of a subject or unit receiving the treatment of interest given the observed baseline covariates (Rosenbaum, 1983). In IPW, each unit's weight equals the reciprocal of the probability of receiving the treatment that the unit received. We utilized the ATT weights (Austin and Small, 2014; Austin and Stuart, 2017), which are defined for the IPW as fixing the treated units' weight at unity, and the control units as $\frac{\hat{e}(x)}{1-\hat{e}(x)}$ (Imbens, 2004). We estimated $\hat{e}(x)$ by using a logistic regression model to regress treatment status on the covariates associated with the treatment.

3 Simulation study

We conducted a series of Monte Carlo simulations to examine the performance of entropy balancing in estimating treatment effects while using the IPW method as a benchmark. We considered continuous, binary, count and time-to-event outcomes. All simulations were done using the R statistical package (R Core Team, 2019).

3.1 Data-generating process

We used a data generation scheme derived from previous studies (Lee et al., 2010; Setoguchi et al., 2008). We randomly generated ten baseline covariates, where each of them $(X_1 - X_{10}) \sim N(0, 1)$. Some pair of covariates were induced with specified levels of dependence. $X_1, X_3, X_5, X_6, X_8, X_9$ were dichotomized. Figure 1 describes the simulation design in terms of the causal relationship of the variables.

As shown in Figure 1, the simulation study aligns with practice reality: X_1, X_2, X_3, X_4 are associated with both treatment and outcome, X_5, X_6, X_7 are predictors of the treatment variable only, while X_8, X_9, X_{10} are predictors of the outcome variable only. The

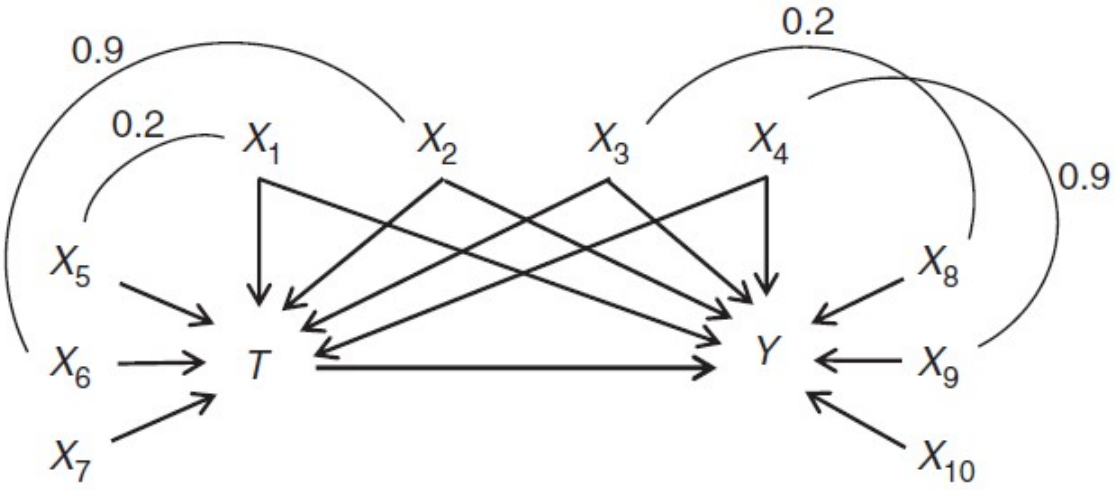

Figure 1: Data structure of the simulation study

treatment status was generated from a Bernoulli distribution: $T_i \sim Ber(p_{i,trt})$ where the probability of treatment selection $p_{i,trt}$ was determined from:

$$log\left(\frac{p_{i,trt}}{1-p_{i,trt}}\right) = \alpha_{0,trt} + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_7 X_7 \quad (8)$$

The coefficients, $\alpha_1, \ldots, \alpha_7$ were based on real-life data utilized in a previous study (Setoguchi et al., 2008), while $\alpha_{0,trt}$ was selected so that the proportion of units who received the treatment (subsequently referred to as prevalence of treatment) was fixed at $\pi = 10\%$, 20%, 30%, 40%, and 50%. We developed the following iterative algorithm which was used to determine the value of $\alpha_{0,trt}$ that induced targeted prevalence π (Amusa et al., 2019a):

(i) We varied values of $\alpha_{0,trt}$ within reason (-3 to 3 in this case), and simulated n units.

(ii) For all the considered $\alpha_{0,trt}$ values, the corresponding individual $p_{i,trt}$ values were computed using (5), while the treatment variables $T_i \sim Ber(p_{i,trt})$ were generated, and the mean of each T_i correspond to π.

(iii) Based on the principle of the law of large numbers: the average of the results obtained from a large number of trials should be close to the expected value, Steps (i) and (ii) were repeated 1000 times to increase the precision of the estimation, and the value of $\alpha_{0,trt}$ which correspond to the desired π is chosen.

For each of the units, we generated an outcome Y_i conditional on T_i, and the seven covariates $(X_1, X_2, X_3, X_4, X_8, X_9, X_{10})$ associated with the outcome. Y_i was generated separately for continuous, binary, count, and time-to-event outcomes.

3.1.1 Continuous outcomes

While we fixed the true treatment effect at $\gamma = 1$, the continuous outcome was generated as

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i + \varepsilon_i \quad (9)$$

where $\varepsilon_i \sim N(0, \sigma^2)$

3.1.2 Binary outcomes

We generated a binary outcome as $Y_i \sim Bernoulli(p_i)$ using a logistic model:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i \quad (10)$$

3.1.3 Count outcomes

We generated a count outcome as $Y_i \sim Poisson(\eta_i)$ using a Poisson model (Amusa et al., 2019b):

$$log(\eta_i) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i \quad (11)$$

3.1.4 Time-to-event outcomes

For time-to-event outcomes, we used a data-generating process described by a previous study (Bender et al., 2005). Survival times t_i are generated as

$$t_i = \left(\frac{-log(U_i)}{\lambda e^{LP}}\right)^{\frac{1}{v}} \quad (12)$$

Where $U_i \sim Uniform(0, 1)$, and the linear predictor, $LP = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_8 + \beta_6 X_9 + \beta_7 X_{10} + \beta_{trt} T_i$ This process generates survival times from a Cox-Weibull distribution. We assumed that all event times are observed for the current analyses.

3.2 Parameter values for data generation

The regression coefficients in the outcome data generation took the values: $\beta_1 = \beta_2 = \beta_3 = log(2), \beta_4 = \beta_5 = \beta_6 = log(1.75)$, and $\beta_7 = log(1.5)$ to reflect very high, high, and moderate effect sizes (Austin et al., 2007; Austin, 2014).

For continuous outcomes, the standard deviation values were fixed at $\sigma = 1$ and 0.5. The conditional treatment effect β_{trt} values were fixed at log (1.5) and log (0.5) for odds ratios, hazard ratios and rate ratios. The chosen values of β_{trt} = log (1.5) and log (0.5) were aimed at reflecting beneficial ($\beta_{trt} > 0$) and adverse ($\beta_{trt} < 0$) treatment effect, respectively. In generating dichotomous outcomes, the β_0 value was set to ensure that the prevalence of the event of interest occurred for approximately 70% of the units. We set v = 2 and $\lambda = 0.000001$ when generating time-to-event outcomes.

Finally, the above data-generating process has randomly generated treatment variable, covariates, and four different outcomes each of size n units, while inducing a conditional treatment effect.

A conditional treatment effect is the average effect, at the individual or unit level, of moving a unit from control to treated group. In contrast, a marginal effect is the average effect, at the population level, of moving the whole population from control to treated group (Greenland, 1987). Since the difference in means is collapsible, the conditional treatment effect coincides with the true marginal treatment effect. However, the other three treatment effects are not collapsible (Austin, 2013; Gail et al., 1984). Thus, for each of the conditional treatment effects (log-odds ratios, log-hazard ratios, and log-rate ratios), we determined their corresponding true marginal treatment effects. Details of this process of obtaining the true marginal treatment effect have been explained elsewhere (Austin, 2013, 2014; Austin and Stuart, 2017). The obtained true marginal treatment effect in the treated population from this process is regarded as the true ATT, for each of the considered effects.

3.3 Statistical analyses in simulated datasets

For a given treatment effect associated with each of the type of outcome considered, we randomly generated 1000 data sets of size 500 using the earlier described data-generating scheme. Using each of the simulated datasets, we separately estimated the different treatment effects, while utilizing each of the ATT weights of entropy balancing and IPW. The treatment effects, γ, were estimated from the following generalized linear model:

$$g(E(Y|T)) = \beta_0 + \gamma T, \tag{13}$$

Where g was considered as the canonical link function for the normal linear model, logistic model, Poisson model, and Cox survival model for estimating the difference in means, odds ratios, hazard ratios, and rate ratios, respectively. We adopted the robust sandwich estimator for estimating the standard errors (Austin and Stuart, 2015; Joffe et al., 2004). We utilized the R-package ebal (Hainmueller, 2014) for implementing entropy balancing.

Let γ_i denote the ith estimated treatment effect using a given method, whereas γ is the true ATT. We then determined the following: Bias $= \frac{1}{1000}\sum_{i=1}^{1000}(\gamma_i - \gamma)$, mean squared error (MSE) $= \frac{1}{1000}\sum_{i=1}^{1000}(\gamma_i - \gamma)^2$. We also examined precision by averaging the model-based standard errors (SE) over the 1000 simulated datasets. Finally, we examined 95% coverage - the proportion of times γ is enclosed in the 95% confidence interval of γ over the simulated datasets.

4 Results

We present the simulation results according to each of the type of estimated treatment effects explained in the earlier Section. We focus on the performance of entropy balancing method, using the IPW method as a threshold for evaluating the results. As a form of

sensitivity analysis, we ran simulations for other sample sizes (n = 300, 1000), but we do not present the results as no qualitative differences were observed in the relative performance of the methods. However, we present results of the two standard deviation values ($\sigma = 1, 0.5$) assumed while estimating difference in means, as well as the two different true ATT values, varied each for odds ratios, hazard ratios, and rates ratios estimation. Altering these parameter values also did not change the conclusions in all the scenarios, except for when rate ratios were estimated.

4.1 Continuous outcomes: Difference in means

Results are summarized in Figures 2 and 3. In terms of bias, Figure 2 shows that both methods produced estimates with very low (near zero) bias. However, EB produced slightly higher biases, except for when the prevalence rate was 10%. For the MSE, EB outperformed IPW across the board (Figure 2). Both methods yielded very similar SE estimates, with the values decreasing with increasing prevalence rates (Figure 3). Though EB produced superior CI coverages - near perfect in most cases, both methods achieved reasonably high 95% CI coverage (Figure 3).

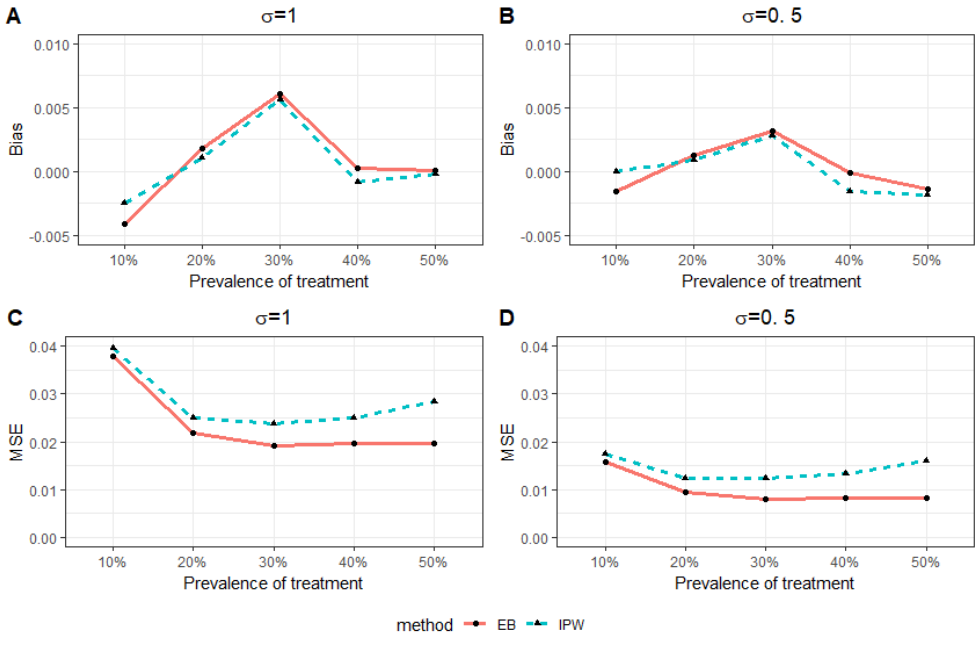

Figure 2: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating difference in means.

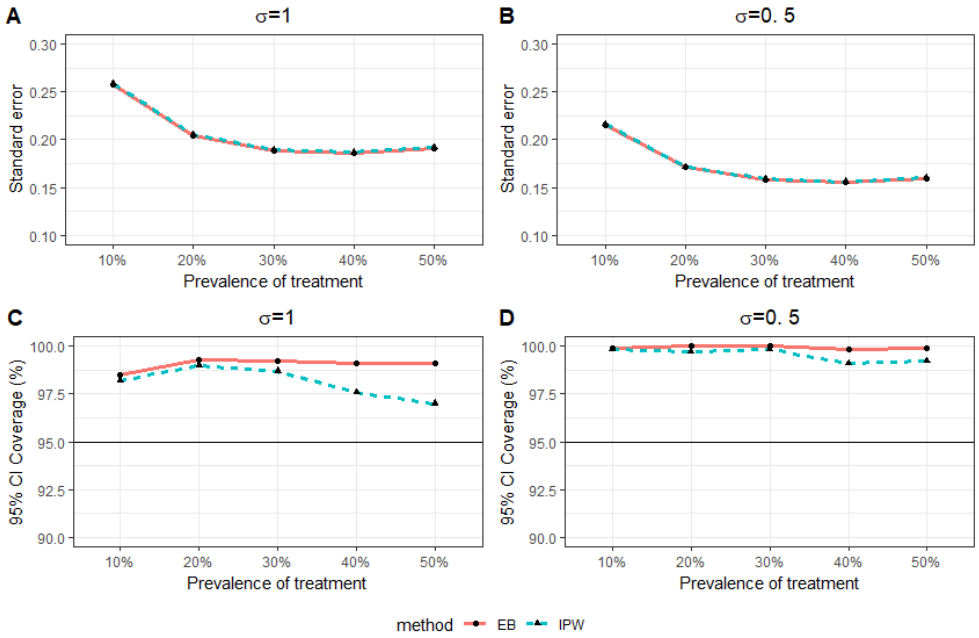

Figure 3: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating difference in means.

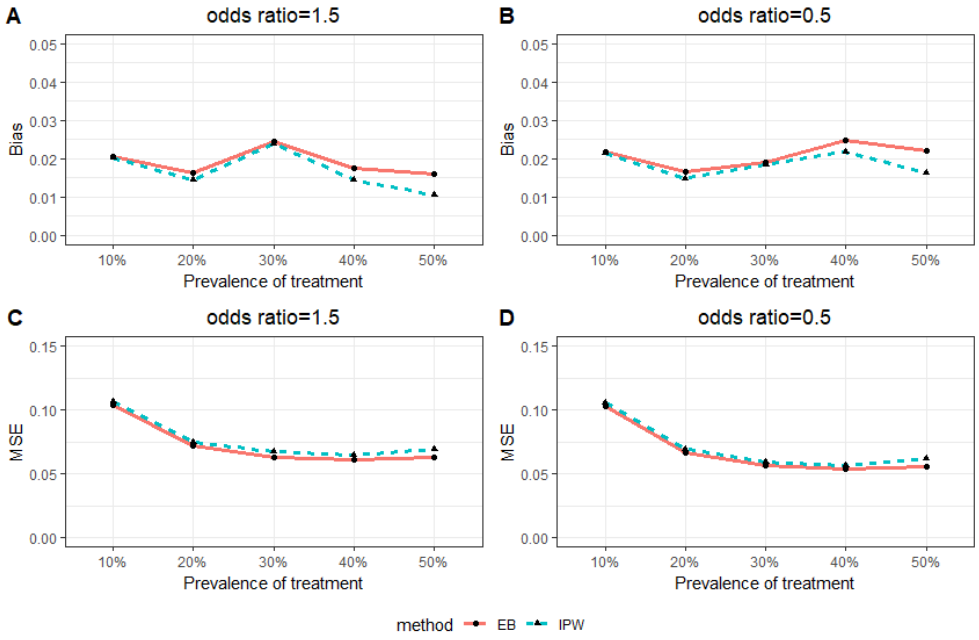

Figure 4: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating odds ratios.

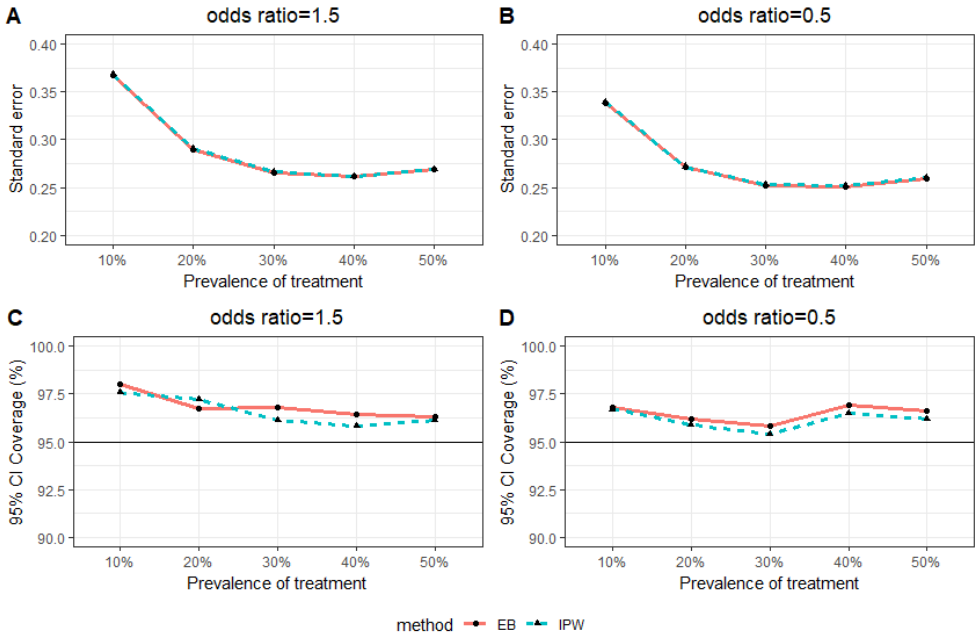

Figure 5: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating odds ratios.

4.2 Binary outcomes: Odds ratios

In terms of bias, Figure 4 shows that EB consistently produced higher biased estimates. For the MSE, EB outperformed IPW across the board, with the values decreasing with increasing prevalence rates (Figure 4). Both methods yielded very similar SE estimates, with the values decreasing with rising prevalence rates (Figure 5). Though EB produced superior CI coverages, both techniques achieved reasonably high 95% CI coverage (Figure 5).

4.3 Count outcomes: Rate ratios

Figure 6 shows that the MSE of both methods increased as the treatment prevalence increased from 10% to 40%. When the conditional rate ratio was positive ($\beta_{trt} > 0$), EB consistently produced estimates with higher bias, higher MSE, and lower 95% CI coverage. However, for ($\beta_{trt} < 0$), EB produced higher bias and MSE estimates only when the treatment prevalence was 20% or lower (Figure 6). The SE estimates were very similar between both methods, with the values decreasing with increasing prevalence rates (Figure 7). Both methods achieved reasonably high 95% CI coverage. Though EB had lower CI coverages when the conditional rate ratio was positive, it is not clear which of them produced higher coverage when the conditional rate ratio was negative (Figure 7).

4.4 Time-to-event outcomes: Hazard ratios

Figure 8 shows that the bias of both methods are not substantially different, except for higher prevalence rates (40% and 50%) where EB produced higher bias estimates. For the MSE, EB consistently outperformed IPW (Figure 8). As shown in Figure 9, the SE estimates were again very similar between both methods, with the values decreasing with increasing prevalence rates. Figure 9 illustrates that when the actual hazard ratio = 0.5, EB produced 95% CIs slightly below the nominal coverage rate at prevalence rates higher than 30%. However, EB provided superior CI coverages overall.

5 Discussion

Propensity score (PS) methods are the most widely used in estimating average treatment effects in observational studies. While the inverse probability of treatment weighting (IPW) method appears to be the most common implementation of PS methods, we introduce entropy balancing – a relatively new, but under-utilized weighting method, despite having nice conceptual properties. This study aims to use Monte Carlo simulations to evaluate the performance of entropy balancing, relative to the traditional IPW, in estimating some standard measures of treatment effect. While focusing on entropy balancing, we summarize our findings, and where necessary, place them in the context of existing literature.

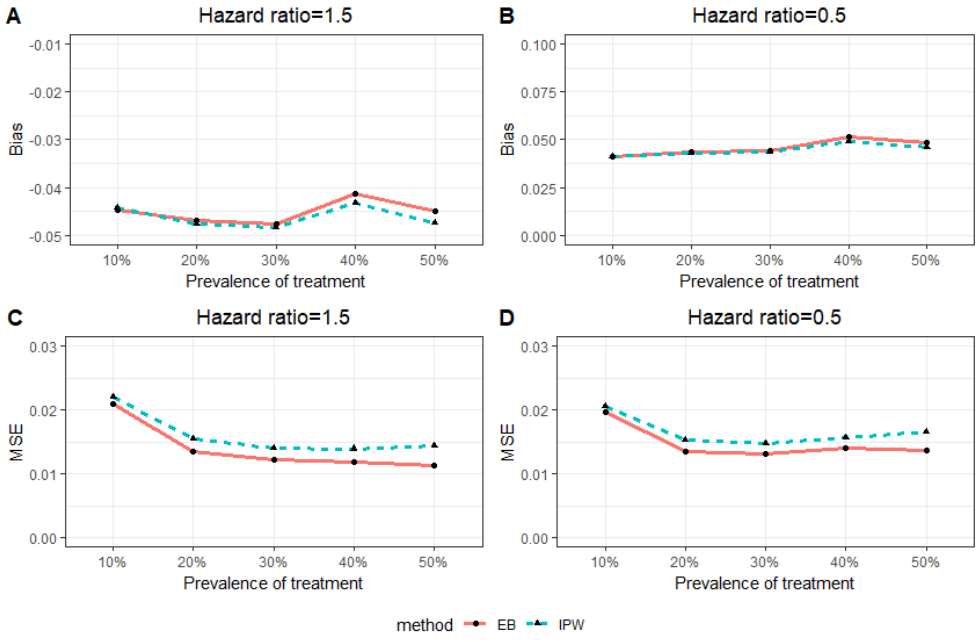

Figure 6: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating hazard ratios.

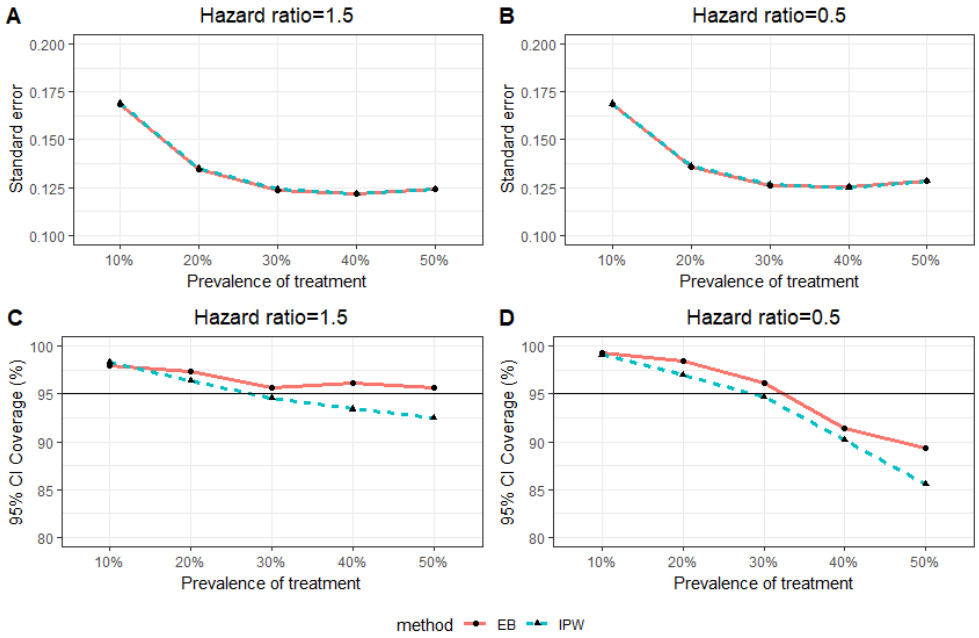

Figure 7: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating hazard ratios.

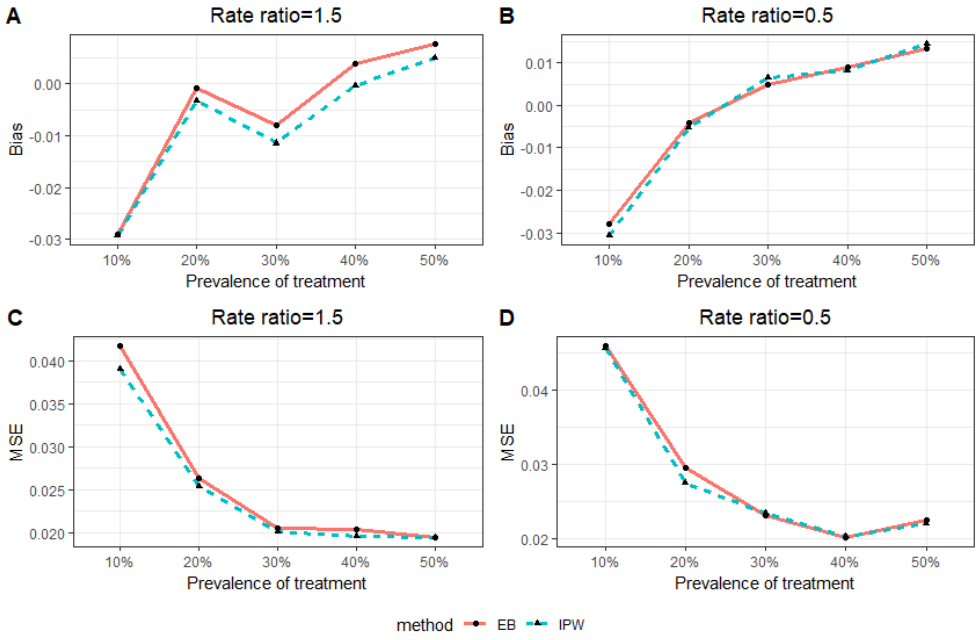

Figure 8: Bias (Top panels) and MSE (Bottom panels) of entropy balancing (EB) and IPW methods for estimating rate ratios.

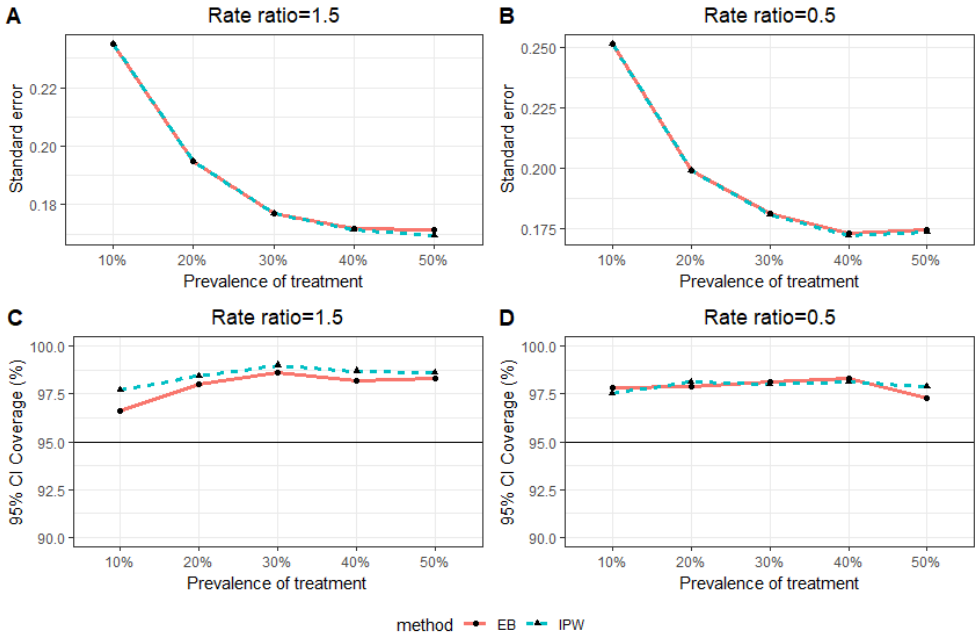

Figure 9: Mean estimated standard error (Top panels) and 95% confidence intervals coverage rates (Bottom panels) of entropy balancing (EB) and IPW methods for estimating rate ratios.

Though both methods performed reasonably well in estimating the various treatment effects considered, we found on average that entropy balancing outperformed IPW for all the considered situations. However, a few exceptions were found: (i) When rate ratios were estimated, entropy balancing tended to produce estimates with slightly higher biases and mean squared errors. Although they considered conditional and not marginal treatment effects, a previous study by Austin (2007) found that conditioning on the propensity score did not substantially introduce bias into the estimation of rate ratios. (ii) The model-based standard errors for both methods were consistently indistinguishable. (iii) In terms of bias, across all the estimated treatment effects, entropy balancing consistently produced more biased estimates. Hence, there is an interesting bias-variance trade-off of the two techniques. However, entropy balancing has the facility to optimize the bias-variance trade-off by tightening the pre-specified tolerance on covariate balance (Harvey et al., 2017). Previous studies (Austin, 2007, 2013; Austin and Stuart, 2017) also support our findings in favour of IPW producing an unbiased estimation of odds ratios and hazard ratios.

A significant strength of this study is in the use of an algorithm which determines the true marginal treatment effect corresponding to a particular conditional treatment effect. Many simulation studies estimated average or marginal treatment effects using a conditional model to relate the outcome with the treatment and associated covariates, even though the estimated effects are not collapsible (i.e. marginal and conditional treatment effects will not coincide) (Austin, 2013; Gail et al., 1984; Greenland, 1987).

For binary outcomes, even though odds ratios are not collapsible and other reasons (Newcombe, 2006), we chose to adopt odds ratios due to its frequent usage in biomedical research.

The limitation of this study is that we did not include censoring in our simulation of time-to-event outcomes. The reason is due to computational simplicity. Allowing the degree of censoring to be another factor in the design of the Monte Carlo simulations would increase the computational burden of the simulations substantially and increase the number of results that would require reporting. However, this may warrant future investigations.

To our knowledge, no previous research had studied the performance of entropy balancing in estimating treatment effects of different types of outcomes, using Monte Carlo simulations. Like any simulation, our simulation results might be limited to the scenarios considered by our simulation data. Therefore, the results cannot be generalized to settings that have not been evaluated.

6 Conclusion

Overall, we found the entropy balancing technique useful and excellent in performance. Entropy balancing merits more widespread adoption for estimating treatment effects of different types when using observational data.

Availability of data and materials

The R codes for a complete reproduction of the results in this study are available from the corresponding author upon request.

Competing interests

The authors declare that they have no competing interests.

References

Adhikary, S. D., Liu, W.-M., Memtsoudis, S. G., Davis III, C. M., and Liu, J. (2016). Body mass index more than 45 kg/m2 as a cutoff point is associated with dramatically increased postoperative complications in total knee arthroplasty and total hip arthroplasty. *The Journal of arthroplasty*, 31(4):749–753.

Amusa, L., Zewotir, T., and North, D. (2019a). Evaluation of subset matching methods: Evidence from a monte carlo simulation study. *American Journal of Applied Sciences*, 16(3):92–100.

Amusa, L., Zewotir, T., and North, D. (2019b). A weighted covariate balancing method for estimating causal effects in case-control studies. *Modern applied science*, 13(4):40–50.

Aria, M., Capaldo, G., Iorio, C., Orefice, C. I., Riccardi, M., and Siciliano, R. (2018). Pls path modeling for causal detection of project management skills: a research field in national research council in italy. *Electronic Journal of Applied Statistical Analysis*, 11(2):516–545.

Austin, P. (2007). The performance of different propensity score methods for estimating marginal odd ratios. *Stat. Med.*, 26:3078–3094.

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine*, 32(16):2837–2849.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6):1057–1069.

Austin, P. C., Grootendorst, P., Normand, S. T., and Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a monte carlo study. *Statistics in medicine*, 26(4):754–768.

Austin, P. C. and Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine*, 33(24):4306–4319.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679.

Austin, P. C. and Stuart, E. A. (2017). Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research*, 26(6):2505–2525.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

Brettschneider, C., Bleibler, F., Hiller, T. S., Konnopka, A., Breitbart, J., Margraf, J., Gensichen, J., Koenig, H. H., and Jena, P. S.-G. (2017). Excess costs of panic disorder with or without agoraphobia in germany - the application of entropy balancing to multiple imputed datasets. *Journal of Mental Health Policy and Economics*, 20:S3–S3.

Carpita, M. and Ciavolino, E. (2017). A generalized maximum entropy estimator to simple linear measurement error model with a composite indicator. *Advances in Data Analysis and Classification*, 11(1):139–158.

Ciavolino, E. and Carpita, M. (2015). The gme estimator for the regression model with a composite indicator as explanatory variable. *Quality & Quantity*, 49(3):955–965.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161.

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3):431–444.

Golan, A. (2018). *Foundations of info-metrics: Modeling, inference, and imperfect information*. Oxford University Press.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American journal of epidemiology*, 125(5):761–768.

Grupp, H., Kaufmann, C., König, H.-H., Bleibler, F., Wild, B., Szecsenyi, J., Herzog, W., Schellberg, D., Schäfert, R., and Konnopka, A. (2017). Excess costs from functional somatic syndromes in germany—an analysis using entropy balancing. *Journal of psychosomatic research*, 97:52–57.

Guo, S., Barth, R., and Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, 28:357–83.

Guo, S. and Fraser, M. (2010). *Propensity score analysis; Statistical methods and applications*. Advanced Quantitative Techniques in the Social Sciences. SAGE Publications.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hainmueller, J. (2014). *ebal: Entropy reweighting to create balanced samples*. R package version 0.1-6.

Harvey, R. A., Hayden, J. D., Kamble, P. S., Bouchard, J. R., and Huang, J. C. (2017). A comparison of entropy balance and probability weighting methods to generalize observational cohorts to a population: a simulation and empirical example. *Pharmacoepidemiology and Drug Safety*, 26(4):368–377.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278.

Hirshberg, D. A. and Zubizarreta, J. R. (2017). On two approaches to weighting in causal inference. *Epidemiology*, 28(6):812–816.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., and Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician*, 58(4):272–279.

Kullback, S. (1959). *Information theory and statistics*. Wiley, New York.

Lee, B., Lessler, J., and Stuart, E. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29:337–346.

Mattke, S., Han, D., Wilks, A., and Sloss, E. (2015). Medicare home visit program associated with fewer hospital and nursing home admissions, increased office visits. *Health Affairs*, 34(12):2138–2146.

Newcombe, R. G. (2006). A deficiency of the odds ratio as a measure of effect size. *Statistics in Medicine*, 25(24):4235–4240.

Parish, W. J., Keyes, V., Beadles, C., and Kandilov, A. (2018). Using entropy balancing to strengthen an observational cohort study design: lessons learned from an evaluation of a complex multi-state federal demonstration. *Health Services and Outcomes*

Research Methodology, 18(1):17–46.

Pearson, J. L., Stanton, C. A., Cha, S., Niaura, R. S., Luta, G., and Graham, A. L. (2014). E-cigarettes and smoking cessation: insights and cautions from a secondary analysis of data from a study of online treatment-seeking smokers. *Nicotine & Tobacco Research*, 17(10):1219–1227.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, P. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R., and Cook, E. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, 17:546–555.