

RESEARCH

Open Access



Assessment of the E-value in the presence of bias amplification: a simulation study

Eric Barrette¹, Lucas Higuera¹ and Kael Wherry^{1*}

Abstract

Background The E-value, a measure that has received recent attention in the comparative effectiveness literature, reports the minimum strength of association between an unmeasured confounder and the treatment and outcome that would explain away the estimated treatment effect. This study contributes to the literature on the applications and interpretations of E-values by examining how the E-value is impacted by data with varying levels of association of unobserved covariates with the treatment and outcome measure when covariate adjustment is applied. We calculate the E-value after using regression and propensity score methods (PSMs) to adjust for differences in observed covariates. Propensity score methods are a common observational research method used to balance observed covariates between treatment groups. In practice, researchers may assume propensity score methods that balance treatment groups across observed characteristics will extend to balance of unobserved characteristics. However, that assumption is not testable and has been shown to not hold in realistic data settings. We assess the E-value when covariate adjustment affects the imbalance in unobserved covariates.

Methods Our study uses Monte Carlo simulations to evaluate the impact of unobserved confounders on the treatment effect estimates and to evaluate the performance of the E-Value sensitivity test with the application of regression and propensity score methods under varying levels of unobserved confounding. Specifically, we compare observed and unobserved confounder balance, odds ratios of treatment vs. control, and E-Value sensitivity test statistics from generalized linear model (GLM) regression models, inverse-probability weighted models, and propensity score matching models, over correlations of increasing strength between observed and unobserved confounders.

Results We confirm previous findings that propensity score methods – matching or weighting – may increase the imbalance in unobserved confounders. The magnitude of the effect depends on the strength of correlation between the confounder, treatment, and outcomes. We find that E-values calculated after applying propensity score methods tend to be larger when unobserved confounders result in more biased treatment effect estimates.

Conclusions The E-Value may misrepresent the size of the unobserved effect needed to change the magnitude of the association between treatment and outcome when propensity score methods are used. Thus, caution is warranted when interpreting the E-Value in the context of propensity score methods.

Keywords Propensity score matching, E-value, Residual confounding, Simulation

Prior presentations: 14th Annual FDA/AdvaMed Medical Device Statistical Issues Conference May 2022, International Conference on Health Policy Statistics January 2023.

*Correspondence:

Kael Wherry
kael.wherry@medtronic.com

¹ Medtronic, Plc, Minneapolis, MN, USA

Background

Observational data or “real-world data” (RWD) are data including administrative healthcare claims, electronic health records (EHR), non-randomized registries, and patient data collected via mobile applications or wearable devices that offer many advantages for research [1]. However, observational data also present challenges for



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

researchers. Most notably, treatments or interventions are rarely randomly assigned outside of clinical trial settings, and patient populations often include everyone treated in the usual course of care without the specific inclusion and exclusion criteria of a clinical trial. Many times, the observable covariates between people who select the intervention and people who do not select the intervention are unbalanced.

Common analytic methods for achieving unbiased treatment effect estimates, such as propensity score matching, inverse probability of treatment weighting (using propensity scores), and regression-based approaches can adjust for differences in observed covariates. Propensity score methods (PSMs) such as matching and weighting have the benefit of producing balance in observable covariates between treated and untreated groups, analogous to a randomized study [2]. In practice, researchers may assume PSM-induced balance between treatment groups across observed characteristics will extend to balance of unobserved characteristics. However, this assumption is not testable and has been shown to not always be true. [3]. Achieving unbiased estimates with PSMs or regression-based methods is predicated on the assumption of “strong ignorability” or ‘unconfoundedness’, that is, that given observed covariates, treatment assignment is independent of the potential outcomes [4]. Unfortunately, with any of these methods it is impossible to directly test if unobserved covariates are related to the treatment and the outcome, not to mention balanced. Moreover, prior research has shown that using PSMs to balance observed covariates can result in more biased treatment effect estimates, compared to non-PSMs, by increasing the imbalance in unobserved covariates.

Numerous approaches, i.e., “sensitivity analyses”, have been proposed to assess the potential impact of unmeasured confounders [5–8]. One technique growing in popularity is the E-value, defined as “the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and outcome, conditional on the measured covariates, to fully explain away a specific treatment” [9]. Unlike many sensitivity tests, the E-value does not require assumptions about the number of unmeasured confounders or their functional form. The E-value is also appealing due to the direct calculation from a risk ratio or an approximation of a risk ratio from other common treatment effect estimates (e.g., odds ratios or hazard ratios). However, if the treatment effect estimate is biased, the effect on the E-value and its subsequent interpretation is not obvious. Our study contributes to the growing literature on the applications and interpretations of E-values, and by extension the sensitivity analysis literature. Specifically, this study sought to answer the question of how the

performance of the E-value is impacted in simulated data with varying levels of association between unobserved covariates and treatment and outcome.

Conceptual background

Our simulation study is tangentially related to the bias amplification literature. That literature considers the effect of conditioning on variables that are associated with treatment but not the outcome (except through treatment) – also known as instrumental variables – or variables that are much more strongly associated with treatment than the outcome – sometimes referred to as “near instruments”. Theoretical results and simulation studies have shown that controlling for an instrumental variable causes bias in treatment effect estimates [10, 11]. Potential bias amplification is an important consideration when designing an observational analysis because it has been shown through simulation studies to occur in a variety of realistic models [3, 12–15]. Our study diverges from the bias amplification literature by considering data with an unobserved covariate associated with treatment only. If this type of covariate was observed, it would be an instrumental variable but in our simulated data, it is not available to the researcher.

Our choice of this data structure is two-fold. First, previous research has used this structure in simulations and found that imbalance in the portion of the variation of the unobserved covariates that affect treatment choice that is independent of the observed covariates is necessary for propensity score-based methods to achieve balance in observed covariates. However, achieving balance also leads to greater imbalance in unobserved covariates and subsequently results in more biased treatment effect estimates [16]. Second, we contend, as did the researchers who used it previously, that this data structure is not uncommon. Consider a hypothetical population of patients with diabetes. The treatment is use of an insulin pump versus multiple daily injections of insulin. The outcome of interest could be a discrete measure of whether blood glucose time in target range was achieved or not. Characteristics associated with both treatment and outcome like age would be observed. Other demographic or socioeconomic characteristics may also be associated with both treatment and outcome but not observed. Finally, there is some other unobserved factor related only to the probability of using an insulin pump, such as physicians’ preference.

Our simulation seeks to assess how the E-value magnitude varies relative to a treatment effect estimate that has varying degrees of bias. It has been shown that the E-value has a nearly linearly monotonic relationship to a treatment effect estimate. Thus, for a given treatment effect estimate value the E-value is always the same, no

matter the research setting, data, or analysis method used [17]. Moreover, the derivation of the E-value assumes that unmeasured covariates are equally related to the treatment and outcome [9]. This is an assumption that has been contested by other researchers as being unlikely in many settings [18]. Using simulated, but realistic, data we are able to vary the strengths of associations in unobserved covariates between treatment and outcome. To provide practical results for practitioners we include commonly used treatment effect estimation methods: regression and PSMs to control for observed covariates.

Methods

We test the relationship between the E-value (and potential conclusions drawn from the E-value) and propensity score methods under varying scenarios of unobserved confounding. Using Monte Carlo methods we simulate a simple dataset including observed and unobserved covariates with varying levels of correlation between treatment and outcome based on the model in Brooks and Ohsfeldt [3]. This published model shows the tradeoffs between balance and bias in PSMs, and offers an appropriate framework to test how E-values handle unobserved confounders in a realistic approximation of observational research. First, we report the estimated odds ratio of the effect of treatment on the outcome relative to a control across various correlation scenarios and propensity score methods (inverse-probability weighted models and propensity score matching, based on the same propensity score). Next, we compare observed and unobserved covariate balance across the simulated scenarios. Finally, to assess potential conclusions about study robustness to unobserved covariates, we evaluate the calculated E-values across correlation and PSM scenarios.

Model

As in Brooks and Ohsfeldt [3], a patient's net utility gain from treatment (Tx) depends on the value of being cured (V), the relative cost of treatment (S), an observed confounder (X_m), and a set of unobserved confounders (X_{u1}, X_{u3}, X_{u4}).

$$Tx = V\beta_T - S + \alpha X_m + \alpha X_{u1} + \alpha X_{u3} + \alpha X_{u4} \quad (1)$$

Parameter α weights how confounders affect treatment decision, and β_T denotes how treatment affects the likelihood of being cured; this is our parameter of interest. The distributions and correlations of Xs are described below. A patient is treated ($T = 1$) if $Tx > 0$, and it is not ($T = 0$) otherwise.

The probability of a patient being cured depends on treatment T , the observed confounder X_m , and a set of unobserved confounders X_{u1}, X_{u2} .

$$\Pr(C) = \frac{\exp(\beta_T T + \beta_m X_m + \beta_{u1} X_{u1} + \beta_{u2} X_{u2})}{1 + \exp(\beta_T T + \beta_m X_m + \beta_{u1} X_{u1} + \beta_{u2} X_{u2})} \quad (2)$$

A patient is cured ($C = 1$) based on a Bernoulli distribution with probability $\Pr(C)$.

Data and simulations

All data are simulated in this study. We used the same distributions in Brooks and Ohsfeldt to make our results comparable to theirs. There is one observed confounder X_m drawn from a uniform [0,1] distribution, while unobserved confounders X_{u1}, X_{u3}, X_{u4} are linear combinations of X_m and μ , a random variable distributed uniform [0,1], weighted by a correlation $\rho \geq 0$.

$$X_{ui} = \rho X_m + (1 - \rho)\mu \quad (3)$$

The remaining unobserved confounder X_{u2} is a linear combination between unobserved confounder X_{u3} and μ .

$$X_{u2} = 0.5X_{u3} + 0.5\mu \quad (4)$$

Note that the unobserved confounder X_{u1} affects both the treatment decision and the probability of cure, and it is the main source of bias in the model. However, the correlation between X_{u2} and X_{u3} introduces an indirect path between treatment decision and the probability of cure. We use these confounders to generate T and C according to Eqs. (1) and (2), respectively. We generate 1,000 random datasets with 10,000 observations each for values of $\rho = \{0, 0.1, 0.2, \dots, 0.9\}$.

Estimation

For each dataset we estimate β_T and its associated risk ratio from a series of generalized linear regressions with C as outcome, binomial family and logit link, and 1) no additional confounders, 2) X_m , or observed confounders only, and 3) X_m, X_{u1}, X_{u2} , or observed and unobserved relevant confounders. For the propensity score methods, we estimate 1) an inverse probability weighted generalized linear regression, weighted by the inverse of a probability of treatment predicted from a probit model with X_m as control, and 2) a 1:1 greedy propensity score matched model with a caliper of 0.001; as a sensitivity analysis, we estimate 1:1 propensity score matching models with less restrictive calipers of 0.1 and 0.01. These methods reflect current practices in observational research, where an association –with and without a causal interpretation– of a treatment with an outcome is estimated using only observed confounders. We also calculate the Standardized Mean Difference (SMD) for X_m, X_{u1}, X_{u2} without any adjustments and with the propensity score methods. Lastly, we use the estimated risk

ratios (RR) to calculate the E-Value as in VanderWeele and Ding [9]:

$$E = RR + (RR \times (1 - RR))^{\frac{1}{2}}$$

Results

Figure 1 shows the SMD of observables (X_m) and unobservables (X_{u1} and X_{u2}) of the results of the Monte Carlo simulations by method (unadjusted, inverse probability of treatment weighting (IPTW), and PSM) and correlation ρ . For the observed confounder X_m , at all correlation levels both IPTW and PSM successfully improve balance with respect to the unadjusted results; however, at higher correlations the balance of IPTW worsens.

Compared to the unadjusted results, the SMD of X_{u1} is higher with IPTW and PSM when the correlation is low. For higher correlations, the SMDs from IPTW and PSM are lower than the unadjusted, but only the SMD with PSM decreases as correlation increases. The SMDs of X_{u2} follow a similar pattern, where the SMDs from IPTW and PSM are higher with respect to the unadjusted results at lower correlations, but these SMDs decrease when correlation increases. Except for the observed confounder X_m , neither IPTW nor PSM achieve SMDs to the informal level of 0.1 to consider the imbalance corrected.

Figure 2 shows the proportion of observations in common support from IPTW and PSM, defined as observations with an overlapping estimated probability of treatment. At lower correlations, few observations are outside the common support, but at higher correlations observations in common support reduce rapidly. Median observations in the common support are lower in PSM than in IPTW at all correlations greater than 0.

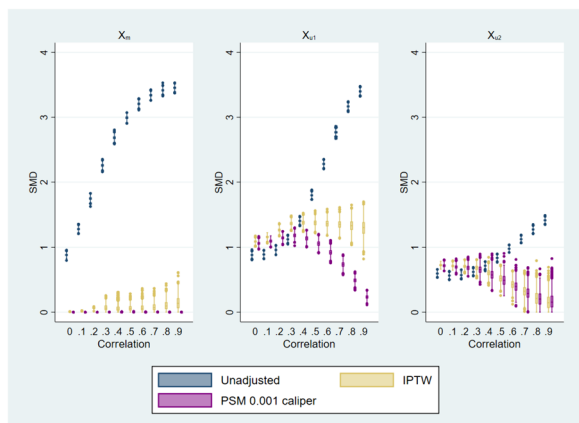


Fig. 1 Balance of observable X_m and unobservables X_{u1} , X_{u2} before and after adjustment. Note: Median, interquartile range box, and outliers of the standardized mean differences after 1,000 simulations. SMD standardized mean difference. IPTW Inverse probability treatment weighting. PSM Propensity score matching

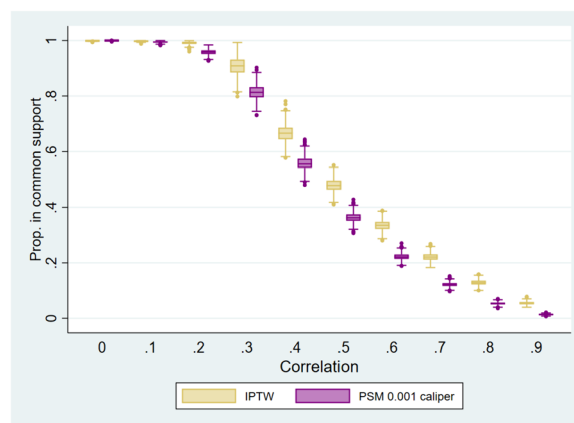


Fig. 2 Proportion of observations in common support. Note: Median interquartile range box, and outliers of the proportion of observations in common support after 1,000 simulations results. OR Odds ratio. IPTW Inverse probability treatment weighting. PSM Propensity score matching

The results of the regression specifications applying IPTW or PSM are reported in Table 1.

In Fig. 3, we plot the estimated coefficients and associated risk ratios of each estimate against the full information regression. The full information regressions correctly estimate β_T at 0.2, with an associated risk ratio of 1.07. The results from the regression without confounders (upper left side of Fig. 3) show an upwardly biased $\hat{\beta}_T$, and this bias increases with a higher correlation. IPTW results show a similar pattern, with an overall biased $\hat{\beta}_T$, and a higher bias with a higher correlation. But there is more variability in the β_T estimate with IPTW compared to the regression specification without confounders. Conversely, the regression with X_m as control (upper right side of Fig. 3) shows a biased $\hat{\beta}_T$ at lower correlation levels, but this bias decreases as the correlation increases. Results from PSM specifications also show a decrease in bias at higher levels of correlation, and a substantially higher variability at higher correlations.

In Fig. 4 we plot median E-values and bias in the estimation of β_T (defined as the difference between β_T and $\hat{\beta}_T$) by level of correlation for each estimation method, compared to the full information regression. There is a positive association between E-values and bias: larger E-values are paired to larger differences between the estimated risk ratio and the true risk ratio. As in the previous results, higher correlations between the observable and unobservable confounders increase the E-values in the regressions without controls and in IPTW specifications and decrease the E-values in the regressions with controls and in PSM. The variability in E-values is the highest in the PSM specifications at higher levels of correlation.

Table 1 Model results, by correlation and estimation method

ρ	All confounders		No confounders		Observed confounder		IPTW		PSM caliper 0.001	
	$\widehat{\beta}_T$	RR	$\widehat{\beta}_T$	RR	$\widehat{\beta}_T$	RR	$\widehat{\beta}_T$	RR	$\widehat{\beta}_T$	RR
0	0.197 (0.057)	1.069 (0.021)	0.310 (0.043)	1.110 (0.016)	0.262 (0.048)	1.092 (0.018)	0.263 (0.049)	1.093 (0.018)	0.267 (0.064)	1.095 (0.024)
0.1	0.202 (0.057)	1.071 (0.020)	0.327 (0.041)	1.117 (0.016)	0.258 (0.049)	1.091 (0.018)	0.261 (0.053)	1.092 (0.019)	0.263 (0.072)	1.094 (0.027)
0.2	0.206 (0.064)	1.072 (0.023)	0.343 (0.042)	1.123 (0.016)	0.258 (0.057)	1.091 (0.021)	0.260 (0.071)	1.091 (0.027)	0.262 (0.090)	1.092 (0.033)
0.3	0.202 (0.067)	1.070 (0.024)	0.358 (0.041)	1.129 (0.016)	0.247 (0.062)	1.086 (0.023)	0.271 (0.108)	1.097 (0.040)	0.264 (0.142)	1.093 (0.053)
0.4	0.198 (0.073)	1.069 (0.027)	0.370 (0.042)	1.133 (0.016)	0.233 (0.069)	1.082 (0.025)	0.287 (0.110)	1.102 (0.041)	0.239 (0.154)	1.084 (0.057)
0.5	0.201 (0.074)	1.070 (0.027)	0.383 (0.043)	1.138 (0.017)	0.224 (0.073)	1.078 (0.027)	0.314 (0.109)	1.112 (0.041)	0.242 (0.173)	1.085 (0.064)
0.6	0.201 (0.080)	1.070 (0.029)	0.393 (0.043)	1.142 (0.017)	0.217 (0.079)	1.076 (0.029)	0.342 (0.108)	1.122 (0.041)	0.246 (0.195)	1.086 (0.072)
0.7	0.200 (0.085)	1.070 (0.031)	0.401 (0.043)	1.145 (0.017)	0.206 (0.085)	1.072 (0.031)	0.366 (0.097)	1.131 (0.038)	0.236 (0.233)	1.081 (0.087)
0.8	0.200 (0.086)	1.070 (0.031)	0.414 (0.043)	1.150 (0.017)	0.203 (0.086)	1.070 (0.031)	0.384 (0.089)	1.138 (0.035)	0.217 (0.309)	1.074 (0.119)
0.9	0.204 (0.084)	1.071 (0.030)	0.426 (0.043)	1.155 (0.017)	0.205 (0.084)	1.071 (0.030)	0.411 (0.070)	1.148 (0.027)	0.235 (0.491)	1.079 (0.188)

Median and standard deviation of coefficients and Odds ratios of 1,000 simulations results. RR: Risk ratio. IPTW: Inverse probability treatment weighting. PSM: Propensity score matching

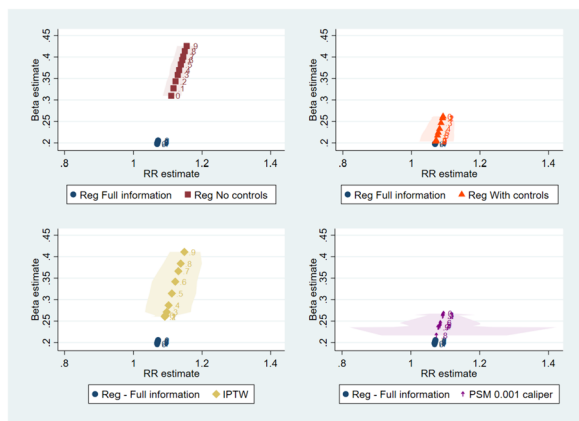


Fig. 3 Coefficient estimates and Risk Ratios. Note: Median (marker) and x-axis percentiles 5 and 95 (shaded area) of 1,000 simulations results. RR Risk ratio. IPTW Inverse probability treatment weighting. PSM Propensity score matching

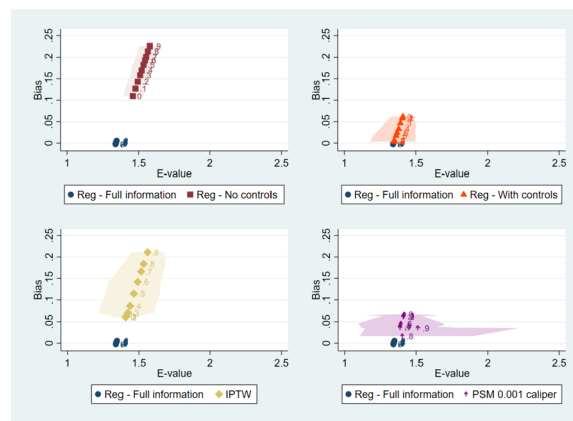


Fig. 4 Bias and E-Value. Note: Median (marker) and x-axis percentiles 5 and 95 of 1,000 simulations results. Bias defined as $\beta_T - \beta_{T,IPTW}$. IPTW Inverse probability treatment weighting. PSM Propensity score matching

Discussion

Controlling for covariates is essential to estimating treatment effects and PSMs are among the most common methodologies available to do so in observational research. Unfortunately, in almost all research studies there will be factors that are unobservable to the researcher. This limitation has motivated the development of numerous tools and best practices for designing,

conducting, and assessing an observational analysis including multiple sensitivity analysis methods that can be applied to provide a measure of study robustness, including sub-cohort analyses, falsification tests, alternate specifications [8, 19–21]. However, the fact remains that balancing covariates in observational analyses does not guarantee balance in unobserved confounders which must be considered when applying sensitivity tests.

Our study examines the performance of a sensitivity analysis to estimate whether unobserved confounders would change the conclusion regarding the treatment effect estimate in the presence of amplification bias. We contribute to a body of literature that examines how methods to control for confounding may actually introduce bias [3, 22, 23]. We first confirm our data generating model results in increased imbalances in unobserved confounders after balancing observed covariates. We find that the treatment effect estimates relative to the true effect vary by specifics of the propensity score method (i.e., matching vs weighting and size of the matching caliper). We also document the effect of increasing strength of correlated unobservables on reducing the size of the final analytic sample. In practice, this has implications for the generalizability of the treatment effect estimate. We then extend our analysis to evaluate the impact of correlation among unobservables on the E-value calculation.

The appropriate application and interpretation of the E-value remains a point of contention in the literature. Besides the simple calculation and minimal assumptions, the E-value is intended to have a straightforward interpretation; a larger E-value indicates a treatment effect estimate is more robust to unmeasured confounding and a smaller E-value indicates more sensitivity to unmeasured confounding. In their review of the use of the E-value, Blum et al. (2020) highlight how E-value results are presented in publications with phrases such as “These results demonstrate that substantial unmeasured confounding would be needed to reduce the observed associations to null” [24]. However, there are no threshold values or formal guidance around appropriate conclusions that can be drawn based on the E-value. Critics of the E-value have suggested users should provide guidance to readers on the interpretation of the E-value in addition to a pre-specified value for an ‘explain away’ threshold [24]. It may be reasonable to place greater or lesser emphasis on the E-value depending on what is already known about unmeasured confounders and a research topic; however, no formal recommendations exist.

Although the purpose of the E-value is not to test for bias, our simulation study demonstrates that its interpretation can be affected by the presence and magnitude of amplification bias. Confirmation that a biased treatment effect estimate is not sensitive to unobserved confounding is not necessarily informative. It may confirm the existence of a causal effect but in many instances the magnitude of the effect also matters. We find that as the estimated treatment effect becomes more biased away from the true treatment effect, the E-value also increases. Thus, in this setting, the E-value incorrectly suggests that it is less likely that an unobserved confounder would change the conclusion

of the analysis when in fact the treatment effect is biased toward finding an effect.

Our findings provide empirical support that an entire observational study protocol—both the main analyses and sensitivity analyses – must be informed by expertise on what is already known about potential unmeasured confounders in the context of a specific research question. Researchers who are aware of potential unobserved factors and possibly even a rough approximation of their magnitudes will be better able to determine the appropriate application of the E-value. This recommendation builds on a small but growing literature regarding E-values best practices. Blum and colleagues’ recommendations from their systematic literature review of E-values: users should provide guidance to readers on the interpretation of the E-value in addition to a pre-specified value for the ‘explain away’ threshold [24]. The recommendation is also consistent with VanderWeele and Mathur [25] who suggest that authors discuss potential unmeasured confounders and compare the E-value with covariate–outcome associations with prior literature. Our results demonstrate why this type of qualitative and quantitative bias assessment is needed.

Our paper has several limitations. First, our simulations used a single data generating process and only varied one aspect of the covariate correlation structure. It is impossible to know how much of an impact PSMs will have on the balance of unobserved covariates in any other data. In particular, our results do not generalize to data without variation of the unobserved covariates that affect treatment choice that is independent of the observed covariates. However, we expect that PSMs will always result in greater imbalance in unobserved covariates associated with only the treatment in settings where there is independent variation. Thus, the treatment effect estimate bias will be greater with PSMs relative to a regression approach in those instances but not for all data. In practice, researchers may consider multiple treatment effect estimation methods.

A second limitation is that our analysis of propensity score methods was limited to only two approaches. There have been numerous advances in PSMs and balancing methodologies more generally that may have different effects on the balance in unobservables [26–28]. The variation in effects and the inability to calculate exactly what the magnitude of the impact is for a study should also give researchers pause in applying a single measure to assess the robustness of their results.

Conclusion

Bias in treatment effect estimates due to imbalance in unobserved confounders may result in the E-value suggesting a spurious confidence in results under various covariate adjustment methodologies.

Abbreviations

EHR	Electronic health record
GLM	Generalized linear model
IPTW	Inverse probability of treatment weighting
PSM	Propensity score methods
RCT	Randomized controlled trial
RWD	Real-world data
RWE	Real-world evidence
SMD	Standardized mean difference

Acknowledgements

We thank the following for comments: Tim Hanson and Patrick Zimmerman, Bryan Dowd, Melissa Garrido, John M Brooks.

Authors' contributions

LH and EB contributed to the development and analysis of the data. LH prepared the tables and figures. LH, EB, and KW contributed to the conception and design of the work, interpretation of the data, and the drafting and editing of the manuscript. LH, EB, and KW have approved the submitted version of the manuscript and agree to be accountable for the contributions of this work.

Funding

The authors are employees of Medtronic, plc.

Availability of data and materials

Datasets available from corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 13 October 2023 Accepted: 3 March 2024

Published online: 28 March 2024

References

- Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence — What Is It and What Can It Tell Us? *N Engl J Med*. 2016;375(23):2293–7.
- Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*. 2010;25(1):1.
- Brooks JM, Ohsfeldt RL. Squeezing the Balloon: Propensity Scores and Unmeasured Covariate Balance. *Health Serv Res*. 2013;48(4):1487–507.
- Angrist JD, Pischke J-S. *Mostly harmless econometrics: An empiricist's companion*: Princeton university press; 2009.
- Rosenbaum PR. Sensitivity analyses informed by tests for bias in observational studies. *Biometrics*. 2023;79(1):475–87.
- Arah OA. Bias analysis for uncontrolled confounding in the health sciences. *Annu Rev Public Health*. 2017;38:23–38.
- Delaney JA, Seeger JD. Sensitivity analysis. Developing a protocol for observational comparative effectiveness research: a user's guide: Agency for Healthcare Research and Quality (US); 2013.
- Zhang X, Stamey JD, Mathur MB. Assessing the impact of unmeasured confounders for credible and reliable real-world evidence. *Pharmacoepidemiol Drug Saf*. 2020;29(10):1219–27.
- VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med*. 2017;167(4):268–74.
- Ding P, VanderWeele T, Robins JM. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*. 2017;104(2):291–302.
- Wooldridge JM. Should instrumental variables be used as matching variables? *Res Econ*. 2016;70(2):232–7.
- Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*. 2011;174(11):1213–22.
- Walker AM. Matching on provider is risky. *J Clin Epidemiol*. 2013;66(8 Suppl):S65–8.
- Ali MS, Groenwold RH, Pestman WR, Belitser SV, Roes KC, Hoes AW, et al. Propensity score balance measures in pharmacoepidemiology: a simulation study. *Pharmacoepidemiol Drug Saf*. 2014;23(8):802–11.
- Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*. 2009;6:1–9.
- King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. *Polit Anal*. 2019;27(4):435–54.
- Localio AR, Stack CB, Griswold ME. Sensitivity analysis for unmeasured confounding: E-values for observational studies. *Ann Intern Med*. 2017;167(4):285–6.
- Ioannidis JP, Tan YJ, Blum MR. Limitations and misinterpretations of E-values for sensitivity analyses of observational studies. *Ann Intern Med*. 2019;170(2):108–11.
- Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JP. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.
- Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report-Part I. *Value Health*. 2009;12(8):1044–52.
- Dreyer NA, Velentgas P, Westrich K, Dubois R. The GRACE checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution. *J Manag Care Pharm*. 2014;20(3):301–8.
- Bhattacharya J, Vogt WB. Do instrumental variables belong in propensity scores? : National Bureau of Economic Research Cambridge. USA: Mass; 2007.
- Daw JR, Hatfield LA. Matching and Regression to the Mean in Difference-in-Differences Analysis. *Health Serv Res*. 2018;53(6):4138–56.
- Blum MR, Tan YJ, Ioannidis JPA. Use of E-values for addressing confounding in observational studies—an empirical assessment of the literature. *Int J Epidemiol*. 2020;49(5):1482–94.
- VanderWeele TJ, Mathur MB. Commentary: Developing best-practice guidelines for the reporting of E-values. *Int J Epidemiol*. 2020;49(5):1495–7.
- Hirano K, Imbens GW, Ridder G. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*. 2003;71(4):1161–89.
- Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*. 2005;61(4):962–73.
- Hainmueller J. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Polit Anal*. 2012;20(1):25–46.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.