Substantially inflated type I error rates if propensity score method is not fixed in advance

Markus Neuhäuser[1*], Julia M. Kraechter[1], Matthias Thielmann[2] and Graeme D. Ruxton[3]

1: RheinAhrCampus, Koblenz University of Applied Sciences, Joseph-Rovan-Allee 2, 53424 Remagen, Germany, Tel. +49 2642 932417, Fax +49 2642 932399, e-mail: neuhaeuser@rheinahrcampus.de

2: Dept. of Thoracic and Cardiovascular Surgery, West German Heart Centre, University Hospital Essen, Essen, Germany

3: School of Biology, University of St Andrews, St Andrews, Scotland, UK

*: corresponding author

Abstract

Propensity scores are often used to adjust for between-group variation in covariates, when individuals cannot be randomised to groups. There is great flexibility in how these scores can be appropriately used. This flexibility might encourage p-value hacking – where several alternative uses of propensity scores are explored and the one yielding the lowest p-value is selectively reported. Such unreported multiple testing must inevitably inflate type I error rates – our focus is on exploring how strong this inflation effect might be. Across three different scenarios, we compared the performance of four different methods. Each taken individually gave type I error rates near the nominal (5%) value, but taking the minimum value of four tests led to actual error rates between 150% and 200% of the nominal value. Hence, we strongly recommend pre-selection of the details of the statistical treatment of propensity scores to avoid risk of very serious over-inflation of type I error rates.

Introduction

In non-interventional and other observational studies, treatments cannot be randomly assigned. As a consequence, groups usually differ in some baseline covariates. In order to adjust for between-group differences in observational studies, statistical methods based on the propensity score have become increasingly popular; see for instance a study about abdominal aortic aneurysm repair [1]. The propensity score is the conditional probability to receive a particular treatment given the observed baseline covariates, see e.g. D'Agostino [2] or Benedetto et al. [3] for more details.

Common techniques using the propensity score are matching, stratification, regression adjustment and inverse probability weighting [2, 3]. According to a review [4], based on studies published in 2011 and 2012, matching was the most commonly applied method: used in 68.9% of studies. Regression adjustment (20.9%), stratification (13.6%) and inverse probability weighting (7.1%) were less often carried out. As the four percentages show, sometimes more than one method is applied in one study. Moreover, even for a single method several different options are available (and used in applications). For instance, when a stratification is applied different numbers of strata might be used [5]. For inverse probability weighting trimming large weights might be useful, but, ”without guidance on the optimal level of trimming, there exists the dangerous potential for trimming being used to artificially achieve a desired result” [6].

For clinical trials the study protocol, including the description of the planned statistical methods, has to be submitted to ethics committees, institutional review boards and/or regulatory authorities before the start of the study. In addition, details of the study, again including some description of the planned statistical methods, are recorded in advance in clinical trial registries. Thus, the statistical analysis is pre-planned and cannot be changed after data are available. For observational studies this is usually not the case. Thus, in the majority of analyses applying propensity scores details are not fixed in advance and it cannot be excluded that some p-hacking happens in some applications.

By “p-hacking” we mean the performance of several alternative statistical tests and the selective reporting of the one yielding the smallest p-value. Of course, such a practice is not acceptable from a statistical point of view and consequently not scientifically sound. Without any adjustment for multiple testing, the error probabilities are not controllable. The aim of this note is to investigate how much the type I error rate is inflated when p-hacking is applied with different propensity score methods. That is, it is clear that p-hacking must inflate type I error rates, our interest is in exploring how strong this effect can be.

Although p-hacking, also known as inflation bias, is difficult to detect [7], quantifying p-hacking is important. Head et al. [7] present empirical evidence that p-hacking is widespread throughout science. However, while p-hacking is probably common, the study of Head et al. [7] suggests that its effect is weak relative to the real effect sizes.

Material and Methods

As mentioned above, there are four common techniques using the propensity score: matching, stratification, regression adjustment and inverse probability weighting. In a simulation study performed with R (version 3.4.0) we applied one variant of each of the four common techniques. We selected variants that individually have type-I error rates near the nominal level (see Tab. 1). To be precise, we used the following variants:

- Stratification based on propensity scores with ten strata (values of both groups combined were used to define approximately equally-sized strata).
- Nearest neighbour 1:1 matching with replacement
- Regression adjustment

- Inverse probability of treatment weighting (IPTW) with stabilized weights and truncation of the largest 1% of weights [6]

In addition to these four methods, the minimum of the p-values of the four methods was used to imitate p-hacking.

Our first simulation was carried out as described by Austin [8], however, with nine covariates X_1 to X_9 in total. The simulated covariates have a multinomial normal distribution with correlation ρ which ranges from 0 to 1 by 0.2. The first six covariates were used to compute the propensity score as follows:

$$p_{treat} = \frac{\exp\,(0.1X_1+0.2X_2+0.3X_3+0.4X_4+0.5X_5+0.6X_6)}{1+\exp\,(0.1X_1+0.2X_2+0.3X_3+0.4X_4+0.5X_5+0.6X_6)}\ .$$

The treatment group was simulated according to a Bernoulli distribution with probability p_{treat}. Three of the six first covariates plus three additional covariates were used to simulate a binary outcome. To be precise the probability p_{out} was computed as

$$p_{out} = \frac{\exp\,(0.4X_1+0.1X_2+0.5X_3+0.3X_7+0.2X_8+0.6X_9)}{1+\exp\,(0.4X_1+0.1X_2+0.5X_3+0.3X_7+0.2X_8+0.6X_9)}\ .$$

Then, the binary outcome was simulated according to a Bernoulli distribution with probability p_{out}. Note that the treatment group does not influence p_{out} because we consider the null hypothesis that there is no difference between the two treatment groups. To analyse the outcome, logistic regression was used with a nominal significance level of $\alpha = 5\%$. For the stratification, a conditional logistic regression model was applied.

For the second simulation the scenario used by Craycroft [9] was utilized. Here, there are three standard normally distributed covariates X_1 to X_3 and two binary covariates X_4 and X_5, both with a success probability of 0.5. Three covariates were used to compute the propensity score:

$$p_{treat} = \frac{\exp{(0.5+X_1+X_3+X_5)}}{1+\exp{(0.5+X_1+X_3+X_5)}} .$$

The probability p_{out} was computed as

$$p_{out} = \frac{\exp{(-1+X_2+X_3+X_4+X_5)}}{1+\exp{(-1+X_2+X_3+X_4+X_5)}} .$$

Thus, one covariate (X_1) influences the treatment allocation only, two covariates (X_2, X_4) influence the binary outcome only, and two further covariates (X_3, X_5) influence both treatment allocation and outcome.

A third simulation is identical to the first simulation with the exception that the simulated outcome was normally distributed. To be precise, the outcome was simulated as $5\,\mathrm{N}(p_{out}+1, p_{out}/2)$. Instead of the logistic regression a linear regression was applied. For the stratification the factor stratum was included in a resulting analysis of covariance model.

For all simulations the correlation between the covariates ranges from 0 to 1 by 0.2. For each scenario, 10000 simulation runs were used to estimate the actual type I error rate. The sample size was 1000 per study for all three simulation models. The R code used for our simulation is available at www.hs-koblenz.de/profilepages/neuhaeuser/programme. When actually performing the propensity score analysis, the covariates and estimated propensity scores could be used. In the R code provided by Schuler [10] the observed covariates are included in the model for matching and IPTW. Here, to harmonize models and to reduce the number of variables in the model the propensity scores are included in the model as opposed to the observed covariates (for matching, IPTW, and, of course, regression adjustment).

In addition to the simulation we consider, as an application, a study investigating patients with diabetes mellitus and triple-vessel disease undergoing coronary artery bypass surgery [11]. In one group (n_1 = 621) the bypass surgery was the primary revascularization procedure, in the other group (n_2 = 128) patients were treated with a previous percutaneous coronary intervention (PCI) before the bypass surgery. Hence, the aim was to determine whether previous PCI has a prognostic impact. The two binary outcome variables are death and occurrence of major adverse cardiac events (MACE), both determined in hospital during index hospitalization. The propensity score was computed using a logistic regression based on 12 covariates [11]. Differences between these covariates disappear when testing in a stratified analysis with ten strata based on the propensity score [see also 5].

Results

Table 1 presents the simulation results. The single methods each have acceptable type I error rates close to the desired nominal significance level α = 5%. In contrast, the simulated p-hacking strategy to select the minimum p-value from the four different methods has unacceptably high actual type I error levels of 7 to 10%. Even in cases where single methods are conservative the minimum p-value strategy has an inflated actual level of approx. 7%.

In order to evaluate the extent of type I error rate inflation, Bradley's [12] liberal criterion is used. Based on this criterion, an actual type I error rate between 0.5α and 1.5α is considered as acceptable. Bradley's liberal criterion has been applied in recent investigations [see e.g. 13, 14]. According to this criterion all four single methods are acceptable, but the minimum p-value's actual type I error rate is, in the majority of situations, outside the limits set by Bradley's liberal criterion (Tab. 1).

When analysing the example study [11], the p-values displayed in Table 2 occurred. Although all p-values are smaller than 0.05, there is a substantial variability in the p-values. For the outcome variable death the largest p-value is 2.3 times larger than the smallest. For MACE this factor is 3.0.

Discussion

In our small simulation study we investigated four common methods using the propensity score. These methods were applied among others by Wendt et al. [15, stratification with ten strata], Lee et al. [16, nearest neighbour 1:1 matching], Doll et al. [17, regression adjustment], and Rosenbloom et al. [18, IPTW with stabilized weights].

Although we performed only four different methods, we could show that the actual type I error rate of the strategy to select the minimum p-value is inflated and unacceptably high, even according to Bradley's liberal criterion. In reality there is much more flexibility available to the data analyst than our study explored. On the one hand, there is much more variety of methods, for each method a suite of modifications are possible. For instance, the number of strata can vary when a stratification is applied, or for the regression adjustment several different regression adjustment models are available. Further, in a real study there is some flexibility and arbitrariness in selection of the covariables used to compute the propensity score. The review of Sanni Ali et al. [4] showed that the execution and reporting of covariate selection is far from optimal.

In the majority of applications balance diagnostics for examining whether the propensity score model has been adequately specified, is applied after fitting a propensity score [4, 19]. This covariate balance was checked and reported in 59.8% of studies included in the review [4] mentioned above. If the desired level of balance is not achieved then the propensity score estimation model is adjusted. As long as the outcomes are not incorporated prior to revising the propensity score model, this will not inflate the type I error rate.

Nevertheless, if the effect on the outcome of interest is considered in covariate selection, the effect of inflated actual type I error rates might be larger in real applications than observed in our simulation study. However, even our approach using just four single methods could demonstrate that the actual type I error rate of the minimum p-value is unacceptably high (according to Bradley's liberal criterion).

Due to the enlarged actual type I error rates the strategy of p-hacking leads to false-positive results and, therefore, can contribute to the reproducibility crisis where scientific studies are impossible to reproduce or replicate. What can be done? On the one hand, the statistical analysis should be planned and documented in advance (including the fine detail of how propensity scores will be calculated and how they will used, including how covariate balance is checked and how the model is adjusted subsequently). Further, data sharing can facilitate exploration of how robust results are to variation in the choices made in the statistical analysis.

Acknowledgement

We would like to thank an anonymous reviewer for helpful comments that substantially improved the manuscript.

Conflict of interest

There was no funding for this study and the authors declare no conflict of interest.

References

1. Piffaretti G, Mariscalco G, Riva F, et al. Abdominal aortic aneurysm repair: long-term follow-up of endovascular versus open repair. Arch Med Sci 2014; 10: 273-82.
2. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998; 17: 2265-81.

3. Benedetto U, Head SJ, Angelini GD, Blackstone H. Statistical primer: propensity score matching and its alternatives. Eur J Cardio-Thorac Surg 2018; 53: 1112-1117.
4. Sanni Ali M, Groenwold RHH, Belitzer SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol 2015; 68: 112-131.
5. Neuhäuser M, Thielmann M, Ruxton GD. The number of strata in propensity score stratification for a binary outcome. Arch Med Sci 2018; 14: 695-700.
6. Lee BK, Lessler J, Stuart EA. Weight Trimming and Propensity Score Weighting. PLOS ONE 2011; 6(3): e18174.
7. Head ML, Holman L, Lanfear R, et al. The extent and consequences of p-hacking in science. PLOS Biology 2015; 13(3): e1002106.
8. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Medical Decision Making 2009; 29: 661-77.
9. Craycroft J. Propensity score methods: a simulation and case study involving breast cancer patients. Master thesis, University of Louisville, 2016.
10. Schuler M. Overview of implementing propensity score analyses in statistical software. In: Pan W & Bai H eds. Propensity score analysis. The Guilford Press, New York, 2015. Pp. 20-46.
11. Thielmann M, Neuhäuser M, Knipp S, et al. Prognostic impact of previous percutaneous coronary intervention in patients with diabetes mellitus and triple-vessel disease undergoing coronary artery bypass surgery. Journal of Thoracic and Cardiovascular Surgery 2007; 134, 470-476.
12. Bradley JV. Robustness? Br J Math Stat Psychol. 1978; 31, 144-152.
13. Nguyen DT, Kim ES, de Gil PR et al. Parametric tests fort wo population menas under normal and non-normal distributions. J Mod Appl Stat Methods 2016; 16, 141-159.

14. Welz A, Ruxton GD, Neuhäuser M. A non-parametric maximum test for the Behrens-Fisher problem. Journal of Statistical Computation and Simulation 2018; 88, 1336-1347.

15. Wendt D, Kahlert, P, Lenze, T, et al. Management of high-risk patients with aortic stenosis and coronary artery disease. Annals of Thoracic Surgery 2013; 95, 599-605

16. Lee SW, Kwon JH, Lee HL, et al. Comparison of tenofovir and entecavir on the risk of hepatocellular carcinoma and mortality in treatment-naïve patients with chronic hepatitis B in Korea: a large-scale, propensity score analysis. Gut (published online first: 31 October 2019. doi: 10.1136/gutjnl-2019-318947).

17. Doll JA, Kaltenbach LA, Anstrom KJ, et al. Impact of copayment reduction intervention on medical persistence and cardiovascular events in hospitals with and without prior medication financial assistance programs. J Am Heart Assoc 2020; e014975.

18. Rosenbloom JI, Rhoades JS, Woolfolk CL, et al. Prostaglandins and caesarean delivery for nonreassuring fetal status in patients delivering small-for-gestational age neonates at term. J Matern Fetal Neonatal Med 2019; 24: 1-7.

19. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research 2011; 46: 399-424.

Tab. 1: Simulated actual type I error rates for the three different simulation models and different correlation coefficients ρ for the nominal significance level α = 0.05

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 1$
Simulation 1						
Stratification	0.050	0.050	0.051	0.048	0.050	0.050
Nearest neighbour matching	0.048	0.050	0.051	0.052	0.055	0.052
Regression adjustment	0.049	0.051	0.051	0.048	0.050	0.052
IPTW	0.048	0.050	0.050	0.048	0.050	0.052
Minimum p-value	0.077	0.079	0.082	0.080	0.083	0.082
Simulation 2						
Stratification	0.054	0.054	0.051	0.049	0.052	0.051
Nearest neighbour matching	0.049	0.048	0.039	0.037	0.039	0.033
Regression adjustment	0.054	0.055	0.049	0.048	0.049	0.047
IPTW	0.054	0.055	0.049	0.048	0.049	0.047
Minimum p-value	0.079	0.083	0.074	0.071	0.072	0.072
Simulation 3						
Stratification	0.051	0.050	0.051	0.049	0.055	0.054
Nearest neighbour matching	0.050	0.055	0.053	0.054	0.064	0.064
Regression adjustment	0.051	0.050	0.049	0.051	0.060	0.058
IPTW	0.050	0.049	0.048	0.050	0.059	0.057
Minimum p-value	0.078	0.082	0.081	0.082	0.094	0.095

Tab. 2: p-values of the different propensity score methods based on the data of Thielmann et al. [11]

Method	Outcome death	Outcome MACE
Stratification	0.0204	0.0157
Nearest neighbour matching	0.0471	0.0475
Regression adjustment	0.0277	0.0260
IPTW	0.0278	0.0273