

Research Article

Corresponding Author:

Dr. Francesco Cottone, Italian Group for Adult Hematologic Diseases (GIMEMA)

Data Center and Health Outcomes Research Unit, Via Benevento 6, Rome, 00161, Italy. Email:

f.cottone@gimema.it; Orcid: 0000-0001-6240-8317.

Propensity score methods and regression adjustment for analysis of non-randomized studies with health-related quality of life outcomes

PS and MLR for non-randomized HRQoL studies

Francesco Cottone¹, Amelie Anota², Franck Bonnetain², Gary S. Collins³ and Fabio Efficace¹

¹Italian Group for Adult Hematologic Diseases (GIMEMA) Data Center and Health Outcomes Research Unit, Rome, Italy

²Quality of Life in oncology clinical research Platform, France;
University Hospital of Besançon, Methodology and Quality of Life in Oncology Unit, Besançon, France

³Centre for Statistics in Medicine, University of Oxford, Oxford, UK

Key points:

- In non-randomized studies with Health-Related Quality of Life outcomes (HRQoL), the combined use of propensity score (PS) based methods and multivariable linear regression (MLR) adjustment might provide more reliable estimates of the average treatment effect on the treated, than either methods used separately.
- Further MLR adjustment on PS-balanced data can account for the possible confounding impact on outcome, of additional post-treatment variables not usable for propensity score estimation.
- The conjoint use of PS methods and MLR might be useful in those HRQoL studies where small sample size and/or poor overlap do not allow to form well-balanced groups by using any PS-based technique.

Word count excluding abstract, tables, figures and references:

2993 words.

Abstract

Purpose

The aim of this study was to investigate the potential added value of combining propensity score (PS) methods with multivariable linear regression (MLR) in estimating the average treatment effect on the treated (ATT) in non-randomized studies with health related quality of life (HRQoL) outcomes.

Methods

We first used simulations to compare the performances of different PS-based methods, either alone or in combination with further MLR adjustment, in estimating ATT. PS-methods were, respectively, optimal pair (OPM) and full (OFM) PS matching, sub-classification on the PS (SBC) and the Inverse Probability of Treatment Weighting (IPTW). We simulated several scenarios, according to different sample sizes, proportions of treated vs untreated subjects and types of HRQoL outcomes. We also applied the same methods to a real clinical data set.

Results

OPM and IPTW provided the closest Type I error to the nominal threshold $\alpha=0.05$ across all scenarios. Overall, both methods showed also lower variability in estimates than SBC and OFM. SBC performed worst, generally providing the highest levels of bias. Further MLR adjustment lessened bias for all methods, however providing higher Type I error for SBC and OFM. In the real case, all methods provided similar ATT estimates except for one outcome.

Conclusions

Our findings suggest that for sample sizes up to $n=200$, OPM and IPTW are to be preferred to OFM and SBC in estimating ATT on HRQoL outcomes. Specifically, OPM performed best in sample sizes of $n \geq 80$, IPTW for smaller sample sizes. Additional MLR adjustment can further improve ATT estimates.

Keywords

Health related quality of life; Multivariable linear regression; Non-randomized studies; Propensity score-based methods; Summative scale outcomes.

1. Introduction

The importance of health-related quality of life (HRQoL) outcomes in medical research is now critical to better inform patient care and to facilitate clinical decision-makings.. Cancer clinical trials, for example, now frequently include HRQoL as an outcome (either primary or secondary) along with more traditional clinical or laboratory ¹. In studies including a HRQoL evaluation, an important aspect is often the estimate of the average treatment effect on the treated (ATT) on patients' self-reported health status and/or symptoms. ATT is the expected difference between observed outcomes of treated subjects and those they would experience had they not been treated (unobserved)[†]. The estimate of ATT can be impaired by systematic imbalances in individuals' characteristics², which might have affected the probability of treatment allocation and/or the outcome of interest. Multivariable linear regression (MLR) and PS-based methods ³ are commonly used to deal with this issue in HRQoL research. MLR is typically used to estimate ATT by a two-steps process, i.e. first estimating the regression coefficients of observed covariates in the untreated group, then combining these estimates with the observed values of covariates in treated individuals to predict their unobserved outcomes. The mean difference between observed and predicted outcomes provides an estimate of ATT.

A PS is the probability of being assigned to a treatment group, conditional to a set of observed pre-treatment variables (covariates). PS-based methods help in balancing pre-treatment differences in the observed covariates between groups, based on the estimated propensity scores. Groups can then be compared to estimate ATT, as they would be in a randomized study. Frequently used PS-based methods are the propensity score matching (PSM), sub-classification (or stratification) on the

[†] ATT differs from the average treatment effect (ATE), which is the expected effect of the treatment across all subjects in the population of interest. When assignment to treatment is randomized, then $ATT=ATE$, as the potential outcomes of all subjects are independent from treatment assignment. In non-randomized settings, this assumption is likely to be violated implying $ATT \neq ATE$.

propensity score (SBC), inverse probability of treatment weighting (IPTW) and the use of estimated PS as an adjustment covariate in a linear regression model. Both MLR and PS-based approaches are able to assess only the overt component of bias in the ATT estimate, i.e. the one stemming from differences in the distribution of observed covariates between the treated and untreated groups⁴. Therefore, the impact of unobserved confounding on ATT estimates should be assessed in applied research. MLR estimates are highly sensitive to different modeling assumptions, functional forms and model specification⁵. Also, in HRQoL studies the outcomes are typically based on summative scales⁶ with skewed and clustered distributions, far from being normally distributed⁷. The prediction of outcomes of the treated subjects, based on the MLR coefficients estimated on the untreated individuals, can be rather inaccurate and provide biased ATT estimates. Indeed, the use of MLR is an issue with HRQoL outcomes^{8,9}.

Propensity score methods can help in reducing some of these problems posed by MLR. When applied to PS-balanced data, the estimates of MLR coefficients in the untreated group would benefit from a covariate distribution closer to that of treated subjects, as this would reduce the bias of outcome predictions in the latter group due to differences in the distribution of covariates. Previous PS-based balance of groups would also allow to correctly estimate ATT as the regression coefficient of a treatment status indicator in a linear regression model performed on the whole sample. That is, MLR could be performed even in the case of insufficient sample size for the implementation of the two-steps MLR approach.

PS-based methods can also benefit from further MLR adjustment on previously PS-balanced data^{10,11}, as it might further lessen the bias in ATT estimates due to residual imbalance in the observed covariates¹²⁻¹⁴. Furthermore, under specific assumptions¹⁵ this allows to account for post-treatment observed variables potentially affecting the outcome, which could not be used to estimate the propensity score. A specific limitation of matching methods in estimating ATT, is that it might bound estimates to a subset of treated subjects, if part of these do not find a match in the untreated

group. This might add bias to ATT estimate, if there were systematic differences in pre-treatment characteristics between matched and unmatched treated individuals[‡].

The conjoint use of PS-based methods and MLR (doubly robust approach) rather than their mutually exclusive implementation, could therefore be a better choice for evaluating ATT in non-randomized HRQoL studies^{5,16,17}, by improving the comparability between groups^{18,19}.

In this work we aim to compare optimal pair matching (OPM), optimal full matching (OFM), SBC and IPTW both on their own and in combination with MLR, by investigating their performances in estimating ATT on HRQoL outcomes in a simulation study. We will not consider the use PS as a variable in linear regression, because this approach is generally not recommended by previous literature²⁰. We will focus on small to moderate sample sizes, as many HRQoL studies are conducted on very small samples or in large studies in which only a few patients have provided HRQoL data. Therefore, we will only consider additional MLR modeling where ATT is estimated as a coefficient of a treatment indicator. We will also apply the described methods to a real case as a motivating example.

[‡] However, this issue might be addressed by testing the null hypothesis of no systematic differences in pre-treatment variables between these subgroups.

2. Methods

The estimation problem can be described according to the potential outcomes framework²¹⁻²³.

Under specific assumptions on treatment assignment³, ATT can be estimated as the difference in outcomes between treated and untreated subjects, conditional to a set of observed covariates.

Formal details about underlying assumptions are provided in Appendix 1 and a more comprehensive presentation about estimation of causal effects can be found in Imbens and Rubin²⁴ and Rubin²⁵. Before estimating ATT, we balanced treated and untreated groups on the propensity scores estimated on a set of observed covariates, respectively by OPM, OFM, SBC and IPTW[ref.]. Balance in observed covariates between matched groups was checked for each variable by calculating the corresponding differences in standardized mean (SMD) between groups before and after balance²⁶. We estimated ATT on balanced groups by either the mean of differences in outcomes between matched pairs in OPM or the weighted difference in means between treated and untreated groups²⁷. We assessed statistical significance by the paired *t*-test when using OPM and a weighted *t*-test with OFM, SBC and IPTW. When performing MLR on previously PS-balanced data, we estimated ATT as the coefficient of a treatment status indicator in a model with all observed variables^{28,29}, denoting such methods as OPM-RA, OFM-RA, SBC-RA and IPTW-RA. Formal details about matching and estimation methods can be found in Appendix 1.

Patient's self-reported outcomes were measured by the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30 (EORTC QLQ-C30)³⁰. QLQ-C30 is an internationally validated HRQOL questionnaire comprising 30 items and providing fifteen outcomes, each represented by a linear scale ranging from 0 to 100, derived from the raw scores of the items by a scoring algorithm. A higher score represents a higher level of functioning and health status/QOL or higher level of symptoms.

HRQoL scales were analyzed as continuous outcomes. The relative performance of methods was evaluated by investigating absolute bias of estimates \widehat{ATT} , $B = E[\widehat{ATT} - ATT]$, root mean squared error, $RMSE = \sqrt{E[(\widehat{ATT} - ATT)^2]}$, type I error and power. These were defined as the proportion of rejections, respectively of the true null hypothesis $H_0: ATT=0$ and of the true alternative hypothesis $H_1: ATT \neq 0$. All tests for statistical significance were two-sided ($\alpha=0.05$). In the real case, we performed additional sensitivity analyses to assess the impact on ATT estimates of unobserved confounding. We used R software V. 3.2.4 to generate and balance data²⁷ and to perform sensitivity analyses³¹. All further analyses were performed using SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA) and the codes are available in Appendix 3.

3. Simulations

Simulations were designed to mimic a typical non-randomized study where the outcomes of treated and untreated individuals are to be compared and different methods can be applied to estimate ATT. Simulated data comprised five covariates (x_1 - x_5) influencing treatment assignment and outcomes and one post-treatment variable (t) affecting outcomes only. Treatment was assigned as based on the propensity scores estimated on x_1 - x_4 by a logistic model (x_5 was considered as an unobserved confounder). Pre and post-treatment HRQoL outcomes were generated by the Partial Credit model³², as the EORTC QLQ-C30 scales Insomnia (SL), Cognitive functioning (CF), Fatigue (FA) and Emotional functioning (EF), reflecting the subjects' specific patterns of covariates. We chose to simulate these scales as each has a specifically skewed and clustered distribution, depending on the different underlying number of items. Data generating process is detailed in Appendix 1.

For each simulated scale, the true ATT was defined as $\delta=|12|$ points. We defined several scenarios based on the simulated data (Table 1), all including a partial overlap between cases and controls in

the distributions of covariates. Simulations were based on samples of treated subjects, each of size r ($r=10, 20, 30, 40$).

[Insert Table 1 here]

For each r -sample, three samples of $k \times r$ controls were drawn, $k=2, 3, 4$, to form samples of overall size $m=(1+k) \times r$, which had a proportion of $\pi=kr/m$ controls per treated individual ($\pi=0.66, 0.75$ and 0.80 , see Table 1). The increase of π corresponded to an increase of non-overlapping controls. For each scenario, we investigated results of 5000 replicates when overall sample size was $n \leq 80$ and 1000 replicates for larger overall sample sizes.

3.1 Results

All methods improved the balance of covariates and the overlap of the estimated propensity scores, with respect to raw data. OFM generally provided the pattern with lowest bias across all scenarios (Figures 1, left column). For all methods but OFM, the increase in the proportion of non-overlapping controls caused an increment in bias for samples up to $n=60$ and less different bias patterns with the increase of sample size. When further adjusted by MLR, all methods showed consistently less biased patterns (Figure 1, right column). However, IPTW-RA showed the best performance for all of the scales for $n < 80$. For larger sample sizes, bias patterns were broadly similar for all methods but SBC, which performed equal to or worse than other method across all scenarios. Further details on bias results are presented in the online Appendix 2, table I.

[Insert Figure 1 here]

With respect to RMSE and regardless further regression adjustment, SBC and OFM-based estimators consistently resulted in the highest variability with respect to other approaches (see

online Appendix 2, table II). Conversely, OPM and IPTW showed patterns of lower RMSE, although IPTW performed better than OPM up to $n=80$, but the same or worse for larger sample sizes. When further adjusted, both OPM-RA and IPTW-RA provided almost identical patterns. When considering type I error at a nominal $\alpha=0.05$ level, OPM and OPM-RA provided the best performance, consistently producing the least variable patterns of Type I errors, also the closest to the nominal $\alpha=0.05$ level (Figure 2). IPTW provided patterns similar to OPM only up to $n=60$, while showing progressively increasing Type I errors with the increment of sample size. Further regression adjustment corresponded to an inflation of Type I error for SBC, leaving OPM and IPTW substantially unaltered. Further details on results for Type I error are presented in the online Appendix 2 (table III).

[Insert Figure 2 here]

With respect to power, SBC performed best than other methods for sample sizes $n \leq 60$, providing the highest pattern (Figure 3, left column). For larger sample sizes, IPTW provided similar patterns and also OPM markedly improved its performance. OFM was the less sensitive method in power change to the sample size increase. After adjustment, SBC-RA showed the highest power, while OPM-RA the lowest (Figure 3, right column). Further details on results for power are presented in the online Appendix 2 (table IV).

[Insert Figure 3 here]

4. Application to real data

Data stem from a prospective non-randomized multicenter study, including women diagnosed with primary breast cancer³³. In this work, 183 women were included in the analyses, of which 155 (84.7%) received high-intensity therapies (chemotherapy and/or radiotherapy), while 28 (15.3%)

received either hormone therapy or no therapy at all. We estimated ATT of concomitant high-intensity therapies vs low-intensity or no therapy, on women post-operative HRQoL, by the same methods described in section 2. Possible confounders which we used to estimate PS for ATT evaluation included variables measured before treatment assignment, i.e. patient's age, tumor extension, comorbidities, weight and previous surgery interventions.

Main characteristics of subjects are shown in Table 2, before and after balancing on the estimated PS. Once balanced, similarity between groups improved markedly in all characteristics at diagnosis, across all methods. For example, the standardized mean difference (SMD) of age at diagnosis decreased from 64.6% to a range from 2.36% to 20.89%, respectively for IPTW and OPM. When performing further linear regression adjustment on PS-balanced data, we considered those covariates with residual SMD $\geq 10\%$ ³⁴. In addition, we included time since diagnosis (TSD) in the MLR model as a posttreatment concomitant variable, as we deemed unlikely that it had been affected by the type of therapy for primary breast cancer. Indeed, we found reasonable that TSD might reflect unobserved pretreatment differences between groups impacting both on treatment assignment and outcomes, e.g. a prompt access to regular screening for tumor prevention. In addition, we also performed MLR adjustment without time since diagnosis, to assess the corresponding sensitivity of ATT estimates¹⁵. Overall, MLR-adjusted ATT estimates (\widehat{ATT}) showed small differences (range: from 0.05 to 1.95) with respect to those not adjusted (Figure 4). Also, estimates did not differ substantially across methods for SL and EF, also with MLR adjustment and regardless the inclusion of TSD. Instead, ATT estimates for FA were markedly larger in matching (OPM and OFM) vs weighting methods (SBC and IPTW), also exceeding the 5-points minimal important difference § (MID). When further adjusting for residual imbalance, the estimated differences for FA fell below the MID, i.e. 4.31 in OPM-RA and 3.88 in OFM-RA. However, when time since diagnosis was not included in MLR

§ A minimal important difference is defined as “the smallest difference in a score...which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management” (Schünemann HJ, Guyatt GH. Commentary--goodbye M(C)ID! Hello MID, where do you come from?. Health Serv Res. 2005;40(2):593-7.)

adjustment, ATT estimates for FA remained still clinically significant, i.e. 6.30 in OPM-RA and 5.61 in OFM-RA).

[Insert Table 2 here]

These changes highlight the potential important impact of further adjustment for concomitant post-treatment variables which could not be considered for PS estimation. All ATT estimates however, were not statistically significant. Based on results of simulations in section 3.1, OPM seem to be the most reliable method in this applied case. Sensitivity analyses for unobserved confounding suggested that, in this study, a unobserved confounder strongly associated to the outcomes would be necessary to make ATT estimates statistically significant different from 0.

[Insert Figure 4 here]

5. Discussion

The conjoint use of PS-based methods and MLR, might be applied in non-randomized HRQoL studies to improve ATT estimates obtained by either methods when used separately, conditional on the collapsibility of the outcome measure ³⁵, e.g. when this is assessed by a mean difference between HRQoL scores. In this work we investigated the relative performance of eight alternative methods in estimating the ATT on HRQoL outcomes, i.e. OPM, OFM, SBC and IPTW, used either with or without further MLR adjustment. Our aim was to propose the possibly fruitful application of doubly robust ATT estimation methods in HRQoL research, highlighting the importance of comparability between groups to improve the accuracy of ATT estimates. To this purpose we performed simulations according to different overall sample sizes, types of HRQoL scales and proportions of

non-overlapping controls. The performances of the methods were evaluated by type I error, power, bias and variability of ATT estimates.

Further linear regression adjustment on matched data can account for residual imbalance left in matched groups, thus lessening the impact of model choice for the estimation of PS, potentially based on all available pretreatment variables related to either treatment assignment, outcome or both^{36,37}. Previous balance of groups on the PS allows to directly perform MLR on the whole sample, to estimate ATT as the coefficient of a treatment indicator, rather than using the estimated MLR coefficients of observed covariates in the untreated group, to predict the individual outcomes in the treated subjects. This would allow the use of MLR to adjust for residual imbalance in those cases when the latter MLR approach is unfeasible, e.g. due to the insufficient subjects/covariates ratio.

A first limitation of this work is that we did not investigate issues in MLR performances (e.g. outliers or missing data) other than the normality assumption of the outcomes' conditional distribution. However, we note that all analyses we performed were based on the same data structure, thus sharing the same possible limitations due to not considered issues. Another limitation is that we only considered cross-sectional comparisons and further research in longitudinal settings would be desirable.

This work has also strengths. First, we considered a variety of propensity score-based methods. In addition, the use of simulated data allowed us to evaluate the performance of such methods in several different scenarios, knowing the true ATT while varying key data features. Also, we compared the methods with respect to Type I error and power, which could be beneficial for design of non-randomized HRQoL studies. For example, a researcher might calculate sample size as based on the power (for a fixed type I error and hypothesized ATT) of a MLR model, including at least all covariates one would use to estimate propensity scores, beside the treatment status indicator. Also, we used a real clinical example to motivate our research, highlighting issues in ATT estimates due to different methods of estimation.

Overall, our findings suggest the conjoint use of PS based methods and MLR to be a useful method to achieve estimates of ATT in HRQoL non-randomized studies, potentially more accurate than those obtained by either methods used separately. Also, in this study IPTW and OPM showed overall better performances than other methods and should be preferred to OFM and SBC in estimating ATT on HRQoL outcomes. Specifically, OPM performed best in sample sizes of $n \geq 80$, while IPTW for smaller sample sizes. Although this work focused on HRQoL scales, we speculate our findings could be reasonably extended to other similar outcomes. Indeed, the distributional characteristics of summative scales and the corresponding issues that ATT estimation methods have to deal with, stem from a widespread ordinal response scheme.

Acknowledgements

The authors sincerely thank Dr. Kathrin Sommer for the outstanding help provided in code programming.

Contributions

Conception and design: FC

Statistical analyses: FC, GSC

Interpretation of results: all authors

Manuscript writing: all authors

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Efficace F, Fayers P, Pusic A, et al. Quality of patient-reported outcome reporting across cancer randomized controlled trials according to the CONSORT patient-reported outcome extension: A pooled analysis of 557 trials. *Cancer*. Sep 15 2015;121(18):3335-3342.
2. Rosenbaum P. *Design of Observational Studies*. Berlin: Springer; 2010.
3. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
4. Heckman JJ, Ichimura H, Todd PE. Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*. 1997;64(4):605-654.
5. Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*. 2007;15(03):199-236.
6. Colton D, Covert R. *Designing and Constructing Instruments for Social Research and Evaluation*: Jossey-Bass; 2015.
7. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health*. 2002;23:151-169.
8. Arostegui I, Nunez-Anton V, Quintana JM. Statistical approaches to analyse patient-reported outcomes as response variables: an application to health-related quality of life. *Stat Methods Med Res*. Apr 2012;21(2):189-214.
9. Basu A, Manca A. Regression estimators for generic health-related quality of life and quality-adjusted life years. *Med Decis Making*. Jan-Feb 2012;32(1):56-69.
10. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci*. Feb 1 2010;25(1):1-21.
11. Zhao Z. Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. *The Review of Economics and Statistics*. 2004;86(1):91-107.
12. Abadie A, Imbens GW. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics*. 2011/01/01 2011;29(1):1-11.
13. Rubin DB. Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association*. 1979;74(366):318-328.
14. Rubin DB, Thomas N. Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *Journal of the American Statistical Association*. 2000;95(450):573-585.
15. Rosenbaum, PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J Roy Statist Soc A*. 1984; 147, 656—666.

16. Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. *Am J Epidemiol*. Apr 1 2011;173(7):761-767.
17. Kreif N, Gruber S, Radice R, Grieve R, Sekhon JS. Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. *Stat Methods Med Res*. Oct 2016;25(5):2315-2336.
18. Cottone F, Efficace F, Apolone G, Collins GS. The added value of propensity score matching when using health-related quality of life reference data. *Stat Med*. Jun 5 2013.
19. Peinemann F, Labeit AM, Thielscher C, Pinkawa M. Failure to address potential bias in non-randomised controlled clinical trials may cause lack of evidence on patient-reported outcomes: a method study. *BMJ Open*. 2014;4(6):e004720.
20. Hade EM, Lu B. Bias associated with using the estimated propensity score as a regression covariate, *Stat Med*. 2014 Jan 15; 33(1): 74–87;
21. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; 66(5):688–701.;
22. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health* 2000; 21:121–145.
23. Holland PW. Statistics and causal inference. *Journal of the American Statistical Association* 1986; 81(396):945–960.
24. Imbens GW, Rubin DB. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press; 2015.
25. Rubin DB. For objective causal inference, design trumps analysis. 2008.
26. Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*. 1985;39(1):33-38.
27. Ho D, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. 2011 *J Stat Soft*;42(8):28.
28. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods*. Dec 2008;13(4):279-313.
29. Stuart EA. Developing practical recommendations for the use of propensity scores: Discussion of ‘A critical appraisal of propensity score matching in the medical literature between 1996 and 2003’ by Peter Austin, *Statistics in Medicine*. *Statistics in Medicine*. 2008;27(12):2062-2065.
30. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. Mar 3 1993;85(5):365-376.
31. Carnegie NB, Harada M and Hill JL. Assessing Sensitivity to Unmeasured Confounding Using a Simulated Potential Confounder. *Journal of Research on Educational Effectiveness*. 2016; 9:3, 395-420]
32. Masters GN. A rasch model for partial credit scoring. *Psychometrika*. June 01 1982;47(2):149-174.
33. Dabakuyo TS, Guillemin F, Conroy T, et al. Response shift effects on measuring post-operative quality of life among breast cancer patients: a multicenter cohort study. *Qual Life Res*. Feb 2013;22(1):1-11.
34. Nguyen TL, Collins GS, Spence J, Daurès JP *et al.*, Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC Med Res Methodol*. 2017 Apr 28;17(1):78.]
35. Hernán MA, Clayton D, Keiding N . The Simpson’s paradox unraveled. *International Journal of Epidemiology*. 2011; 40(3), 780–785
36. 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. *Stat Med*. Feb 20 2007;26(4):734-753.

37. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol.* Jun 15 2006;163(12):1149-1156.

Table 1. Simulation framework: overall sample size by treated and untreated individuals.

Treated individuals	Proportion of untreated individuals		
	$\pi=0.66$	$\pi=0.75$	$\pi=0.80$
10	30	40	50
20	60	80	100
30	90	120	150
40	120	160	200

Legend: The table shows the overall sample size m for all simulations run on each type of summative scale, according to the number r of treated individuals and the corresponding proportion π of untreated individuals in the whole sample. To illustrate, the sample sized $m=80$ was made up of $r=20$ treated and $\pi \cdot m=0.75 \cdot 80=60$ controls.

Table 2. Pretreatment characteristics and balance diagnostics

Variable	High intensity therapies (n=155)	Low intensity/no therapies (n=28)	SMD (%) before matching	SMD(%) after OPM	SMD (%) after OFM	SMD(%) after SBC	SMD (%) after IPTW
Age at diagnosis (Mean, SD)	56.16 (10.64)	63.38 (11.17)	64.60	20.89	14.03	8.42	2.36
Weight (Mean, SD)	69.59 (15.51)	63.25 (10.15)	62.46	12.32	6.73	15.43	-1.96
Local tumor			81.00	13.62	1.73	17.74	-0.32
No	44 (28.39)	2 (7.14)					
Yes	111 (71.61)	26 (92.86)					
Comorbidity			45.69	14.06	0.80	20.82	-1.17
1	119 (76.77)	15 (53.57)					
>1	36 (23.23)	13 (46.43)					
Previous surgery			19.74	8.55	3.67	7.41	2.67
No	46 (29.68)	6 (21.43)					
Yes	109 (70.32)	22 (78.57)					
Time since diagnosis	55.68 (47.14)	49.36 (27.61)	22.91	21.86	31.27	17.47	13.75

Abbreviations: SD, standard deviation, SMD, standardized mean difference, OPM, optimal pair matching, OFM, optimal full matching, SBC, sub-classification, IPTW, inverse probability of treatment weighting.

Figure 1. Absolute bias of ATT estimates for each estimating method, on summative scales with one to four items.

Legend: the figure represents estimates of ATT based on matched data, either with or without further adjustment by multivariable linear regression, including all observed covariates plus a posttreatment concomitant variable.

Abbreviations: OPM, optimal pair matching; OFM, optimal full matching; SBC, sub-classification on the propensity scores; IPTW, inverse probability of treatment weighting;

Figure 2. Type I error of ATT estimates for each estimating method, on summative scales with one to four items

Legend: the figure represents estimates of ATT based on matched data, either with or without further adjustment by multivariable linear regression, including all observed covariates plus a posttreatment concomitant variable. The continuous line represents the nominal level of type I error ($\alpha=0.05$).

Abbreviations: OPM, optimal pair matching; OFM, optimal full matching; SBC, sub-classification on the propensity scores; IPTW, inverse probability of treatment weighting.

Figure 3. Power of ATT estimates for each estimating method, on summative scales with one to four items

Legend: the figure represents estimates of ATT based on matched data, either with or without further adjustment by multivariable linear regression, including all observed covariates plus a posttreatment concomitant variable.

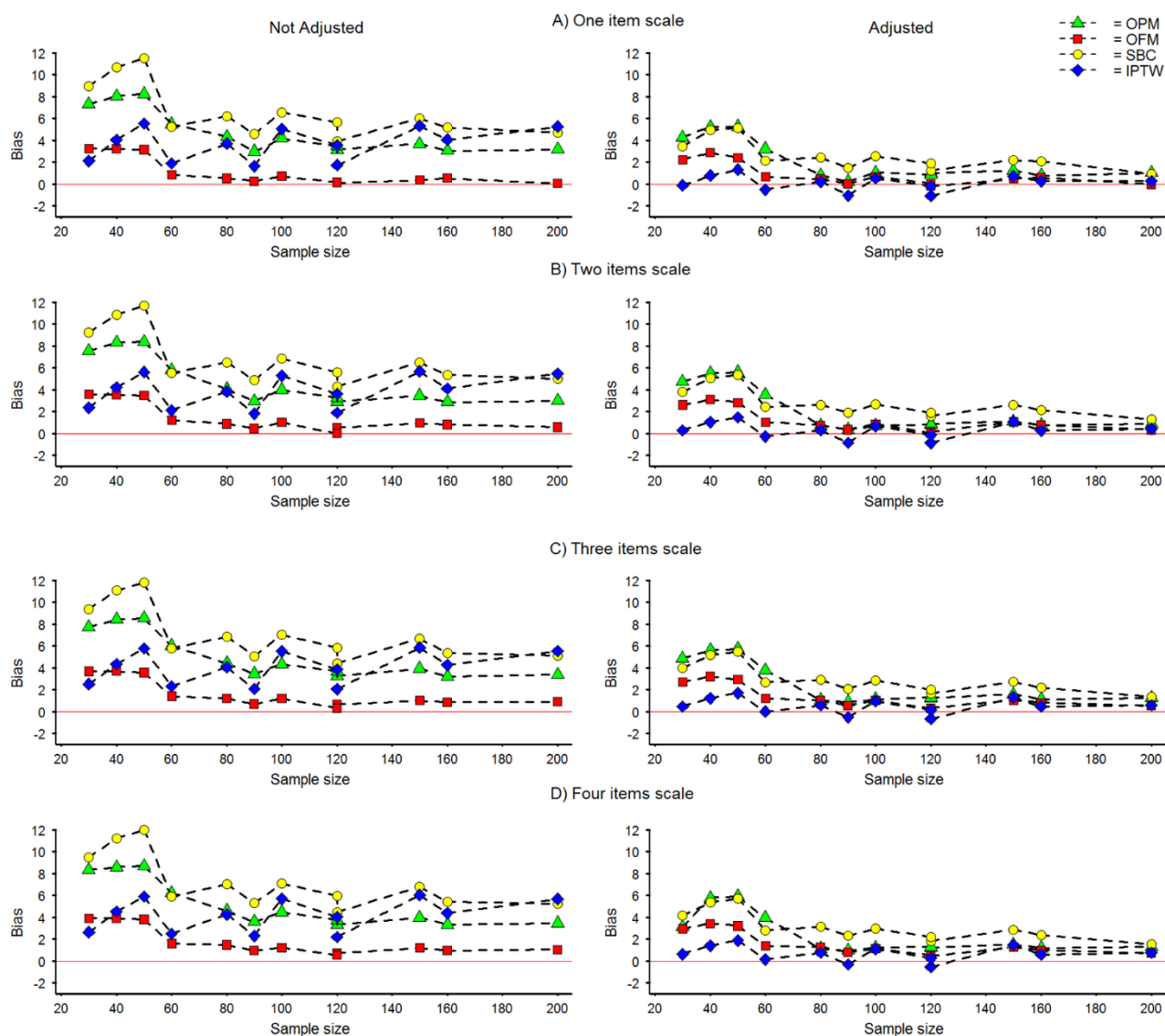
Abbreviations: OPM, optimal pair matching; OFM, optimal full matching; SBC, sub-classification on the propensity scores; IPTW, inverse probability of treatment weighting.

Figure 4. Point estimates and 95% CIs of ATT on HRQoL, low vs high-intensity therapies.

Legend: the figure represents estimates of ATT based on matched data, either with or without further adjustment by multivariable linear regression, including observed covariates with residual imbalance $\geq 10\%$, time since diagnosis.

Abbreviations: OPM, optimal pair matching; OPM-RA, OPM plus regression adjustment; OFM, optimal full matching; OFM-RA, OFM plus regression adjustment; SBC, sub-classification; SBC-RA, SBC plus regression adjustment; IPTW, inverse probability of treatment weighting; IPTW-RA, IPTW plus regression adjustment.

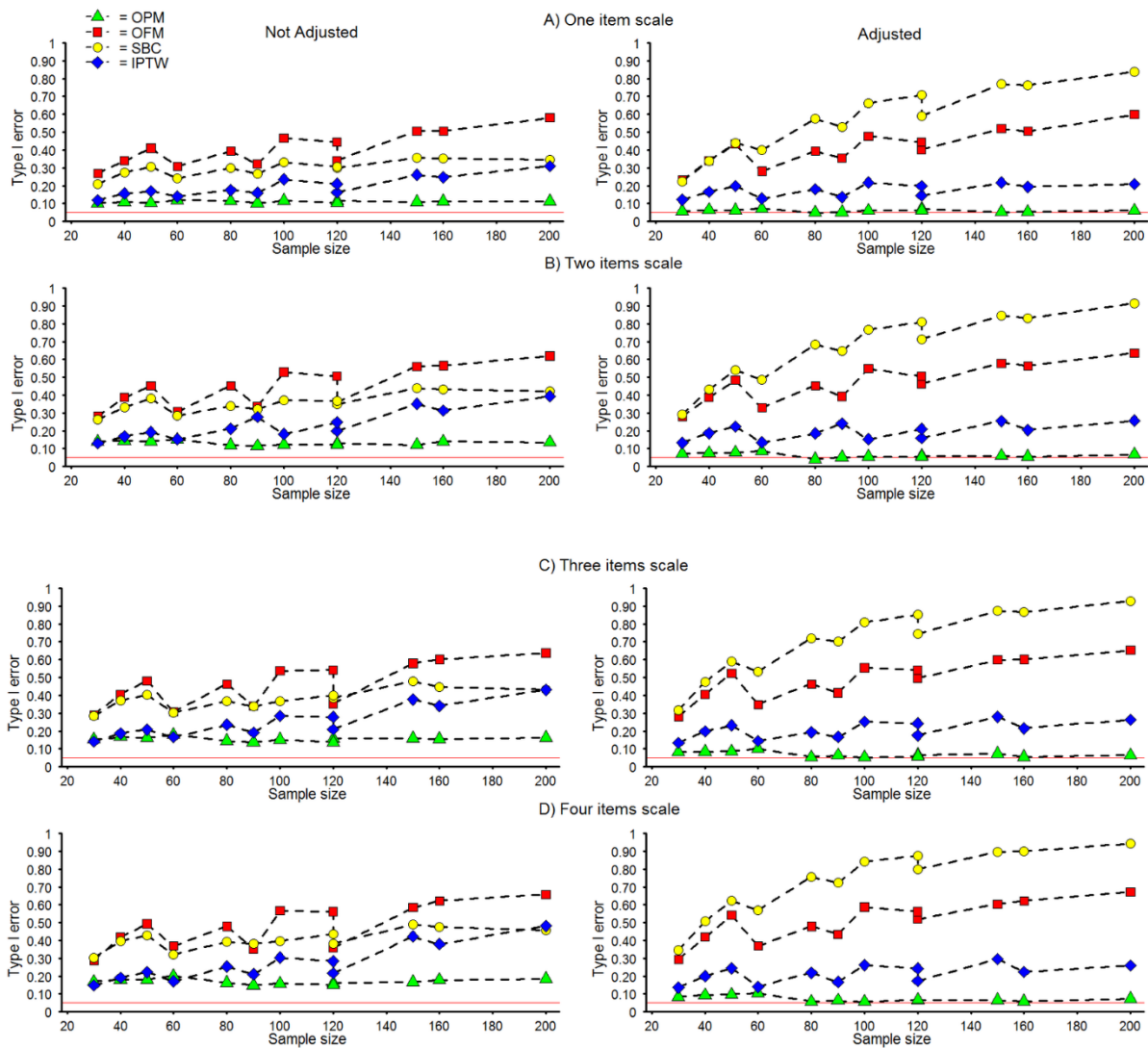
Figure 1. Absolute bias of ATT estimates for each estimating method, on summative scales with one to four items.



Legend: the figure represents estimates of ATT based on matched data, either with or without further adjustment by multivariable linear regression, including all observed covariates plus a posttreatment concomitant variable.

Abbreviations: OPM, optimal pair matching; OFM, optimal full matching; SBC, sub-classification on the propensity scores; IPTW, inverse probability of treatment weighting;

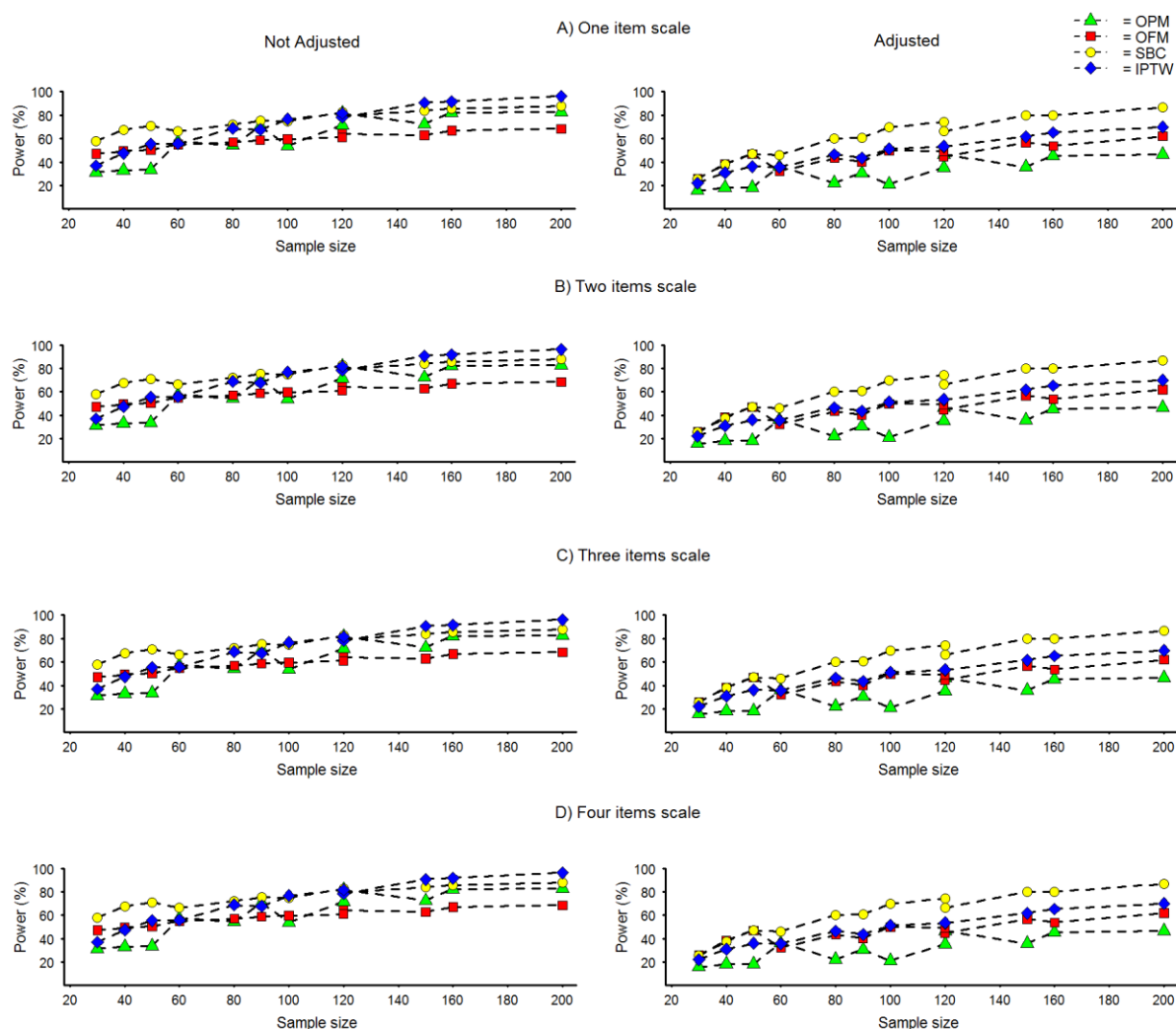
Figure 2. Type I error of ATT estimates for each estimating method, on summative scales with one to four items.



Legend: the figure represents estimates of ATT based on matched data, either with or without further adjustment by multivariable linear regression, including all observed covariates plus a posttreatment concomitant variable.

Abbreviations: OPM, optimal pair matching; OFM, optimal full matching; SBC, sub-classification on the propensity scores; IPTW, inverse probability of treatment weighting;

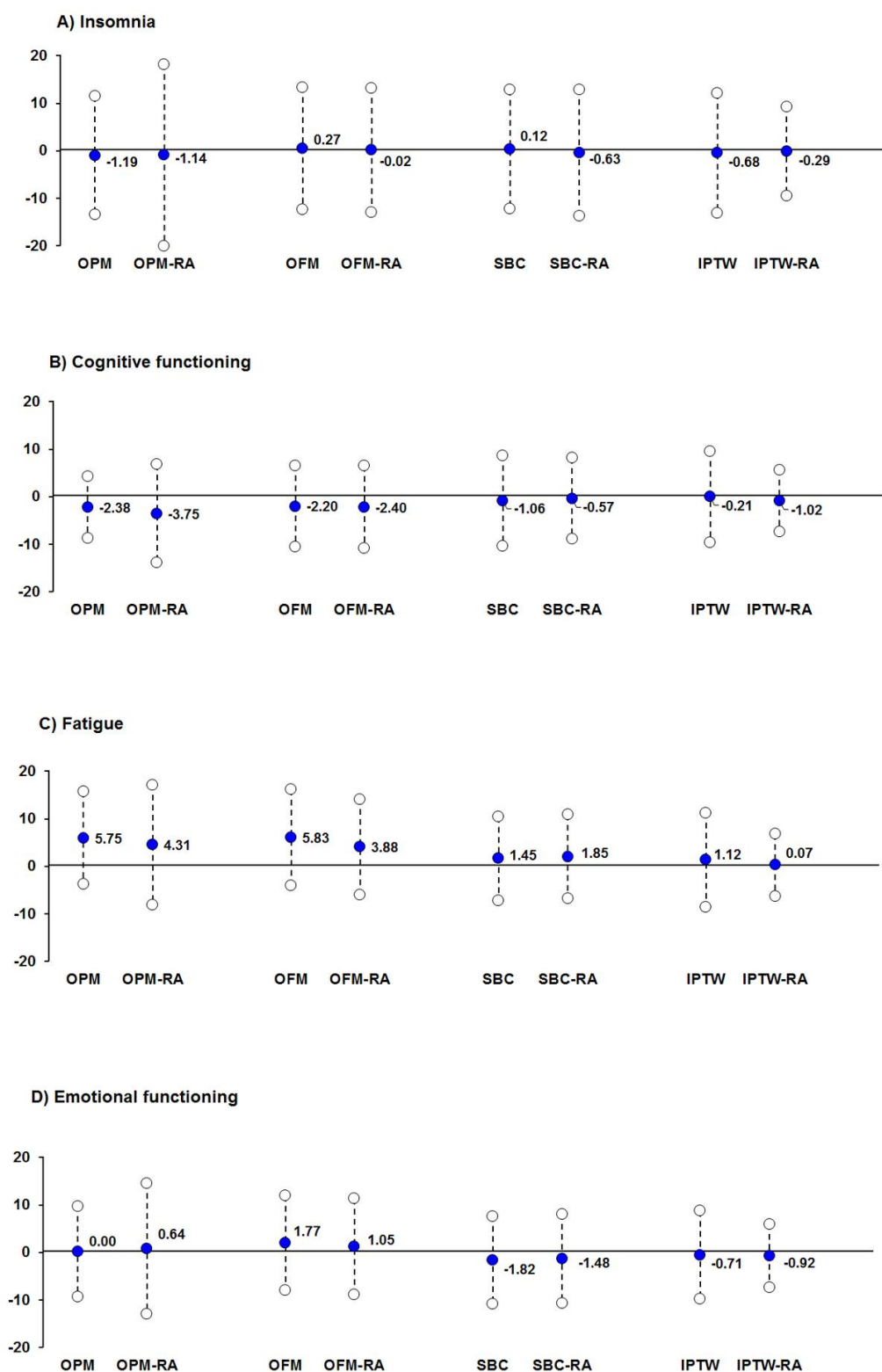
Figure 3. Power of ATT estimates for each estimating method, on summative scales with one to four items.



Legend: the figure represents estimates of ATT based on matched data, either with or without further adjustment by multivariable linear regression, including all observed covariates plus a posttreatment concomitant variable.

Abbreviations: OPM, optimal pair matching; OFM, optimal full matching; SBC, sub-classification on the propensity scores; IPTW, inverse probability of treatment weighting;

Figure 4 Estimates and 95% CIs of ATT on HRQoL, low vs high-intensity therapies.



Legend: the figure represents estimates of ATT based on matched data, either with or without further adjustment by multivariable linear regression, including observed covariates with residual imbalance $\geq 10\%$, time since diagnosis.

Abbreviations: OPM, optimal pair matching; OPM-RA, OPM plus regression adjustment; OFM, optimal full matching; OFM-RA, OFM plus regression adjustment; SBC, sub-classification; SBC-RA, SBC plus regression adjustment; IPTW, inverse probability of treatment weighting; IPTW-RA, IPTW plus regression adjustment.

APPENDIX 1

Theoretical framework

ATT is defined as

$$E[\delta] = E[Y_1 - Y_0 | D = 1] = E(Y_1 | D = 1) - E(Y_0 | D = 1),$$

where D is the treatment status and $E(Y_1 | D = 1)$, $E(Y_0 | D = 1)$ are, respectively, the expected observed and potential HRQoL outcomes in the treated individuals. In non-randomized studies and under the following assumptions

$$E(Y_0 | \mathbf{X}, D = 1) = E(Y_0 | \mathbf{X}, D = 0) \quad (1)$$

$$0 < Pr(D = 1 | \mathbf{X}) < 1 \quad (2),$$

ATT can be estimated as

$$E[\delta] = E_{X|D=1}[E(Y_1 - Y_0 | \mathbf{X}, D = 1)],$$

where \mathbf{X} is a set of observed confounders. Assumption (1) implies that the expected unobservable outcomes of treated subjects are the same to the (expected) observable outcomes of those untreated individuals, conditional on a set of observed covariates. This might also hold in presence of treatment effect modification. Assumption (2) is required to ensure that a potential control might exist for each treated individual. However, it does not imply that each treated subject actually finds an untreated match.

Statistical methods

I. Optimal Pair Matching (OPM). Propensity scores were estimated by

$$\text{logit}(p_{treat}|\alpha, \mathbf{x}) = \alpha_0 + \alpha' \mathbf{x}, (3)$$

where \mathbf{x} is the vector of observed confounders. Each individual was matched to one treated subject by optimal pair matching minimizing the overall distance between the estimated propensity scores (EPS) in each matched pair. ATT was then estimated by the mean of differences in outcomes between matched pairs.

II. Optimal Full Matching (OFM). Propensity scores were estimated by (3). Untreated subjects were matched to treated individuals by optimal full matching. In this setting, OFM assigned one treated individual to at least one untreated subject (1:k matching), by minimizing the overall distance in weighted average of EPS between treated and untreated subjects, within each matched set. The ATT was then estimated by the weighted difference in means between the treated and untreated individuals. Treated units were assigned weight 1, while each control within each matched subset were assigned the same subset-specific weight, as to reflect the corresponding treated/controls proportion.

III. Sub-classification on the propensity scores (SBC)

All subjects were classified into four strata, based on the quartiles of the propensity scores estimated as in (3). We chose this number of strata to make SBC feasible in small sample sizes. Weights were defined as to reflect the proportions of treated units in each stratum.

IV. Inverse probability of treatment weighting (IPTW) using the propensity scores

Each subject is assigned a weight

$$w_i = d_i + \frac{(1 - d_i)e(\mathbf{x}_i)}{1 - e(\mathbf{x}_i)},$$

where d_i is the treatment status indicator and $e(\mathbf{x}_i)$ the estimated propensity score as in (3). ATT was estimated by the weighted difference in mean outcomes between the treated and untreated individuals.

V. **Optimal Pair Matching plus Regression Adjustment (OPM-RA).** Further adjustment was performed on the set of matched individuals as defined in item I, using the linear regression model

$$y^* = \beta_0 + \beta_{treat}d + \boldsymbol{\beta}'\mathbf{x} + u \quad (4)$$

VI. **Optimal Full Matching plus Regression Adjustment (OFM-RA).** Further adjustment was performed on the set of matched individuals as defined in item II, using the linear regression model (4) on individuals, weighted as in item II.

VII. **Sub-classification on the propensity scores plus regression adjustment (SBC-RA).**

The MLR model (4) was applied to the set of weighted individuals as defined in item III.

VIII. **IPTW plus Regression Adjustment (IPTW-RA)**

The MLR model (4) was applied to the set of weighted individuals as defined in item IV.

Data-generating process

We generated a population of $10^6/2$ treated and $10^6/2$ untreated individuals, each defined by an observed pattern of pretreatment (baseline) characteristics, x_1 - x_5 and the time from diagnosis to HRQoL evaluation (t), independently sampled from distinct random variables, X_1, X_4 drawn from Beta distributions, X_2, X_3, X_5 from binomials, t from a uniform. A baseline continuous latent trait θ was generated for each individual:

$$\theta_i = 3 \cdot (P_{treat} | x_i), \quad (1)$$

where P_{treat} was the probability of being assigned to treatment as estimated by a logit model that included all of the baseline covariates. Each individual was then assigned to treatment according to a Bernoulli distribution with parameter P_{treat} . This way, the individuals' latent trait was linked to their specific pattern of covariates, reflecting the systematic baseline differences between treated and untreated individuals.

The latent trait θ_i^t at HRQoL assessment (T_1) was generated as $\theta_i^t = \theta_i \times u_i$, $U \sim (0,2)$, representing the actual health status of the untreated and the potential health status of the treated had they not been treated. The actual latent trait of treated individuals at T_1 was generated by $\theta_i^{td} = \theta_i + u_i$, $U \sim (0,1.2)$, representing a treatment effect which caused a deterioration of their health status.

HRQoL outcomes were simulated in two steps. We used the Partial Credit Model (PCM) from item response theory to generate individual responses to different items, each with four possible ordinal response categories (e.g., “Not at all”=1, “A little bit”=2, “Quite a bit”=3, “Very much”=4). The PCM allows a continuous latent trait θ to be mapped on a discrete ordinal scale, modeling the probability of the individual i to choosing response category k , ($k=1, \dots, m_j$), for the item j among m_j possible responses, given its specific latent trait and the m_j-1 category difficulty parameters δ_{jk} for the item j . In this setting, difficulty parameters were set as the quartiles of θ_i^t , respectively $\delta_{j1}=0.002$, $\delta_{j2}=0.287$ and $\delta_{j3}=2.922$ for all j s, with $j, k=1, \dots, 4$. Overall, we simulated four rating items

each with four possible responses. Raw responses were then scored to scales ranging from 0 to 100, according to the standard scoring algorithm of the EORTC QLQ-C30 questionnaire v.3 (<http://groups.eortc.be/qol/manuals>). Outcomes based on one and three items reflect the two symptom scales in the real case (“Insomnia” and “Fatigue”, respectively) Outcomes based on two and four items reflect the functional scales (“Cognitive functioning” and “Emotional functioning”, respectively).

Partial credit model

The latent trait θ_i^t at HRQoL assessment (T_1) was generated as $\theta_i^t = \theta_i \times u_i$, $U \sim (0,2)$, representing the actual health status of the untreated and the potential health status of the treated had they not been treated. The actual latent trait of treated individuals at T_1 was generated by $\theta_i^{td} = \theta_i + u_i$, $U \sim (0,1.2)$, representing a treatment effect which caused a deterioration of their health status.

HRQoL outcomes were simulated in two steps. We used the Partial Credit Model (PCM) from item response theory to generate individual responses to different items, each with four possible ordinal response categories (e.g., “Not at all”=1, “A little bit”=2, “Quite a bit”=3, “Very much”=4). The PCM allows a continuous latent trait θ to be mapped on a discrete ordinal scale, modeling the probability of the individual i to choosing response category k , ($k=1, \dots, m_j$), for the item j among m_j possible responses, given its specific latent trait and the m_j-1 category difficulty parameters δ_{jk} for the item j . In this setting, difficulty parameters were set as the quartiles of θ_i^t , respectively $\delta_{j1}=0.002$, $\delta_{j2}=0.287$ and $\delta_{j3}=2.922$ for all j s, with $j, k=1, \dots, 4$.

APPENDIX 2

Table I. Absolute bias of the average treatment effect on the treated estimate with sample size for each estimating method, on summative scales with one to four items.

Sample size	OPM	OPM-RA	OFM	OFM-RA	SBC	SBC-RA	IPTW	IPTW-RA	π
1 item scale									
30	7.282	4.266	3.254	2.226	8.930	3.453	2.109	-0.122	0.66
40	8.014	5.216	3.222	2.870	10.693	4.910	4.018	0.765	0.75
50	8.214	5.241	3.149	2.413	11.498	5.103	5.538	1.309	0.80
60	5.470	3.216	0.867	0.665	5.235	2.103	1.889	-0.523	0.66
80	4.321	0.799	0.518	0.464	6.199	2.418	3.710	0.210	0.75
90	2.980	0.232	0.269	0.012	4.581	1.498	1.627	-1.055	0.66
100	4.179	1.038	0.715	0.668	6.530	2.570	5.032	0.522	0.80
120	3.131	0.988	0.121	-0.181	3.926	1.235	1.733	-1.084	0.66
120	3.470	0.885	0.160	0.032	5.633	1.870	3.551	-0.203	0.75
150	3.657	1.232	0.363	0.476	6.036	2.190	5.310	0.706	0.80
160	3.049	0.786	0.530	0.590	5.184	2.084	4.064	0.255	0.75
200	3.163	1.023	0.060	-0.043	4.707	0.960	5.261	0.256	0.80
2 items scale									
30	-7.546	-4.750	-3.594	-2.617	-9.212	-3.804	-2.358	-0.290	0.66
40	-8.306	-5.502	-3.539	-3.110	-10.859	-5.050	-4.210	-1.043	0.75
50	-8.388	-5.591	-3.475	-2.839	-11.678	-5.336	-5.621	-1.472	0.80
60	-5.809	-3.558	-1.229	-1.044	-5.556	-2.445	-2.122	0.264	0.66
80	-4.048	-0.743	-0.882	-0.750	-6.495	-2.616	-3.816	-0.304	0.75
90	-2.953	-0.354	-0.465	-0.339	-4.861	-1.902	-1.824	0.856	0.66
100	-3.981	-0.708	-1.013	-0.866	-6.850	-2.676	-5.269	-0.689	0.80
120	-2.897	-0.853	-0.519	-0.190	-4.271	-1.603	-1.896	0.862	0.66
120	-3.243	-0.809	-0.009	-0.048	-5.599	-1.865	-3.629	0.133	0.75
150	-3.456	-1.094	-0.963	-1.052	-6.518	-2.613	-5.670	-1.077	0.80
160	-2.875	-0.776	-0.827	-0.781	-5.327	-2.147	-4.099	-0.260	0.75
200	-2.993	-0.881	-0.585	-0.337	-5.015	-1.301	-5.448	-0.424	0.80
3 items scale									
30	7.714	4.858	3.688	2.705	9.341	3.952	2.517	0.454	0.66
40	8.418	5.618	3.710	3.206	11.057	5.195	4.344	1.220	0.75
50	8.546	5.735	3.568	2.931	11.816	5.485	5.776	1.690	0.80
60	6.018	3.781	1.420	1.215	5.762	2.646	2.309	0.009	0.66
80	4.419	1.070	1.189	1.022	6.857	2.923	4.067	0.584	0.75
90	3.448	0.878	0.701	0.551	5.054	2.058	2.075	-0.526	0.66
100	4.328	1.118	1.199	1.021	7.035	2.846	5.510	0.951	0.80
120	3.258	1.227	0.658	0.331	4.368	1.672	2.056	-0.659	0.66
120	3.687	1.310	0.311	0.307	5.808	2.033	3.857	0.144	0.75
150	3.927	1.578	1.008	1.055	6.674	2.697	5.849	1.310	0.80
160	3.217	1.126	0.864	0.807	5.373	2.202	4.249	0.468	0.75
200	3.357	1.260	0.889	0.529	5.121	1.352	5.528	0.580	0.80
4 items scale									
30	-8.312	-3.114	-3.898	-2.920	-9.500	-4.148	-2.626	-0.625	0.66
40	-8.559	-5.747	-3.918	-3.428	-11.214	-5.367	-4.522	-1.404	0.75
50	-8.682	-5.922	-3.799	-3.210	-11.969	-5.693	-5.895	-1.876	0.80
60	-6.243	-3.948	-1.596	-1.366	-5.898	-2.777	-2.487	-0.149	0.66
80	-4.574	-1.117	-1.473	-1.288	-7.043	-3.121	-4.257	-0.753	0.75
90	-3.588	-0.958	-0.95	-0.793	-5.295	-2.311	-2.302	0.307	0.66
100	-4.464	-1.223	-1.229	-1.138	-7.115	-2.968	-5.696	-1.110	0.80
120	-3.321	-1.239	-0.712	-0.446	-4.489	-1.814	-2.186	0.534	0.66
120	-3.760	-1.306	-0.546	-0.500	-5.975	-2.192	-4.020	-0.272	0.75
150	-3.958	-1.511	-1.192	-1.265	-6.804	-2.864	-6.039	-1.475	0.80
160	-3.328	-1.178	-0.955	-0.981	-5.446	-2.372	-4.375	-0.603	0.75
200	-3.445	-1.287	-1.046	-0.718	-5.224	-1.543	-5.666	-0.765	0.80

Abbreviations: OPM, optimal pair matching; OPM-RA, OPM plus regression adjustment; OFM, optimal full matching; OFM-RA, OFM plus regression adjustment; SBC, sub-classification; SBC-RA, SBC plus regression adjustment; IPTW, inverse probability of treatment weighting; IPTW-RA, IPTW plus regression adjustment.

Legend: π , proportion of untreated individuals.

Table II. Root mean square error of the average treatment effect on the treated estimate with sample size for each estimating method, on summative scales with one to four items

Sample size	OPM	OPM-RA	OFM	OFM-RA	SBC	SBC-RA	IPTW	IPTW-RA	π
1 item scale									
30	10.969	11.920	14.53	13.968	12.092	11.919	9.543	11.580	0.66
40	11.489	12.638	14.613	14.407	13.306	12.507	9.594	11.684	0.75
50	11.598	12.297	14.475	14.268	13.878	12.670	9.705	11.182	0.80
60	7.787	7.719	11.334	10.173	9.312	9.456	6.967	8.131	0.66
80	7.231	8.141	11.208	10.130	9.958	9.919	7.161	8.105	0.75
90	5.159	6.181	10.042	8.978	8.040	7.642	5.999	6.901	0.66
100	7.169	8.056	11.184	10.253	9.996	10.028	7.663	8.126	0.80
120	4.998	5.634	8.726	7.821	7.458	7.097	5.111	5.820	0.66
120	5.641	6.425	9.320	8.625	8.281	8.040	6.148	6.742	0.75
150	5.564	6.317	9.741	8.719	8.506	8.135	6.639	6.350	0.80
160	4.837	5.499	8.740	8.082	7.827	7.461	5.732	5.793	0.75
200	4.885	5.457	8.316	7.720	7.241	7.043	6.130	5.370	0.80
2 items scale									
30	9.678	9.927	12.344	11.499	11.000	9.734	7.831	9.306	0.66
40	10.306	10.401	12.412	11.846	12.35	10.362	8.256	9.510	0.75
50	10.337	10.236	12.310	11.785	13.001	10.641	8.422	9.145	0.80
60	7.037	6.400	9.276	8.246	8.145	7.705	5.836	6.567	0.66
80	6.156	6.326	9.354	8.360	8.764	7.905	6.097	6.479	0.75
90	4.361	4.862	8.067	7.158	7.135	6.470	5.046	5.639	0.66
100	6.037	6.218	9.139	8.209	8.819	8.219	6.801	6.588	0.80
120	4.080	4.336	7.281	6.526	6.545	5.857	4.382	4.871	0.66
120	4.673	5.123	7.631	6.883	7.327	6.485	5.373	5.516	0.75
150	4.789	5.079	8.428	7.453	8.038	7.097	6.405	5.484	0.80
160	4.102	4.312	7.114	6.515	6.939	6.044	5.163	4.684	0.75
200	4.125	4.428	6.868	6.348	6.631	5.951	5.900	4.491	0.80
3 items scale									
30	9.343	9.237	11.511	10.587	10.720	8.960	7.316	8.573	0.66
40	9.877	9.527	11.412	10.841	12.094	9.635	7.735	8.703	0.75
50	9.994	9.473	11.466	10.856	12.782	9.954	8.104	8.449	0.80
60	6.805	5.933	8.580	7.485	7.834	7.065	5.440	5.990	0.66
80	5.886	5.630	8.761	7.663	8.663	7.455	5.918	5.855	0.75
90	4.470	4.515	7.401	6.504	6.783	5.958	4.694	5.162	0.66
100	5.797	5.620	8.516	7.571	8.531	7.651	6.692	6.035	0.80
120	4.057	3.981	6.716	6.008	6.256	5.500	4.152	4.537	0.66
120	4.658	4.678	7.045	6.387	7.204	6.149	5.245	5.180	0.75
150	4.850	4.560	7.563	6.765	7.796	6.595	6.337	5.134	0.80
160	3.990	3.885	6.500	5.945	6.691	5.605	5.083	4.341	0.75
200	4.051	3.952	6.525	6.003	6.462	5.529	5.872	4.236	0.80
4 items scale									
30	9.236	6.342	10.944	9.996	10.601	8.555	6.990	8.109	0.66
40	9.723	9.128	11.000	10.367	12.038	9.289	7.494	8.272	0.75
50	9.825	9.125	10.899	10.366	12.751	9.593	7.927	8.051	0.80
60	6.826	5.748	8.208	7.108	7.671	6.757	5.272	5.687	0.66
80	5.769	5.344	8.457	7.318	8.653	7.282	5.890	5.623	0.75
90	4.396	4.221	7.175	6.238	6.767	5.797	4.570	4.859	0.66
100	5.658	5.265	7.976	7.123	8.376	7.267	6.648	5.780	0.80
120	3.987	3.733	6.368	5.635	6.036	5.156	3.962	4.109	0.66
120	4.535	4.297	6.728	6.029	7.137	5.848	5.156	4.873	0.75
150	4.790	4.340	7.243	6.516	7.713	6.425	6.440	4.918	0.80
160	3.959	3.574	6.178	5.621	6.555	5.356	5.036	4.123	0.75
200	4.042	3.747	6.244	5.687	6.298	5.218	5.931	4.013	0.80

Abbreviations: OPM, optimal pair matching; OPM-RA, OPM plus regression adjustment; OFM, optimal full matching; OFM-RA, OFM plus regression adjustment; SBC, sub-classification; SBC-RA, SBC plus regression adjustment; IPTW, inverse probability of treatment weighting; IPTW-RA, IPTW plus regression adjustment.

Legend: π , proportion of untreated individuals.

Table III. Type I error of the average treatment effect on the treated estimate with sample size for each estimating method, on summative scales with one to four items

Sample size	OPM	OPM-RA	OFM	OFM-RA	SBC	SBC-RA	IPTW	IPTW-RA	π
1 item scale									
30	0.100	0.058	0.269	0.232	0.210	0.224	0.118	0.121	0.66
40	0.107	0.063	0.340	0.341	0.272	0.340	0.156	0.166	0.75
50	0.104	0.062	0.410	0.433	0.306	0.440	0.169	0.198	0.80
60	0.119	0.071	0.309	0.281	0.242	0.401	0.140	0.127	0.66
80	0.114	0.049	0.394	0.394	0.298	0.576	0.175	0.180	0.75
90	0.102	0.051	0.321	0.354	0.268	0.529	0.160	0.136	0.66
100	0.116	0.061	0.467	0.478	0.331	0.663	0.235	0.217	0.80
120	0.114	0.070	0.341	0.403	0.300	0.589	0.162	0.146	0.66
120	0.104	0.062	0.444	0.444	0.306	0.708	0.210	0.198	0.75
150	0.109	0.052	0.506	0.519	0.357	0.769	0.261	0.217	0.80
160	0.113	0.053	0.506	0.505	0.352	0.761	0.248	0.193	0.75
200	0.112	0.060	0.581	0.598	0.345	0.839	0.310	0.209	0.80
2 items scale									
30	0.140	0.071	0.282	0.280	0.263	0.293	0.130	0.133	0.66
40	0.144	0.075	0.387	0.387	0.333	0.431	0.168	0.186	0.75
50	0.141	0.078	0.452	0.484	0.382	0.539	0.192	0.224	0.80
60	0.150	0.087	0.306	0.329	0.284	0.486	0.153	0.135	0.66
80	0.119	0.041	0.452	0.452	0.338	0.685	0.212	0.185	0.75
90	0.114	0.052	0.338	0.393	0.320	0.648	0.278	0.240	0.66
100	0.123	0.055	0.529	0.548	0.371	0.765	0.182	0.153	0.80
120	0.129	0.060	0.359	0.463	0.350	0.714	0.199	0.159	0.66
120	0.123	0.055	0.505	0.505	0.368	0.810	0.248	0.211	0.75
150	0.121	0.060	0.560	0.578	0.440	0.846	0.350	0.255	0.80
160	0.140	0.055	0.565	0.564	0.431	0.830	0.314	0.205	0.75
200	0.135	0.067	0.619	0.636	0.422	0.914	0.394	0.257	0.80
3 items scale									
30	0.153	0.082	0.290	0.281	0.283	0.317	0.141	0.134	0.66
40	0.167	0.083	0.406	0.406	0.371	0.476	0.186	0.198	0.75
50	0.162	0.085	0.481	0.525	0.404	0.591	0.208	0.232	0.80
60	0.179	0.100	0.307	0.348	0.302	0.532	0.166	0.144	0.66
80	0.145	0.053	0.463	0.463	0.367	0.721	0.235	0.194	0.75
90	0.137	0.063	0.340	0.414	0.339	0.701	0.191	0.167	0.66
100	0.152	0.052	0.538	0.554	0.368	0.811	0.285	0.252	0.80
120	0.158	0.065	0.352	0.497	0.387	0.743	0.209	0.175	0.66
120	0.136	0.058	0.542	0.542	0.400	0.851	0.279	0.243	0.75
150	0.159	0.072	0.579	0.599	0.478	0.875	0.377	0.280	0.80
160	0.155	0.054	0.601	0.601	0.446	0.866	0.341	0.215	0.75
200	0.163	0.064	0.637	0.653	0.433	0.927	0.431	0.263	0.80
4 items scale									
30	0.168	0.084	0.287	0.294	0.304	0.347	0.148	0.136	0.66
40	0.181	0.093	0.418	0.419	0.398	0.509	0.191	0.199	0.75
50	0.180	0.097	0.493	0.541	0.428	0.621	0.222	0.244	0.80
60	0.201	0.106	0.368	0.368	0.321	0.568	0.171	0.140	0.66
80	0.163	0.057	0.479	0.479	0.393	0.755	0.253	0.219	0.75
90	0.149	0.063	0.353	0.435	0.381	0.725	0.211	0.167	0.66
100	0.158	0.056	0.567	0.587	0.398	0.841	0.304	0.262	0.80
120	0.162	0.066	0.359	0.519	0.383	0.800	0.216	0.173	0.66
120	0.150	0.068	0.562	0.562	0.434	0.875	0.283	0.243	0.75
150	0.166	0.066	0.584	0.603	0.491	0.896	0.423	0.296	0.80
160	0.178	0.058	0.622	0.621	0.477	0.898	0.379	0.224	0.75
200	0.185	0.073	0.657	0.672	0.456	0.941	0.482	0.260	0.80

Abbreviations: OPM, optimal pair matching; OPM-RA, OPM plus regression adjustment; OFM, optimal full matching; OFM-RA, OFM plus regression adjustment; SBC, sub-classification; SBC-RA, SBC plus regression adjustment; IPTW, inverse probability of treatment weighting; IPTW-RA, IPTW plus regression adjustment.

Legend: π , proportion of untreated individuals.

Table IV. Power of the average treatment effect on the treated estimate with sample size for each estimating method, on summative scales with one to four items

Sample size	OPM	OPM-RA	OFM	OFM-RA	SBC	SBC-RA	IPTW	IPTW-RA	π
1 item scale									
30	0.312	0.157	0.471	0.259	0.580	0.255	0.370	0.220	0.66
40	0.330	0.183	0.494	0.384	0.676	0.378	0.474	0.309	0.75
50	0.337	0.184	0.505	0.469	0.710	0.469	0.554	0.361	0.80
60	0.573	0.363	0.550	0.323	0.662	0.461	0.558	0.356	0.66
80	0.544	0.224	0.570	0.433	0.722	0.599	0.688	0.465	0.75
90	0.704	0.306	0.588	0.402	0.752	0.609	0.676	0.436	0.66
100	0.537	0.211	0.596	0.498	0.748	0.699	0.767	0.513	0.80
120	0.817	0.467	0.641	0.448	0.796	0.665	0.782	0.531	0.66
120	0.715	0.354	0.610	0.492	0.830	0.743	0.816	0.536	0.75
150	0.724	0.359	0.629	0.566	0.840	0.797	0.906	0.616	0.80
160	0.823	0.454	0.669	0.537	0.855	0.799	0.917	0.652	0.75
200	0.829	0.466	0.685	0.619	0.880	0.865	0.962	0.700	0.80
2 items scale									
30	0.481	0.247	0.549	0.319	0.744	0.343	0.512	0.288	0.66
40	0.507	0.279	0.579	0.445	0.810	0.483	0.629	0.401	0.75
50	0.517	0.286	0.602	0.548	0.840	0.583	0.712	0.463	0.80
60	0.793	0.555	0.641	0.395	0.813	0.579	0.730	0.465	0.66
80	0.727	0.320	0.648	0.505	0.858	0.726	0.857	0.569	0.75
90	0.885	0.500	0.484	0.484	0.881	0.736	0.828	0.563	0.66
100	0.733	0.298	0.677	0.589	0.884	0.793	0.906	0.626	0.80
120	0.960	0.653	0.524	0.524	0.913	0.783	0.911	0.683	0.66
120	0.894	0.514	0.693	0.583	0.929	0.851	0.927	0.675	0.75
150	0.900	0.513	0.737	0.637	0.938	0.869	0.980	0.756	0.80
160	0.952	0.665	0.778	0.613	0.956	0.873	0.973	0.806	0.75
200	0.954	0.663	0.789	0.674	0.954	0.930	0.997	0.836	0.80
3 items scale									
30	0.577	0.317	0.599	0.356	0.810	0.395	0.586	0.342	0.66
40	0.600	0.343	0.622	0.482	0.870	0.548	0.710	0.457	0.75
50	0.607	0.354	0.638	0.588	0.890	0.645	0.787	0.520	0.80
60	0.884	0.668	0.680	0.437	0.872	0.648	0.804	0.542	0.66
80	0.818	0.386	0.704	0.544	0.908	0.785	0.901	0.662	0.75
90	0.952	0.606	0.738	0.512	0.929	0.794	0.897	0.653	0.66
100	0.817	0.391	0.730	0.603	0.928	0.853	0.943	0.703	0.80
120	0.991	0.781	0.787	0.584	0.949	0.831	0.947	0.752	0.66
120	0.952	0.648	0.763	0.619	0.964	0.887	0.966	0.734	0.75
150	0.959	0.642	0.786	0.662	0.971	0.905	0.994	0.820	0.80
160	0.992	0.769	0.838	0.659	0.972	0.924	0.992	0.858	0.75
200	0.989	0.774	0.839	0.699	0.976	0.949	1.000	0.886	0.80
4 items scale									
30	0.747	0.556	0.619	0.385	0.845	0.431	0.630	0.375	0.66
40	0.662	0.388	0.656	0.507	0.905	0.578	0.756	0.493	0.75
50	0.674	0.407	0.675	0.611	0.922	0.679	0.826	0.558	0.80
60	0.926	0.740	0.714	0.476	0.899	0.685	0.841	0.576	0.66
80	0.886	0.438	0.719	0.574	0.926	0.829	0.921	0.691	0.75
90	0.976	0.665	0.766	0.551	0.945	0.825	0.921	0.696	0.66
100	0.869	0.427	0.758	0.645	0.940	0.881	0.965	0.729	0.80
120	0.994	0.822	0.821	0.615	0.958	0.873	0.964	0.792	0.66
120	0.978	0.694	0.803	0.662	0.971	0.904	0.976	0.774	0.75
150	0.968	0.694	0.814	0.682	0.981	0.913	0.995	0.866	0.80
160	0.998	0.834	0.856	0.683	0.981	0.938	0.994	0.882	0.75
200	0.994	0.817	0.871	0.709	0.982	0.967	1.000	0.916	0.80

Abbreviations: OPM, optimal pair matching; OPM-RA, OPM plus regression adjustment; OFM, optimal full matching; OFM-RA, OFM plus regression adjustment; SBC, sub-classification; SBC-RA, SBC plus regression adjustment; IPTW, inverse probability of treatment weighting; IPTW-RA, IPTW plus regression adjustment.

Legend: π , proportion of untreated individuals.

APPENDIX 3

```

/*****
/*      OPTIMAL PAIR MATCHING
*****/

/*MACROS*/

%macro means(data,var,class,dig);
ods trace on;
proc means mean std stderr sum p95 q1 median q3 min max n nmiss range
MAXDEC=&dig data=&data;
class &class;
var &var;
ods output summary=&var&class;
run;
ods trace off;
%mend means;

/*SETTING FIRST ROW OF DATASET FOR OUTPUT RESULTS*/

data temp;
INPUT ordsamp true_eff

/*1*/ match_noadj_est  match_noadj_se  match_noadj_lowb match_noadj_upb
match_noadj_incl match_noadj_p  match_noadj_pow
/*2*/ match_est match_se match_lowb match_upb match_incl match_p match_pow
datalines ;
. . . . .
. . . . .
run;

data shell;
set temp;
run;

%macro scen_sim(samplesize,nsamp,scr,true_eff);

data temp&samplesize;
set temp;
run;

/*IMPORT MATCHED DBs*/

PROC IMPORT OUT= WORK.comptot
            DATAFILE= <file path>
            DBMS=TAB REPLACE;
            GETNAMES=YES;
            DATAROW=2;
RUN;

data comptot;
set comptot;
Y1=&scr;
run;
```

```

%macro simul(nsamp);

%do j=1 %to &nsamp;
%let ordsamp=&j;

data resout_sampsize&sampsize;
set shell;
run;

data resout_sampsize&sampsize;
set resout_sampsize&sampsize;
ordsamp=&ordsamp;
run;

/*****
/*ESTIMATING TREATMENT EFFECT
*****/

data comp&j;
set comptot;
where ordsamp=&ordsamp;
Yl=&scr;
run;

/* 1. matched but not adjusted*/

proc ttest data =comp&j;
weight weights;
class treat;
var Yl;
ods output  ConfLimits=est_sbcYl&j    TTests=psbcYl&j;
run;

data est_sbcYl&j;
set est_sbcYl&j;
if class="Diff (1-2)"
then MeanrevYl=-1*Mean;
run;

data _null_;
set est_sbcYl&j ;
if Method="Satterthwaite" then
call symput ('match_noadj_est',MeanrevYl);
if Method="Satterthwaite" then
call symput ('match_noadj_lowb',UpperCLMean ); /*as the mean is of opposite
sign*/
if Method="Satterthwaite" then
call symput ('match_noadj_upb',LowerCLMean); /*as the mean is of opposite sign*/
run;

data _null_;
set psbcYl&j ;
if Method="Satterthwaite" then
call symput ('wsrp',Probt);
run;

```

```

data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
match_noadj_est=&match_noadj_est;
match_noadj_lowb=&match_noadj_lowb*(-1); /*as original value needs to be
inverted*/
match_noadj_upb=&match_noadj_upb*(-1); /*as original value needs to be
inverted*/
match_noadj_p=&wsrp;
if &wsrp<0.05 & &wsrp^=. then match_noadj_pow=1;
if &wsrp>=0.05 then match_noadj_pow=0;
run;

```

/*2. REGRESSION ON MATCHED DATA with ALL OBSERVED VARIABLES*/

```

proc glm data=comp&j;
weight weights;
model yl=treat dur age1 male education_high1 z1/ CLPARM;
ods output ParameterEstimates=pars&j;
run;
quit;

```

```

data _null_;
set pars&j;
if Parameter="treat" then
call symput ('match_est',Estimate);
call symput ('match_lowb',LowerCL);
call symput ('match_upb',UpperCL);
call symput ('wsrp',Probt);
run;

```

```

data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
match_est=&match_est;
match_lowb=&match_lowb;
match_upb=&match_upb;
match_p=&wsrp;
if &wsrp<0.05 & &wsrp^=. then match_pow=1;
if &wsrp>=0.05 then match_pow=0;
run;

```

```

proc append base=temp&samplesize data=resout_sampsize&samplesize;
run;
%end;
%mend simul;

```

```
%simul(&nsamp)
```

```

data outdb_sample&samplesize;
set temp&samplesize;
if ordsamp=. then delete;
run;
/*COMPUTING BIAS*/

```

```

data outdb_sample&samplesize;
set outdb_sample&samplesize;
match_noadj_bias=match_noadj_est-&>true_eff;
match_bias=match_est-&>true_eff;

```

```
run;
```

```
data outdb_sample&samplesize;  
set outdb_sample&samplesize;  
match_noadj_rmse=sqrt((match_noadj_est-&true_eff)*(match_noadj_est-&true_eff));  
match_rmse=sqrt((match_est-&true_eff)*(match_est-&true_eff));  
run;
```

```
PROC EXPORT DATA= work.outdb_sample&samplesize  
            OUTFILE= <file path>  
            DBMS=TAB REPLACE;  
RUN;
```

```
/* OVERALL RESULTS */
```

```
options orientation=portrait;  
filename myrtf <file path> ;  
ods rtf body=myrtf startpage=off style=styles.utex;
```

```
ods rtf text="Estimates";
```

```
%means(outdb_sample&samplesize,match_noadj_est,,3)  
%means(outdb_sample&samplesize,match_est,,3)
```

```
ods rtf text="Coverage probabilities C.I. 95%";
```

```
%means(outdb_sample&samplesize,match_noadj_incl,,3)  
%means(outdb_sample&samplesize,match_incl,,3)
```

```
ods rtf text="Bias";
```

```
%means(outdb_sample&samplesize,match_noadj_bias,,3)  
%means(outdb_sample&samplesize,match_bias,,3)
```

```
ods rtf text="Root mean squared error ";
```

```
%means(outdb_sample&samplesize,match_noadj_rmse,,3)  
%means(outdb_sample&samplesize,match_rmse,,3)
```

```
ods rtf text="Power at 0.05 alpha level";
```

```
%means(outdb_sample&samplesize,match_noadj_pow,,3)  
%means(outdb_sample&samplesize,match_pow,,3)
```

```
ods rtf close;
```

```
%mend scen_sim;
```

```
%scen_sim(<n treated>,<overall sample size>,<score>,<true_ATT>)
```

```

/*****
/*      OPTIMAL FULL MATCHING
*****/

/*SETTING FIRST ROW OF DATASET FOR OUTPUT RESULTS*/

data temp;
INPUT ordsamp true_eff

/*1*/ match_noadj_est  match_noadj_se  match_noadj_lowb match_noadj_upb
match_noadj_incl match_noadj_p  match_noadj_pow
/*2*/ match_est match_se match_lowb match_upb match_incl match_p match_pow
datalines ;
. . . . .
. . . . .
run;

data shell;
set temp;
run;

%macro scen_sim(samplesize,nsamp,scr,true_eff);

data temp&samplesize;
set temp;
run;

/*IMPORT MATCHED DBs*/

PROC IMPORT OUT= WORK.comptot
            DATAFILE= <file path>
            DBMS=TAB REPLACE;
            GETNAMES=YES;
            DATAROW=2;
RUN;

data comptot;
set comptot;
Y1=&scr;
run;

%macro simul(nsamp);

%do j=1 %to &nsamp;
%let ordsamp=&j;

data resout_sampsize&samplesize;
set shell;
run;

data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
ordsamp=&ordsamp;
run;

```

```

/*****
/*ESTIMATING TREATMENT EFFECT
*****/

data comp&j;
set comptot;
where ordsamp=&ordsamp;
Yl=&scr;
run;

/* 1. matched but not adjusted*/

proc ttest data = comp&j;
weight weights;
class treat;
var Yl;
ods output  ConfLimits=est_fullYl&j    TTests=pfullYl&j;
run;

data est_fullYl&j (Compress=Yes);
set est_fullYl&j;
if class="Diff (1-2)"
then MeanrevYl=-1*Mean;
run;

data _null_;
set est_fullYl&j;
if Method="Pooled" then
call symput ('match_noadj_est',MeanrevYl);
if Method="Pooled" then
call symput ('match_noadj_lowb', UpperCLMean); /*as the mean is of opposite
sign*/
if Method="Pooled" then
call symput ('match_noadj_upb',LowerCLMean); /*as the mean is of opposite sign*/
run;

data _null_;
set pfullYl&j;
call symput ('wsrp',Probt);
run;
data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
match_noadj_est=&match_noadj_est;
match_noadj_lowb=&match_noadj_lowb*(-1); /*as original value needs to be
inverted*/
match_noadj_upb=&match_noadj_upb*(-1); /*as original value needs to be
inverted*/
match_noadj_p=&wsrp;
if &wsrp<0.05 & &wsrp^=. then match_noadj_pow=1;
if &wsrp>=0.05 then match_noadj_pow=0;
run;

/*2. REGRESSION ON MATCHED DATA with ALL OBSERVED VARIABLES*/

proc glm data=comp&j;
weight weights;
model yl=treat dur age1 male  education_high1  z1/ CLPARM;

```



```
ods output ParameterEstimates=pars&j;
run;
quit;
```

```
data _null_;
set pars&j;
if Parameter="treat" then
call symput ('match_est',Estimate);
call symput ('match_lowb',LowerCL);
call symput ('match_upb',UpperCL);
call symput ('wsrp',Probt);
run;
```

```
data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
ordsamp=&ordsamp;
match_est=&match_est;
match_lowb=&match_lowb;
match_upb=&match_upb;
match_p=&wsrp;
if &wsrp<0.05 & &wsrp^=. then match_pow=1;
if &wsrp>=0.05 then match_pow=0;
run;
```

```
proc append base=temp&samplesize data=resout_sampsize&samplesize;
run;
%end;
%mend simul;

%simul(&nsamp)
```

```
data outdb_sample&samplesize;
set temp&samplesize;
if ordsamp=. then delete;
run;
```

```
/*COMPUTING BIAS*/
```

```
data outdb_sample&samplesize;
set outdb_sample&samplesize;
match_noadj_bias=match_noadj_est-&>true_eff;
match_bias=match_est-&>true_eff;
run;
```

```
data outdb_sample&samplesize;
set outdb_sample&samplesize;
match_noadj_rmse=sqrt((match_noadj_est-&>true_eff)*(match_noadj_est-&>true_eff));
match_rmse=sqrt((match_est-&>true_eff)*(match_est-&>true_eff));
run;
```

```
PROC EXPORT DATA= work.outdb_sample&samplesize
OUTFILE= <file path>
DBMS=TAB REPLACE;

RUN;
```

```

/* OVERALL RESULTS */

options orientation=portrait;
filename myrtf <file path> ;
ods rtf body=myrtf startpage=off style=styles.utext;

ods rtf text="Estimates";

%means(outdb_sample&samplesize,match_noadj_est,,3)
%means(outdb_sample&samplesize,match_est,,3)

ods rtf text="Coverage probabilities C.I. 95%";

%means(outdb_sample&samplesize,match_noadj_incl,,3)
%means(outdb_sample&samplesize,match_incl,,3)

ods rtf text="Bias";

%means(outdb_sample&samplesize,match_noadj_bias,,3)
%means(outdb_sample&samplesize,match_bias,,3)

ods rtf text="Root mean squared error ";

%means(outdb_sample&samplesize,match_noadj_rmse,,3)
%means(outdb_sample&samplesize,match_rmse,,3)

ods rtf text="Power at 0.05 alpha level";

%means(outdb_sample&samplesize,match_noadj_pow,,3)
%means(outdb_sample&samplesize,match_pow,,3)

ods rtf close;

%mend scen_sim;

%scen_sim(<n treated>,<overall sample size>,<score>,<>true_ATT>)

```

```

/*****
/*      SUB-CLASSIFICATION ON THE PROPENSITY SCORES
*/
*****/

/*SETTING FIRST ROW OF DATASET FOR OUTPUT RESULTS*/

data temp;
INPUT ordsamp true_eff

/*1*/ match_noadj_est  match_noadj_se  match_noadj_lowb match_noadj_upb
match_noadj_incl match_noadj_p  match_noadj_pow
/*2*/ match_est match_se match_lowb match_upb match_incl match_p match_pow
datalines ;
. . . . .
. . . . .
run;

data shell;
set temp;
run;

%macro scen_sim(samplesize,nsamp,scr,true_eff);

data temp&samplesize;
set temp;
run;

/*IMPORT MATCHED DBs*/

PROC IMPORT OUT= WORK.comptot
            DATAFILE= <file path>
            DBMS=TAB REPLACE;
            GETNAMES=YES;
            DATAROW=2;
RUN;

data comptot;
set comptot;
Yl=&scr;
run;

%macro simul (nsamp);

%do j=1 %to &nsamp;
%let ordsamp=&j;

```

```

data resout_sampsize&samplesize;
set shell;
run;

data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
ordsamp=&ordsamp;
run;

/*****
/* /*ESTIMATING TREATMENT EFFECT */
*****/

data comp&j;
set comptot;
where ordsamp=&ordsamp;
Yl=&scr;
run;

/* 1. matched but not adjusted*/

proc ttest data =comp&j;
weight weights;
class treat;
var Yl;
ods output  ConfLimits=est_sbcYl&j    TTests=psbcYl&j;
run;

data est_sbcYl&j;
set est_sbcYl&j;
if class="Diff (1-2)"
then MeanrevYl=-1*Mean;
run;

data _null_;
set est_sbcYl&j ;
if Method="Satterthwaite" then
call symput ('match_noadj_est',MeanrevYl);
if Method="Satterthwaite" then
call symput ('match_noadj_lowb',UpperCLMean ); /*as the mean is of opposite
sign*/
if Method="Satterthwaite" then
call symput ('match_noadj_upb',LowerCLMean); /*as the mean is of opposite sign*/
run;

data _null_;
set psbcYl&j ;
if Method="Satterthwaite" then
call symput ('wsrp',Probt);
run;

data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
match_noadj_est=&match_noadj_est;
match_noadj_lowb=&match_noadj_lowb*(-1); /*as original value needs to be
inverted*/
match_noadj_upb=&match_noadj_upb*(-1); /*as original value needs to be
inverted*/
match_noadj_p=&wsrp;
if &wsrp<0.05 & &wsrp^=. then match_noadj_pow=1;

```

```

if &wsrp>=0.05 then match_noadj_pow=0;
run;

```

```

/*2. REGRESSION ON MATCHED DATA with ALL OBSERVED VARIABLES*/

```

```

proc glm data=comp&j;
weight weights;
model yl=treat dur agel male education_high1 z1/ CLPARM;
ods output ParameterEstimates=pars&j;
run;
quit;

```

```

data _null_;
set pars&j;
if Parameter="treat" then
call symput ('match_est',Estimate);
call symput ('match_lowb',LowerCL);
call symput ('match_upb',UpperCL);
call symput ('wsrp',Probt);
run;

```

```

data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
match_est=&match_est;
match_lowb=&match_lowb;
match_upb=&match_upb;
match_p=&wsrp;
if &wsrp<0.05 & &wsrp^=. then match_pow=1;
if &wsrp>=0.05 then match_pow=0;
run;

```

```

proc append base=temp&samplesize data=resout_sampsize&samplesize;
run;
%end;
%mend simul;

```

```

%simul(&nsamp)

```

```

data outdb_sample&samplesize;
set temp&samplesize;
if ordsamp=. then delete;
run;

```

```

/*COMPUTING BIAS*/

```

```

data outdb_sample&samplesize;
set outdb_sample&samplesize;
match_noadj_bias=match_noadj_est-&>true_eff;
match_bias=match_est-&>true_eff;
run;

```

```

data outdb_sample&samplesize;
set outdb_sample&samplesize;

```

```

match_noadj_rmse=sqrt((match_noadj_est-&true_eff)*(match_noadj_est-&true_eff));
match_rmse=sqrt((match_est-&true_eff)*(match_est-&true_eff));
run;

```

```

PROC EXPORT DATA= work.outdb_sample&samplesize
            OUTFILE= <file path>
            DBMS=TAB REPLACE;

RUN;

```

```

/* OVERALL RESULTS */

```

```

options orientation=portrait;
filename myrtf <file path>;
ods rtf body=myrtf startpage=off style=styles.utext;

```

```

ods rtf text="Estimates";

```

```

%means(outdb_sample&samplesize,match_noadj_est,,3)
%means(outdb_sample&samplesize,match_est,,3)

```

```

ods rtf text="Coverage probabilities C.I. 95%";

```

```

%means(outdb_sample&samplesize,match_noadj_incl,,3)
%means(outdb_sample&samplesize,match_incl,,3)

```

```

ods rtf text="Bias";

```

```

%means(outdb_sample&samplesize,match_noadj_bias,,3)
%means(outdb_sample&samplesize,match_bias,,3)

```

```

ods rtf text="Root mean squared error ";

```

```

%means(outdb_sample&samplesize,match_noadj_rmse,,3)
%means(outdb_sample&samplesize,match_rmse,,3)

```

```

ods rtf text="Power at 0.05 alpha level";

```

```

%means(outdb_sample&samplesize,match_noadj_pow,,3)
%means(outdb_sample&samplesize,match_pow,,3)

```

```

ods rtf close;

```

```

%mend scen_sim;

```

```

%scen_sim(<n treated>,<overall sample size>,<score>,<true_ATT>)

```

```

/*****
/*      INVERSE PROBABILITY OF TREATMENT WEIGHTING      */
*****/

/*SETTING FIRST ROW OF DATASET FOR OUTPUT RESULTS*/

data temp;
INPUT ordsamp true_eff

/*1*/ match_noadj_est  match_noadj_se  match_noadj_lowb match_noadj_upb
match_noadj_incl match_noadj_p  match_noadj_pow
/*2*/ match_est match_se match_lowb match_upb match_incl match_p match_pow
datalines ;
. . . . .
. . . . .
run;

data shell;
set temp;
run;

%macro scen_sim(samplesize,nsamp,scr,true_eff);

/*IMPORT MATCHED DBs*/

PROC IMPORT OUT= WORK.comptot
            DATAFILE= <file path>
            DBMS=TAB REPLACE;
            GETNAMES=YES;
            DATAROW=2;
RUN;
data comptot;
set comptot;
Y1=&scr;
run;

%macro simul(nsamp);

%do j=1 %to &nsamp;
%let ordsamp=&j;

data resout_sampsize&samplesize;
set shell;
run;

data resout_sampsize&samplesize;
set resout_sampsize&samplesize;
ordsamp=&ordsamp;
run;

```

```

/*****
/* ESTIMATING TREATMENT EFFECT */
*****/

data comp&j;
set comptot;
where ordsamp=&ordsamp;
run;

/* 1. matched but not adjusted*/

data comp_iptw;
set comp&j;
wIPW= treat+(((1-treat)*ptr)/(1-ptr));
run;

%sort(comp_iptw,treat,1);
proc ttest data =comp_iptw;
weight wIPW;
class treat;
var Y1;
ods output  ConfLimits=est_iptwY1&j  TTests=piptwY1&j;
run;

data est_iptwY1&j;
set est_iptwY1&j;
if class="Diff (1-2)"
then MeanrevY1=-1*Mean;
run;

data _null_;
set est_iptwY1&j;
if Method="Satterthwaite" then
call symput ('match_noadj_est',MeanrevY1);
if Method="Satterthwaite" then
call symput ('match_noadj_lowb',UpperCLMean ); /*as the mean is of opposite
sign*/
if Method="Satterthwaite" then
call symput ('match_noadj_upb',LowerCLMean); /*as the mean is of opposite sign*/
run;

data _null_;
set piptwY1&j;
if Method="Satterthwaite" then
call symput ('wsrp',Probt);
run;

data resout_sampsize&samplesize ;
set resout_sampsize&samplesize;
match_noadj_est=&match_noadj_est;
match_noadj_lowb=&match_noadj_lowb*(-1); /*as original value needs to be
inverted*/
match_noadj_upb=&match_noadj_upb*(-1); /*as original value needs to be
inverted*/
if ordsamp=&ordsamp & &>true_eff>match_noadj_lowb & &>true_eff< match_noadj_upb
then match_noadj_incl=1;
else match_noadj_incl=0;
match_noadj_p=&wsrp;
if &wsrp<0.05 & &wsrp^=. then match_noadj_pow=1;
if &wsrp>=0.05 then match_noadj_pow=0;
run;

```



```
/*2. REGRESSION ON MATCHED DATA with ALL OBSERVED VARIABLES*/
```

```
data comp_iptw;  
set comp&j;  
wIPW= treat+(((1-treat)*ptr)/(1-ptr));  
run;
```

```
proc glm data=comp_iptw;  
weight wIPW;  
model Y1= treat dur agel male education_high1 z1/ CLPARM;  
ods output ParameterEstimates=pars1&j;;  
run;  
quit;
```

```
data _null_;  
set pars1&j;;  
if parameter='treat' then  
call symput ('par1',Estimate);  
if parameter='treat' then  
call symput ('match_lowb',LowerCL);  
if parameter='treat' then  
call symput ('match_upb',UpperCL);  
if parameter='treat' then  
call symput ('pall',Probt);  
run;
```

```
data resout_sampsize&samplesize;  
set resout_sampsize&samplesize;  
ordsamp=&ordsamp;  
match_est=&par1;  
match_lowb=&match_lowb;  
match_upb=&match_upb;  
if ordsamp=&ordsamp & &true_eff>match_lowb & &true_eff< match_upb  
then match_incl=1;  
else match_incl=0;  
match_p=&pall;  
if &pall<0.05 & &pall^=. then match_pow=1;  
if &pall>=0.05 then match_pow=0;  
run;
```

```
proc append base=temp&samplesize data=resout_sampsize&samplesize;  
run;  
%end;  
%mend simul;
```

```
%simul(&nsamp)
```

```
data outdb_sample&samplesize;  
set temp&samplesize;  
if ordsamp=. then delete;  
run;
```

```

/*COMPUTING BIAS*/

data outdb_sample&samplesize;
set outdb_sample&samplesize;
match_noadj_bias=match_noadj_est-&true_eff;
match_bias=match_est-&true_eff;
run;

data outdb_sample&samplesize;
set outdb_sample&samplesize;
match_noadj_rmse=sqrt((match_noadj_est-&true_eff)*(match_noadj_est-&true_eff));
match_rmse=sqrt((match_est-&true_eff)*(match_est-&true_eff));
run;

PROC EXPORT DATA= work.outdb_sample&samplesize
            OUTFILE= <file path>
            DBMS=TAB REPLACE;
RUN;

/* OVERALL RESULTS */

options orientation=portrait;
filename myrtf <file path>;
ods rtf body=myrtf startpage=off style=styles.utext;

ods rtf text="Estimates";

%means(outdb_sample&samplesize,match_noadj_est,,3)
%means(outdb_sample&samplesize,match_est,,3)

ods rtf text="Coverage probabilities C.I. 95%";

%means(outdb_sample&samplesize,match_noadj_incl,,3)
%means(outdb_sample&samplesize,match_incl,,3)

ods rtf text="Bias";

%means(outdb_sample&samplesize,match_noadj_bias,,3)
%means(outdb_sample&samplesize,match_bias,,3)

ods rtf text="Root mean squared error ";

%means(outdb_sample&samplesize,match_noadj_rmse,,3)
%means(outdb_sample&samplesize,match_rmse,,3)

ods rtf text="Power at 0.05 alpha level";

%means(outdb_sample&samplesize,match_noadj_pow,,3)
%means(outdb_sample&samplesize,match_pow,,3)

ods rtf close;

%mend scen_sim;

%scen_sim(<n treated>,<overall sample size>,<score>,<true_ATT>)

```