American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2018. This work is written by (a) US Government employee(s) and is in the public domain in the US.

Vol. 187, No. 8
DOI: 10.1093/aje/kwy049
Advance Access publication:
March 15, 2018

## Practice of Epidemiology

# Relative Performance of Propensity Score Matching Strategies for Subgroup Analyses

**Shirley V. Wang*, Yinzhu Jin, Bruce Fireman, Susan Gruber, Mengdong He, Richard Wyss, HoJin Shin, Yong Ma, Stephine Keeton, Sara Karami, Jacqueline M. Major, Sebastian Schneeweiss, and Joshua J. Gagne**

* Correspondence to Dr. Shirley V. Wang, Division of Pharmacoepidemiology and Pharmacoeconomics at Brigham and Women's Hospital Department of Medicine, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (e-mail: swang1@bwh.harvard.edu).

*Initially submitted August 2, 2017; accepted for publication March 5, 2018.*

Postapproval drug safety studies often use propensity scores (PSs) to adjust for a large number of baseline confounders. These studies may involve examining whether treatment safety varies across subgroups. There are many ways a PS could be used to adjust for confounding in subgroup analyses. These methods have trade-offs that are not well understood. We conducted a plasmode simulation to compare relative performance of 5 methods involving PS matching for subgroup analysis, including methods frequently used in applied literature whose performance has not been previously directly compared. These methods varied as to whether the overall PS, subgroup-specific PS, or no rematching was used in subgroup analysis as well as whether subgroups were fully nested within the main analytical cohort. The evaluated PS subgroup matching methods performed similarly in terms of balance, bias, and precision in 12 simulated scenarios varying size of the cohort, prevalence of exposure and outcome, strength of relationships between baseline covariates and exposure, the true effect within subgroups, and the degree of confounding within subgroups. Each had strengths and limitations with respect to other performance metrics that could inform choice of method.

propensity score matching; simulation; subgroup analyses

Abbreviations: ACE, angiotensin-converting enzyme; ASAMD, average standardized absolute mean difference; ATT, average treatment effect in the treated; CIDA + PS, Cohort Identification and Descriptive Analysis + Propensity Score; PS, propensity score.

The United States Food and Drug Administration's Sentinel system is a national program for medical-product safety monitoring involving a large distributed network of health-care databases and a suite of routine querying tools (1, 2). These tools enable semiautomated prospective and one-time assessments with multivariable-adjustment via propensity score (PS) matching and other adjustment strategies. The Cohort Identification and Descriptive Analysis + Propensity Score (CIDA + PS) matching tool was extensively validated for fidelity of code to intended design, and it has been tested with numerous known positive and negative associations. The tool is now being used for several postapproval-safety surveillance activities (3–5). One of the key strengths of Sentinel's distributed database is the ability to include large numbers of patients in an analysis, which can potentially support a number of subgroup analyses. Routine surveillance activities often involve evaluation of prespecified subgroups because the Food and Drug Administration is interested in examining treatment effects in defined subgroups as well as in all patients for whom compared treatments are alternatives (6–8).

Currently CIDA + PS uses the PS estimated in the full cohort to perform matching for the main analysis and rematching in prespecified subgroups. However, there are multiple ways that a PS could be used for subgroup analyses. A systematic review found that PS matching approaches for subgroup analyses compared in methods papers did not include evaluation of commonly used subgroup PS matching methods in applied studies (9). These methods have trade-offs that are not well understood.

The objective of this study was to compare the relative performance of alternative methods for using PS matching in subgroup analysis, including methods used in applied



Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019

literature whose performance has not previously been directly compared.

## METHODS

We evaluated 5 PS matching strategies for subgroup analysis:

A. Use overall PS to match in full cohort; use same PS to rematch subgroups (ignore matching from full cohort).
B. Use overall PS to match in full cohort; use same PS to rematch subgroups (restrict to those matched in full cohort).
C. Fit separate PS within unique combinations of subgroup strata; match within strata, and aggregate for main analysis.
D. Use overall PS to match within unique combinations of subgroup strata; aggregate to create the full matched cohort.
E. Use overall PS to match for main analysis; break matches; conduct subgroup analyses without additional adjustment.

These were selected from a prior literature review (9) of methods and applied papers that used PS matching for subgroup analyses. Only strategies A and B were compared in the methods literature (10–14). Strategies C–E were reported in the applied literature review, although strategy E was the most commonly reported method of doing subgroup analysis after PS matching, used in over one-third of papers included in the review.

In order to evaluate PS matching strategies for subgroup analysis, we designed a "plasmode" simulation (details in Web Table 1, available at https://academic.oup.com/aje). In contrast to simulation where all variables are generated based on probability distributions, the complexity and correlation of covariates observed in a real data set are retained in plasmode simulation (15). However, the magnitude of the true effect size and confounding are still specified by the investigator. We conducted a plasmode simulation in order to retain realistic correlation among covariates in our evaluation of PS matching methods for subgroup analysis.

### Base cohort for simulations

The base cohort for simulation was extracted from a large administrative health-care claims database converted to the Sentinel Common Data Model via CIDA + PS (version 5.0). We chose to use the protocol for evaluating the known positive relationship between angiotensin-converting enzyme (ACE) inhibitors versus β-blockers on risk of angioedema used in prior Sentinel evaluations because it was a thoroughly studied and uncontroversial example (6–8). The base cohort included new users of ACE inhibitors or β-blockers between January 1, 2009, and December 31, 2012, enrolled with medical and prescription benefits for at least 183 days (30-day gaps allowed) without exposure to either study drug of interest or diagnosis of angioedema (*International Classification of Diseases, Ninth Revision, Clinical Modification*, code 995.1) in any care setting. Follow-up began on the date of initiation and continued until any censoring criteria were met. Censoring criteria included the outcome of interest (angioedema diagnosis in any care setting), discontinuation or switching of study drugs, disenrollment, 365 days elapsed, or end of study period. Baseline covariates were assessed in the 183 days prior to the index date and included age, sex, and a history of allergic reactions, diabetes mellitus, heart failure, ischemic heart disease, and use of prescription nonsteroidal antiinflammatory drugs. Additional details of the protocol are available on the Sentinel website (8).

### Generating simulated data with known truth

Each of our simulation scenarios maintained the observed baseline covariate structure by resampling from the observed covariate matrix rather than generating each covariate from a probability distribution. In order to reduce the run time for simulations, as well as mimic a moderate-sized cohort, the base scenario used a random sample of 50,000 from the observed covariate vectors. We generated exposure based on a true PS model with intercept and coefficients selected such that the proportion exposed was the same as observed in the real data. We fitted a Cox proportional hazards model for the outcome and a Cox model for censoring in the data with simulated exposure and observed baseline risk factors. The baseline hazard of the outcome model was modified so that the incidence of the outcome was 0.01. We then modified the observed coefficient for treatment effect to generate survival curves for each patient based on simulated exposure. For some scenarios, we also added coefficients for interactions between exposure and a subgroup variable when generating survival curves. For each patient, the simulated outcome was 1 if the survival time was less than the censoring time; otherwise it was set to 0. The magnitude of the true conditional hazard ratio was set to 2.0. However, due to confounding, the observed hazard ratio in unadjusted analyses was 3.0. Example simulation code is available in Web Appendix 1.

We then investigated 12 simulation scenarios, each varying 1 or more parameters from the base scenario (Web Tables 1 and 2). Confounding that varied according to subgroup was generated by including terms for interaction between baseline characteristics and the prespecified subgroup variables in the true PS model when generating the exposure data. Heterogeneity in the true hazard ratio was generated by including interactions between exposure and the subgroup variables in the outcome model. We generated 500 bootstrap samples under the specifications for each simulation scenario.

For each of the simulated scenarios, we varied the number of prespecified subgroups to evaluate (1 or 3) as well as whether the predefined PS model included interactions (none versus all pairwise interactions between subgroup variables and other covariates in the model included) (Web Table 2). The subgroups we considered were age (3 categories), sex, and history of heart failure.

We applied each of the 5 PS matching strategies for subgroup analysis outlined above (Web Table 2). For each of the strategies, PSs were generated using logistic regression with exposure as the dependent variable and confounders as independent variables. We implemented nearest neighbor, 1:1 matching using SAS-based macros from a publically available PharmacoEpi Toolbox (16). Each PS model included all true confounders as main terms. However, in some scenarios the true PS varied by subgroup and in others it did not. We evaluated performance for each scenario using PS models fitted with and without subgroup interaction terms.

Strategies C and D required matching within subgroup and then aggregating to form the cohort for the main analysis. We



Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019

implemented strategies C and D as if the prespecified subgroup was age, sex, or heart failure separately (i.e., single, prespecified subgroup) as well as simultaneously (i.e., 3 prespecified subgroups in the analysis). The former required estimation within 2 subgroup strata for sex and heart failure or 3 subgroup strata for age. The latter required fitting a PS within 12 unique strata formed by the subgroups (Table 1). Because strategy C involves matching on stratum-specific PS, a separate PS was estimated within each unique combination of subgroups prior to matching. In contrast, for strategy D, the overall PS was used to match within each unique subgroup combination. Using different prespecified subgroups for these strategies produced different aggregated cohorts for the main analysis.

We estimated overall and subgroup-specific hazard ratios by fitting a Cox proportional hazards model regressing time to event on treatment in a 1:1-matched population. The models did not condition on matched set.

### Performance metrics

We evaluated balance in unmatched and matched cohorts via the average standardized absolute mean difference (ASAMD) for baseline covariates as well as the *C* statistic. A larger ASAMD indicates greater imbalance. For considering the standardized differences between compared exposure groups to indicate meaningful imbalance, a rule of thumb sets the threshold at values over 0.10 (17). *C* statistics are a measure of concordance that range between 0.5 and 1 (18). The further the *C* statistic is from 0.5, the worse the covariate balance between exposure groups. A model containing baseline characteristics as independent variables and exposure status as the dependent variable with a *C* statistic of 0.5 indicates poor discriminative ability or good balance because baseline characteristics do not predict exposure status better than chance.

We also evaluated bias, precision, and coverage of the parameter estimate in the unconditional Cox proportional hazards model for each PS-matching subgroup strategy. Bias was measured as the difference between the estimated feasible sample average treatment effect in the treated (ATT, from the model coefficient) and the simulated true counterfactual ATT in the matched treated population based on the true outcome model. These were evaluated overall as well as within each subgroup. Values further from 0.0 reflect greater bias. Precision was measured by the model-based standard error from unconditional Cox proportional hazard models. Coverage was calculated as the proportion of bootstrap samples for which the feasible sample ATT and 95% confidence intervals from outcome models contained the simulated true counterfactual ATT.

Some approaches can result in different numbers of patients or outcomes included in the main analysis compared with subgroup analyses, such that the numbers in matched subgroup analyses do not sum to the number of outcomes in the main analysis. We evaluated the consistency of main and subgroup populations for each PS matching strategy by comparing the number of exposed patients or outcomes included in the matched main analyses with the sum included across subgroup analyses.

**Table 1.** Subgroup Strata to Fit Propensity Scores for Strategy C[a]

| Stratum | Sex | Age Group | History of Heart Failure |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 1 | 1 |
| 5 | 0 | 2 | 0 |
| 6 | 0 | 2 | 1 |
| 7 | 1 | 0 | 0 |
| 8 | 1 | 0 | 1 |
| 9 | 1 | 1 | 0 |
| 10 | 1 | 1 | 1 |
| 11 | 1 | 2 | 0 |
| 12 | 1 | 2 | 1 |

[a] Strategy C: Fit separate propensity score within unique combinations of subgroup strata, match within strata, and aggregate for main analysis.

## RESULTS

The cohort extracted from a large administrative claims health-care database included 385,649 new initiators of ACE inhibitors and 274,977 new initiators of β-blockers (Table 2). β-blocker initiators tended to have fewer comorbid conditions than ACE-inhibitor initiators, with standardized differences greater than 0.2 for several baseline covariates (data not shown). The *C* statistic of 0.67 indicated modest discrimination between new initiators of ACE inhibitors and β-blockers based on the characteristics included in the PS. This separation was also observed in the PS distribution (Figure 1).

With respect to subgroup variables, β-blocker initiators tended to be older than ACE-inhibitor initiators, and a larger proportion were female. The proportion of ACE-inhibitor initiators with prior heart failure was 1.3% compared with 4.5% for β-blocker initiators. The distribution of subgroup characteristics was retained via bootstrap resampling with replacement for the simulation scenarios.

Details about the specifications for the 12 simulated cohort scenarios based on this extracted cohort are available in Web Tables 1 and 2.

### Balance: ASAMD

Figures 2–4 shows the distribution of ASAMD of baseline covariates in 500 bootstrap samples for the base simulation scenario (scenario 0). The ASAMD in the full unmatched cohort was above 0.20, and the ASAMD within unmatched age subgroups was over 0.15 (Figure 2A). After matching on PS with and without interactions (Figure 2B and 2C), regardless of subgroup matching strategy, the ASAMD was near 0.0 within the matched full cohort and subgroup analyses. These ASAMD results were similar for subgroups based on sex and baseline heart-failure status, although the ASAMD was larger and had greater variability for the low-prevalence subgroup with heart failure at baseline (Figures 3 and 4).

The results from the baseline scenario were paralleled in scenarios 1–9, which varied size of the cohort, prevalence of



Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019

**Table 2.** Cohort of New Initiators of Angiotensin-Converting Enzyme Inhibitors and β-Blockers (Unmatched), Using Data From a Large Administrative Health-Care Claims Database Converted to the Sentinel Common Data Model, United States, 2009–2012

| Characteristic | Primary Analysis | | | | | |
|---|---|---|---|---|---|---|
| | ACE Inhibitor | | | β-Blocker | | |
| | No. of Patients | % | Mean (SD) | No. of Patients | % | Mean (SD) |
| Study cohort | | | | | | |
| No. of patients matched | 385,649 | 100.0 | | 274,977 | 100.0 | |
| No. of events while on therapy | 1,008 | 0.3 | | 327 | 0.1 | |
| Person time at risk, days | | | 230.3 (265.0) | | | 215.2 (260.6) |
| Patient characteristic | | | | | | |
| Age, years | | | 56.3 (7.6) | | | 58.1 (8.9) |
| Age group, years | | | | | | |
| 45–54 | 185,315 | 48.1 | | 113,555 | 41.3 | |
| 55–64 | 161,137 | 41.8 | | 116,237 | 42.3 | |
| 65–99 | 39,197 | 10.2 | | 45,185 | 16.4 | |
| Female sex | 174,168 | 45.2 | | 141,859 | 51.6 | |
| Recorded history | | | | | | |
| Combined comorbidity score | | | −0.1 (1.1) | | | 0.4 (1.7) |
| Allergic reactions | 37,538 | 9.7 | | 31,648 | 11.5 | |
| Diabetes | 84,545 | 21.9 | | 43,579 | 15.8 | |
| Heart failure | 4,943 | 1.3 | | 12,261 | 4.5 | |
| Ischemic heart disease | 19,036 | 4.9 | | 45,988 | 16.7 | |
| Recorded use of NSAIDs | 52,241 | 13.5 | | 38,307 | 13.9 | |
| Health-service utilization intensity | | | | | | |
| No. of unique generics dispensed | | | 3.6 (3.6) | | | 4.5 (4.3) |
| No. of filled prescriptions | | | 8.2 (10.0) | | | 10.6 (12.0) |
| No. of inpatient hospital encounters | | | 0.1 (0.3) | | | 0.2 (0.5) |
| No. of nonacute institutional encounters | | | 0.0 (0.7) | | | 0.1 (1.4) |
| No. of emergency room encounters | | | 0.2 (1.4) | | | 0.3 (1.7) |
| No. of ambulatory encounters | | | 5.3 (7.1) | | | 8.3 (10.5) |
| No. of other ambulatory encounters | | | 0.2 (1.3) | | | 0.4 (1.8) |
| Mahalanobis distance | | | 0.389 | | | |

Abbreviations: ACE, angiotensin-converting enzyme; NSAID, nonsteroidal antiinflammatory drug; SD, standard deviation.

exposure or outcome, and presence of treatment-effect heterogeneity. Scenarios 10–12 included confounding that varied according to subgroup. The pattern of ASAMD in these scenarios showed that, although balance was better when interactions were included in the PSs, all approaches resulted in ASAMD that was well below 0.05, regardless of the subgroup matching strategy (Web Figures 1K, 2K, and 3K). However, the worst-performing strategy was strategy E, which split the 1:1-matched cohort into subgroup strata and analyzed without further matching within subgroup. A full set of figures for each scenario is available in Web Figures 1–3.

### Balance: *C* statistic

In the base simulation scenario (scenario 0), the mean *C* statistic in the full, unmatched cohort was 0.74 and the mean *C* statistic within unmatched age subgroups ranged between 0.69 and 0.77. The pattern of results was very similar to the pattern for ASAMD after matching on PS with and without interactions, for sex and heart-failure subgroups, and across simulated scenarios.

### Bias: difference between estimated feasible sample ATT and counterfactual ATT

In the base simulation scenario (scenario 0), the mean bias in the full, unmatched cohort was approximately 0.5 in the full cohort and age subgroups (Web Figure 4A, left panel). After matching on PS with and without interactions (middle and right panels), regardless of subgroup matching strategy, the bias was near 0.0 within the matched full cohort and subgroup analyses. These bias results were similar for subgroups based on sex and baseline heart-failure status; however, there was greater variability for the low-prevalence subgroup with heart failure prior to



Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019

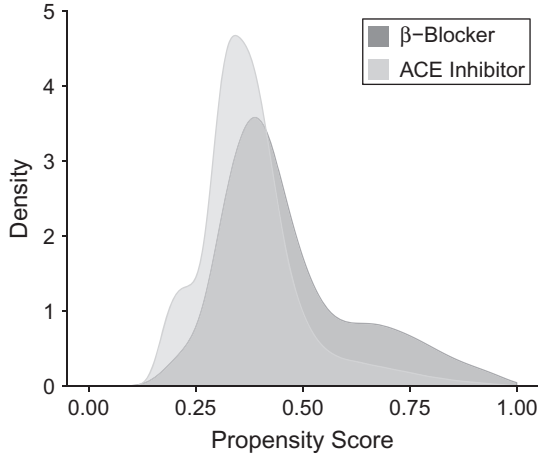


**Figure 1.** Propensity score distribution for the observed cohort of angiotensin-converting enzyme (ACE)-inhibitor and β-blocker initiators, using data from a large administrative health-care claims database converted to the Sentinel Common Data Model, United States, 2009–2012. The *C* statistic for the propensity score was 0.67.

baseline (Web Figures 5A and 6A). The results from the baseline scenario were paralleled in scenarios 1–9. Although balance was marginally worse for subgroup matching strategies in scenarios 10–12, where true confounding varied by subgroup and the PS did not include interactions, these residual imbalances had negligible impact on bias (Web Figures 4K, 5K, and 6K). A full set of figures for each scenario are available in Web Figures 4–6.

### Precision: standard error

In the base simulation scenario (scenario 0), the mean model-based standard error in the full unmatched cohort was approximately 0.1 in the full cohort and up to 0.2 in age subgroups. As expected, in the matched cohorts, the standard errors overall and in subgroups were larger. However, matching on the PS with and without interactions (middle and right panels in web figures), regardless of subgroup matching strategy, produced nearly identical standard errors. The precision was similar for subgroups based on sex and baseline heart-failure status, but there was greater variability for the low-prevalence subgroup with heart failure prior to baseline. The results from the baseline scenario were paralleled in scenarios 1–12.

### Coverage

In unmatched data, the coverage was around 0.0. The coverage for unmatched data within strata of age, sex, or lack of prior heart failure tended to lie in the 0.2–0.5 range. Coverage in the low-prevalence subgroup with heart failure at baseline was higher, in some scenarios over 0.95 prior to matching. In matched data, the coverage largely hovered just

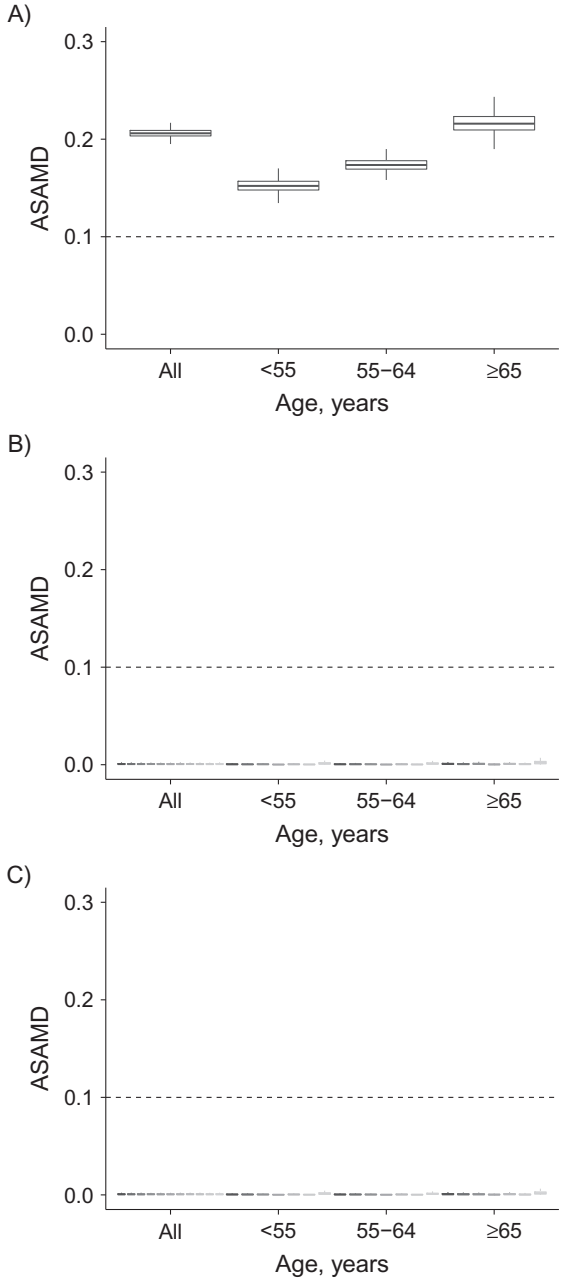


**Figure 2.** Distribution of average standardized absolute mean difference (ASAMD) for the baseline scenario broken down by age group (500 bootstrap samples), using simulated data. The ASAMD is presented for the unmatched cohort, a cohort 1:1 matched on a propensity score from a model that included only main terms for predictors, and a cohort 1:1 matched on a propensity score from a model that included interactions. A) ASAMD for unmatched cohorts; B) ASAMD for cohorts 1:1 matched on a propensity score from a model without interactions between predictors; C) ASAMD for cohorts 1:1 matched on a propensity score from a model with interactions between age, gender, heart failure status, and other baseline characteristics. For (B) and (C), the ASAMD is shown after applying each propensity-score matching strategy from Web Table 2. The strategies A, B, C1_1, C1_2, C1_3, D1_1, D1_2, D1_3, and E are shown in order from left to right.

Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019

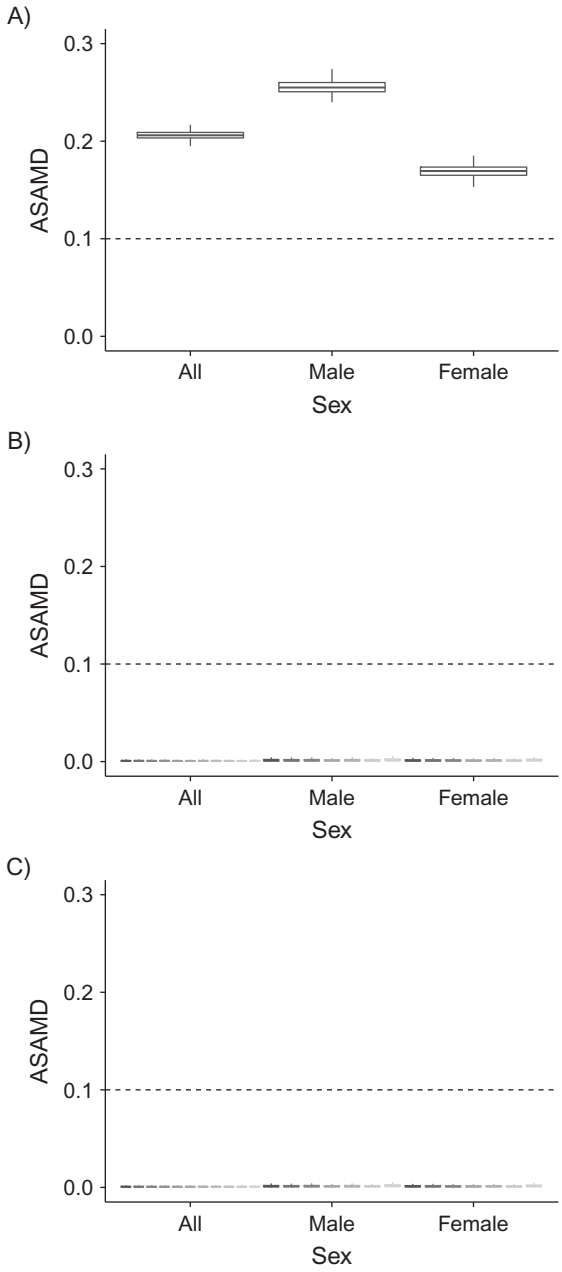


**Figure 3.** Distribution of average standardized absolute mean difference (ASAMD) for the baseline scenario broken down by sex (500 bootstrap samples), using simulated data. The ASAMD is presented for the unmatched cohort, a cohort 1:1 matched on a propensity score from a model that included only main terms for predictors, and a cohort 1:1 matched on a propensity score from a model that included interactions. A) ASAMD for unmatched cohorts; B) the ASAMD for cohorts 1:1 matched on a propensity score from a model without interactions between predictors; C) ASAMD for cohorts 1:1 matched on a propensity score from a model with interactions between age, gender, heart failure status, and other baseline characteristics. For (B) and (C), the ASAMD is shown after applying each propensity score matching strategy from Web Table 2. The strategies A, B, C1_1, C1_2, C1_3, D1_1, D1_2, D1_3, and E are shown in order from left to right.

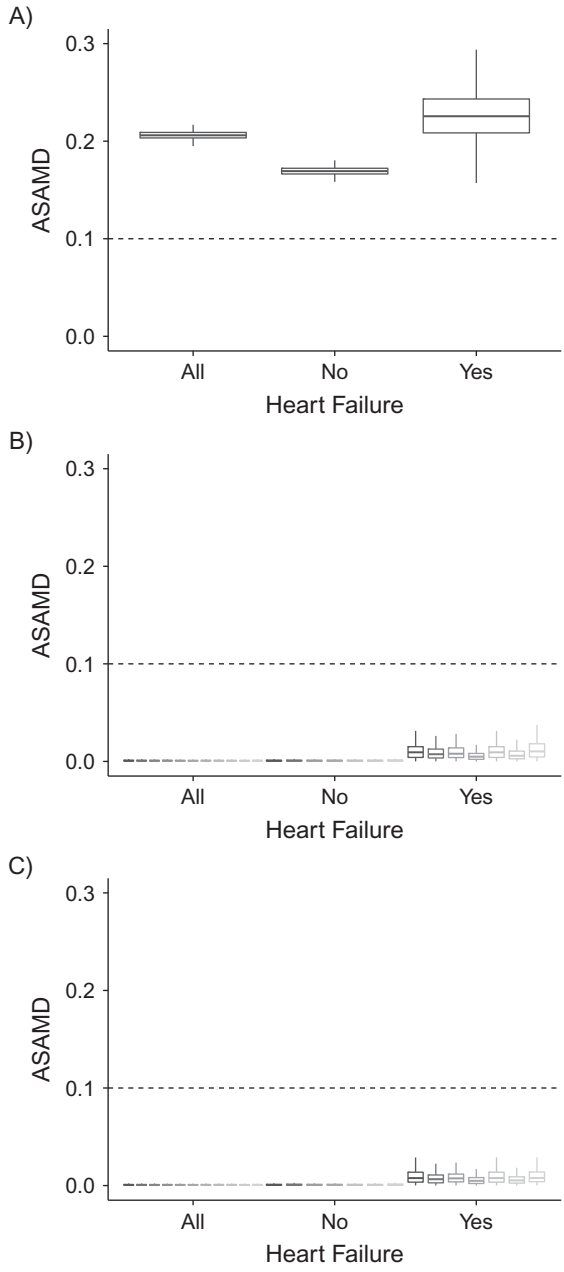


**Figure 4.** Distribution of average standardized absolute mean difference (ASAMD) for the baseline scenario broken down by heart failure status at baseline (500 bootstrap samples), using simulated data. The ASAMD is presented for the unmatched cohort, a cohort 1:1 matched on a propensity score from a model that included only main terms for predictors, and a cohort 1:1 matched on a propensity score from a model that included interactions. A) ASAMD for unmatched cohorts; B) ASAMD for cohorts 1:1 matched on a propensity score from a model without interactions between predictors; C) ASAMD for cohorts 1:1 matched on a propensity score from a model with interactions between age, gender, heart failure status, and other baseline characteristics. For (B) and (C), the ASAMD is shown after applying each propensity score matching strategy from Web Table 2. The strategies A, B, C1_1, C1_2, C1_3, D1_1, D1_2, D1_3, and E are shown in order from left to right.

Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019

**Table 3.** Strengths and Limitations of Alternative Methods Using Propensity Score Matching for Subgroup Analysis

| Strategy | Strengths | Limitations |
|---|---|---|
| A[a] | Main analysis uses best available matches from identified cohort.<br>Generally has larger matched overall cohort and subgroup sample sizes (difference may be slight). | Subgroups are not fully nested within main analysis matched cohort (best matches may include members in different subgroup strata). |
| B | Subgroups are fully nested within main analysis matched cohort. | Size of subgroups will often be smaller than strategy A (difference may be slight). |
| C | Subgroups are fully nested within main analysis matched cohort. | If post hoc subgroups are evaluated, these would not be nested within original main analysis. Fully nested post hoc subgroups would involve rematching to create a new main analysis cohort.<br>More likely to have convergence issues. |
| D | Subgroups are fully nested within main analysis matched cohort. | If post hoc subgroups are evaluated, these would not be nested within original main analysis. Fully nested post hoc subgroups would involve rematching to create a new main analysis cohort. |
| E | No need to rematch within subgroups. | Tended to have worst performance on several metrics in multiple scenarios. |

[a] Strategy A is already available in the Cohort Identification and Descriptive Analysis + Propensity Score matching tool (2).

below 0.95 across scenarios, with a few exceptions. Coverage was very similar across the evaluated PS matching strategies within matched subgroups. Model-based standard errors for the treatment effect were smaller than the standard deviations of treatment-effect estimates (overall and within subgroup) from the 500 simulated data sets in each scenario. While not the focus of this paper, appropriate measures of variability in estimates of 1:1-matched time-to-event analyses in settings of rare outcomes should be further evaluated—for example, use of robust standard errors or otherwise accounting for variability from estimating the PS.

### Concordance of subgroup and main analysis populations

PS matching strategies C, D, and E forced the same individuals to contribute to the main and subgroup analyses. PS matching strategies A and B resulted in some exposed patients being included in the main analysis that were not included in the subgroup analysis. PS matching strategy A also included exposed patients who were matched in the subgroup analyses but not the main analysis. Nevertheless, even when the exact individuals in the matched analysis varied, the total proportion of exposed patients who were matched in the main and subgroup analyses for each PS matching strategy was similar across simulation scenarios (Web Figure 7A).

The pattern of matched outcomes that contributed to the main and subgroup analyses was similar to the pattern for exposed patients (Web Figure 7B).

### Broken matches

We evaluated the proportion of matched sets where the 1:1-matched pairs in the main analyses belonged to different subgroup strata. Across scenarios, the level of discordancy in subgroup membership was less than 2%. The proportion of matched pairs with members in different subgroup strata was highest in scenarios with a small cohort (scenario 1), rare exposure (scenario 3), and where confounding varied by subgroup (scenarios 10–12).

### Convergence

The only PS matching strategy with convergence warnings from fitting the PS was strategy C, which involved fitting the PS within unique strata formed by the prespecified subgroups. When heart failure was prespecified as the sole subgroup for strategy C, 14% of the models fitted in the group with heart failure at baseline had convergence warnings in scenario 1 (small cohort), and 41% had convergence warnings or failure in scenario 4 (strong discrimination). When 3 subgrouping variables were prespecified, between 0.2% and 83.0% of the unique strata within scenarios had convergence warnings or failure.

## DISCUSSION

We conducted a plasmode simulation to compare the performance of 5 methods for using PS matching for subgroup analysis, including those not previously evaluated. We examined numerous factors that could influence balance, bias, and precision of effect estimation. In general, the matching strategies were affected similarly on those metrics across 12 scenarios. However, when confounding or treatment effect varied across subgroups, balance metrics were modestly improved when interactions by subgroup were included in the PS. Prior studies comparing a subset of the evaluated strategies had similar findings (10–14).

Comparing the relative strengths and limitations of the methods we evaluated against the PS matching method for subgroup analysis already available with CIDA + PS matching (strategy A) (Table 3), we noted that although strategy E did not require rematching within subgroup, it tended to have worse performance on several metrics. Strategies B, C, and D had subgroups that were nested within the main analysis cohort, but strategy C



Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019

frequently had convergence issues, and strategy B may have smaller subgroups than strategy A. For strategies C and D, if there are post hoc subgroup analyses, those would not be nested unless a new main analysis cohort was created by aggregating matches within the new unique combinations of subgroup strata. The new main analysis cohort could include different numbers of matches and outcomes from the original. Strategy C, which involved matching on a PS model fitted within unique strata of prespecified subgroups, was the only method that had warnings about convergence when fitting the PS. A major limitation of this strategy is that with more prespecified subgroups, the unique subgroup strata could become quite small, resulting in issues with convergence of models for PS estimation.

While performance was quite similar across the 5 compared methods in terms of balance, bias, and precision, other factors merit consideration when choosing which method to implement. All but one of the compared methods required fitting only 1 PS model, which reduces concern about running into convergence issues. Some methods involved matching for the main analysis and rematching of patients for subgroup analysis, resulting in different exposed patients or outcomes being included in the matched main versus subgroup analyses. While this did not have an impact on balance, bias, or precision, the number of exposed patients or outcomes across subgroups sometimes did not sum up to the total in the main analysis. Other methods ensured that subgroups were fully nested in the main analytical cohort by matching within subgroups and aggregating for the main analysis. However, if the investigators identify additional subgroups to evaluate after running the preplanned analyses, then using some subgroup matching strategies would result in a different main analysis cohort, because the main analysis cohort is created by aggregating matched sets within subgroups. For other strategies, using rematching within subgroups, irrespective of the main analysis matches, the patients matched in the subgroup analysis may differ from those matched in the main analysis.

Prior simulation studies have evaluated use of overall versus subgroup-specific PS in situations where the true PS varied by subgroup, but they have not compared the performance of methods frequently used in applied studies (10–14). These studies have focused on covariate balance as a metric rather than bias, precision, and coverage, or have evaluated use of PS adjustment in the outcome model rather than matching. To our knowledge, our simulation was the first to compare the performance of 5 PS matching methods for subgroups commonly implemented in the literature. We evaluated performance of these matching methods in a variety of plausible scenarios that included variation in exposure and outcome prevalence, the type of heterogeneity in effect on a relative scale, and whether confounding varied by subgroup. We considered numerous metrics when evaluating the relative performance of the 5 evaluated methods, including balance of baseline covariates, bias, precision, coverage, concordance of subgroup and main analysis populations, convergence issues, and broken matches in subgroup analyses. These metrics provided insight into statistical issues as well as logistical concerns with implementation and interpretation of results.

Our study has some limitations. The simulation involved only 7 binary covariates as potential confounders. In real studies, there can be dozens or hundreds of binary, categorical, and continuous covariates. The generally comparable performance of strategy E in our simulation may be related to the low prevalence of broken matches observed in our simulated scenarios. We theorize that the low prevalence of broken matches with strategy E would not be observed in a higher-dimensional PS that includes more covariates (or continuous covariates). While our simulations included a range of plausible scenarios that might be encountered in applied database studies, more extreme scenarios in terms of sample size, strength of confounding across subgroups, or other factors could result in greater divergence in performance of the evaluated subgroup PS matching strategies. The performance of subgroup matching methods in other scenarios could be evaluated in future work.

The methods evaluated for using PS to match for subgroup analyses in this simulation study are all commonly used in the literature and performed similarly in terms of balance, bias, and precision in scenarios varying size of the cohort, prevalence of exposure and outcome, strength of relationships between baseline covariates and exposure, the true effect within subgroups, and the degree of confounding within subgroups. However each had strengths and limitations with respect to other performance metrics that could inform choice of method. Future work could evaluate whether or how prevalence of broken matches correlates with bias and precision of strategy E.

In theory, using subgroup-specific PS to match within subgroups would be a preferred strategy. However, in practice, small subgroup sizes or numerous prespecified subgroups will preclude this approach. Using an overall PS (with interaction terms) to match within subgroups and then aggregating to form the main analytical cohort is more feasible analytically and ensures that subgroups sum to the overall matched cohort. That said, each of the 5 subgroup PS-matching methods evaluated were reasonable alternatives with similar performance in terms of balance, bias, and precision in our simulations. Because none of the subgroup PS-matching methods was clearly superior in terms of balance, bias, or precision, the decision of which to use can be context-dependent. For some investigators, it will be more important to have subgroups that are fully nested in the main analysis than to have main analyses based on finding the best overall matches (independent of subgroups). For others, the reverse will be true. The choice of matching strategy may also depend on the size of the study population, how many subgroups have been prespecified, and the how strongly confounding is expected to differ within subgroup strata.

## ACKNOWLEDGMENTS

Author affiliations: Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Shirley V. Wang, Yinzhu Jin, Mengdong He, Richard Wyss, Sebastian Schneeweiss, Joshua J. Gagne); Kaiser Permanente Vaccine Study Center, Kaiser Permanente Northern California, Oakland, California (Bruce Fireman); Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Susan Gruber, HoJin Shin); Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland (Yong Ma, Stephine Keeton); and Office of



Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019

Pharmacovigilance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland (Sara Karami, Jacqueline M. Major).

This work was supported by the US Food and Drug Administration (grant HHSF223200910006I).

This article reflects the views of the authors and should not be construed to represent the Food and Drug Administration's views or policies.

S.V.W. was principal investigator on an investigator-initiated grant from Novartis Pharmaceuticals Corporation to Brigham and Women's Hospital and was a consultant to Aetion, Inc., for unrelated work. J.J.G. was principal investigator on an investigator-initiated grant from Novartis Pharmaceuticals Corporation to Brigham and Women's Hospital and was a consultant to Aetion, Inc., and Optum, Inc., for unrelated work. S.S. is consultant to WHISCON, LLC, and to Aetion, Inc., a software manufacturer in which he also owns equity; he is principal investigator of investigator-initiated grants to the Brigham and Women's Hospital from Bayer, Genentech, and Boehringer Ingelheim for unrelated work. The other authors report no conflicts.

## REFERENCES

1. Platt R, Carnahan RM, Brown JS, et al. The US Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 1):1–8.
2. Sentinel Initiative. *Querying Tools: Overview of Functionality and Technical Documentation*. Silver Spring, MD: US Food and Drug Administration; 2017. https://www.sentinelinitiative.org/sites/default/files/SurveillanceTools/RoutineQuerying/Sentinel-Routine_Querying_System-Documentation_5.1.0.pdf. Accessed December 12, 2017.
3. Gagne JJ, Wang SV, Rassen JA, et al. A modular, prospective, semi-automated drug safety monitoring system for use in a distributed data environment. *Pharmacoepidemiol Drug Saf*. 2014;23(6):619–627.
4. Gagne JJ, Wang SV, Rassen JA, et al. *Mini-Sentinel Methods: Developing, Implementing, and Testing a Program for High-Dimensional Propensity Score Adjustment in the Mini-Sentinel Distributed Data Environment*. Silver Spring, MD: US Food and Drug Administration; 2013. https://www.sentinelinitiative.org/sites/default/files/Methods/Mini-Sentinel_High-Dimensional-Propensity-Score-Adjustment_0.pdf. Accessed June 12, 2017.
5. Connolly JG, Maro JC, Wang SV, et al. *Mini-Sentinel Methods: Developments, Applications, and Methodological Challenges to the Use of Propensity Score Matching Approaches in FDA's Sentinel Program*. Silver Spring, MD: US Food and Drug Administration; 2016. https://www.sentinelinitiative.org/sites/default/files/Methods/Sentinel-Methods_PSM-Approaches-in-Sentinel.pdf. Accessed June 12, 2017.
6. Toh S, Reichman ME, Houstoun M, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. *Arch Intern Med*. 2012;172(20):1582–1589.
7. Gagne JJ, Han X, Hennessy S, et al. Successful comparison of US Food and Drug Administration Sentinel analysis tools to traditional approaches in quantifying a known drug-adverse event association. *Clin Pharmacol Ther*. 2016;100(5):558–564.
8. Toh D, Hennessy S, Reichman ME, et al. Drugs that act on the renin-angiotensin-aldosterone system (angiotensin converting enzyme inhibitors, angiotensin receptor blockers, aliskiren) and angioedema. Silver Spring, MD: US Food and Drug Administration; 2012. https://www.sentinelinitiative.org/drugs/assessments/drugs-act-renin-angiotensin-aldosterone-system-angiotensin-converting-enzyme. Accessed May 12, 2017.
9. Wang SV, He M, Jin Y, et al. A review of the performance of different methods for propensity score matched subgroup analyses and a summary of their application in peer-reviewed research studies. *Pharmacoepidemiol Drug Saf*. 2017;26(12):1507–1512.
10. Rassen JA, Glynn RJ, Rothman KJ, et al. Applying propensity scores estimated in a full cohort to adjust for confounding in subgroup analyses. *Pharmacoepidemiol Drug Saf*. 2012;21(7):697–709.
11. Kreif N, Grieve R, Radice R, et al. Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Med Decis Making*. 2012;32(6):750–763.
12. Radice R, Ramsahai R, Grieve R, et al. Evaluating treatment effectiveness in patient subgroups: a comparison of propensity score methods with an automated matching approach. *Int J Biostat*. 2012;8(1):25.
13. Green KM, Stuart EA. Examining moderation analyses in propensity score methods: application to depression and substance use. *J Consult Clin Psychol*. 2014;82(5):773–783.
14. Girman CJ, Gokhale M, Kou TD, et al. Assessing the impact of propensity score estimation and implementation on covariate balance and confounding control within and across important subgroups in comparative effectiveness research. *Med Care*. 2014;52(3):280–287.
15. Franklin JM, Schneeweiss S, Polinski JM, et al. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
16. Rassen JA, Huang DM, Schneeweiss W, et al. Pharmacoepidemiology Toolbox. Boston, MA. http://www.drugepi.org/dope-downloads/#Pharmacoepidemiology%20Toolbox. Published August 19, 2017. Accessed January 15, 2017.
17. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083–3107.
18. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.

Downloaded from https://academic.oup.com/aje/article-abstract/187/8/1799/4937537 by guest on 06 September 2019