Comparison of Cardinality Matching and Propensity Score Matching for Causal Inference in Observational Research

Stephen P Fortin (stephenfortin12@gmail.com)
Janssen R&D, LLC, Raritan, NJ,

Stephen S Johnston
Johnson & Johnson, New Brunswick

Martijn J Schuemie
Janssen R&D, LLC, Raritan, NJ,

Research article

Keywords: cardinality matching, propensity score matching, causal inference, residual bias, systematic error, sample size, balance, hypertension, ACEI, diuretic

Posted Date: October 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-94412/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on August 21st, 2021. See the published version at https://doi.org/10.1186/s12874-021-01365-z.

Abstract

Background: Cardinality matching (CM), a novel matching technique, finds the largest matched sample meeting prespecified balance criteria thereby overcoming limitations of propensity score matching (PSM) associated with limited covariate overlap, which are especially pronounced in studies with small sample sizes. The current study compares CM and PSM in terms of post-match sample size, covariate balance and residual confounding at progressively smaller sample sizes.

Methods: To evaluate CM and PSM within a comparative cohort study of new users of angiotensin-converting enzyme inhibitor (ACEI) and thiazide or thiazide-like diuretic monotherapy identified from a U.S. insurance claims database. Candidate covariates included patient demographics, and all observed prior conditions, drug exposures and procedures. Propensity scores were calculated using LASSO regression, and candidate covariates with non-zero beta coefficients in the propensity model were defined as matching covariates for use in CM. One-to-one matching was performed using progressively tighter parameter settings. Covariate balance was assessed using standardized mean differences. Hazard ratios were estimated using unconditional Cox models for negative control outcomes perceived as unassociated with treatment (e.g., true hazard ratio of 1). Residual confounding was assessed using the expected systematic error of the empirical null distribution of negative control effect estimates compared to the ground truth. Analyses were repeated within 10%, 1% and 0.5% subsample groups.

Results: A total of 172,117 patients (ACEI: 129,078; thiazide: 43,039) met the study criteria. Compared to PSM, CM was associated with increased sample retention except for analyses failing to converge to a matched sample. Although PSM achieved balance across all matching covariates within the full study population, substantial matching covariate imbalance was observed within the 1% and 0.5% subsample groups. Meanwhile, CM achieved matching covariate balance across all analyses. PSM was associated with better candidate covariate balance within the full study population. Otherwise, both matching techniques achieved comparable candidate covariate balance and expected systematic error.

Conclusion: CM found the largest matched sample meeting prespecified balance criteria while achieving comparable candidate covariate balance and residual confounding. At smaller sample sizes, CM achieved superior matching covariate balance. We recommend CM as an alternative to PSM in studies with small sample sizes.

Background

Randomization tends to produce comparable study groups in terms of both observed and unobserved covariates in controlled experimentation. Unfortunately, random assignment of treatment is conspicuously absent from observational studies[1]. In the absence of randomization, differences in covariate distributions between study groups may prevent valid statistical inference from data[2]. As such, a key component in the design of observational studies includes addressing the presence of confounding covariates to reduce study bias using statistical methods such as matching[3,4,5].

While propensity score matching (PSM) is the most ubiquitous matching technique for causal inference in observational research, the technique is subject to limitations. First, PSM is susceptible to substantial bias, large variance in estimates and poor sample retention in studies with limited overlap of covariate distributions between study groups[4,6,7]. Second, due to limited degrees of freedom, restrictions on the number of matching covariates used may be necessary to avoid model over-parameterization and overfitting[8]. These limitations are especially pronounced in studies with small sample sizes.

A novel matching method, cardinality matching (CM), uses recent advancements in integer programming to find the largest matched sample meeting a set of prespecified balance criteria[4]. For instance, CM solves for the optimal (i.e., largest) matched sample subject to investigator-defined constraints on the maximum standardized mean difference of covariates between study groups. By matching directly on the original covariates rather than propensity scores, CM handles issues of limited overlap of covariate distributions and maximizes sample size retention while meeting covariate balance criteria[4].

The current study compares the performance of CM and PSM in an observational study of new users of angiotensin converting enzyme inhibitor (ACEI) vs. thiazide or thiazide-like diuretic monotherapy. Both matching techniques are evaluated in terms of post-match sample size, candidate and matching covariate balance and residual confounding at progressively smaller sample sizes and more stringent parameter settings.

Methods

Study design and data source

We conducted a retrospective comparative new-user cohort study in the IBM® MarketScan® Commercial Claims and Encounters Database (CCAE), which primarily consists of de-identified, patient-level health data from over 142 million individuals enrolled in employer-sponsored health insurance plans in the United States. The CCAE database includes adjudicated health insurance claims (inpatient, outpatient, and prescription) and enrollment data from large employers and health plans who provide private insurance coverage. Data were standardized to the Observational Health and Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5.3, which maps international coding systems into standard vocabulary concepts[9].

The CCAE database consists of de-identified healthcare records. In the United States, retrospective analyses of the CCAE data are considered exempt from informed consent and institutional review board (IRB) approval as dictated by Title 45 Code of Federal Regulations, Part 46 of the United States, specifically 45 CFR 46.104 (d)(4).

Study population

We identified new users of ACEI and thiazide or thiazide-like diuretic monotherapy between October 1, 2015 and January 1, 2017. For each patient, we defined the *index* as the date of first drug exposure.

The study was limited to patients with a minimum of 365 days of continuous enrollment in the database prior to index. We required patients to have a recorded diagnosis for hypertension at or within 365 days prior to index (see Supplemental Appendix A for a list of codes used to query the database). As described in Suchard et al., new users were defined as patients whose first observed treatment for hypertension was ACEI or thiazide or thiazide-like diuretic monotherapy[10]. Patients with exposure to any other active ingredient listed within the five primary drug classes for the treatment of hypertension in the 2017 American College of Cardiology/American Heart Association (ACC/AHA) guidelines (i.e., ACEI, thiazide or thiazide-like diuretics, angiotensin receptor blockers, dihydropyridine calcium channel blockers, non-dihydropyridine calcium channel blockers) within 365 days prior or 7 days post-index were excluded[10,11].

Example outcome of interest

We examined the safety outcome of angioedema, which was identified from diagnoses recorded on inpatient and emergency room healthcare claim records. Patients with a recorded diagnosis for angioedema at or within any time prior to index were excluded from the study.

Time-at-risk

The time-at-risk window was defined based on the intention-to-treat principle, and patients were followed from day 1 post-index to the earliest of July 31, 2019 or end of continuous observation in the database[12]. Analyses were limited to patients with a minimum time-at-risk of 1 day.

Patient demographic and clinical characteristics

We measured patient demographics at index including age, grouped into categories in 5-year increments; sex; and index year and month. Patient clinical characteristics included all observed condition, drug exposure, measurement and observation codes occurring within a long-term or short-term window (i.e., at or within 365 or 30 days prior to index, respectively). Furthermore, we measured all observed drug exposures occurring within the time-at-risk window. All drug exposures were grouped at both the ingredient-level and according to the Anatomical Therapeutic Chemical (ATC) classification system. Patient comorbidities were measured using the Charlson Comorbidity Index (CCI)[13]. Finally, we measured the following disease severity and risk scores: Diabetes Complications Severity Index (DCSI), $CHADS_2$ score, and CHA_2DS_2-VASc score[14,15,16]. The CCI, DCSI, $CHADS_2$ score and CHA_2DS_2-VASc score were measured based on all observed conditions occurring prior to the end of the time-at-risk window.

Large-scale propensity score matching

Candidate covariates were defined as all aforementioned patient demographic and clinical characteristics, and heuristic feature selection was used to identify candidate covariates with a frequency greater than 0.1%. We developed propensity models using LASSO regression with 10-fold cross-validation for hyperparameter tuning including all candidate covariates identified through heuristic feature selection, and propensity scores were calculated using the propensity model[17]. New users of ACEI and thiazide or thiazide-like diuretic monotherapy were matched at a 1:1 ratio using greedy matching enforcing a caliper of 0.10 and 0.20 of the pooled standard deviation of the logit of propensity scores in two separate analyses. To facilitate comparisons between CM and PSM, we defined matching covariates as candidate covariates with non-zero beta coefficients in the propensity model.

Cardinality matching

Heuristic feature selection of candidate covariates was performed as previously described with one notable exception: due to memory constraints associated with CM, in analyses using the full study population, the heuristic feature selection used a frequency threshold of 2% instead of 0.1%. Specifically, CM failed to converge to a matched sample due to insufficient memory while attempting to match on approximately 220 million data points (172,117 patients and 1,237 matching covariates). The frequency threshold used within all subsample group analyses was consistent between CM and PSM.

Matching covariates – covariates used in the CM - were empirically selected; propensity scores were estimated as previously described and matching covariates were defined as candidate covariates with non-zero beta coefficients in the LASSO propensity model. CM utilizes advancements in optimization algorithms to solve for the largest sample size meeting prespecified balance criteria (e.g., maximum standardized mean difference [SMD] of matching covariates)[4]. We performed CM using the following prespecified balance criteria in four separate analyses: exact marginal distributional balance (i.e., fine balance; SMD = 0) and maximum SMD of 0.01, 0.05 and 0.10 of matching covariates between study groups.

All analyses were performed using an Amazon Web Services (AWS) Virtual Private Cloud (VPCx) m4.4xlarge Elastic Compute Cloud (EC2) instance. This instance included 16 2.3 GHz Intel® Xeon® vCPUs, 64 GiB of memory and a dedicated Elastic Block Storage (EBS) bandwidth of 2000 Mbps. Furthermore, all analyses were performed in R version 3.6.3 using the Health Analytics Data-to-Evidence Suite (HADES), Gurobi™ solver and cardmatch library[18].

Evaluation of post-match sample size

We evaluated patient retention in the matched samples after CM and PSM based on post-match sample size.

Evaluation of post-match covariate balance

The performance of CM and PSM were compared in terms of post-match covariate balance. In order to determine the level of balance achieved within covariates indirectly and directly adjusted during matching, candidate and matching covariate balance, respectively, were assessed separately. SMDs, as defined by Rosenbaum et. al (see Eq. 1), were used to assess the post-match balance of candidate and matching covariates; specifically,

$\text{SMD} = (\bar{x}_{\text{treatment}} - \bar{x}_{\text{comparator}}) / s_p$

where $\bar{x}_{treatment}$ and $\bar{x}_{comparator}$ represent the post-match covariate mean of treatment and comparator group, respectively, and s_p represents the pre-match covariate pooled standard deviation[19]. An absolute SMD less than 0.10 was considered balanced.

Evaluation of post-match residual confounding

Residual study bias due to unmeasured potential confounders and systematic error may still exist subsequent to CM or PSM[20,21]. To quantify the magnitude of residual study bias, we included a total of 105 negative control outcomes in our experiment believed to be caused by neither ACEIs nor thiazide or thiazide-like diuretics, which, therefore, have a true hazard ratio equal to 1[20]. These negative control outcomes were identified through a data-rich algorithm and manual clinical review (see Supplemental Appendix B for a list of negative control outcomes used in the current study)[22]. Hazard ratios were estimated for negative control outcomes as well as the example outcome of interest (angioedema) using unconditional Cox proportional hazards models in the matched samples

Comparing the estimated hazard ratios of the negative control outcomes to the ground truth (of no effect) provides insight into residual study bias. We assume the observed log hazard ratio ($^\wedge\theta_i$) depends on the log of the true effect size (θ_i), which is assumed to be 0, plus a systematic error component (β_i), and let τ_i denote the standard error corresponding to θ_i. Furthermore, we assume β_i to be distributed following a normal distribution with parameters μ and σ^2, which we estimate using the observed estimates (i.e., $^\wedge\theta_i$) of negative control outcomes[20]. In summary, we assume:

$^\wedge\theta_i \sim N\ (\theta_i + \beta i,\ \tau^2{}_i\ )$, and

$\beta_i \sim N\ (\mu,\ \sigma^2)$

To summarize the empirical null distribution into a single measure we computed the expected systematic error (ESE), defined as the expected absolute systematic error based on the estimated null distribution parameters:

$ESE = E(|\beta_i|)$

Given a finite number of negative control outcomes and uncertainty in estimated hazard ratios due to limited sample size, the distribution parameters and, therefore, the ESE come with uncertainty, which we quantified using Markov-Chain Monte Carlo and expressed as 95% credible intervals.

Analyses of angioedema outcome

Unconditional Cox proportional hazards models were used to compare the safety outcome of angioedema between study groups in the matched samples. All hazard ratio (HR) estimates, 95% confidence intervals (CI) and p-values were calibrated to incorporate the uncertainty expressed in the empirical null distribution of negative control outcomes[20,23]. We considered a two-sided calibrated pvalue < 0.05 to be statistically significant. For reference, we further examined uncalibrated effect estimates.

Analyses of subsample groups

All aforementioned analyses, with the exception of analyses of the angioedema outcome, were repeated in a series of progressively smaller subsample groups, including a 10%, 1% and 0.5% subsample group. The 10%, 1% and 0.5% subsample groups included 5, 50 and 100 subsample draws, respectively. Each subsample draw was performed by random sampling without replacement from the study population stratified by study comparison group.

Within each subsample draw, candidate covariates were defined as all aforementioned patient demographics and clinical characteristics observed within the respective subsample draw, and filtered using the aforementioned frequency thresholds. Propensity scores were estimated within each draw as previously described, and matching covariates were defined as those candidate covariates with non-zero beta coefficients in the propensity model for that draw. As such, a distinct set of candidate covariates and matching covariates were identified for each subsample draw; however, within individual subsample draws, candidate and matching covariates were consistent between CM and PSM.

For each subsample group, we assessed the average post-match sample size of their respective subsample draws. Meanwhile, candidate and matching covariate balance were assessed based on the SMD of covariates across all subsample draws within each subsample group considered jointly. Hazard ratios for negative control outcomes were estimated independently within each subsample draw using unconditional Cox proportional hazards models. Residual confounding was assessed based on the ESE of the empirical null distribution of negative control outcomes, which was derived from hazard ratio estimates considered jointly across all subsample draws within each subsample group. Analyses of the angioedema outcome were not performed across subsample groups due to insufficient occurrence of the outcome.

Results

Post-match sample size

The study inclusion criteria were met by 172,117 patients in the CCAE database, of which 129,078 (75.0%) and 43,039 (25.0%) were new users of ACEI and thiazide or thiazide-like monotherapy, respectively. Each subsample draw for the 10%, 1% and 0.5% subsample groups included 17,210 (ACEI: 12,907; thiazide or thiazide-like diuretic: 4,303), 1,720 (ACEI: 1290; thiazide or thiazide-like diuretic: 430) and 860 (ACEI: 645; thiazide or thiazide-like diuretic: 215) patients, respectively.

The average post-match sample size across all analyses is shown in Fig. 1. In the full study population, CM failed to converge to an optimal solution while requiring fine balance of matching covariates but was able to match every patient in the thiazide or thiazide-like diuretic group to a patient in the ACEI group (matched sample size = 86,078) at all other prespecified balance criteria. The use of more stringent balance criteria and a tighter caliper was associated with a slight reduction in post-match patient retention in CM and PSM, respectively, within subsample group analyses. With the exception of CM requiring fine balance of matching covariates, CM was associated with greater sample size retention as compared to PSM.

Post-match matching covariate balance

In the full study population, 1,237 matching covariates were identified by LASSO regression for analyses using PSM. Due to memory constraints associated with CM at larger sample sizes, the frequency threshold of heuristic feature selection used to limit candidate covariates considered during LASSO regression was increased from 0.1–2% for analyses within the full study population using CM, which led to the identification of 717 matching covariates. An average of 210.6 (standard deviation [sd] = 43.7), 42.0 (sd = 19.6) and 23.2 (sd = 9.3) matching covariates were identified by LASSO regression across all subsample draws within the 10%, 1% and 0.5% subsample groups, respectively.

Figure 2 depicts the SMD of matching covariates across all analyses, and summary statistics on the average absolute SMD of matching covariates are available in Supplemental Appendix C. As evidenced by absolute SMDs greater than 0.10, significant matching covariate imbalance existed prior to matching. After CM, no imbalanced matching covariates were observed within either the full study population or any subsample group. Furthermore, more stringent prespecified balance criteria were associated with a reduction in the average SMD of matching covariates; and CM requiring exact marginal distributional balance achieved perfect balance (e.g., SMD = 0, sd = 0) of matching covariates. While PSM achieved balance across all matching covariates within the full study population and 10% subsample group, the average frequency and proportion of post-match matching covariate imbalance within the 1% and 0.5% subsample groups were: caliper = 0.10, 8.2 (19.6%) and 7.6 (32.7%), respectively; and caliper = 0.20, 8.4 (19.9%) and 7.5 (32.5%), respectively.

Post-match candidate covariate balance

A total of 50,391 candidate covariates were observed in the full study population. Due to a decrease in sample size, fewer candidate covariates were observed among subsample groups. The average number of candidate covariates observed within the 10%, 1% and 0.5% subsample groups was 26,696.8 (sd = 467.4), 11,644.2 (sd = 441.9) and 8,581.1 (sd = 436.1), respectively.

The SMD of candidate covariates before matching and across all analyses is shown in Fig. 3. Overall, candidate covariate imbalance was negatively correlated with sample size. In the full study population, no imbalanced covariates were observed post-PSM (see Supplemental Appendix D). Similarly, PSM was associated with a small, albeit non-significant, improvement in the average SMD of candidate covariates in the full study population as compared to CM (see Supplemental Appendix E). Nevertheless, comparable improvements in candidate covariate balance were achieved by both matching techniques within each subsample group.

Post-match residual confounding

The expected systematic error (ESE) prior to matching and subsequent to CM and PSM within the full study population and each subsample group is shown in Fig. 4. Overall, ESE was negatively correlated with sample size; a significant increase in pre-match ESE was observed between the full study population (ESE = 0.12; 95% CI: [0.03, 0.23]) and 0.5% subsample group (ESE = 0.51; 95% CI: [0.37, 0.66]).

As compared to the pre-match sample, both matching techniques were associated with a substantial decrease in ESE. Furthermore, the post-match reduction in ESE was most pronounced in analyses with smaller sample sizes. Specifically, CM and PSM were associated with a significant decrease in ESE relative to the pre-match sample across most analyses within the 1% and 0.5% subsample groups (e.g., 0.5% subsample group: pre-match, ESE = 0.51 [95% CI: (0.37, 0.66)]; PSM with caliper = 0.10, ESE = 0.10 [95% CI: (0.02, 0.28)]; and CM with maximum SMD = 0.01, ESE = 0.11 [95% CI: (0.02, 0.27)]).

Analyses of angioedema outcome

Results from analyses of the safety outcome of angioedema between new users of ACEI vs. thiazide and thiazide-like monotherapy within the full study population are presented in Fig. 5. As compared to thiazide or thiazide-like monotherapy, ACEI monotherapy was found to be associated with a significant increase in the risk of angioedema across all analyses (calibrated $p < 0.05$), and calibrated HR estimates did not significantly differ between CM and PSM. Furthermore, CM was associated with a slight decrease in the standard error of calibrated HR estimates relative to PSM. Similar trends were observed among uncalibrated effect estimates.

Discussion

In this applied comparison of CM and PSM among new users of ACEI vs. thiazide and thiazide-like diuretic monotherapy, CM found the largest matched sample meeting prespecified balance criteria. The performance of both matching techniques was assessed at progressively smaller sample sizes. While both matching techniques achieved similar candidate covariate balance, CM was associated with improved matching covariate balance in analyses with smaller sample sizes. Furthermore, CM was associated with improved patient retention as compared to PSM translating to slight improvements in the precision of effect estimates. Finally, CM and PSM were associated with similar improvements in residual confounding, which was assessed based on the ESE of empirical null distribution of negative control outcomes.

Prior literature comparing CM and PSM is limited. In a study examining the impact of earthquakes on electoral outcomes in Chile, Visconti et al. describe the performance of both matching techniques. Before matching, the study included a total of 172 observations. As compared to PSM, CM was associated with a decrease in both post-match sample size (108 vs. 154) and, as evidenced by a SMD greater than 0.10, matching covariate imbalance (0 vs. 13 out of 18 imbalanced matching covariates)[4]. Similarly, in a

Monte Carlo simulation study, Resa et al. found CM to systematically select the largest sample size meeting a set of prespecified balance criteria[24].

Consistent with prior literature, as evidenced by a SMD less than 0.10, CM achieved balance for all matching covariates across all analyses. While PSM achieved balance of all matching covariates in analyses with larger sample sizes (e.g., full study population and 10% subsample groups), the matching technique was associated with substantial matching covariate imbalance in analyses with smaller sample sizes (e.g., the 1% and 0.5% subsample groups). Furthermore, CM was associated with improved sample retention across all analyses with the exception of fine balance within the study population, which failed to converge to an optimal solution, indicating that the achievement of prespecified balance criteria was not mutually exclusive to superior sample size retention.

Both candidate covariate imbalance and ESE were negatively correlated with sample size. As compared to the pre-match sample, improvements in candidate covariate balance were achieved with either matching technique; however, PSM achieved better candidate covariate balance in analyses with larger sample sizes. That being said, it is important to note that fewer matching covariates were used with CM as compared to PSM (717 vs. 1,237) in analyses within the full study population due to memory limitations associated with CM. The pre-match ESE of the full study population was significantly higher as compared to the 1% and 0.5% subsample groups indicating increased baseline residual confounding at smaller sample sizes. Reductions in residual confounding were comparable between matching techniques and especially pronounced in analyses with smaller sample sizes (e.g., the 1% and 0.5% subsample groups) as evidenced by a significant post-match decrease in the ESE relative to the pre-match sample. These findings may indicate both matching techniques are comparable in reducing residual confounding stemming from imbalances in unmeasured or otherwise unadjusted covariates.

The current study also found pre-match ESE to be significantly higher within analyses of smaller sample size. This finding may indicate the presence of an additional source of bias within the pre-match sample. Specifically, the increase in systemic bias may be due to a failure to meet the normality assumption on the likelihood distribution of Cox proportional hazards models at smaller sample sizes prior to matching. Additional research is necessary to explore this hypothesis.

Calibrated hazard ratio estimates were similar in direction and magnitude across all analyses within the study population indicating ACEI monotherapy was associated with a significant increase in the risk of angioedema as compared to thiazide or thiazide-like monotherapy. However, as compared to PSM, CM was associated with a slight reduction in the standard error of estimates. Similar trends were observed among uncalibrated analyses. The increased precision of effect estimates may be due to the improved sample retention observed with CM.

Limitations

The current study was subject to limitations. First, due to memory constraints, the identification of matching covariates through LASSO regression within the full study population was limited to covariates with a minimum frequency of 2% for CM and 0.1% for PSM. As such, the performance of CM as compared to PSM in addressing potential confounding within studies of large sample sizes may have been underestimated. Nevertheless, this highlights practical limitations of CM in large-scale studies associated with limitations in computing power; CM failed to converge due to memory constraints using a dataset containing approximately 220 million data points (172,117 observations and 1,237 matching covariates) but successfully converged using a dataset containing approximately 120 million data points (172,117 observations and 717 matching covariates). These practical limitations may be overcome with access to more powerful computing resources.

Second, the use of negative control experiments limited analyses to subsample groups with a pre-match sample size sufficient to ensure the observation of negative control outcomes after matching. The current study addressed this limitation by considering the joint results of analyses across multiple subsample draws within subsample groups, and the smallest subsample group contained 860 patients. Nevertheless, the relative performance of CM may be improved in studies with even smaller sample sizes, which are more likely to suffer from issues of limited covariate overlap and potential model over-parameterization.

Conclusion

The current study compared the performance of CM and PSM in terms of post-match sample size, covariate balance and residual confounding. CM found the largest matched sample meeting prespecified balance criteria thereby achieving superior sample retention and, in analyses with smaller sample sizes, improved matching covariate balance as compared with PSM. Candidate covariate balance and residual bias were comparable between matching techniques. These findings support the use of CM as an alternative to PSM for causal inference in observational research with small sample sizes where balance on a specific set of matching covariates is desired. Further research is necessary to compare the performance of CM and PSM in studies where empirical covariate selection may not be possible due to limited sample size or availability of data.

List Of Abbreviations

Abbreviation	Definition
AMA	American Heart Association
ACEI	Angiotensin-converting enzyme inhibitor
AHA	American Heart Association
ATC	Average Treatment Effect on the Controls
ATE	Average Treatment Effect
CCAE	IBM® MarketScan® Commercial Claims and Encounters Database
CCI	Charlson Comorbidity Index
CDM	Common Data Model
CI	Confidence interval
CM	Cardinality matching
DCSI	Diabetes Complications Severity Index
ESE	Expected systematic error
HADES	Health Analytics Data-to-Evidence Suite
HR	Hazard Ratio
LEGEND-HTN	Large-Scale Evidence Generation and Evaluation across a Network of Databases hypertension
MCMC	Markov Chain Monte Carlo
OHDSI	Observational Health and Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PSM	Propensity score matching
sd	Standard deviation
SMD	Standardized mean difference

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The datasets generated and analyzed during the current study are not publicly available as they were obtained from IBM under a proprietary data use agreement but are available from the corresponding author on reasonable request. The code used to generate and analyze the datasets for the current study are available in the github, https://github.com/ohdsi-studies/EvaluatingCardinalityMatching.

Competing interests: SPF, SSJ and MS are employees of Johnson & Johnson and Janssen R&D, LLC, a subsidiary of Johnson & Johnson.

Funding: Funding for this study was provided by Johnson & Johnson.

Authors' contributions: SPF analyzed and interpreted the data, and was a major contributor in writing the manuscript. MS was a major contributor in development of study methods. All authors read and approved the final manuscript.

Acknowledgements: Not applicable

References

1. Armitage P. The role of randomization in clinical trials. Stat Med. 1982;1(4):345-352. doi:10.1002/sim.4780010412
2. Dorn HF. Philosophy of inferences from retrospective studies. Am J Public Health Nations Health. 1953;43(6 Pt 1):677-683. doi:10.2105/ajph.43.6_pt_1.677
3. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. Sankhya, Ser. A. 1973;35:417–446
4. Visconti G, Zubizarreta J. Handling Limited Overlap in Observational Studies with Cardinality Matching. Observational Studies 2018; 4:217-249.
5. Rosenbaum P.R. Design of observational studies.New York, NY: Springer-Verlag; 2010.
6. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009; 96 (1), pp. 187–199.
7. Rothe C. Robust confidence intervals for average treatment effects under limited overlap. Econometrica. 2017; 85: 645-660. doi: 10.3982/ECTA13141.
8. Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. BMC Med Res Methodol. 2012;12:70. doi:10.1186/1471-2288-12-70.
9. OHDSI (2019). The Book of OHDSI: Observational Health Data Sciences and Informatics. OHDSI.
10. Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet. 2019;394(10211):1816-1826. doi:10.1016/S0140-6736(19)32317-7.
11. ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. 2018;138(17):e426-e483. doi:10.1161/CIR.0000000000000597.

12. Montori VM, Guyatt GH. Intention-to-treat principle. CMAJ. 2001 (165):1339-1341.
13. Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives, J Clin Epidemiol, 1993, vol. 46 10(pg. 1075-1079).
14. Young BA, Lin E, Von Korff M, et al. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. Am J Manag Care. 2008;14(1):15-23.
15. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA. 2001;285(22):2864-2870. doi:10.1001/jama.285.22.2864.
16. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-272. doi:10.1378/chest.09-1584.
17. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. Int J Epidemiol. 2018;47(6):2005-2014. doi:10.1093/ije/dyy120
18. HADES: Health analytics data-to-evidence suite. https://ohdsi.githu.io/Hades/. Accessed Aug 4, 2020.
19. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28:3083–107.
20. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med. 2014;33(2):209-218. doi:10.1002/sim.5925.
21. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Robust empirical calibration of p-values using observational data. Stat Med. 2016;35(22):3883-3888. doi:10.1002/sim.6977.
22. Voss AE, Boyce RD, Ryan PB, van der Lei J, Rijbbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. J Biomed Inform. 2017; 66: 72-81.
23. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci U S A*. 2018;115(11):2571-2577. doi:10.1073/pnas.1708282114
24. Resa M, Zubizarreta JR. Evaluation of subset matching methods and forms of covariate balance. Stat Med. 2016;35(27):4961-4979. doi:10.1002/sim.7036

Figures

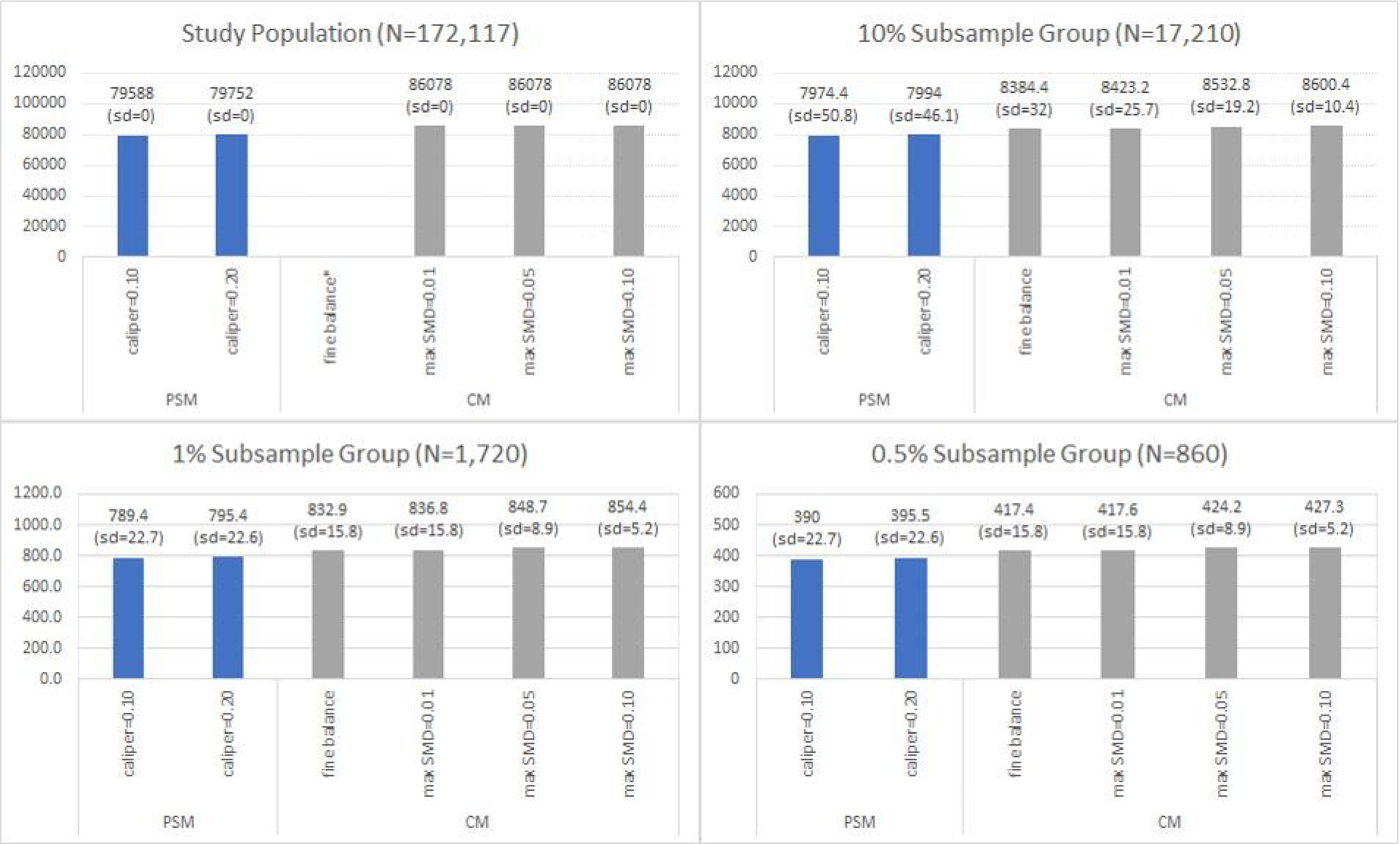


Figure 1

Average post-propensity score matching (PSM) and cardinality matching (CM) sample sizes across all analyses Fine balance: exact marginal distributional balance * Failed to converge to a matched sample

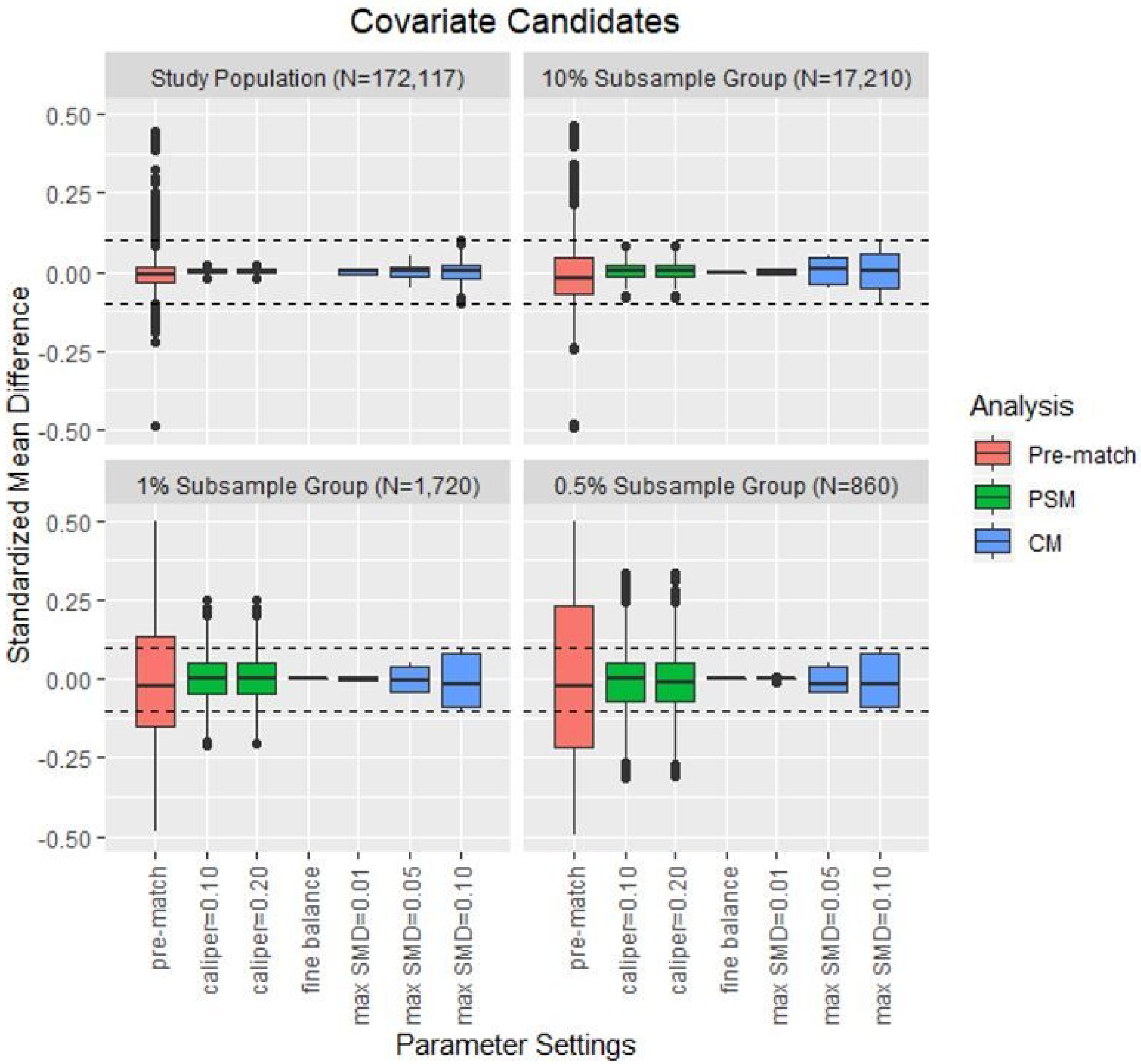


Figure 2

Standardized mean differences of matching covariates post-propensity score matching (PSM) and cardinality matching (CM) Fine balance: exact marginal distributional balance * Data represent standardized mean difference of matching covariates across all subsample draws within each subsample group considered jointly

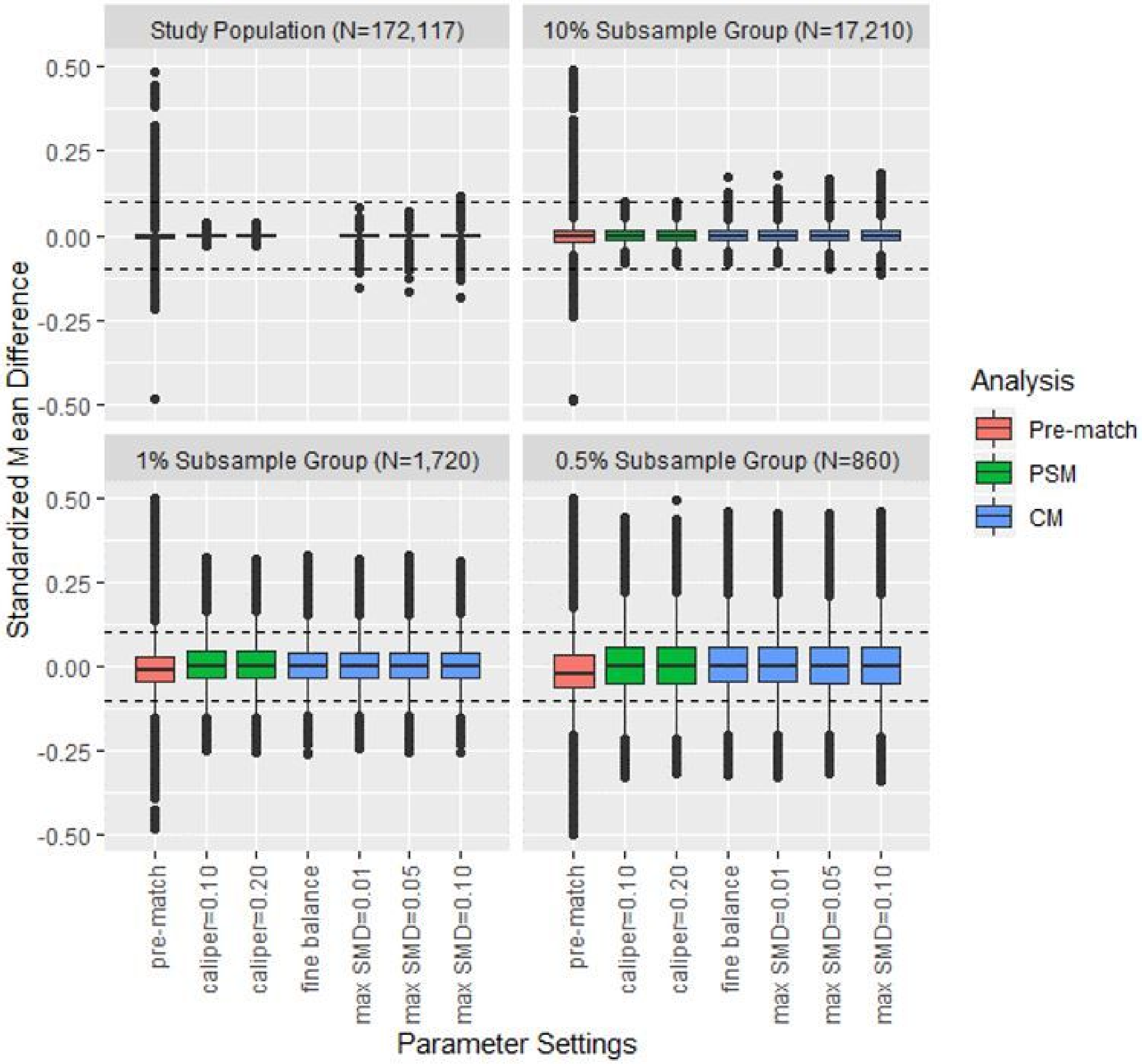
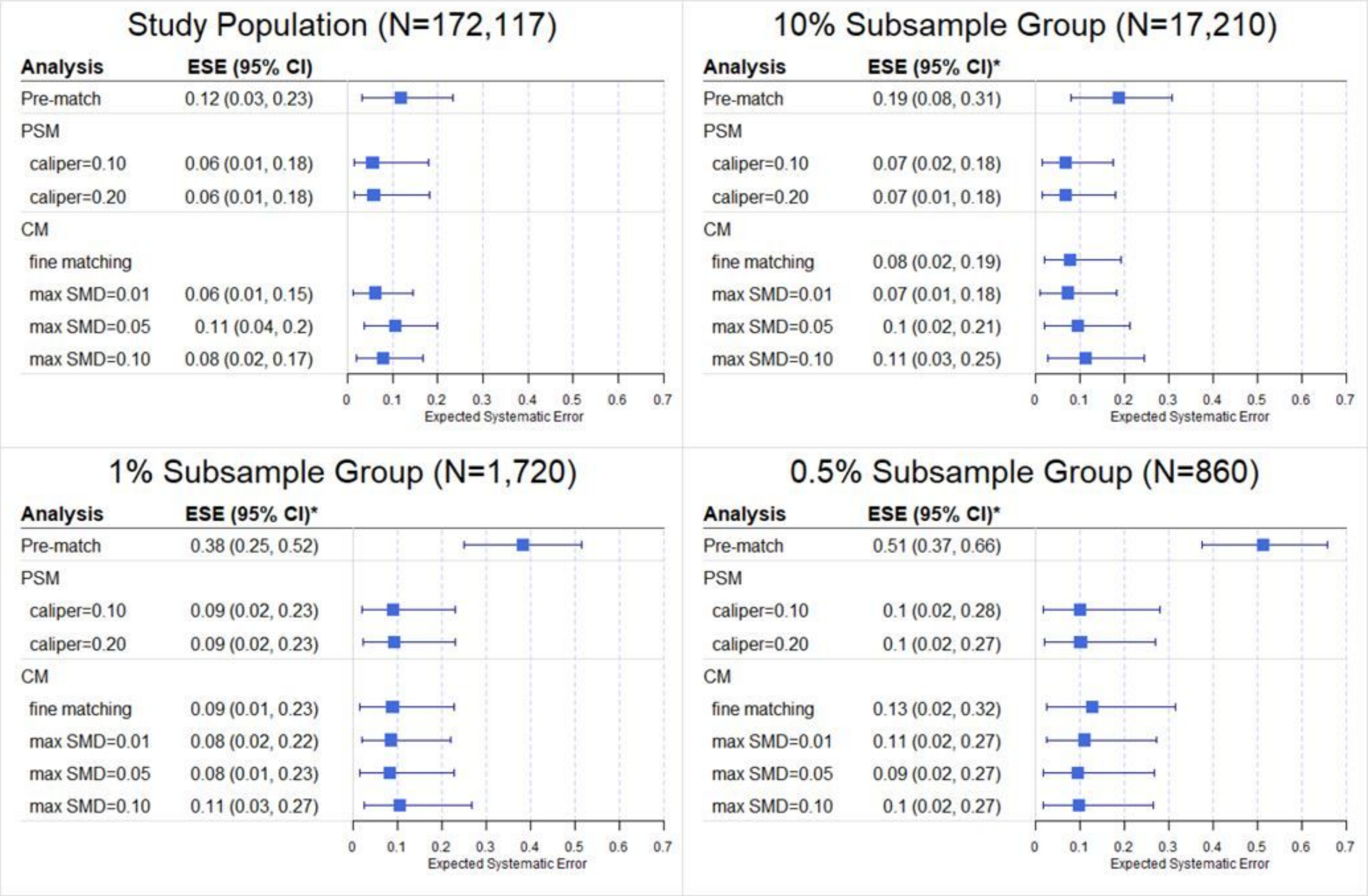


Figure 3

Standardized mean differences of candidate covariates post-propensity score matching (PSM) and cardinality matching (CM) Fine balance: exact marginal distributional balance * Data represent standardized mean difference of matching covariates across all subsample draws within each subsample group considered jointly



Study Population (N=172,117)

Analysis	ESE (95% CI)
Pre-match	0.12 (0.03, 0.23)
PSM	
caliper=0.10	0.06 (0.01, 0.18)
caliper=0.20	0.06 (0.01, 0.18)
CM	
fine matching	
max SMD=0.01	0.06 (0.01, 0.15)
max SMD=0.05	0.11 (0.04, 0.2)
max SMD=0.10	0.08 (0.02, 0.17)

10% Subsample Group (N=17,210)

Analysis	ESE (95% CI)*
Pre-match	0.19 (0.08, 0.31)
PSM	
caliper=0.10	0.07 (0.02, 0.18)
caliper=0.20	0.07 (0.01, 0.18)
CM	
fine matching	0.08 (0.02, 0.19)
max SMD=0.01	0.07 (0.01, 0.18)
max SMD=0.05	0.1 (0.02, 0.21)
max SMD=0.10	0.11 (0.03, 0.25)

1% Subsample Group (N=1,720)

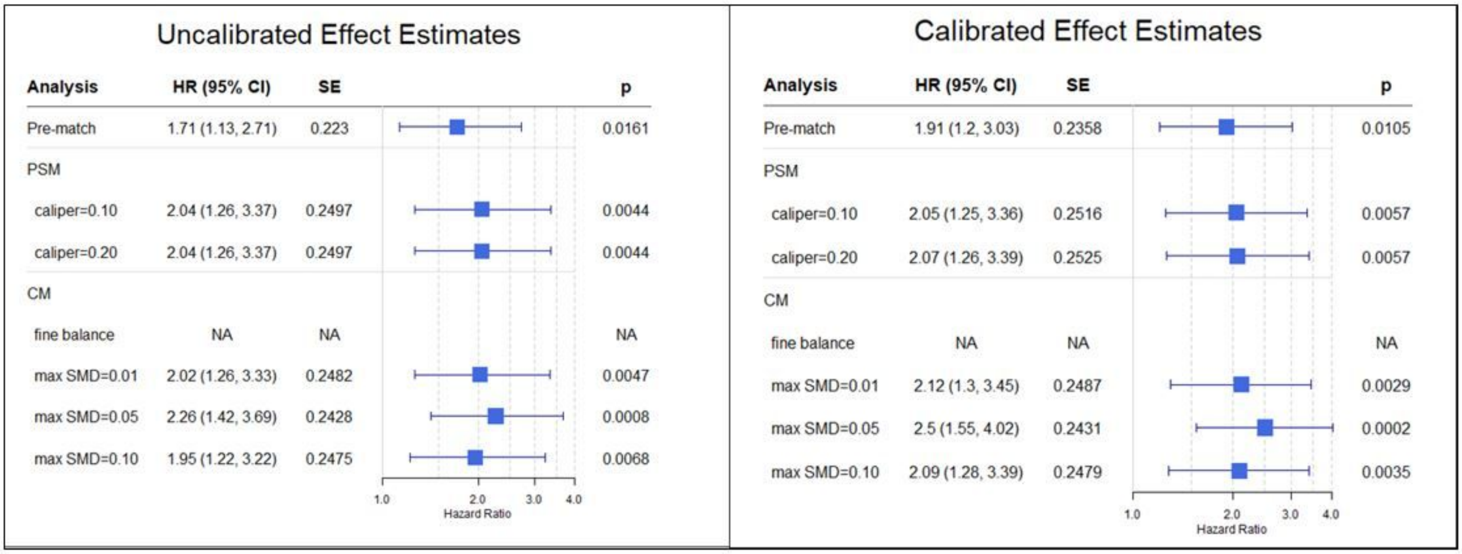
Analysis	ESE (95% CI)*
Pre-match	0.38 (0.25, 0.52)
PSM	
caliper=0.10	0.09 (0.02, 0.23)
caliper=0.20	0.09 (0.02, 0.23)
CM	
fine matching	0.09 (0.01, 0.23)
max SMD=0.01	0.08 (0.02, 0.22)
max SMD=0.05	0.08 (0.01, 0.23)
max SMD=0.10	0.11 (0.03, 0.27)

0.5% Subsample Group (N=860)

Analysis	ESE (95% CI)*
Pre-match	0.51 (0.37, 0.66)
PSM	
caliper=0.10	0.1 (0.02, 0.28)
caliper=0.20	0.1 (0.02, 0.27)
CM	
fine matching	0.13 (0.02, 0.32)
max SMD=0.01	0.11 (0.02, 0.27)
max SMD=0.05	0.09 (0.02, 0.27)
max SMD=0.10	0.1 (0.02, 0.27)

Figure 4

Expected systematic error from negative control experiments post-propensity score matching (PSM) and cardinality matching (CM) ESE: expected systematic error; fine balance: exact marginal distributional balance * Data represent the expected systematic error of empirical null distribution of negative control outcomes across all subsample draws within each subsample group considered jointly



Uncalibrated Effect Estimates

Analysis	HR (95% CI)	SE	p
Pre-match	1.71 (1.13, 2.71)	0.223	0.0161
PSM			
caliper=0.10	2.04 (1.26, 3.37)	0.2497	0.0044
caliper=0.20	2.04 (1.26, 3.37)	0.2497	0.0044
CM			
fine balance	NA	NA	NA
max SMD=0.01	2.02 (1.26, 3.33)	0.2482	0.0047
max SMD=0.05	2.26 (1.42, 3.69)	0.2428	0.0008
max SMD=0.10	1.95 (1.22, 3.22)	0.2475	0.0068

Calibrated Effect Estimates

Analysis	HR (95% CI)	SE	p
Pre-match	1.91 (1.2, 3.03)	0.2358	0.0105
PSM			
caliper=0.10	2.05 (1.25, 3.36)	0.2516	0.0057
caliper=0.20	2.07 (1.26, 3.39)	0.2525	0.0057
CM			
fine balance	NA	NA	NA
max SMD=0.01	2.12 (1.3, 3.45)	0.2487	0.0029
max SMD=0.05	2.5 (1.55, 4.02)	0.2431	0.0002
max SMD=0.10	2.09 (1.28, 3.39)	0.2479	0.0035

Figure 5

Hazard ratio for angioedema between new users of ACEI vs. thiazide or thiazide-like diuretic monotherapya ACEI: angiotensin-converting enzyme inhibitor; HR: hazard ratio; CI: confidence interval; SE: standard error of the natural log of the hazard ratio a Uncalibrated effect estimates (left panel) and calibrated effect estimates (right panel) for analyses performed within the full study population (N=172,117). Calibrated hazard ratio, estimates, confidence intervals, standard errors and p-values based on the empirical null distribution of negative control outcomes

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalAppendices.docx