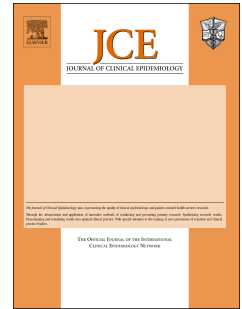


Accepted Manuscript

Magnitude and Direction of Missing Confounders had Different Consequences on Treatment Effect Estimation in Propensity Score Analysis

Tri-Long Nguyen, Gary S. Collins, Jessica Spence, Charles Fontaine, Jean-Pierre Daurès, P.J. Devereaux, Paul Landais, Yannick Le Manach.



PII: S0895-4356(16)30346-8

DOI: [10.1016/j.jclinepi.2017.04.001](https://doi.org/10.1016/j.jclinepi.2017.04.001)

Reference: JCE 9365

To appear in: *Journal of Clinical Epidemiology*

Received Date: 12 August 2016

Revised Date: 19 January 2017

Accepted Date: 4 April 2017

Please cite this article as: Nguyen T-L, Collins GS, Spence J, Fontaine C, Daurès J-P, Devereaux PJ, Landais P, Le Manach. Y, Magnitude and Direction of Missing Confounders had Different Consequences on Treatment Effect Estimation in Propensity Score Analysis, *Journal of Clinical Epidemiology* (2017), doi: 10.1016/j.jclinepi.2017.04.001.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Magnitude and direction of missing confounders had different consequences on treatment effect estimation in propensity score analysis

Tri-Long Nguyen,¹ Gary S. Collins,² Jessica Spence,³ Charles Fontaine,⁴ Jean-Pierre Daurès,⁵ P.J. Devereaux,⁶ Paul Landais,⁷ Yannick Le Manach.⁸

¹ Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, EA2415, Montpellier University, Montpellier, France; Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada.

² Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, United Kingdom.

³ Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada.

⁴ Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, EA2415, Montpellier University, Montpellier, France.

⁵ Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, EA2415, Montpellier University, Montpellier, France.

⁶ Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Perioperative Medicine and Surgical Research Unit & Departments of Clinical Epidemiology and Biostatistics and Medicine, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada.

⁷ Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, EA2415, Montpellier University; Department of Biostatistics, Clinical Research and Medical Informatics, Nîmes University Hospital, France.

⁸ Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada.

Corresponding author:

Yannick LE MANACH, MD, PhD;

Departments of Anesthesia & Clinical Epidemiology and Biostatistics; Michael DeGroote School of Medicine; Faculty of Health Sciences, McMaster University. Hamilton, Ontario, Canada.

Population Health Research Institute; David Bradley Cardiac, Vascular and Stroke Research Institute; Perioperative Medicine and Surgical Research Unit; Hamilton, Ontario, Canada.

Phone: (905) 527-7327; email: yannick.lemanach@phri.ca

Author contributions:

- **Study concept and design:** Nguyen, Collins, Le Manach.
- **Analysis and interpretation:** Nguyen, Collins, Le Manach.
- **Drafting of the manuscript:** Nguyen, Collins, Le Manach.
- **Critical revision of the manuscript for important intellectual content:** Nguyen, Collins, Spence, Fontaine, Daurès, Devereaux, Landais, Le Manach.

Financial support:

None.

Conflict of interest:

The authors declare no conflict of interest.

Acknowledgements:

We would like to thank Dr Jennifer de Beyer (Centre for Statistics in Medicine, University of Oxford, UK) for proof reading the manuscript. We warmly thank Sarah Kabani for her erudite expertise in editing the manuscript.

ABSTRACT

Objective: Propensity score (PS) analysis allows an unbiased estimate of treatment effects, but assumes that all confounders are measured. We assessed the impact of omitting confounders from a PS analysis on clinical decision-making.

Study Design and Setting: We conducted Monte Carlo simulations on hypothetical observational studies based on virtual populations and on the population from a large randomized trial (CRASH-2). In both series of simulations, PS analysis was conducted with all confounders and with omitted confounders, which were defined to have different strengths of association with the outcome and treatment exposure. After inverse probability of treatment weighting, we calculated the absolute risk differences and numbers needed to treat (NNT).

Results: In both series of simulations, omitting a confounder that was moderately associated with the outcome and exposure led to negligible bias on the NNT scale. The bias induced by omitting strongly positive confounding variables remained below 15 patients to treat. Major bias and reversed effects were found only when omitting highly prevalent, strongly negative confounders that were similarly associated with the outcome and exposure with odds ratios greater than 4.00 (or < 0.25). This omission was accompanied by a substantial decrease in analysis power.

Conclusion: The omission of strongly negative confounding variables from a PS analysis can lead to incorrect clinical decision-making. However, omitting these variables also decreases the analysis power, which may prevent the reporting of significant but misleading effects.

KEYWORDS

Causal inference; propensity score; confounding bias.

INTRODUCTION

Propensity score (PS) analysis is a popular method for causal inference in observational studies [1, 2]. When conducting a randomized controlled trial (RCT) is impractical for technical or ethical reasons [3], PS analysis can be used to estimate treatment effects and guide clinician decisions.

Sometimes called the “balancing score,” the PS is defined as the conditional probability of receiving the treatment of interest given a set of measured covariates [4]. The PS can be used to design a framework in which the covariates are balanced across treatment groups, as in an RCT. To provide unbiased estimates of causal effect, PS analysis assumes that all confounding variables are measured [4]. If confounders are omitted, PS analysis is exposed to confounding bias that can either exaggerate (positive confounding) or underestimate (negative confounding) the apparent treatment effect [5]. Austin *et al.* showed that PS models should only include true confounders to optimally balance treatment groups and remove confounding [6]. However, recent systematic reviews have highlighted the lack of discussion on PS modeling in the literature [1, 2]. A large proportion of published observational studies do not explicitly state the type of variables included in their PS models. Clearly distinguishing between true confounders and other variables is difficult in practice and recording all true confounders remains a challenge. Thus, the omission of confounders in PS analysis may pose a threat to current studies.

Although we know that unmeasured confounders may introduce bias into a PS analysis [4], the effect of this bias on clinical decision-making has not yet been explored. We hypothesized that the effect of the bias would vary according to the direction (positive or negative confounding) and the strength of the association between the unmeasured confounders and either the outcome or treatment exposure.

We present two series of Monte Carlo simulations, based on virtual populations and on a published data set. We evaluate how the direction and strength of confounding affects medical decision-making when confounders are omitted from a PS analysis.

ACCEPTED MANUSCRIPT

METHODS

We conducted two series of Monte Carlo simulations, one on fictive data and one on a real-world data set collected in a large RCT, CRASH-2 [7, 8].

Simulations on fictive data

We generated a fictive data set (sample size = 2,000 units) that mimicked real perioperative settings [9, 10], by using an approach similar to Setoguchi's method [11]. We created 14 random standard normal variables. These base covariates were designed to take into account different degrees of correlation, as shown in *Figure 1*. We dichotomized these variables on cut-off values that allowed us to obtain realistic covariates, the prevalence of which agreed with perioperative recorded data (e.g. comorbidity, physiological and biological status). We added a 15th variable that was a combination of the other variables and mimicked the cardiac risk index established by Lee *et al.* [12]. The data set was thus composed of 15 explanatory variables (W_1 - W_{15}). The treatment (Z) and outcome (Y), both binary, were defined by logistic models (*Appendix*). The model coefficients were set for an expected treatment exposure, $p(Z)$, of 0.35-0.45 and an expected outcome prevalence, $p(Y)$, of 0.08-0.12.

We used five true confounders (W_1 , W_2 , W_7 , W_{11} and W_{13}). We focused on the effect of omitting W_7 , a binary variable that mimicked the hypertension status. W_7 was designed for a high population exposure: $p(W_7) = 0.48$. We defined seven values for β_7 and α_7 , which are the odds ratio for the exposure and outcome, respectively: 0.17, 0.25, 0.5, 1, 2, 4 and 6. Combining these values gave 49 scenarios, which represented different associations between W_7 and the outcome and exposure.

The treatment was designed to have a protective conditional effect, with an odds ratio of 0.60 (log-odds ratio of -0.51), in alignment with published perioperative studies [13-15]. As the

odds ratio has been criticized for its poor clinical meaning [16], we computed the true average treatment effect (ATE) as an absolute risk difference (ARD):

$$p(Y_1) - p(Y_0),$$

where $p(Y_1)$ is the expected outcome prevalence if every subject was treated and $p(Y_0)$ is the expected outcome prevalence if no subjects were treated [17]. Using this definition, the ATE only depended on the outcome model, not the exposure. The true ATE therefore had seven values, driven by the seven α_7 values that resulted in seven outcome models. We defined the intercept α_0 for each value of α_7 so that every outcome model had an ATE of -0.04 (number needed to treat: 25), allowing the outcome models to be directly compared.

For each scenario, we estimated the ARD by PS analysis using the inverse probability of treatment weighting [18]. The PS was estimated by a logistic regression that included all 15 explanatory variables (W_1 - W_{15}) as main effects. We recognize that including instrumental variables inflates the bias and believe this model to be realistic [2] rather than optimal [6, 19, 20]. We then removed W_7 from the analysis and defined the weights for creating pseudo-populations as $W = 1/PS$ in the treated and $W = 1/(1-PS)$ in the control units to estimate the ATE [18]. We calculated the ARD on the weighted samples.

To assess the clinical meaning of the estimates, we computed the number needed to treat (NNT) to prevent one outcome as $NNT = -1/ARD$. We calculated the empirical power of each analysis as the percentage of statistically significant estimates ($P < 0.05$). We reported the bias, expressed on the ARD and the NNT scale, induced by omitting W_7 as the difference between the true ATE and the estimate obtained from the PS analysis that omitted W_7 .

The analysis was conducted 1,000 times on each scenario, giving a total of 49,000 simulations.

Illustrative example: simulations on CRASH-2 data

We used the open-access data from the CRASH-2 trial as an illustrative example [7, 8] (data set available at <https://ctu-app.lshtm.ac.uk/freebird/>). CRASH-2 was a large multicenter RCT conducted in 40 countries that recruited 20,211 patients. It assessed the effect of tranexamic acid (TXA) *versus* placebo on all-cause mortality within 28 days in trauma patients with significant hemorrhage [7]. It reported an occurrence of death of 14.5% in the treatment group *versus* 16.0% in the control group [7], giving an estimated ATE, calculated as the ARD, of -0.015 (NNT = 67 patients).

We only used complete cases from the CRASH-2 data set (95% of the initial sample, $N = 19,128$) that recorded age, systolic blood pressure (SBP), respiratory rate (RR), heart rate (HR), the Glasgow coma score (GCS), central capillary refill time (CC) and penetrating injury (PIN). Perel *et al.* developed a prognostic model [21] (available at www.sealedenvelope.com/trials/crash2model/) on the CRASH-2 data set and externally validated it on data from the Trauma Audit Research Network. We used this model to calculate the expected outcome prevalence if every subject was treated, $p(Y_1)$, and the expected prevalence if no subjects were treated, $p(Y_0)$ [17]. The expected ATE, calculated as $p(Y_1) - p(Y_0)$, should be similar to the result presented in the CRASH-2 trial. As the effect was calculated without considering the time of TXA or placebo administration [7], we dropped the interaction terms between TXA and the hours since injury from the prognostic model and slightly modified the TXA coefficient (*Appendix*).

To first check the assumption that this model provided an ATE that agreed with the reported ARD, we simulated 1,000 RCTs by bootstrapping the initial sample and randomly assigning the treatment following a binomial distribution with probability 0.5. Using this model, we computed the outcome probability, $p(Y)$, in the treated and control groups. The outcome occurred when $p(Y)$ was greater than $u \sim U(0,1)$. The simulated median mortality rate was

15.2% across all of the samples, with 14.7% in the treated and 16.3% in the controls (ADR = -0.016), which agreed with the original CRASH-2 report [7].

Next, given that this model agreed with a realistic scenario, we used it to simulate a series of observational studies. Since treatment is not allocated at random in observational studies, we mimicked clinical decision-based allocations by attributing TXA to each patient according to their clinical characteristics. In 1,000 independent bootstrapped samples (*i.e.* 1,000 simulated observational studies), we administered TXA to trauma patients who were more likely to be: in shock (SBP < 75 mmHg, HR > 110 beats/min and CC > 2 s); be elderly; live in high-income countries. The corresponding true PS model is reported in the Appendix.

All of the variables in this model were binary, except Age, which was continuous and linear: MIC = 1 if the patient lived in a middle-income country, LIC = 1 if the patient lived in a low-income country, SBP_{binary} = 1 if SBP < 75 mmHg, HR_{binary} = 1 if HR > 110 beats/min and CC_{binary} = 1 if CC > 2 s. According to the outcome model, we had five true confounders (MIC, LIC, SBP, HR and Age) and one treatment predictor (CC). TXA was actually administered when $p(\text{TXA})$ was greater than $u \sim U(0,1)$. The coefficients were chosen to set the expected treatment exposure, $p(\text{TXA})$, to 0.60.

For each sample, we derived a PS model that included all six covariates. We then successively removed either MIC, LIC, Age or SBP from the model. Based on the estimated PS, we created pseudo-populations using the inverse probability of treatment weighting and estimated the ATE [18]. The ARD and NNT were calculated using the weighted samples. Each simulation was performed 1,000 times.

RESULTS

Simulations on fictive data

Across all of the simulations, the median treatment frequency was 41.0% and the median outcome prevalence was 10.0%. Despite their different α_7 values, which represented the outcome odds ratio (0.17, 0.25, 0.5, 0, 2, 4 and 6), the scenarios had very similar ARD empirical reference values (-0.041, -0.042, -0.041, -0.040, -0.040, -0.040 and -0.041, respectively, with NNT = 24, 24, 24, 25, 25, 25 and 24, respectively).

As reported in *Table 1*, PS analysis resulted in an unbiased ARD when all of the confounders were measured. If a covariate was not measured, PS analysis only led to an unbiased effect if the covariate had no association with either the outcome or exposure, *i.e.*, it was not a true confounder.

Omitting a true confounder from the PS analysis led to biased estimates, as expected. As depicted in *Figure 2*, the magnitude and direction of the bias depended on the association between the confounding variable and the outcome and treatment exposure. The bias was particularly strong when the unmeasured confounder was strongly associated with both the outcome and exposure. If the confounder had opposite associations with the outcome and exposure (for example, as a protective factor for the exposure, but a risk factor for the outcome, or *vice-versa*), its omission exaggerated the apparent effect, taking it further from null (positive confounding). In contrast, if the confounder was similarly associated with the outcome and exposure, omitting a negative confounder attenuated the effect, bringing it closer to null. We observed Simpson's paradox in extreme cases [22], a phenomenon in which the treatment effect was reversed from protective to harmful.

Table 2 shows the estimated effects on the NNT scale. Interestingly, the bias on the NNT scale (*Figure 3*) was weaker if a positive confounder was not measured, with a maximum bias

of $NNT = 14$ patients. *Table 3* shows that omitting a positive confounder increased the analysis power, whereas omitting a negative confounder decreased it. Again, the magnitude of this variation was proportional to the strength of the confounding variable (*Figure 4*).

Additional simulations were used to explore how population exposure to the unmeasured confounder could influence results (*Figure 5*). We considered the worst case, which omitted a negative confounder with outcome and treatment odds ratios of 6.00, and varied its prevalence in the population. The bias due to its omission was notable only when the prevalence was greater than 10.0%.

Simulations on CRASH-2 data

Across all of the simulations on the real-world data, the empirical reference value of ARD was -0.016 (NNT: 63), which agrees with the estimates reported in the CRASH-2 trial (ARD = -0.015 and NNT = 67) [7]. As reported in *Table 4*, the estimated ARD was unbiased when no confounders were omitted. Again, the effect of the unmeasured confounders on the PS analysis depended on the direction and strength of their association with the outcome and treatment exposure. There were four unmeasured confounders: MIC (moderately associated with both TXA and mortality), LIC (strongly associated with TXA and moderately associated with mortality), Age (moderately associated with TXA and strongly associated with mortality) and SBP (strongly associated with both TXA and mortality). MIC and LIC were positive confounders, while Age and SBP were negative confounders. The bias in the estimates had a negligible clinical effect when MIC, LIC or Age were omitted, as the median NNT was at most only 14 patients to treat greater than the reference value (*Table 4*). Due to the inverse scale and the low ATE, there was a large variance in the NNT. The opposite effect was seen when SBP was omitted from the analysis and the treatment effect was estimated to be harmful. However, the power for detecting this biased estimated effect was low (*Table 4*).

The simulations using real-world data confirmed the results obtained on fictive data. Omitting confounders from the PS analysis led to biased estimates. This bias had a negligible clinical significance if the omitted confounder was positive, but was clinically meaningful if the unmeasured confounder was negative and strongly associated with both the treatment exposure and outcome. The estimated effects were generally statistically insignificant, even when the bias was clinically meaningful.

DISCUSSION

We have shown that omitting confounders from a PS analysis can lead to biased treatment effects. However, this bias is only clinically meaningful if the unmeasured confounders are strongly associated with both the outcome and treatment exposure. When negative confounders are omitted, the ATE estimator is biased towards the null, so the analysis has a decreased power and may miss a statistically significance.

PS analysis is based on the assumption that all confounders are measured [4]. However, it is difficult in practice to make a complete record of all confounders, as some may not be known. Confounding bias is therefore a concern in observational studies. We found that the magnitude of this bias was often clinically negligible in terms of NNT. When omitted confounders were moderately associated with the outcome and treatment exposure, the estimated NNT differed from the reference value by only a few subjects. The difference was smaller than those generally reported in meta-analyses of RCTs [23-25]. We showed that the magnitude of the bias increased notably when the unmeasured confounders were strongly associated with the outcome and treatment exposure. This magnitude differed with the direction of the bias.

Depending on their relationship with the exposure and outcome, confounders can lead to an apparent treatment effect that is attenuated (negative confounding) or exaggerated (positive confounding) compared to the true treatment effect [5]. In clinical practice, treating patients who have a lower risk for the outcome (or not treating patients who have a greater risk of the outcome occurring) does not make sense if the treatment is expected to be protective. Therefore, we hypothesize that negative confounders are more common than positive confounders in perioperative observational studies, though this assumption does not hold for pharmacoepidemiological studies, in which the treatment is often suspected to cause harm. We found that omitting positive confounders did not greatly change the NNT, whereas

omitting negative confounders caused much greater NNT changes. In the extreme case of effect attenuation, omitting a major negative confounder reversed the treatment effect, a phenomenon known as Simpson's paradox [22]. This phenomenon was accompanied by a substantial decrease in the analysis power and in the probability of considering a biased treatment effect to be significant. However, incorrect conclusions will still be drawn from these results and published in the medical literature. The PS must be specified as correctly as possible to reduce other sources of bias. Misspecification can reduce covariates balance and amplify confounding bias [26-28].

The strength of the association between a confounder and the exposure on the odds ratio scale is equivalent to the factor Γ described by Rosenbaum [29]. Factor Γ quantifies the gap between treatment attribution due to a confounder and random treatment assignment. We found that omitting confounding variables led to major bias only if the treatment attribution due to the unmeasured confounders was very different from random assignment. This condition was necessary but not sufficient, as the unmeasured confounders also had to be strongly associated with the outcome to cause a meaningful bias. We also showed that the magnitude of the bias depended on the population exposure to the strong confounding variables. Although we hypothesize that researchers are likely to know about any highly prevalent, strong confounders, goodness-of-fit tests [30] and graphical assessment of calibration curves [31] cannot be used to detect unmeasured confounders.

Confounding bias is a well-known problem [4], but its clinical meaning has been poorly explored. Brooks and Ohsfeldt compared regression and PS analysis, and found that the latter could exacerbate bias in estimates due to unmeasured confounders [32]. We did not compare PS analysis with regression analysis. PS analysis and RCTs allow a marginal effect to be estimated, whereas regression analysis provides a conditional effect. Comparing PS analysis and logistic regression was inappropriate in this study, as these two effect estimates are not

collapsible on the odds ratio scale, while they are on the absolute risk difference scale [33-35]. Logistic regression provides an odds ratio, which has also been criticized for its poor clinical meaning [16]. As bias on an odds ratio scale is hardly interpretable, simulations comparing PS analysis with logistic regression would not explain the effect on clinical decision-making. For all these reasons, we estimated the treatment effect as an absolute risk difference, which allowed us to report a number needed to treat. We agree with Brooks and Ohsfeldt that non-measurement of confounders leads to bias in estimated effects [32]. However, we draw different conclusions about the clinical impact of this bias, as we have shown that the confounding bias is only clinically meaningful under particular conditions.

Our study should be interpreted in light of some limitations. We explored only one method of PS analysis (inverse probability of treatment weighting), which allowed us to estimate ATE. We did not assess the effect of unmeasured confounders on the average treatment effect in the treated estimand, which can be also be obtained by inverse probability of treatment weighting, full matching or matching analysis [26, 36, 37]. In our simulations, we assessed the impact of omitting one variable in the PS model. We did not explore the scenario in which multiple variables with moderate associations were unmeasured. Our study also only took into account treatments with moderate effects on binary outcomes. We hypothesize that the magnitude of the bias in NNT scale would be weaker if the treatment effect was stronger, because it would be further from null. In addition to the NNT, other comprehensive measures of health benefit, such as the quality-adjusted life years, should be considered in clinical decision making [38]. We did not consider interaction terms between the treatment and the omitted confounder. In addition, we focused on the treatment effect estimation and did not report the covariate balance in the weighted samples.

Although our simulations on fictive data was designed to mimic large cohorts from the perioperative literature [9, 10], we cannot be sure that our virtual populations were realistic.

We also only defined linear, additive models for the treatment exposure (*i.e.* the true PS model) and the outcome (*i.e.* the true outcome model), even though non-linearity and non-additivity exist in practice. We overcame these limitations by working on published data, choosing the CRASH-2 data set as an illustrative case.

In conclusion, we have shown that the bias induced by omitting confounders in PS analysis is clinically meaningful under certain conditions. Omitting confounders can greatly affect clinical decision-making if the unmeasured confounders are negative and strongly associated with both the outcome and treatment exposure. However, omitting these confounders decreases the power of the analysis, which reduces the chance that these misleading conclusions will be considered significant and thus be reported.

REFERENCES

- [1] Ali MS, Groenwold RH, Belitser SV, Pestman WR, Hoes AW, Roes KC, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *J Clin Epidemiol*. 2015;68:112-21.
- [2] Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary JY, Porcher R. Propensity scores in intensive care and anaesthesiology literature: a systematic review. *Intensive care medicine*. 2010;36:1993-2003.
- [3] Rossi PH, Freeman HE. A systematic approach. Sage Publications, Inc, Newbury Park. 1993.
- [4] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
- [5] Gail MH, Benichou J, eds. Encyclopedia of epidemiologic methods. New York: Wiley. 2000.
- [6] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine*. 2007;26:734-53.
- [7] CRASH-2 Collaborators, Shakur H, Roberts I, Bautista R, Caballero J, Coats T, et al. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial. *Lancet*. 2010;376:23-32.
- [8] CRASH-2 Collaborators. Effect of tranexamic acid in traumatic brain injury: a nested randomised, placebo controlled trial (CRASH-2 Intracranial Bleeding Study). *Bmj*. 2011;343:d3795.

- [9] Vascular Events In Noncardiac Surgery Patients Cohort Evaluation Study Investigators, Devereaux PJ, Chan MT, Alonso-Coello P, Walsh M, Berwanger O, et al. Association between postoperative troponin levels and 30-day mortality among patients undergoing noncardiac surgery. *Jama*. 2012;307:2295-304.
- [10] Botto F, Alonso-Coello P, Chan MT, Villar JC, Xavier D, Srinathan S, et al. Myocardial injury after noncardiac surgery: a large, international, prospective cohort study establishing diagnostic criteria, characteristics, predictors, and 30-day outcomes. *Anesthesiology*. 2014;120:564-78.
- [11] Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*. 2008;17:546-55.
- [12] Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation*. 1999;100:1043-9.
- [13] POISE Study Group, Devereaux PJ, Yang H, Yusuf S, Guyatt G, Leslie K, et al. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet*. 2008;371:1839-47.
- [14] Schouten O, Boersma E, Hoeks SE, Benner R, van Urk H, van Sambeek MR, et al. Fluvastatin and perioperative events in patients undergoing vascular surgery. *The New England journal of medicine*. 2009;361:980-9.
- [15] Pearse RM, Harrison DA, MacDonald N, Gillies MA, Blunt M, Ackland G, et al. Effect of a perioperative, cardiac output-guided hemodynamic therapy algorithm on outcomes following major gastrointestinal surgery: a randomized clinical trial and systematic review. *Jama*. 2014;311:2181-90.

- [16] Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol*. 1994;47:881-9.
- [17] Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *J Clin Epidemiol*. 2010;63:2-6.
- [18] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*. 2004;23:2937-60.
- [19] Pearl J. Invited commentary: understanding bias amplification. *Am J Epidemiol*. 2011;174:1223-7; discussion pg 8-9.
- [20] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163:1149-56.
- [21] Perel P, Prieto-Merino D, Shakur H, Clayton T, Lecky F, Bouamra O, et al. Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. *Bmj*. 2012;345:e5166.
- [22] Simpson EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*. 1951;13: 238–41.
- [23] Moiniche S, Romsing J, Dahl JB, Tramer MR. Nonsteroidal antiinflammatory drugs and the risk of operative site bleeding after tonsillectomy: a quantitative systematic review. *Anesthesia and analgesia*. 2003;96:68-77, table of contents.
- [24] Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading. *Bmj*. 1999;318:1548-51.

- [25] Kalesan B, Pilgrim T, Heinimann K, Raber L, Stefanini GG, Valgimigli M, et al. Comparison of drug-eluting stents with bare metal stents in patients with ST-segment elevation myocardial infarction. *European heart journal*. 2012;33:977-87.
- [26] Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical methods in medical research*. 2015.
- [27] Wyss R, Ellis AR, Brookhart MA, Girman CJ, Jonsson Funk M, LoCasale R, et al. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *Am J Epidemiol*. 2014;180:645-55.
- [28] Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *Am J Epidemiol*. 2015;181:108-19.
- [29] Rosenbaum PR. Design sensitivity in observational studies. *Biometrika*. 2004;11:153-64.
- [30] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiology and drug safety*. 2005;14:227-38.
- [31] Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in medicine*. 2014;33:517-35.
- [32] Brooks JM, Ohsfeldt RL. Squeezing the balloon: propensity scores and unmeasured covariate balance. *Health services research*. 2013;48:1487-507.

- [33] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *International journal of epidemiology*. 2008;37:1142-7.
- [34] Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125:761-8.
- [35] Miettinen OS, Cook EF. Confounding: essence and detection. *Am J Epidemiol*. 1981;114:593-603.
- [36] Austin PC, Stuart EA. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research*. 2015.
- [37] Austin PC, Stuart EA. Optimal full matching for survival outcomes: a method that merits more widespread use. *Statistics in medicine*. 2015;34:3949-67.
- [38] Chong CA, Tomlinson G, Chodirker L, Figdor N, Uster M, Naglie G, et al. An unadjusted NNT was a moderately good predictor of health benefit. *J Clin Epidemiol*. 2006;59:224-33.

FIGURE LEGENDS

Figure 1. Variable definitions and data generation of virtual populations. We used five true confounders (W_1 , W_2 , W_7 , W_{11} and W_{13}), seven potential confounders (W_3 , W_4 , W_5 , W_9 , W_{10} , W_{12} and W_{15}), one treatment predictor (W_6) and two unrelated variables (W_8 , W_{14}). Arrows represent causal effects. Arcs represent correlations and are accompanied by the correlation coefficients.

Figure 2. Bias in estimated absolute risk difference (ARD) after omitting a confounder (W_7), according to its strength of association with the outcome and treatment exposure. The diameters of the circles are proportional to the reported bias.

Figure 3. Bias in the estimated number needed to treat (NNT) after omitting a confounder (W_7), according to its strength of association with the outcome and treatment exposure. The diameters of the circles are proportional to the reported bias. When the effect is reversed, the number needed to harm is calculated.

Figure 4. Variation in the empirical power after omitting a confounder (W_7), according to its strength of association with the outcome and treatment exposure. The diameters of the circles are proportional to the variation.

Figure 5. Estimated treatment effects and power regarding the population exposure to the unmeasured confounder (W_7). For each exposure, we repeated the PS analysis 1,000 times. Shaded areas indicate reversed effects, where the number needed to harm is calculated instead of the NNT. ARD, absolute risk difference; NNT, number needed to treat; PSA, propensity score analysis.

Table 1. Estimated absolute risk difference (ARD), according to the association strength of a confounding variable (W_7) with the outcome and treatment exposure. Medians are reported [2.5th; 97.5th percentiles].

<i>Propensity score analysis with all confounders</i>								
		Odds ratio for the treatment exposure						
		<i>0.17</i>	<i>0.25</i>	<i>0.50</i>	<i>1.00</i>	<i>2.00</i>	<i>4.00</i>	<i>6.00</i>
Odds ratio for the outcome	<i>6.00</i>	-0.042 [-0.071; -0.011]	-0.040 [-0.069; -0.014]	-0.040 [-0.067; -0.015]	-0.041 [-0.066; -0.015]	-0.040 [-0.066; -0.015]	-0.041 [-0.067; -0.013]	-0.041 [-0.068; -0.012]
	<i>4.00</i>	-0.040 [-0.07; -0.009]	-0.040 [-0.067; -0.013]	-0.040 [-0.063; -0.014]	-0.040 [-0.065; -0.014]	-0.040 [-0.066; -0.014]	-0.040 [-0.066; -0.011]	-0.039 [-0.069; -0.01]
	<i>2.00</i>	-0.041 [-0.067; -0.013]	-0.041 [-0.069; -0.014]	-0.041 [-0.066; -0.013]	-0.041 [-0.064; -0.016]	-0.041 [-0.066; -0.013]	-0.039 [-0.066; -0.012]	-0.040 [-0.068; -0.01]
	<i>1.00</i>	-0.040 [-0.066; -0.012]	-0.041 [-0.065; -0.014]	-0.040 [-0.065; -0.014]	-0.040 [-0.063; -0.016]	-0.040 [-0.066; -0.015]	-0.040 [-0.069; -0.011]	-0.041 [-0.068; -0.012]
	<i>0.50</i>	-0.040 [-0.067; -0.011]	-0.041 [-0.068; -0.014]	-0.040 [-0.066; -0.016]	-0.040 [-0.065; -0.015]	-0.040 [-0.066; -0.012]	-0.041 [-0.068; -0.012]	-0.040 [-0.069; -0.013]
	<i>0.25</i>	-0.041 [-0.071; -0.013]	-0.040 [-0.066; -0.014]	-0.041 [-0.065; -0.015]	-0.042 [-0.067; -0.016]	-0.042 [-0.068; -0.015]	-0.041 [-0.07; -0.014]	-0.040 [-0.07; -0.011]
	<i>0.17</i>	-0.040 [-0.067; -0.012]	-0.040 [-0.067; -0.014]	-0.040 [-0.067; -0.014]	-0.041 [-0.066; -0.013]	-0.041 [-0.068; -0.014]	-0.041 [-0.067; -0.012]	-0.041 [-0.069; -0.011]
<i>Propensity score analysis omitting a confounder</i>								
		Odds ratio for the treatment exposure						
		<i>0.17</i>	<i>0.25</i>	<i>0.50</i>	<i>1.00</i>	<i>2.00</i>	<i>4.00</i>	<i>6.00</i>
Odds ratio for the outcome	<i>6.00</i>	-0.092 [-0.118; -0.067]	-0.081 [-0.108; -0.056]	-0.061 [-0.088; -0.036]	-0.041 [-0.067; -0.015]	-0.018 [-0.044; 0.006]	0.001 [-0.026; 0.027]	0.012 [-0.014; 0.04]
	<i>4.00</i>	-0.080 [-0.107; -0.055]	-0.072 [-0.095; -0.047]	-0.057 [-0.081; -0.03]	-0.040 [-0.066; -0.014]	-0.023 [-0.049; 0.003]	-0.007 [-0.033; 0.02]	0.002 [-0.024; 0.028]
	<i>2.00</i>	-0.062 [-0.086; -0.036]	-0.058 [-0.084; -0.033]	-0.049 [-0.074; -0.023]	-0.041 [-0.065; -0.015]	-0.031 [-0.056; -0.006]	-0.022 [-0.049; 0.003]	-0.019 [-0.043; 0.008]
	<i>1.00</i>	-0.040 [-0.065; -0.014]	-0.041 [-0.065; -0.015]	-0.040 [-0.065; -0.014]	-0.040 [-0.063; -0.017]	-0.041 [-0.065; -0.015]	-0.040 [-0.066; -0.013]	-0.041 [-0.064; -0.015]

0.50	-0.018 [-0.046; 0.006]	-0.024 [-0.048; 0.001]	-0.031 [-0.056; -0.006]	-0.041 [-0.066; -0.016]	-0.049 [-0.076; -0.022]	-0.057 [-0.083; -0.033]	-0.062 [-0.087; -0.037]
0.25	-0.001 [-0.029; 0.028]	-0.008 [-0.032; 0.018]	-0.025 [-0.049; 0.003]	-0.042 [-0.067; -0.015]	-0.059 [-0.084; -0.034]	-0.074 [-0.101; -0.05]	-0.083 [-0.108; -0.057]
0.17	0.009 [-0.017; 0.037]	-0.001 [-0.027; 0.026]	-0.020 [-0.047; 0.007]	-0.041 [-0.066; -0.013]	-0.062 [-0.089; -0.033]	-0.082 [-0.107; -0.055]	-0.092 [-0.119; -0.067]

Table 2. Estimated number needed to treat to prevent one outcome (NNT), according to the association strength of a confounding variable (W_7) with the outcome and treatment exposure. Medians are reported [2.5th; 97.5th percentiles]. When the effect is reversed, the NNT is not available (NA) and a number needed to harm should be calculated.

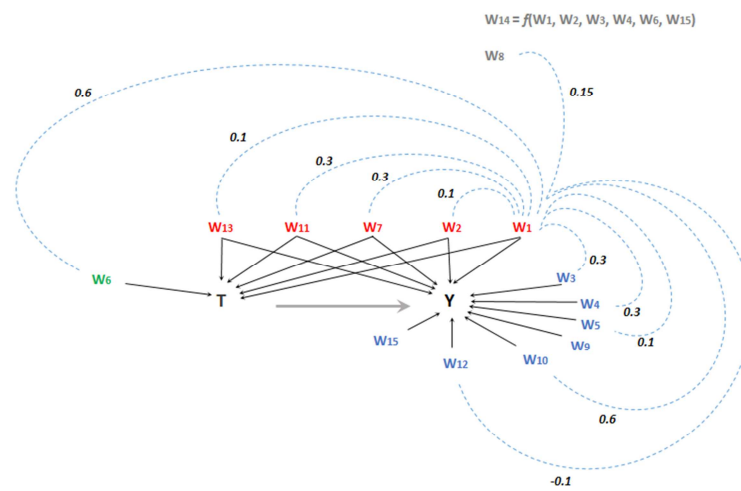
<i>Propensity score analysis with all confounders</i>								
		Odds ratio for the treatment exposure						
		0.17	0.25	0.50	1.00	2.00	4.00	6.00
Odds ratio for the outcome	6.00	24 [14; 88]	25 [14; 73]	25 [15; 65]	24 [15; 68]	25 [15; 63]	24 [15; 74]	24 [14; 80]
	4.00	25 [14; 93]	25 [15; 76]	25 [16; 71]	25 [15; 71]	25 [15; 71]	25 [15; 88]	26 [14; 83]
	2.00	24 [15; 78]	24 [15; 71]	25 [15; 74]	25 [16; 64]	25 [15; 77]	26 [15; 78]	25 [14; 88]
	1.00	25 [15; 76]	25 [15; 71]	25 [15; 73]	25 [16; 61]	25 [15; 66]	25 [14; 85]	24 [14; 79]
	0.50	25 [15; 86]	25 [15; 72]	25 [15; 64]	25 [15; 62]	25 [15; 80]	25 [15; 77]	25 [15; 76]
	0.25	24 [14; 76]	25 [15; 67]	24 [15; 65]	24 [15; 60]	24 [15; 64]	24 [14; 68]	25 [14; 83]
	0.17	25 [15; 81]	25 [15; 72]	25 [15; 67]	24 [15; 76]	24 [14; 71]	24 [14; 74]	24 [14; 79]
<i>Propensity score analysis omitting a confounder</i>								
		Odds ratio for the treatment exposure						
		0.17	0.25	0.50	1.00	2.00	4.00	6.00
Odds ratio for the outcome	6.00	11 [8; 15]	12 [9; 18]	16 [11; 28]	24 [15; 62]	48 [NA; 440]	NA [NA; 1299]	NA [NA; 571]
	4.00	13 [9; 18]	14 [11; 21]	18 [12; 33]	25 [15; 73]	42 [NA; 429]	51 [NA; 921]	NA [NA; 1162]
	2.00	16 [12; 28]	17 [12; 30]	20 [13; 44]	25 [15; 65]	32 [17; 141]	42 [NA; 399]	47 [NA; 593]
	1.00	25 [15; 70]	24 [15; 66]	25 [15; 68]	25 [16; 60]	25 [15; 63]	25 [15; 74]	24 [16; 67]
	0.50	46 [NA; 449]	40 [NA; 415]	32 [18; 142]	25 [15; 63]	20 [13; 45]	17 [12; 30]	16 [12; 27]
	0.25	36 [NA; 972]	60 [NA; 804]	39 [NA; 315]	24 [15; 65]	17 [12; 30]	13 [10; 20]	12 [9; 18]
	0.17	NA [NA; 1036]	36 [NA; 1122]	45 [NA; 427]	25 [15; 76]	16 [11; 30]	12 [9; 18]	11 [8; 15]

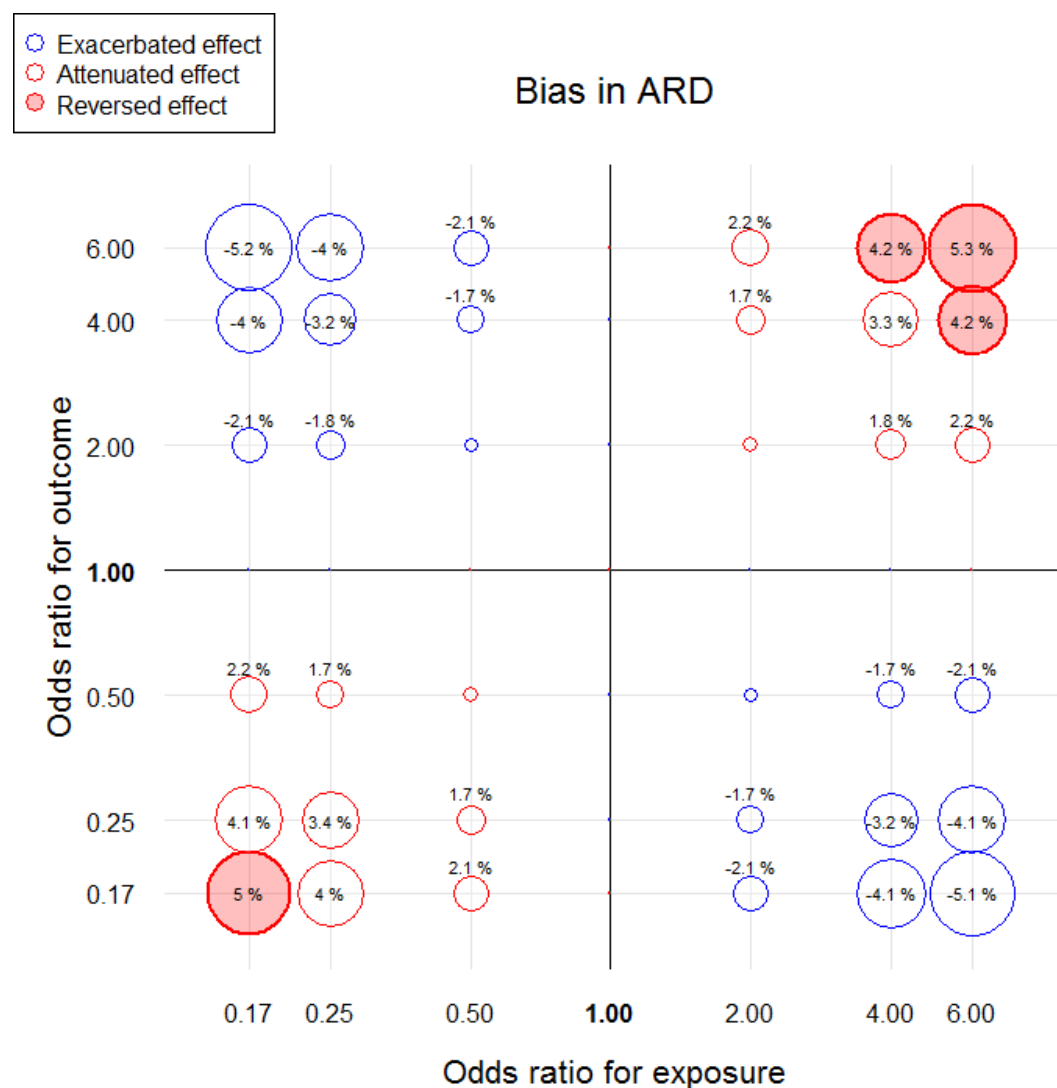
Table 3. Empirical power (%) of analysis, according to the association strength of a confounding variable (W_7) with the outcome and treatment exposure (propensity score analysis including W_7 | propensity score analysis omitting W_7).

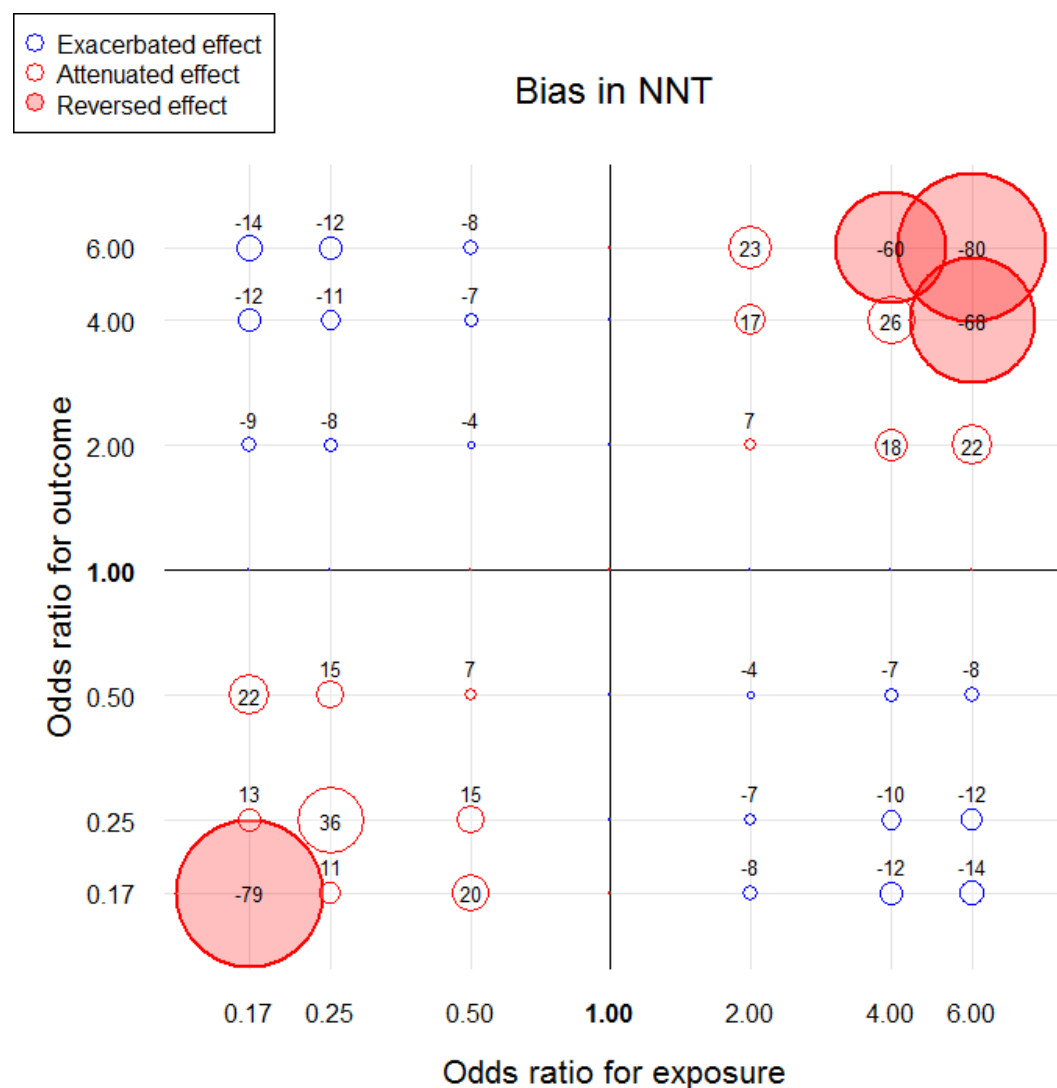
		Odds ratio for the exposure						
		<i>0.17</i>	<i>0.25</i>	<i>0.50</i>	<i>1.00</i>	<i>2.00</i>	<i>4.00</i>	<i>6.00</i>
Odds ratio for the outcome	<i>6.00</i>	75.6 100.0	77.8 100.0	80.4 99.3	86.1 84.5	83.4 24.6	80.1 4.2	78.2 11.1
	<i>4.00</i>	74.2 100.0	76.7 100.0	82.5 99.1	84.2 83.1	83.6 38.3	79.1 8.1	75.2 4.4
	<i>2.00</i>	79.6 99.8	80.8 99.6	83.0 95.7	85.2 85.0	84.5 66.1	78.5 38.6	75.3 27.7
	<i>1.00</i>	78.0 85.0	82.9 86.7	84.4 86.0	85.3 85.2	85.4 86.3	79.1 84.9	77.4 86.3
	<i>0.50</i>	78.2 30.4	81.5 44.5	86.4 67.3	87.7 86.9	81.8 95.4	80.1 99.1	75.0 99.8
	<i>0.25</i>	80.3 6.3	81.0 7.4	85.4 43.8	87.0 86.7	84.6 99.0	80.6 100.0	74.3 100.0
	<i>0.17</i>	77.5 9.5	82.8 5.1	83.9 32.1	84.7 84.8	82.8 99.3	76.9 100.0	73.9 100.0

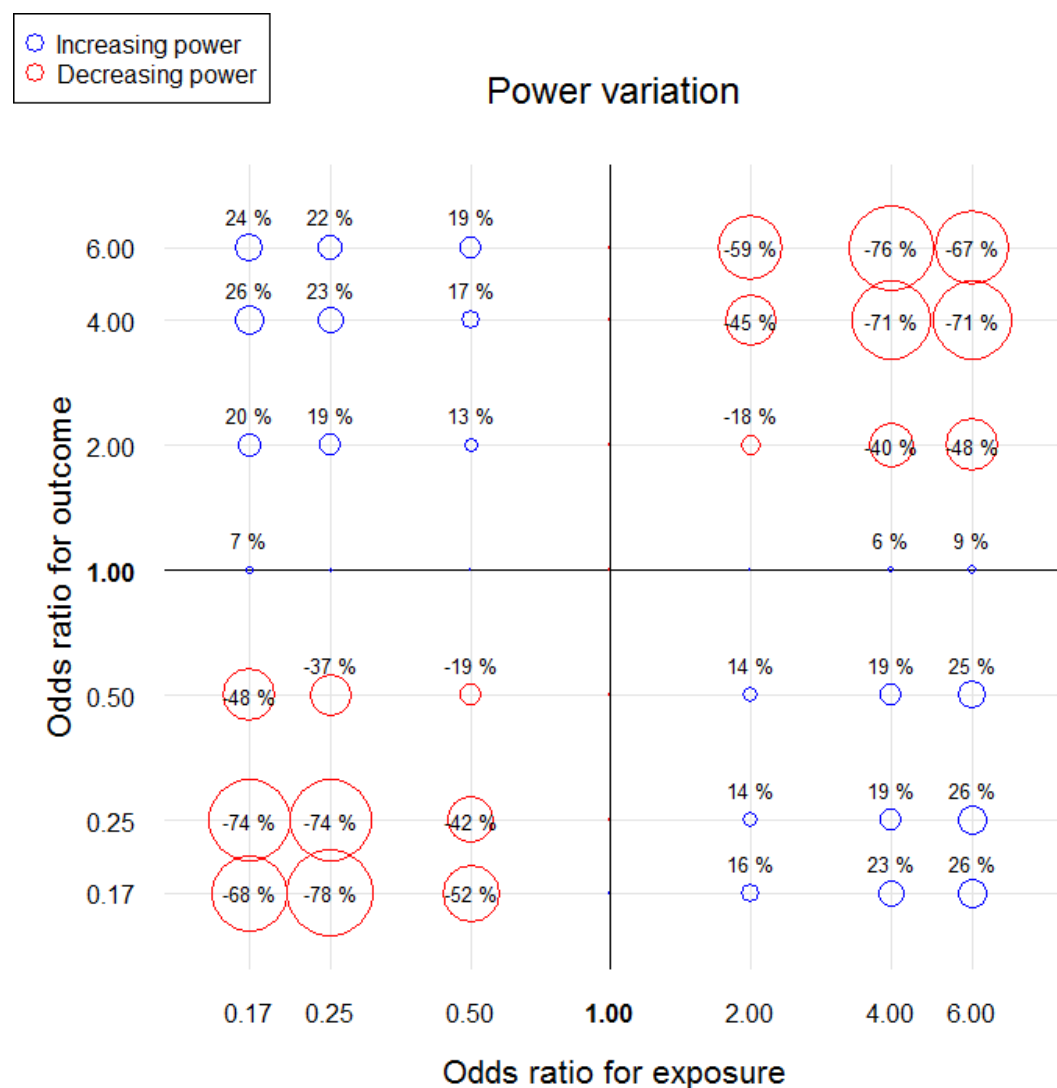
Table 4. Estimated effect of tranexamic acid in a propensity score analysis of the CRASH-2 data. Medians [2.5th; 97.5th percentiles] are reported. The empirical reference absolute risk difference (ARD) is -0.016 and the reference number needed to treat (NNT) is 63. The lower limit of the NNT is not available (NA) for some scenarios as the ARD is greater than zero. In these circumstances, a number needed to harm should be calculated.

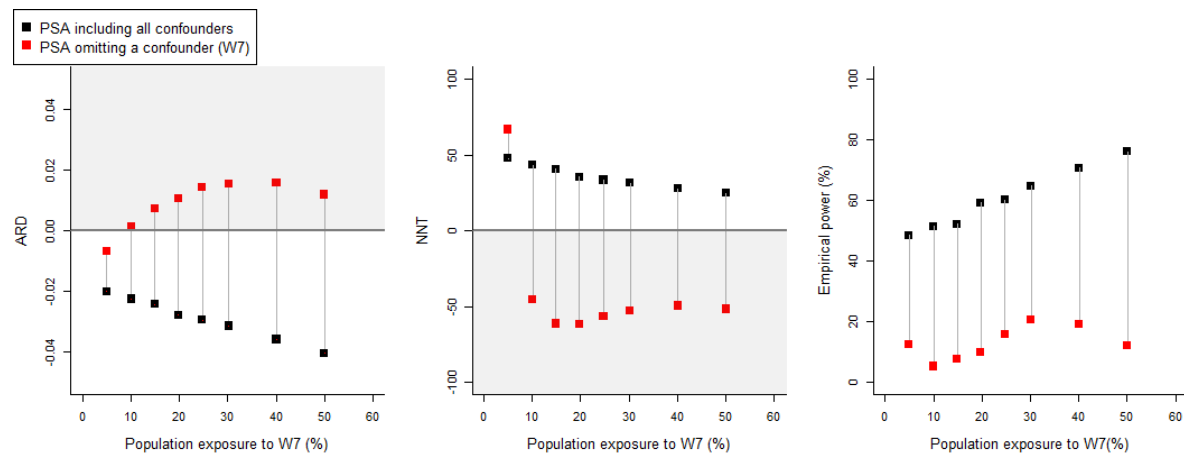
Unmeasured confounder	ARD	NNT	Empirical power (%)
None	-0.016 [-0.035; 0.001]	60 [NA; 489]	40
Middle-income country	-0.017 [-0.035; 0.000]	59 [21; 445]	44
Low-income country	-0.017 [-0.036; -0.001]	56 [25; 409]	48
Age	-0.010 [-0.028; 0.006]	77 [NA; 770]	18
Systolic blood pressure	0.008 [-0.004; 0.020]	NA [NA; 1219]	22











What is new?

- 1) Omitting positive confounders from a propensity score analysis does not greatly change the estimated number needed to treat.
- 2) Omitting negative confounders from a PS analysis can cause much greater changes to the estimated number needed to treat.
- 3) Omitting negative confounders biases the estimator towards the null; the analysis has a decreased power and may miss a statistically significance.

Impact of unmeasured confounders in propensity score analysis: threat to clinical decision making? A Monte Carlo simulation study.

Tri-Long Nguyen, Gary S. Collins, Jessica Spence, Charles Fontaine, Jean-Pierre Daurès, P.J. Devereaux, Paul Landais, Yannick Le Manach.

Appendix

Monte Carlo simulations on fictive data

The treatment (Z) and outcome (Y), both binary, were defined by logistic models (*i.e.* the true PS model and the true outcome model, respectively):

$$\text{logit}[p(Z)] = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_6 W_6 + \beta_7 W_7 + \beta_{11} W_{11} + \beta_{13} W_{13},$$

$$\begin{aligned} \text{logit}[p(Y)] = & \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_7 W_7 + \alpha_9 W_9 + \alpha_{10} W_{10} + \alpha_{11} W_{11} \\ & + \alpha_{12} W_{12} + \alpha_{13} W_{13} + \alpha_{15} W_{15} + \gamma_Z Z. \end{aligned}$$

The treatment was assigned to a subject ($Z = 1$) if $p(Z)$ was greater than a random number from the uniform distribution $U(0,1)$. The subjects experienced the outcome ($Y = 1$) if $p(Y)$ was greater than $u \sim U(0,1)$.

Supplementary Table 1. Variable definitions and coefficients for data generation. The α and β coefficients define the true outcome model and true propensity score model, respectively.

Variable name	Variable type	Related perioperative variable	α coefficient	β coefficient
<i>Intercept</i>			-3.35 to -1.60	-1.40 to 0.34
W ₁	Binary	<i>Coronary artery disease</i>	0.70	0.70
W ₂	Binary	<i>Chronic renal failure</i>	0.64	0.63
W ₃	Binary	<i>Diabetes</i>	0.31	0
W ₄	Binary	<i>Chronic heart failure</i>	0.30	0
W ₅	Binary	<i>Chronic obstructive pulmonary disease</i>	0.43	0
W ₆	Binary	<i>History of stroke</i>	0	0.65
W ₇	Binary	<i>Hypertension</i>	-1.79 to 1.79	-1.79 to 1.79
W ₈	Binary	<i>Obesity</i>	0	0
W ₉	Binary	<i>History of cancer</i>	0.42	0
W ₁₀	Binary	<i>Peripheral vascular disease</i>	0.43	0
W ₁₁	Continuous	<i>Age (per year)</i>	0.01	0.01
W ₁₂	Continuous	<i>Preoperative hemoglobin (per g/dL)</i>	0.01	0
W ₁₃	Continuous	<i>Preoperative eGFR (per mL/min)</i>	-0.02	-0.01
W ₁₄	Ordinal (5 levels)	<i>Revised cardiac risk index (per level)</i>	0	0
W ₁₅	Ordinal (3 levels)	<i>Type of surgical procedure (per level)</i>	0.59	0

Simulations on CRASH-2 data

The outcome was defined by the following logistic model:

$$\begin{aligned} \text{Logit}[p(\text{Mortality})] = & -4.462 + 0.762*MIC + 1.278*LIC + 0.316*AGE - 0.092*AGE^2 \\ & + 0.016*AGE^3 - 0.107*SBP + 0.025*SBP^2 + 0.001*SBP^3 - 0.306*RR + 0.391*RR^2 \\ & - 0.070*RR^3 - 0.032*HR + 0.016*HR^2 - 0.180*GCS + 0.010*GCS^2 - 0.099*PIN \\ & - 0.081*PIN*GCS - 0.014*PIN*GCS^2 - 0.173*TXA. \end{aligned}$$

The treatment allocation was defined by the following logistic model:

$$\begin{aligned} \text{Logit}[p(TXA)] = & -0.5 - 1*MIC - 2*LIC + 2*SBP_{\text{binary}} + 1.5*HR_{\text{binary}} + 1.5*CC_{\text{binary}} + \\ & 0.01*Age. \end{aligned}$$