Author's response to reviews

Title: Statistical power in parallel group point exposure studies with time-to-event outcomes: An empirical comparison of the performance of randomized controlled trials and the inverse probability of treatment weighting (IPTW) approach

Authors:

Peter C Austin (peter.austin@ices.on.ca)
Tibor Schuster (tibor.schuster@mcri.edu.au)
Robert R Platt (robert.platt@mcgill.ca)

Version:3**Date:**13 August 2015

Author's response to reviews: see over

Response to reviewers' comments. Note: Reviewers' comments are in bold type, while the authors' responses are in regular type.

Reviewer 1

The study is a good exercise in using simulation to study statistical power. However, the study results are not generalizable, because of two reasons: firstly, the data structure of this simulation study is not realistic; secondly, the IPTW method used in this study has been proved to be less desirable. The details areas following:

Major compulsory revisions---data structure and methods of the simulation study:

1. In the simulation study, all subjects were "followed until the event was observed". This almost never happens in observational studies. Most observational studies have a certain study time period. At the end of the study, subjects either had events or were censored. Censoring, informative and non-informative, plays an important role in power and sample size consideration.
We agree that censoring (informative or non-informative) plays an important role in power and sample size consideration. However, the objective of the simulations was not to examine the statistical power of each method in isolation. Instead, it was to compare the power of an IPTW design with that of a similarly-structured RCT. Thus, the important issue was to compare the two designs when there was an equal amount of censoring (or an equal number of observed events). For computational simplicity, we elected to design our simulations that censoring was not present. We can see no rationale for believing that the relative differences in statistical power would be affected by the degree of censoring. In the limitation paragraph of the Discussion, we highlight this as a limitation of the study (page 21, lines 434-442). However, we suggest that the effect on the conclusions should be minimal. In the revised version of the manuscript, we examined 585 different scenarios. Allowing the degree of censoring to be another factor in the design of the Monte Carlo simulations would increase substantially the computational burden of the simulations and increase the quantity of results that would require reporting. For example, if the degree of censoring was allowed to take on three different values, we would then examine and report on 1,755 different scenarios.

2. In the simulation study, all variables are positively associated with both treatment assignment and outcome, which hardly occurs in real observational studies. With all covariate coefficients (alphas and betas) being positive, including treatment effect, the bias in Table 1 is artificially high. Even without the simulation study, one would have known that the more confounding, the more bias. On the other hand, in observational studies, there are variables that positively associated with treatment but negatively associated with outcome and vice versa. So, given the same c-statistic on the treatment assignment model, different covariate coefficients may result in different bias levels, and in turns, lead to different power and type I error.
Thank you for this suggestion. We have modified the data-generating process so that two of the variables have a negative effect on treatment-selection while having a positive effect on outcomes (page 10, line 196; page 11, line 218 and lines 221-224). The point of Table is not to

demonstrate that with more confounding there is more bias. Rather, it was to illustrate the degree of bias that was present in the observational study in which IPTW was being used.

3. The authors used a dated IPTW method, with normal weight and without trimming. There are two types of IPTW, the normal weight and the stabilized weight. It has been shown that the stabilized weight is more efficient (Hernan and Robins, 2004), because it is less extreme than the normal weight. Imbens and Rubin (2013) also showed that weight trimming reduces bias in the treatment effect estimator. So the results from this study do not apply where more advance methods are used.
Thank you for this comment. We have revised the simulations and analysis methods to use stabilized weights (page 12, line 245 to page 13, line 249).

Major compulsory revisions---study results:

4. Figures 2 to 5 show that IPTW has higher power than RCT in some situations. Is there a reason that observational study can have more power in RCT?
We have added a paragraph providing several potential explanations for these observations (page 19, starting at line 395).

5. Figure 6 states the obvious: 1) observational studies have lower power, 2) statistical power increases with the sample size balance of two treatment groups, 3) statistical power increases with the true hazard ratio. These are well-established facts, and do not need to be reinforced by simulation studies.
The purpose of these figures was not to illustrate these well-established factors. Instead, the purpose was to illustrate that, across a wide range of scenarios, the RCT design had greater statistical power than an IPTW analysis of a similar study. The reviewer asserts that IPTW has been previously described as inefficient. However, most of these comparisons were made with respect to other novel estimators, such as doubly robust methods or in the context of using IPTW to account for missing data (e.g., Seaman, Shaun R et al. "Combining Multiple Imputation and Inverse-Probability Weighting." *Biometrics* 68.1 (2012): 129–137.; http://www.ipc-undp.org/evaluation/aula7-ps_outros/Basu%20et%20al%20-%20Propensity%20Scores%20in%20Non-Linear%20Response%20Models.pdf). However, our simulation study was intended to clarify potential disadvantages of using IPTW methods in situations where (likely more cost-intensive) RCT designs could be implemented. Contrary to the first assertion of the reviewer, it is not obvious that an IPTW analysis of an observational study has lower power than an analysis of a similar RCT. Furthermore, given the popularity of conventional IPTW methods, in which power calculations are not feasible, we wanted to assess whether the power of a similarly-structured RCT provided an estimate of the power of the corresponding observational study.

6. The ANOVA model also states the obvious. The statistical power (dependent variable in the ANOVA model) is generated under various simulation parameters (independent variable in the ANOVA model). It is not surprising that a saturated model with main and interactive effects of simulation parameters interprets 97% variance in the power. This model does not provide helpful information in addition to Figures 2 to 5.
We have deleted these analyses from the revised manuscript.

7. The statement is misleading that the ANOVA model helps predicting the statistical power for a study using IPTW. The ANOVA model is constructed under many assumptions of the data structure. It does not apply to any study where the covariates have different associations with treatment assignment and outcome, or censoring occurs.
The ANOVA model and analyses have been removed from the revised manuscript.

8. The conclusions about statistical power and type I error are misleading, too. Again, the conclusions do not apply to other studies with different data structure. This study is a good exercise. But simulation 780 scenarios does not help the power analysis in any real study. In real studies, the confounders and their association with treatment and outcome, the censoring mechanism, the rate of events can be different. The combinations of these parameters are countless. Power analysis should be specifically performed for each study with its unique condition not generalized from some simulation study run on absolutely different conditions.
We have removed these analyses from the revised manuscript, in keeping with our focus on differences in statistical power between the two study designs.

Reviewer 2

Major compulsory revisions

1) Line 177, 'The simulations in the current study were designed to examine the impact of the following four factors'. For any study, estimating power requires that the investigator have a good understanding of the variance of the estimator being applied. While the simulations provided give a rough picture, it leaves out some important details. For example, the survival curve itself will have a strong effect on variance, as will the distribution for the conditional probabilities of treatment (which is slightly addressed with the variants in strength of the treatment-selection process). Simulations and results should be updated to reflect that. In particular, it would be of value to vary the outcome model such that different survival curves will result. Summaries of the treatment selection model or propensity scores should also be included. Values of alpha_treat, alpha_AUC, and beta_treat should be either stated or summarized if too numerous to cite.
The revised version of the manuscript examined 585 different scenarios. Adding a new factor (the outcome model) would substantially increase the number of scenarios. This would increase the computational burden of the simulations and make the results more difficult to present. Given the large number of scenarios that we have examined, we have not incorporated new factors in the Monte Carlo simulations. The value of alpha_AUC is now explicitly described by a formula described in the given reference (page 11, line 208). We have also summarized the values of alpha_treat and beta_treat in the revised manuscript (page 11, line 212; page 12, lines 224-225).

2) Line 195, 'For each subject, the probability of treatment selection was determined from the following logistic model'. These authors do not address positivity violations or near violations. Their treatment model parameter estimates are chosen such that this potential issue is minimized, if even present. As this is a large cause of exploding estimator variance,

the reviewer recommends that they provide scenarios concerning this. As in practice violations or near violations are likely to occur, these authors' study provides, at most, an optimistic situation that can only be hoped for.
Thank you for this interesting comment. One of the assumptions for causal research using IPTW methods is that the positivity assumption has been satisfied (i.e., that all subjects have a non-zero probability of receiving each of the active treatment and the control treatment). A setting in which the positivity assumption was violated would require special attention and careful thought about how to proceed. We are reluctant to incorporate scenarios in which the positivity assumption is violated, as this would be a setting in which one would be wary about employing IPTW methods. The positivity assumption must be satisfied in RCTs, otherwise random treatment allocation would be contradictory (one could not assign a patient to a treatment for which they had a zero probability of receiving it (e.g., due to contraindications)). Since the RCT design was the main comparator in our simulation study, validity of the positivity assumption is well-justified throughout all considered simulation settings.

3) Line 272, for simulating RCT data, a 'survival time was generated for each subjects which is affected by both treatment and a subset of the baseline covariates'; it is fine to be affected by treatment; however, it might not be fine to be affected by covariates, as the covariates impact may not be balanced between the two group in certain simulation scenarios per the manuscript. But, in RCT, due to randomization, those baseline covariates' distributions should be balanced in theory, and most likely so in reality if the trial size is large, and therefore baseline covariates should not have impact on survival time in the simulations for this research. The simulation setting may violate this principle and thus the conclusions from the related scenarios might not be correct.
We disagree with the reviewer's comment that in an RCT baseline covariates should not have an impact on survival time. Even if baseline covariates are perfectly balanced across treatment groups in an RCT, one would still expect that baseline covariates would influence the outcome. For instance, in a large RCT examining the effect of statin lipid-lowering therapy on survival in patients with a heart attack (acute myocardial infarction), one would anticipate that older patients would have worse survival, while younger patients would have improved survival. Even if patient age was perfectly balanced across the arms of the trial, age remains a prognostically-important variable in terms of post-heart attack survival. Many RCTs routinely publish the results of an analysis examining the effects of baseline covariates on the study outcome – indicating that treatment is not the only variable that influences the outcome.

4) Line 292, 'simulation data in which subjects were not subject to censoring'. Since without censoring, the survival analysis results and results from proportion based endpoints may be essentially the same, then it might not be needed to use survival analysis. That said given no censoring, the RCT and IPTW observational study data can be analyzed by proportion based statistics, such as CMH test, etc., for which it might be possible to study the issue through closed formula, rather than via Monte Carlo simulations. If survival analysis is to be used, then, for both RCT and observational studies, possible informative censoring and time dependant confounding issue should be considered for this research.
We have discussed a similar comment made by the first reviewer (see above).

Minor essential revisions

1) Line 99, 'In an IPTW analysis, subjects are weighted by the inverse of the probability of receiving the treatment that was actually received'. IPTW can be conducted with a static intervention, where the investigator only considers the probability of a specified treatment, as opposed to the actual treatment received. The reviewer suggests changing this to avoid confusion, perhaps to instead say 'In a typical MSM analysis'.
The current study does not consider a typical MSM (marginal structural models) analysis. MSMs are employed in settings with time-varying exposures and time-varying confounding variables. We have modified the text to clarify this issue (page 6, line 98).

2) Line 100, 'In this synthetic, weighted sample, treatment assignment is unconfounded'. The authors should update this to say that treatment is unconfounded in a correctly specified model, as it is one of the assumptions of the IPTW approach.
The text has been modified to address this comment (page 6, lines 100-102).

3) Line 237, 'The inverse probability of treatment weights (IPTWs) are defined as'. The authors fail to consider the use of stabilized weights, which can sometimes (highly) affect the variance of estimation. Doing so should result in weights of 1 for the observational approach in the absence of confounding, corresponding to equivalent results with the RCT approach.
We have revised the manuscript to use stabilized weights, as suggested by both reviewers (page 12, line 245 to page 13, line 249).

4) Line 290, 'The reason for this choice is that statistical power in an RCT is related to the number of observed events'. Reviewer suggest authors update this to say 'statistical power in survival analysis in general', in order to avoid confusion.
The manuscript has been revised to address this suggestion (page 15, lines 298-299).

5) Line 301, 'From the table, one observes that the magnitude of the bias in the estimated crude hazard ratio in the observational data increased with the c-statistic of the treatment-selection model'. It is suggested that the authors use standardize weights, as this result should be mitigated with its use.
We have revised the methods of the paper to use stabilized weights (page 12, line 245 to page 13, line 249).