Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2016 April ; 25(4): 453–461. doi:10.1002/pds.3983.

Evaluation of Propensity Scores, Disease Risk Scores, and Regression in Confounder Adjustment for the Safety of Emerging Treatment with Group Sequential Monitoring

Stanley Xu, PhD¹, Susan Shetterly, MS¹, Andrea J. Cook, PhD², Marsha A. Raebel, PharmD¹, Sunali Goonesekera, MS³, Azadeh Shoaibi, PhD, MHS⁴, Jason Roy, PhD⁵, and Bruce Fireman, MS⁶

¹Institute for Health Research, Kaiser Permanente Colorado, Denver, CO

²Biostatistics Unit, Group Health Research Institute, Seattle, WA

³Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA

⁴Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD

⁵Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA

⁶Division of Research, Kaiser Permanente Northern California, Oakland, CA

Abstract

Purpose—The objective of this study is to evaluate regression, matching and stratification on propensity score (PS) or disease risk score (DRS) in a setting of sequential analyses where statistical hypotheses are tested multiple times.

Methods—In a setting of sequential analyses, we simulated incident users and binary outcomes with different confounding strength, outcome incidence, and the adoption rate of treatment. We compared type I error rate, empirical power, and time to signal using the following confounder adjustments: 1) regression; 2) treatment matching (1:1, 1:4) on PS or DRS, and 3) stratification on PS or DRS. We estimated PS and DRS using lookwise and cumulative methods (all data up to the current look). We applied these confounder adjustments in examining the association between NSAIDS and bleeding.

Results—PS and DRS methods had similar empirical power and time to signal. However DRS methods yielded type I error rates up to 17% for 1:4 matching and 15.3% for stratification methods when treatment and outcome were common and confounding strength with treatment was stronger. When treatment and outcome were not common, stratification on PS and DRS and regression yielded 8–10% type I error rates and inflated empirical power. However when outcome and treatment were common both regression and stratification on PS outperformed other matching methods with type I error rates close to 5%.

Corresponding Author: Stanley Xu, PhD, Institute for Health Research, Kaiser Permanente Colorado, PO Box 378066, Denver, CO 80237, Phone: (303) 614-1210, Fax: (303) 614-1265, stan.xu@kp.org.

Conflict of interest: None declared

Conclusions—We suggest regression and stratification on PS when the outcomes and/or treatment is common, and use of matching on PS with higher ratios when outcome or treatment are rare or moderately rare.

Keywords

propensity score; disease risk score; matching; stratification; group sequential analyses

INTRODUCTION

Appropriate confounder adjustment is critical in post-market surveillance because patients are not randomized as in clinical trials. Monitoring newly marketed treatments can be particularly challenging due to patient, physician, or system-related factors such as low uptake in the early years of treatment availability, rare occurrence of adverse events, and inherent differences between patients treated early and those treated later. Thus, it is crucial when monitoring the safety of a new treatment to adjust for differences between treated and untreated individuals and consider treatment evolution over time.

Propensity scores (PS), the probability of receiving treatment given a set of known covariates, have been widely used in estimating treatments effects in observational studies.^{1–8} Through matching or stratification on PS, a less biased estimate of treatment effect can be obtained because covariate balance between treated and untreated subjects is improved. It has been shown in a cross-sectional setting that stratification on PS produced less biased estimates than the logistic regression estimates when there were seven or fewer events per confounder and results were comparable when there were more events per confounder.⁹ Although PS techniques are available for confounder adjustment in post-market surveillance, PS models generally perform poorly when few individuals are treated (e.g., in early years of medical product availability).^{10,11} Consequently, utility of PS in estimating risks of new treatment in early monitoring may be limited.

Recent literature suggests that applying disease risk scores (DRS) to balance confounders between treated and untreated patients may be particularly useful during early monitoring of a new treatment when exposure is low because DRS estimation does not depend on the uptake of the new treatment.^{1,12} Unlike PS that balance measured confounders between treatment groups, DRS balance disease risk by estimating the probability of occurrence of adverse events as a function of measured confounders and treatment status if both treated and untreated individuals are used.^{10,11,13} However, using DRS may result in inflated type I error rates (>5%) when the association of measured confounders with treatment is strong.^{14,15} Further, systematic evaluation of the pros and cons of using PS and DRS over long term monitoring of a new treatment is incomplete.

Thus we conducted simulation studies to compare PS and DRS methods, including matching, stratification, and regression in monitoring safety of an emerging treatment, while adjusting for multiple confounders in a sequential testing framework. We considered several factors including the uptake of a new treatment over time, baseline rates of adverse events, the association between treatment and confounders, the association between treatment and

outcomes, and two methods of estimating PS and DRS in sequential testing settings. We additionally examined these confounding adjustment methods in a real data example.

METHODS

Confounder adjustments

We focused on contrasting confounder adjustments using PS or DRS and also examined results from multiple regression adjustments. As a general overview, PS were estimated from a logistic regression model with the treatment status (Z) as the dependent variable; DRS were estimated from a logistic regression model with the outcome (Y) as the dependent variable; while multiple regression adjustments used a logistic regression model that included confounders as covariates. We provide additional details on confounder adjustments using PS or DRS below and expand these further in the Appendix.

One intent of active surveillance is to repeatedly assess over time whether the treatment of interest is associated with elevated risk of a particular outcome (i.e., sequential analyses). For a current assessment (or look), the risk adjustment score (i.e. PS or DRS) is estimated for those entering within the current look (e.g., year). There are choices as to what data are used to fit the current look's adjustment scores. For this study we explored two approaches¹⁶: 1) lookwise estimation which uses incident users within the current look and; 2) cumulative estimation which uses the cumulative population up to the current look k . Details of these two methods can be found in Appendix A.1.

Stratification and matching adjustments were completed at each sequential time period. We performed 1:1 and 1:4 matching on the closest values of the score. Only incident users in the current look were eligible to be matched and matched groups remained static in subsequent looks. The quality of the matches can be affected by the maximum permitted difference in scores (i.e., caliper). In the simulation study, we matched to the nearest score without a caliper because it was not feasible to choose an optimal caliper for each simulated dataset. To examine the potential effect of not using calipers in the simulations, we completed sensitivity analyses where the number of treatments remained the same, but the entire population increased tenfold to have more similar controls for matching. In the real data example a caliper of <0.05 was used in matching.

We also performed stratification analyses with score deciles. At each look, each subject was assigned to a stratum based on his/her score; the subject remained in the stratum throughout subsequent sequential analyses.

To examine the association between outcome and treatment, we used conditional logistic regression for matching and stratification methods for both PS and DRS.

Methods for group sequential estimation

We maintained an overall one-sided significance level of 0.05 in the sequential analyses using the Lan-Demets Group Sequential approach.¹⁷ A set of upper boundaries were obtained for sequential analyses. Details can be found in Appendix A.2.

SIMULATION STUDY

We performed simulations to compare confounder adjustment methods under various scenarios in which the frequency of treatment and outcome ranged from moderately rare to common (i.e., for a sample size of 5,000, about 50 per year for moderately rare treatment; about 50 per year for moderately rare outcome; about 750 per year for common treatment; about 220 per year for common outcome). Confounders were measured at entry and were included in estimating PS and DRS with lookwise and cumulative estimating methods.

Simulation of treatment

We simulated whether an individual received the treatment of interest ($Z=1$) or comparator treatment ($Z=0$) using a logistic regression model (1) below. We included age as a continuous variable (x_1) with a mean of 50 and a standard deviation of five, gender (x_2) with 70% of one gender, presence of acute disease (x_3), presence of chronic disease (x_4), and six other covariates ($x_5, \dots, x_{10}$) as binary variables, each being assigned a 5% prevalence. We defined monitoring time (t for year) as discrete times 1 to 10; t was included in the model (1) as a continuous variable for simulating treatment status.

$$\text{Logit}(Z=1|x_1, x_2, x_3 \dots) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_{10} x_{10} + \beta_{11} t + \beta_{12} x_4 t \quad (1)$$

where β_0 is the intercept, $\beta_1, \beta_2, \dots, \beta_{10}$ are coefficients for confounders $x_1, \dots, x_{10}$, and β_{11} is the coefficient for time t . Inclusion of the term $\beta_{11} t$ with positive β_{11} allows the uptake of a treatment to steadily increase over time. We also simulated the situation where sicker or healthier individuals tended to receive treatment earlier by including the interaction between chronic disease (x_4) and time (t).

Simulation of outcome

At each of the 10 discrete times, we simulated the outcome ($y=0$ or 1) using model (2).

$$\text{Logit}(y=1|Z, x_1, x_2, x_3 \dots) = \theta_0 + \theta_z Z + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots + \theta_{10} x_{10} \quad (2)$$

where θ_0 is the intercept, $\theta_1, \dots, \theta_{10}$ are the coefficients for confounders $x_1, \dots, x_{10}$, and θ_z is the parameter for the treatment.

Parameters for simulating treatment and outcome

Details of parameter coefficients are in Table 1. We simulated three different confounding strength scenarios. In the first, the strengths of the associations between confounders $x_1, \dots, x_4$ and treatment and between confounders $x_1, \dots, x_4$ and outcome were the same. In the second, confounders $x_1, \dots, x_4$ were more strongly associated with treatment. In the third, confounders $x_1, \dots, x_4$ were more strongly associated with outcome. To make valid comparisons of type I error rates and empirical power, we chose different intercepts for different confounding strengths so that the frequencies of treatments and outcomes were

similar. Representative examples of frequencies of moderately rare treatment and outcome as well as common treatment and outcome are shown in Table 2. We simulated 1000 replicates, each consisting of 5,000 per year for 10 years.

Evaluation measures

We evaluated the different confounder adjustment methods using the following metrics:

Type I error rate (false positive rate) was calculated as the percentage of simulated datasets that rejected the null hypothesis $\theta_z=0$ under which data were simulated at the significance level of 0.05.

Power to detect an association was calculated as the percentage of datasets that rejected the null hypothesis $\theta_z=0$ when simulations were performed under the true alternative $\theta_z=0.69$ (i.e., odds ratio (OR)=2).

Time to signal detection was determined from the start of monitoring (i.e., year one in simulation) to the time when the test statistic exceeded the boundary.

AN EXAMPLE

We additionally examined PS, DRS and regression adjustment methods in a real data example of severe gastrointestinal (GI) bleeding outcomes within six months of a new prescription of a non-selective anti-inflammatory drugs (NSAID) or a Cyclo-Oxygenase-2 (COX-2) inhibitor. Three health care sites contributed data from 2008 through 2011. We excluded persons with prior history of GI bleeding and captured age, gender, comorbidities and prescriptions for anti-coagulant drugs or drugs that reduce stomach acid as potential confounders. We present risk ratios for non-selective NSAIDs vs COX-2 inhibitors.

RESULTS

Simulation results

We observed no difference in type I error rates, empirical power or time to signal between lookwise and cumulative approaches in estimating PS and DRS. Thus, we only present the results for the cumulative approach.

Comparing PS and DRS, we generally observed no difference between PS and DRS approaches (Figures 1–3). Exceptions occurred in some type I error comparisons where DRS methods yielded higher type I errors than PS methods when treatment was common and confounding strength with treatment and outcome was not equal (Figure 1). For example, when treatment was common and outcome was moderately rare (panel C of Figure 1), the type I error rate for 1:1 matching on PS was 3.8% for PS while it was 7.2% for DRS, and when confounding strength was stronger with treatment, the type I error rate for 1:4 matching was 9.2% for PS while it was 12.2% for DRS; when confounding strength was stronger with outcome, the type I error rate for 1:4 matching was 8.6% for PS while it was 11.1% for DRS. When both treatment and outcome were common (panel D of Figure 1), type I error rates of DRS method was also higher than those of PS method. When

confounding strength was stronger with treatment than with outcome, the type I error rate for 1:1 matching on PS was 4.7% for PS while it was 11.2% for DRS; the type I error rate for 1:4 matching was 10.3% for PS while it was 17.0% for DRS. When confounding strength was stronger with outcome than with treatment, the type I error rate for 1:1 matching on PS was 5.7% for PS while it was 8.0% for DRS; the type I error rate for 1:4 matching was 8.3% for PS while it was 10.9% for DRS (panel D of Figure 1). When stratification was used, we also observed higher type I error rates of DRS than those of PS when confounding strength was not equal.

Figure 2 shows that PS and DRS yielded comparable power. When outcomes were common power is 100% for both PS and DRS (Panel D of Figure 2). Figure 3 shows time to signal results. Times to signal were shorter as treatments and outcomes became common but within analysis strategy PS and DRS provided similar times to signal. In addition, we did not observe an advantage for PS over DRS when treatment was common and outcome was moderately rare; nor did we observe an advantage for DRS over PS when outcome was common and treatment was moderately rare.

In regard to matching ratios, when treatment and outcome were moderately rare, 1:4 matching had type I error rates close to the nominal value 0.05 while 1:1 matching had deflated type I error rates (panel A of Figure 1). Similarly, 1:4 matching always had higher empirical power and shorter time to signal than 1:1 matching (panel A of Figures 2 and 3). In the common treatment and outcome setting, 1:4 matching yielded type I error rates much higher than the nominal value 0.05 if confounding strength with treatment was stronger than with outcome. Because a more limited control pool will exist in the common treatment setting, we completed sensitivity analyses utilizing a larger population base. In the simulations where the number of treatments remained the same, but the entire population increased tenfold, type I error rates were reduced: from 4.7% to 3.4% for 1:1 matching on PS, from 10.3% to 6.4% for 1:4 matching on PS, from 11.2% to 5.6% for 1:1 matching on DRS, and from 17.0% to 10.8% for 1:4 matching on DRS.

In the moderately rare treatment and outcome setting, stratification yielded higher type I error rates and inflated empirical power compared to 1:4 matching on PS or DRS (Figures 1 and 2). In the common treatment and outcome settings, when confounding strength with treatment was stronger, stratification on PS yielded a type I error rate close to 0.05 while 1:4 matching on PS yielded higher type I error rates; both stratification and 1:4 matching on DRS yielded inflated type I error rates (Panel D of Figure 1). In other confounding strength scenarios, stratification on PS or DRS had type I error rates close to 0.05.

As was observed in the PS stratified analyses, regression yielded higher type I error rates and inflated empirical power when the treatment of interest and the outcome were moderately rare. In the common outcome and treatment setting regression method had appropriate type I errors (Panel D of Figure 1). Regression method had similar time to signal to stratification results in all settings (Figure 3).

We also evaluated these confounder adjustment methods when treatment and/or outcome were rare. Details can be found in Appendix A.3.

Example results

In the NSAID and bleeding real data example, we captured 2,688,965 incident NSAID users of whom 2–3% used a COX-2 drug and about 0.14% had GI bleeds. Similar to simulation results, we observed no difference for lookwise versus cumulative estimation methods. Neither stratification on PS nor stratification on DRS signaled. Significant association was detected in the 1:4 matching on DRS in the second year (OR=1.32 non-selective NSAID vs COX-2) while 1:4 matching on PS didn't signal until the third year (OR=1.23).

DISCUSSION

In this study, we observed higher type I error rates when using DRS approaches compared to PS methods for confounder adjustment when confounding strength was not equal, particularly when treatment and outcome were common, and when the confounding strength with treatment was stronger. Our results did not substantially differ when PS and DRS were estimated using the lookwise approach versus the cumulative approaches. These results are consistent with findings of a cross-sectional observational study by Arbogast and Ray¹³ and with findings of others that type I error rates are inflated with DRS when the association of measured confounders with treatment is strong.^{14,15} Our additional simulations where the number of treatments remained the same, but the entire population was tenfold increased showed reduction of type I errors, indicating the inflation of type I error rate was partially due to the unavailability of optimal comparators for matching or stratification.

Contrary to our expectation, DRS methods did not have advantages over PS when the outcome was common and the treatment was moderately rare, nor did PS methods show advantages when the outcome was moderately rare and treatment was common. Because the cumulative approach used data up to the current year to estimate PS or DRS scores, it accommodated more confounders and could produce scores with less uncertainty compared to lookwise approach. However, as the process of matching and stratification uses only predicted PS/DRS scores and does not typically account for the uncertainty of these scores¹⁸, we did not observe substantial differences between lookwise and cumulative approaches in our study. Empirical power increased with higher matching ratios, especially when both treatments and outcomes were moderately rare. A high matching ratio is able to preserve more cases but in real data applications, a higher matching ratio may not be feasible if there is a limited pool of appropriate controls.

PS approaches have been advocated by many because a single PS can be used for multiple outcomes, however different outcomes may require different exclusion criteria or variable adjustment methods that may limit duplicate use of PS. Similarly, a single DRS model may be advantageous when studying multiple treatments but only if the population for the DRS model is appropriate for the differing treatments. Although available, methods for creating PS are not trivial when there are multiple treatments or multiple treatment levels. In such instances, DRS methods may be advantageous because a single DRS model can be fit by including the categorical treatment variable.

Stratification keeps all cases in the analytic datasets, but it may not be the optimal method to adjust for confounders in sequential analyses of observational studies. We observed that

stratification based on either PS or DRS scores resulted in inflated type I error rates and empirical power, especially when both outcomes and treatments were moderately rare. However, when the outcome and/or treatment are more common stratification on PS showed valid type I error and was advantageous over matching methods and similar to regression.

A limitation of this study is that in simulation, models reflect the simulated data and so are appropriately specified whereas, in real data settings, the correct model specifications are typically unknown. PS and DRS are more likely to yield inconsistent results when important confounders are either missing or unknown to a greater degree in one model than in the other. In addition, the simulations did not apply a caliper when matching on PS or DRS. We also did not consider time-varying covariates in adjusting for confounders. For long term outcomes, covariates measured upon treatment initiation may have changed, which may influence the results of DRS method and regression methods. Finally, we did not simulate the situation where historical data could be used in building DRS model.

Strengths of this study include the use of consistent analytic methods and the utilization of the same information including number of subjects, number of outcomes and confounders in the comparisons of PS and DRS. In addition, we purposely calculated the upper boundaries using eligible subjects and number of cases, and then applied the same boundaries in comparing all confounding adjustment approaches so that the observed differences in type I error rate, empirical power, and time to signal were not due to different signal criteria.

In conclusion, our results suggest PS 1:4 matching is preferable to stratification to adjust for potential confounders in the moderately rare outcome and treatment setting, but stratification on PS and regression may be preferable in the more common outcome or treatment setting. In situations where treatment and outcomes are not common, we recommend 1:4 rather than 1:1 matching whenever feasible. We also recommend performing analyses using both PS and DRS methods if possible since model misspecifications are likely in real world settings and may impact results. Although we obtained similar results for cumulative and lookwise approaches, we recommend using the cumulative approach in safety monitoring of emerging treatments because the lookwise approach may require exclusion of important confounders due to non-convergence in building PS or DRS models because of few outcomes or treatments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by Mini-Sentinel which is funded by the FDA through Department of Health and Human Services Contract Number HHSF223200910006I. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the FDA. We thank other members of the Confounder Adjustment Workgroup for their contributions: Eric Frimpong and Brad McEvoy. Xu was also supported by NIH/NCRR Colorado CTSI Grant Number UL1 RR025780.

Abbreviations

PS propensity score

DRS	disease risk score
FDA	US Food and Drug Administration
NSAIDs	non-selective anti-inflammatory drugs
GI	gastrointestinal
COX-2	Cyclo-Oxygenase-2

References

1. Schneeweiss S, Gagne JJ, Glynn RJ, Ruhl M, Rassen JA. Assessing the Comparative Effectiveness of Newly Marketed Medications: Methodological Challenges and Implications for Drug Development. *Clinical Pharmacology and Therapeutics*. 2011; 90(6):777–790. [PubMed: 22048230]
2. D’Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*. 1998; 17:2265–81. [PubMed: 9802183]
3. Perkins SM, Tu W, Underhill MG, et al. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiol Drug Saf*. 2000; 9:93–101. [PubMed: 19025807]
4. Hirano K, Imbens G. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcomes Res Methodol*. 2001; 2:259–78.
5. Allen-Ramey FC, Duong PT, Goodman DC, et al. Treatment effectiveness of inhaled corticosteroids and leukotriene modifiers for patients with asthma: an analysis from managed care data. *Asthma Proc*. 2003; 24:43–51. [PubMed: 12635577]
6. Lipkovic I, Adams DH, Mallinckrodt C, et al. Evaluating dose response from flexible dose clinical trials. *BMC Psychiatry*. 2008; 8:1–9. [PubMed: 18173833]
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–55.
8. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984; 79:516–24.
9. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003; 158:280–287. [PubMed: 12882951]
10. Arbogast PG, Kaltenbach L, Ding H, et al. Adjustment for multiple cardiovascular risk factors using a summary risk score. *Epidemiology*. 2008; 19(1):30–37. [PubMed: 18091000]
11. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res*. 2009; 18(1):67–80. [PubMed: 18562398]
12. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiol Drug Saf*. 2012; 21(Suppl 2):138–147. [PubMed: 22552989]
13. Arbogast PG, Ray WA. Performance of Disease Risk Scores, Propensity Scores, and Traditional Multivariable Outcome Regression in the Presence of Multiple Confounders. *Am J Epidemiol*. 2011; 174(5):613–620. [PubMed: 21749976]
14. Pike MC, Anderson J, Day N. Some insights into Miettinen’s multivariate confounder score approach to case-control study analysis. *Epidemiol Community Health*. 1979; 33(1):104–106.
15. Cook EF, Goldman L. Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *J Clin Epidemiol*. 1989; 42(4):317–324. [PubMed: 2723692]
16. Li L, Kulldorff M, Nelson JC, et al. A Propensity Score-Enhanced Sequential Analytic Method for Comparative Drug Safety Surveillance. *Statistics in Biosciences*. 2011; 3:45–62.
17. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70:659–63.

18. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. *Statistic in Medicine*. 2009; 28(1):94–112. DOI: 10.1002/sim.3460
19. Miettinen OS. Stratification by a multivariate confounder score. *Am J Epidemiol*. 1976; 104:609–620. [PubMed: 998608]
20. Harrell, FE. Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis. Springer-Verlag; New York: 2001.
21. Pocock SJ. Interim Analyses for randomized clinical-trials - The Group Sequential Approach. *Biometrics*. 1982; 38:153–62. [PubMed: 7082757]

Key points

- DRS methods had higher type I error rates than PS methods when treatment was common and confounding strength with treatment and outcome was not equal.
- DRS methods did not have clear advantages over PS when the treatment was early in uptake (e.g. moderately rare), nor did PS methods show clear advantages when the outcome was not common.
- When treatment and outcome were not common, stratification on PS and DRS and regression yielded 8–10% type I error rates and inflated empirical power.
- When outcome and treatment were common both regression and stratification on PS outperformed matching methods with type I error rates close to 5%.

Figure 1.

Type I error rates (%) with cumulative estimation method for computing PS and DRS under $OR=1$ (θ_z). ■ Equal confounding strength; ■ stronger confounding strength with treatment than with outcome; ▨ stronger confounding strength with outcome than with treatment.

Figure 2.

Empirical power (%) with cumulative estimation method for computing PS and DRS under $OR=2$ ($\theta_z = 0.693$). ■ equal confounding strength; ■ stronger confounding strength with treatment than with outcome; ■ stronger confounding strength with outcome than with treatment.

Figure 3. Mean time to signal ($\pm$ 1 standard deviation) with cumulative estimation method for computing PS and DRS under $OR=2$ ($\theta_z=0.693$). ■ equal confounding strength; ■ stronger confounding strength with treatment than with outcome; ▨ stronger confounding strength with outcome than with treatment.

Table 1

Parameter coefficients for different confounding strength

Parameters	Confounding strength scenarios					
	1 (equal strength)	2 (more toward treatment)		3 (more toward outcome)		
	Coefficients for the treatment probabilities	Coefficients for the outcome probabilities	Coefficients for the treatment probabilities	Coefficients for the outcome probabilities	Coefficients for the treatment probabilities	Coefficients for the outcome probabilities
Age (x_1)	0.002	0.002	0.01	0.002	0.002	0.01
Gender (x_2)	0.2	0.2	1.0	0.2	0.2	1.0
Acute Disease (x_3)	0.15	0.15	0.75	0.15	0.15	0.75
Chronic Disease (x_4)	0.15	0.15	0.75	0.15	0.15	0.75
Time (t)	0.1	NA	0.1	NA	0.1	NA
Chronic Disease time *	0.2	NA	0.2	NA	0.2	NA
x_5	0.1	0.1	0.1	0.1	0.1	0.1
x_6	0.1	0.1	0.1	0.1	0.1	0.1
x_7	0.1	0.1	0.1	0.1	0.1	0.1
x_8	0.1	0.1	0.1	0.1	0.1	0.1
x_9	0.1	0.1	0.1	0.1	0.1	0.1
x_{10}	0.1	0.1	0.1	0.1	0.1	0.1
treatment	NA	0, 0.69 *	NA	0, 0.69	NA	0, 0.69

* 0 for null hypothesis; 0.69 for OR=2.0.

Average number of outcomes ($\bar{y}$) and average number of participants receiving the treatment of interest (n) across years for two scenarios: 1) moderately rare treatment and moderately rare outcome and 2) common treatment and outcome scenarios under the null hypothesis (OR=1)

Table 2

year	Moderately rare treatment and moderately rare outcome						Common treatment and common outcome					
	Scenario 1*			Scenario 2			Scenario 3			Scenario 1		
	$\bar{y}$	n	n^-	$\bar{y}$	n	n^-	$\bar{y}$	n	n^-	$\bar{y}$	n	n^-
1	48	36	31	50	31	50	50	31	50	218	508	213
2	47	39	33	50	34	50	50	34	50	218	553	212
3	47	43	38	49	37	50	50	37	50	217	606	213
4	47	48	41	50	41	50	50	41	50	218	656	214
5	47	52	45	49	45	50	50	45	50	217	718	213
6	47	58	50	50	50	50	50	50	50	219	779	214
7	48	64	55	50	56	50	50	56	50	219	848	214
8	47	70	60	50	60	50	50	60	50	218	919	212
9	47	78	67	50	67	50	50	67	50	217	1000	214
10	47	86	74	50	74	50	50	74	50	218	1078	214

* scenario 1: equal strength; scenario 2: more toward treatment; scenario 3: more toward outcome.