Songklanakarin J. Sci. Technol.
43 (2), 406-413, Mar. - Apr. 2021

http://www.sjst.psu.ac.th

Original Article

An empirical exploration of entropy balancing in estimating treatment effects: Insights from simulation and two applied biomedical studies

Lateef Amusa*, Temesgen Zewotir, and Delia North

Department of Statistics, School of Mathematics, Statistics and Computer Science, University of Kwazulu-Natal, Durban, South Africa

Received: 10 May 2019; Revised: 30 January 2020; Accepted: 18 February 2020

Abstract

We present entropy balancing – a relatively new technique for estimating treatment effects, which has been under-utilised in the applied biomedical literature. Our objective is to share our experiences learned from using entropy balancing in non-experimental studies, via Monte Carlo simulations and two empirical examples. We used the inverse probability of treatment weighting method for benchmarking the performance of entropy balancing. Entropy balancing had remarkably superior performance in terms of covariate balance and efficient estimation of treatment effects. Entropy balancing does not require extensive tweaking of the propensity score specification to achieve the optimal covariate balance. Instead, it directly incorporates covariate balance into the weight function that is applied to the sample units. Notably, the excellent performance of entropy balancing is not only on the first moment (mean) but also on higher-order moments. However, we also highlighted the situations where entropy balancing may fail or not have optimal performance. Entropy balancing merits more widespread adoption in biomedical research.

Keywords: entropy balancing, Monte Carlo simulation, observational studies, propensity score weighting, treatment effect

1. Introduction

A focal objective of medical and health research is to estimate the causal effect of a treatment or intervention on an outcome variable. For randomised experiments, making such inferences are regularly clear and straightforward. However, when treatment assignment is complicated by confounders, as in the case of observational studies, such inferences regarding the treatment effects require more sophisticated methodology (Zagar, Kadziola, Lipkovich, & Faries, 2017). Comparisons between treatment groups (treated versus control) can be biased when the groups lack sufficient balance; in other words, have substantially different distributions of relevant covariates.

A commonly used nonparametric balancing strategy is weighting. Weighting, in this context, is the generation of balancing weights, which, when applied to the sample of units in each treatment group, matches the covariate distribution of the treatment group of interest. Comparisons are then made between the weighted outcomes. Weighting methods have taken centre stage in efficiently estimating treatment effects when treatment assignment is confounded with background covariates. There are two general weighting approaches in causal inference: One does not directly make covariate balance its primary objective - it focuses on modelling the data to get probabilities, from which weights that reduce the imbalance to some considerable extent can be obtained. The other approach, also known as the covariate balancing approach, directly uses some minimisation algorithm to choose weights that perfectly balance the covariates, subject to some specified constraints (Chan, Yam, & Zhang, 2016; Hainmueller, 2012; Imai & Ratkovic, 2014; Li, Morgan, & Zaslavsky, 2018; Wong & Chan, 2017).

The literature on weighting methods which agree with the first general weighting approach described above has been dominated by the inverse probability of treatment weighting (IPW), originating from survey research (Crump, Hotz, Imbens, & Mitnik, 2009; Hirano & Imbens, 2001; Hirano, Imbens, & Ridder, 2003; Imbens, 2004). IPW method is the most common method in applied research and among practitioners, especially in the medical and health sciences

*Corresponding author
Email address: amusasuxes@gmail.com

(Austin & Stuart, 2015). However, like every other propensity score (PS) technique, IPW relies heavily on the correct specification of the PS model. Slight misspecification of the PS model will result in a substantial bias of the estimated treatment effects (Kang & Schafer, 2007). It takes a highly skilled user to specify what is close to a correct PS model; consequently, the iterative tweaking of PS models until measured baseline covariates are balanced can be quite tedious. Despite this cycle of attempting to fit the correct PS model, achieving a sufficient level of covariate balance can occasionally be elusive and additional imbalances may be introduced.

In this paper, we present the entropy balancing (EB) procedure - an optimisation-based weighting method that shares the spirit of the first general weighting approach described above. Entropy balancing (Hainmueller, 2012) particularly works remarkably well in achieving covariate balance. Relative to the other optimisation-based weighting methods, we were interested in entropy balancing because it is less time-consuming and straightforward to implement.

We aim to provide an intensive exploration of the use of entropy balancing in medical and health studies. An in-depth search from the Web of Science Core Collection, excluding methodology-based articles, identified 170 published articles that utilised entropy balancing, and only a few of them (26.19%) were in the medical and health sciences; a majority of the applications of entropy balancing have been in the social sciences. A few of these applications in the medical and health literature can be found in (Adhikary, Liu, Memtsoudis, Davis III, & Liu, 2016; Brettschneider *et al.*, 2017; Grupp *et al.*, 2017; Mattke, Han, Wilks, & Sloss, 2015; Pearson *et al.*, 2014). Using the IPW method as a benchmark, the performance of entropy balancing was examined via Monte Carlo simulations, modelling situations typical of the medical and health sciences. Finally, we illustrate the application of entropy balancing with two empirical case studies, exploring changes in its various parameters, as well as its effect on achieving balance on the measured baseline covariates, further focusing also on accuracy in estimating treatment effects.

2. Materials and Methods

The dichotomous treatment variable T_i for the *i*th unit is coded 1 and 0 for treated and control groups, while $X_i = (X_{i1}, X_{i2}, \dots, X_{ik})$ are a set of pre-treatment covariates. Y is the vector of observed outcomes.

In this section, while estimating the average treatment effect among the treated (ATT), we briefly describe the entropy balancing and inverse probability of treatment weighting (IPW), for adjusting the inherent non-randomisation of treatments that characterises an observational study.

2.1 Entropy balancing

Entropy balancing searches for a unique set of weights for the control group units, 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 entropy balancing technique utilises a maximum-entropy reweighting scheme to directly incorporate covariate balance in terms of means or/and higher-order moments into the weight function. Entropy balancing can, therefore, guarantee perfect covariate handling, as well as maximum retention of information (Parish, Keyes, Beadles, & Kandilov, 2018). After assuming initially uniform weights w_i, the control group weights are estimated directly from pre-specified constraints as a log-linear function of the known moment conditions, by the following reweighting scheme:

$$\min_{w_i} \mathrm{H}(\mathrm{w}) = \min_{w_i} \sum_{i|T=0} w_i \log\left(w_i/q_i\right)$$

subject to the following balance and normalization constraints:

$$\sum_{i|T=0} w_i \, c_{ri}(X_i) = m_r, \text{ for r = 1, ..., R}$$

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

$$w_i \geq 0 \quad \text{for all } i,$$

where $q_i = \frac{1}{n_0}$ is a vector of the base weights, and m_r represents a set of R balance constraints imposed on the moments of the reweighted control group.

The three constraints ensure that: (i) a pre-specified level of balance, specified in terms of the *r*th moment, is achieved for all covariates; (ii) the solution weights must be normalised to sum to one (though other constants may be used); and (iii) only weights with positive values are allowed. See (Amusa, Zewotir, & North, 2019; Hainmueller & Xu, 2013) for more details on entropy balancing.

2.2 Inverse probability treatment weighting

The propensity score is defined as the probability of a subject or unit being assigned to the treated group conditional on the observed baseline covariates. In IPW, each unit weights equal the reciprocal of the probability of receiving the treatment that the unit received. We utilised the ATT weights, which are defined as fixing the treated units' weight at unity, and the control units' weight as the odds of the estimated propensity scores (Imbens, 2004). We calculated the propensity scores by using a logistic regression model to regress treatment status on the covariates associated with the treatment.

2.3 Simulation study

We conducted a set of Monte Carlo simulation experiments to examine the performance of entropy balancing, relative to IPW method. We made the simulations typical of biomedical studies by considering binary outcomes (Austin, Manca, Zwarenstein, Juurlink, & Stanbrook, 2010). All simulations were done using the R statistical package (version 3.5.1.). Entropy balancing was performed with the R-package *ebal* (version 0.1-6) (Hainmueller, 2014).

We followed the simulation structure of previous studies (Lee, Lessler, & Stuart, 2010; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008). Figure 1 gives the causal structure of the simulations. We made our simulations to vary based on two factors: (i) Sample size: n = 500, 2000; (ii) the proportion of units assigned to the treated group (prevalence of treatment), which was fixed at 25%, 33%, 50%, and 67%, corresponding to treated: control units' ratio of 1:3, 1:2, 1:1, and 2:1, respectively.

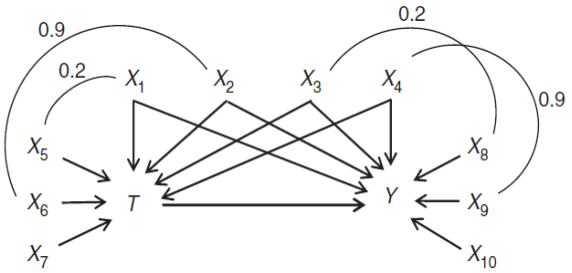

Figure 1. Data structure of the simulation study, where $X_1, X_3, X_5, X_6, X_8, X_9, Y$ are binary

For each of the considered scenarios, we simulated 1000 datasets and obtained ATT weights for both entropy balancing and IPW, and estimated treatment effects from the weighted regressions of Y on T. The treatment effect estimates (estimated ATT) were then averaged over the simulation runs, denoted by $\bar{\hat{\gamma}}$. The true ATT is denoted by γ.

In terms of performance assessment, we utilised the absolute standardised mean difference (ASMD) to measure covariate balance (McCaffrey *et al.*, 2013). For outcome estimation, we computed the bias, mean squared error (MSE), model-based standard errors, and 95% confidence interval (CI) coverage of the estimated treatment effects. CI coverage is defined as the proportion of times the estimated confidence intervals contain the specified parameter value (Burton *et al.*, 2006).

2.4 Simulation results

We present the results for the simulation study according to each of the performance metrics explained in the earlier section. We emphasise more the results of the entropy balancing method while using the IPW method results as a benchmark.

Some authors suggested that ASMD values above 0.1 may be indicative of covariate imbalance (Mamdani *et al.*, 2005; Normand *et al.*, 2001). As shown in Figure 2, both EB and IPW methods performed remarkably well in reducing covariate imbalances, as they both achieved ASMD values well below the threshold of 0.1. However, EB outperformed the IPW method, with marginal improvements observed for the smaller treatment prevalences (25% and 33%), but a substantial outperformance was evidenced for the higher treatment prevalences (50% and 67%). Additionally, EB produced better balance for large sample size (n=2000), with ASMDs achieving a perfect balance (ASMD=0) on almost all of the covariates, across the rates of treatment prevalence.

In terms of bias and MSE, Figure 3 shows that EB produced less biased estimates across the range of scenarios. EB resulted in estimates with substantially lower MSE values. Furthermore, there was no apparent effect of the prevalence of treatment on either the bias or the MSE. As shown in Figure 3, the two methods produced very similar standard errors. Superior 95% CI coverages were observed for EB when the sample size was relatively large (n = 2000). However, for a relatively smaller sample size (n = 500), EB produced higher coverage for higher treatment prevalence (50% and 67%), as shown in Figure 3.

2.5 Empirical examples

We compared the relative performance of entropy balancing and inverse probability of treatment weighting using two biomedical datasets. Throughout, we compared performance based on covariate balance and outcome estimation. We present and discuss results from two distinct examples, which illustrate different experiences we have had with entropy balancing as it compares to IPW.

We applied diagnostics for assessing the covariate balance in the data weighted by entropy balancing (EB), with

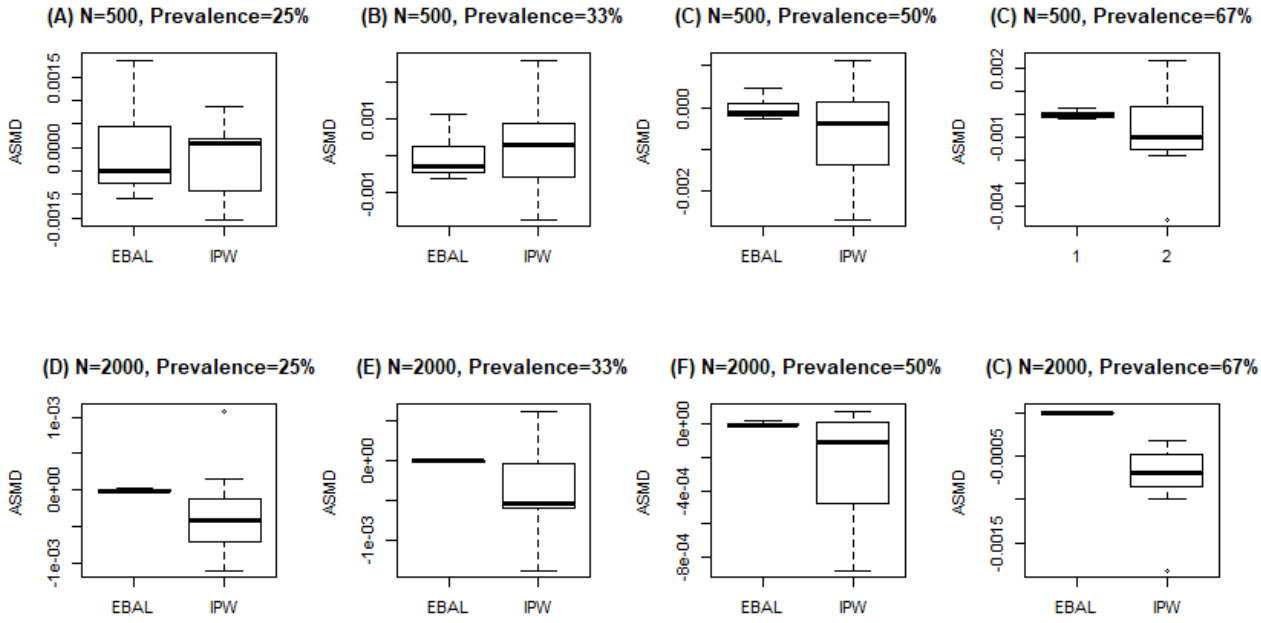

Figure 2. Boxplots for the absolute standardized mean difference for covariates

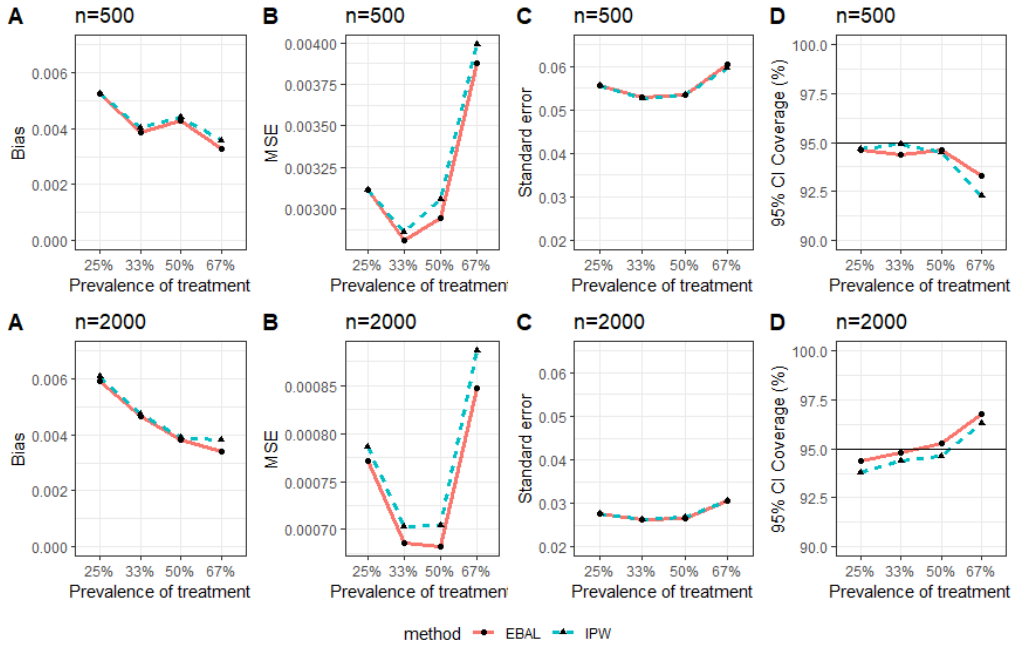

Figure 3. Bias (Panel A), MSE (Panel B), standard errors (Panel C), and 95% CI coverage (Panel D) of estimated treatment effects

performance evaluated relative to IPW. We did not restrict balance to means only, but also investigated variance and the empirical distribution of continuous covariates. For balance on the means alone, we considered the ASMD. For balance on higher-order moments, we adopted variance ratios, for which some authors (Rubin, 2001) recognize values close to 1 as okay, and Kolmogorov-Smirnov (KS) statistic close to 0 is considered satisfactory (Ali *et al.*, 2015). For entropy balancing, we considered the first and second moments on which covariate balance is desired. We attempted to include constraints up to the third moments, as well as interactions between pairs of continuous covariates, but the EB algorithm did not converge.

Next, we estimate the average treatment effect for those who received the treatment. Alongside the IPW, we applied weights to the outcome modelling from entropy balancing, when moment constraints included the means only, denoted by *ebal1*, as well as with added variance constraints, denoted by *ebal2*. Using logistic regression to regress the outcomes, we adopted the risk difference, as suggested by clinical commentators (Cook & Sackett, 1995; Laupacis, Sackett, & Roberts, 1988; Sackett, Deeks, & Altman, 1996; Schechtman, 2002), as the estimate of interest. The model incorporated the weights induced by entropy balancing and IPW. Standard errors of the weighted estimators were estimated using the sandwich-type variance estimators.

2.6 Case study 1: RHC study

We further explored the entropy balancing technique by analysing observational data (Murphy & Cluff, 1990), to study the effectiveness of right heart catheterization (RHC) for critically ill patients. A few influential studies have analysed the data, using different adjustment methods (Connors *et al.*, 1996; Crump *et al.*, 2009; Hirano & Imbens, 2001; Li *et al.*, 2018; Rosenbaum, 2012). In brief, the dataset comprises information on 5735 patients, 2184 of them were treated with RHC (T_i=1) within 24 hours of admission, while the remaining 3551 were not (T_i=0). The outcome of interest was mortality at 30 days of hospitalization. Full details of this data, including the variable description and its summary statistics, have been published elsewhere (Connors *et al.*, 1996; Hirano & Imbens, 2001).

2.6.1 Results

The RHC dataset is unique in terms of its characterization by a bountiful number of covariates, with a moderately sized treatment prevalence (38.1%). Figure 4 provides information on the balance on covariates. Before any weighting, the data had a high degree of covariate imbalance. The weighting methods produced ASMD values that did not exceed the vertical line 0.1 threshold superimposed in Figure 4. Entropy balancing achieved a perfect covariate balance (ASMD=0) on all the covariates, while there were still some noticeably non-zero ASMDs after applying IPW. As expected, variance ratios were all virtually 1 when moment constraints of entropy balancing included the second moment. Even when moment constraints included only the mean, entropy balancing still achieved variance ratios close to one about 71.4% of the time. IPW did poorly on variance ratios, as it increased the values for some covariates, thereby making things worse. As measured by the KS statistic, both methods performed remarkably well on the empirical distribution of the continuous covariates. However, about 67% of the time, entropy balancing produced KS values indicative of better balance, when moment constraints included the second moment; while, when it contained only the first moment, entropy balancing marginally outperformed IPW about 52% of the time.

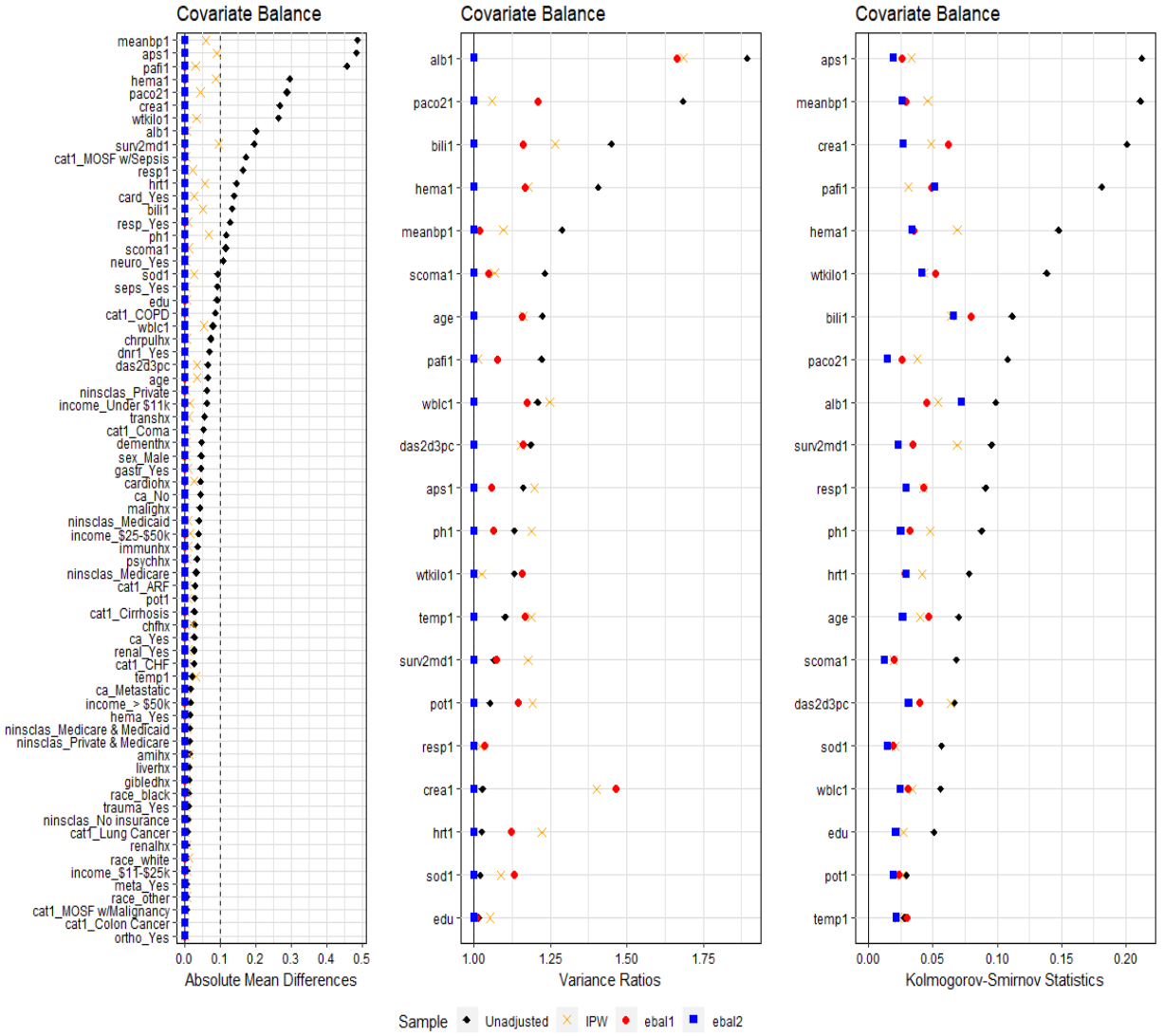

Figure 4. Assessment of covariate balance for the RHC dataset
Note*:* ebal1: entropy balance on the 1st moment; ebal2: entropy balance on the 2nd moment

The causal treatment effects estimated using these stated methods are shown in Table 1. All the considered estimators produced qualitatively similar estimates that are statistically significant at the 0.01 level, which indicate that applying RHC leads to a higher mortality rate. These results agree with the substantive conclusions made in previous studies (Connors *et al.*, 1996; Crump *et al.*, 2009; Li *et al.*, 2018). Estimators based on the entropy balancing had smaller confidence interval lengths (0.0731 for *ebal1* and 0.0767 for *ebal2*) than the corresponding ones based on the IPW (0.0787).

Table 1. Causal effect estimation of RHC

Methods	Estimate	CI	P-value
Unweighted	0.051	0.025─0.076	<0.0001
IPW	0.056	0.017─0.096	0.0051
ebal1	0.072	0.035─0.108	0.0001
ebal2	0.057	0.019─0.096	0.0034

Note: CI: Confidence Interval; ebal1: entropy balance on the 1st moment; ebal2: entropy balance on the 2nd moment

2.7 Case study 2: Lindner study

Randomised controlled trials of percutaneous coronary intervention (PCI) have demonstrated the efficacy of increasing survival of patients. The Ohio Heart Health Center (OHHC) operators at the Lindner Christ Hospital in Cincinnati, Ohio carried out an observational study in 1997. In brief, the Lindner dataset comprises information on 996 patients who received an initial Percutaneous Coronary Intervention (PCI) at the health facility at that time. The treated group are patients who received the PCI with an additional abciximab (abcix) treatment (an expensive, high-molecular-weight IIb/IIIa cascade blocker), while the control group are those who received the PCI alone. Covariates include an indicator for recent acute myocardial infarction (acutemi), left ventricle ejection fraction (ejecfrac), height, number of vessels involved

in initial PCI (ves1proc), an indicator for coronary stent insertion (stent), gender (female), and diabetic indicator (diabetic). An indicator of survival at six months (sixMonthSurvive) is the outcome variable of interest. Further details of this dataset and its analysis have been published elsewhere (Abdia, Kulasekera, Datta, Boakye, & Kong, 2017; Kereiakes *et al*., 2000).

2.7.1 Results

The Lindner dataset is unique in terms of its characterization by a relatively smaller number of covariates, with an unusually high treatment prevalence (70%). Figure 5 provides information on the covariate balance. Before any weighting, there were some meaningful differences in the means and variances for some of the covariates. IPW and entropy balancing both reduced these imbalances, and achieved an ASMD well below the threshold of 0.1, as superimposed in Figure 5. Entropy balancing achieved a perfect covariate balance (ASMD=0), as well as variance ratios close to one for all the covariates. However, there were still some noticeably non-zero ASMDs after applying IPW, as well as a suboptimal performance in terms of variance ratios. As measured by the KS statistic, both methods performed remarkably well on the empirical distribution of the continuous covariates. In all of the covariates, entropy balancing (both of first and second moments) produced KS values indicative of better balance.

The estimated causal treatment effects of abciximab, using both methods, are shown in Table 2. All the considered estimators produced qualitatively similar estimates that are statistically significant at the 0.05 level. There is thus evidence of the effectiveness of the abciximab treatment increasing survival rates of patients. The same conclusion was made by Kereiakes *et al.* (2000). Estimators based on the entropy balancing had smaller confidence interval lengths (0.128 for *ebal1* and 0.124 for *ebal2*) than the corresponding ones based on the IPW (0.149).

Table 2. Causal effect estimation of abciximab

Methods	Estimate	CI	P-value
Unweighted	0.035	0.013─0.056	0.0017
IPW	0.079	0.004─0.153	0.0394
ebal1	0.071	0.007─0.135	0.0288
ebal2	0.070	0.008─0.132	0.0267

Note: CI: Confidence Interval; ebal1: entropy balance on the 1st moment; ebal2: entropy balance on the 2nd moment

3. Discussion

PS weighting methods have conventionally been used to estimate treatment effects in the presence of confounding factors. In this study, we used simulations and two empirical examples to highlight our experiences with using entropy balancing, and its performance relative to the traditional inverse probability of treatment weighting method. We are motivated by the under-utilization of the entropy balancing technique in the biomedical sciences, in reducing bias and estimating causal treatment effects in observational studies, despite its increased usage and successful application in the social sciences. We chose a simulation structure that mimics what is common in most biomedical studies and empirical datasets that have been analysed by some previous studies representative of a clinical application.

Though both entropy balancing and IPW methods provided adequate covariate balance in the considered scenarios, we found that entropy balancing outperformed IPW for all the considered scenarios. For entropy balancing, covariate balance was improved as treatment prevalence increased. Both methods improved covariate balance for larger sample sizes. There was also no evidence of treatment prevalence on the bias and MSE of estimated effects.

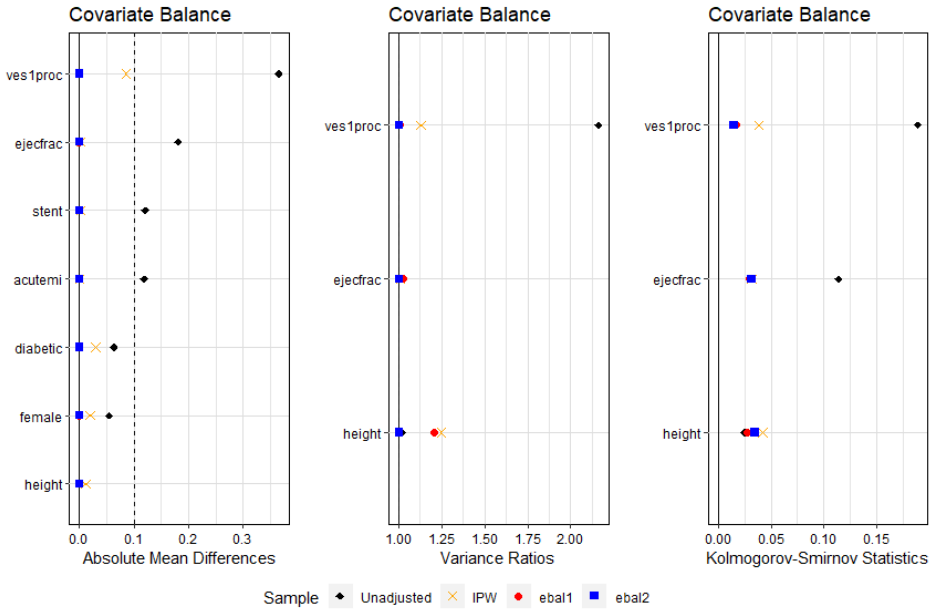

Figure 5. Assessment of covariate balance for the Lindner dataset
Note: ebal1: entropy balance on the 1st moment; ebal2: entropy balance on the 2nd moment

Our empirical applications showed that IPW worsened covariate balance on a few covariates. This could have been remedied by iteratively tweaking the PS model until the desired covariate balance is achieved. However, unlike the entropy balancing method, there is no guarantee that this tedious and exhausting process of PS model specification will help IPW produce the desired covariate balance.

We assessed the following situations, which did not allow convergence of the entropy balancing algorithms: (i) smaller sample sizes (less than 300) for treatment prevalence rates higher than 33%, in the simulations; (ii) including the 3rd moments in the moment constraints for the case study; and (iii) including pairs of interaction of continuous covariates for the case study. The above findings agree with the caution given by Hainmueller (2012), in light of potential situations, depending on the data, that may prevent convergence of the entropy balancing algorithm. Furthermore, even though previous studies like (Zagar *et al.*, 2017) stated that the presence of interaction effects might improve the performance of the entropy balancing, but the interaction effects are not always feasible for a large number of covariates as we have experienced with our second empirical example.

To our knowledge, no previous study had explored entropy balancing using Monte Carlo simulations, with binary outcomes. As with any simulation, our simulation results might be limited to the scenarios considered by our simulation data; therefore, the results cannot be generalised to settings that have not been evaluated. Another limitation of entropy balancing is that it does not address unmeasured confounding, which is still a vexing problem in observational studies.

4. Conclusions

Overall, we found the entropy balancing technique useful, with excellent performance, and one that is frequently less tedious than the traditional inverse probability of treatment weighting approach. Entropy balancing merits more widespread adoption for estimating the effects of treatment, especially in the medical and health sciences, when using observational data.

References

Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, *59*(5), 967-985.

Adhikary, S. D., Liu, W.-M., Memtsoudis, S. G., Davis III, C. M., & 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.

Ali, M. S., Groenwold, R. H., Belitser, S. V., Pestman, W. R., Hoes, A. W., Roes, K. C., . . . Klungel, O. H. (2015). Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: A systematic review. *Journal of Clinical Epidemiology*, *68*(2), 122-131.

Amusa, L., Zewotir, T., & North, D. (2019). Examination of entropy balancing technique for estimating some standard measures of treatment effects: a simulation study. *Electronic Journal of Applied Statistical Analysis*, *12*(2), 491-507.

Austin, P. C., Manca, A., Zwarenstein, M., Juurlink, D. N., & Stanbrook, M. B. (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology*, *63*(2), 142-153.

Austin, P. C., & 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.

Brettschneider, C., Bleibler, F., Hiller, T. S., Konnopka, A., Breitbart, J., Margraf, J., . . . 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.

Chan, K. C. G., Yam, S. C. P., & Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(3), 673-700.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., . . . Califf, R. M. (1996). The effectiveness of right heart catheterization in the initial care of critically III patients. *Jama*, *276*(11), 889-897.

Cook, R. J., & Sackett, D. L. (1995). The number needed to treat: a clinically useful measure of treatment effect. *BMJ*, *310*(6977), 452-454.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187-199.

Grupp, H., Kaufmann, C., König, H.-H., Bleibler, F., Wild, B., Szecsenyi, J., . . . Konnopka, A. (2017). Excess costs from functional somatic syndromes in Germany—An analysis using entropy balancing. *Journal of Psychosomatic Research*, *97*, 52-57.

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. In.

Hainmueller, J., & Xu, Y. (2013). Ebalance: A Stata package for entropy balancing. *Journal of Statistical Software*, *54*(7).

Hirano, K., & 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.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161-1189.

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*(1), 243-263.

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

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*(4), 523-539.

Kereiakes, D. J., Obenchain, R. L., Barber, B. L., Smith, A., McDonald, M., Broderick, T. M., . . . Hattemer, C. R. (2000). Abciximab provides cost-effective survival advantage in high-volume interventional practice. *American heart journal, 140*(4), 603-610.

Laupacis, A., Sackett, D. L., & Roberts, R. S. (1988). An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine, 318*(26), 1728-1733.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*, 337-346.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association, 113*(521), 390-400.

Mamdani, M., Sykora, K., Li, P., Normand, S.-L. T., Streiner, D. L., Austin, P. C., . . . Anderson, G. M. (2005). Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ, 330*(7497), 960-962.

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

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine, 32*(19), 3388-3414.

Murphy, D. J., & Cluff, L. E. (1990). The SUPPORT study. *Journal of Clinical Epidemiology, 43*, V-X.

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology, 54*(4), 387-398.

Parish, W. J., Keyes, V., Beadles, C., & 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 Outcomes Research Methodology, 18*(1), 17-46.

Pearson, J. L., Stanton, C. A., Cha, S., Niaura, R. S., Luta, G., & 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 and Tobacco Research, 17*(10), 1219-1227.

Rosenbaum, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics, 21*(1), 57-71.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169-188.

Sackett, D. L., Deeks, J. J., & Altman, D. G. (1996). Down with odds ratios! *Evidence-based Medicine, 1*(6), 164.

Schechtman, E. (2002). Odds ratio, relative risk, absolute risk reduction, and the number needed to treat—which of these should we use? *Value in Health, 5*(5), 431-436.

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

Wong, R. K., & Chan, K. C. G. (2017). Kernel-based covariate functional balancing for observational studies. *Biometrika, 105*(1), 199-213.

Zagar, A. J., Kadziola, Z., Lipkovich, I., & Faries, D. E. (2017). Evaluating different strategies for estimating treatment effects in observational studies. *Journal of Biopharmaceutical Statistics, 27*(3), 535-553.