Evaluating the performance of Austin's standardized difference heuristic in observational cohort studies with varying sample size

Mitchell Conover, Azza Shoaibi, Joshua Ide, Martijn Schuemie

Background

One common approach to assessing potential confounding by measured covariates in observational causal studies is to compare covariate distributions between exposed and unexposed patients, either before or after adjusting for confounding. The absolute standardized mean difference (ASMD) is a frequently used statistical measure, originally asserted by Austin et al. A long-standing practice has been to define any covariate with ASMD≥0.10 as a potentially problematic confounder. This cut-point was established during an era of observational research where covariates adjusted for in analyses comprise a small, manually curated set of suspected confounders.[1] Due to increased use of secondary health data and large-scale/high-dimensional adjustment approaches, researchers increasingly assess large covariate sets including hundreds of variables. In this context, requiring all variables be to balance with ASMD<0.10 may be unnecessarily conservative. In studies with small sample sizes, chance imbalances are likely to occur for variables that have no meaningful relationship with the outcome and do not reflect problematic confounding. This study aims to assess the ability of Austin's standardized difference statistic to accurately identify problematic covariate imbalance in observational cohort studies of various sample sizes.

Methods

We implemented an applied methods evaluation within two previously-investigated[2] clinical settings: 1) angiotensin-converting enzyme inhibitors (ACEi) vs. thiazide/thiazide-like diuretics (TZD), where a priori we expect less confounding, and 2) ACEi vs. beta-blockers (BB), where a priori we expect more confounding. Within randomly-selected sub-samples, we evaluated ASMD diagnostics in various analytical scenarios adjusting for highly-dimensional covariate sets, using the empirical null distribution to indicate systematic bias.

The study was conducted using data from three nation-wide U.S. administrative claims databases which have been mapped to the OMOP Common Data Model (version 5.3.1):

1. Optum's Clinformatics® Extended Data Mart – Date of Death (Optum DOD)[3],
2. IBM Health MarketScan® Commercial Claims and Encounters Database (IBM CCAE), and
3. IBM Health MarketScan® Medicare Supplemental and Coordination of Benefits Database (IBM MDCR).

The study population included adult (≥18) patients who received at least one eligible prescription for an exposure drug (ACEi, BB, or TZD) and who are observable in each database for at least one year prior to the index date (first observed instance of drug exposure). Patients were required to have a hypertension diagnosis at any point before/on the index date. Cohort exit was defined as the earliest: occurrence of an outcome event, end of exposure, death, loss or disenrollment from the database, or date of last data collection. We separately analyzed a set of 135 randomly-selected sub-samples from the original target/comparator cohorts as follow: 5 20% samples, 10 10% samples, 20 5% samples and 100 1%

samples. For each sub-sample proportion, we present the average of two ASMD statistics: 1) the maximum ASMD for the most imbalanced covariate within that sub-sample and the proportion of covariates with ASMD≥0.10.

We estimated causal effects corresponding to a set of 123 negative control outcomes, (i.e. events that we are relatively confident have no causal relationship with the exposures assessed). Thus, in an analysis with no little-to-no sources of systematic bias, we expect effect estimates to tightly cluster around a value of one, with an approximately normal distribution. Austin's StdzDiff heuristic was used to identify problematic covariate imbalance within three analytic scenarios: 1) a crude analysis: with no analytic adjustment for confounding (crude), 2) 1:1 propensity score matching using data-driven, large-scale regularized logistic regression (PS-matched), and 3) an analysis using a randomly simulated exposure variable (pseudo-random). Estimates of the propensity score used to adjust comparisons were modeled in the full/pooled cohort and the 1:1 matching procedure was applied within each of the population sub-samples.

In the full pooled cohort, Cox proportional hazards models were used to estimate hazard ratios (HRs) between target and comparator treatment cohorts for the risk of each negative control outcome. HR estimates and empirical null distributions were generated within the combined cohort, which combines crude/matched populations produced in each sub-sample. To quantitatively summarize the empirical null distribution, we used the expected systematic error (ESE) statistic and its 95% confidence intervals, which is an aggregate summary of the individual estimates corresponding to each negative control outcome that generates a value between 0 and 1.

Results

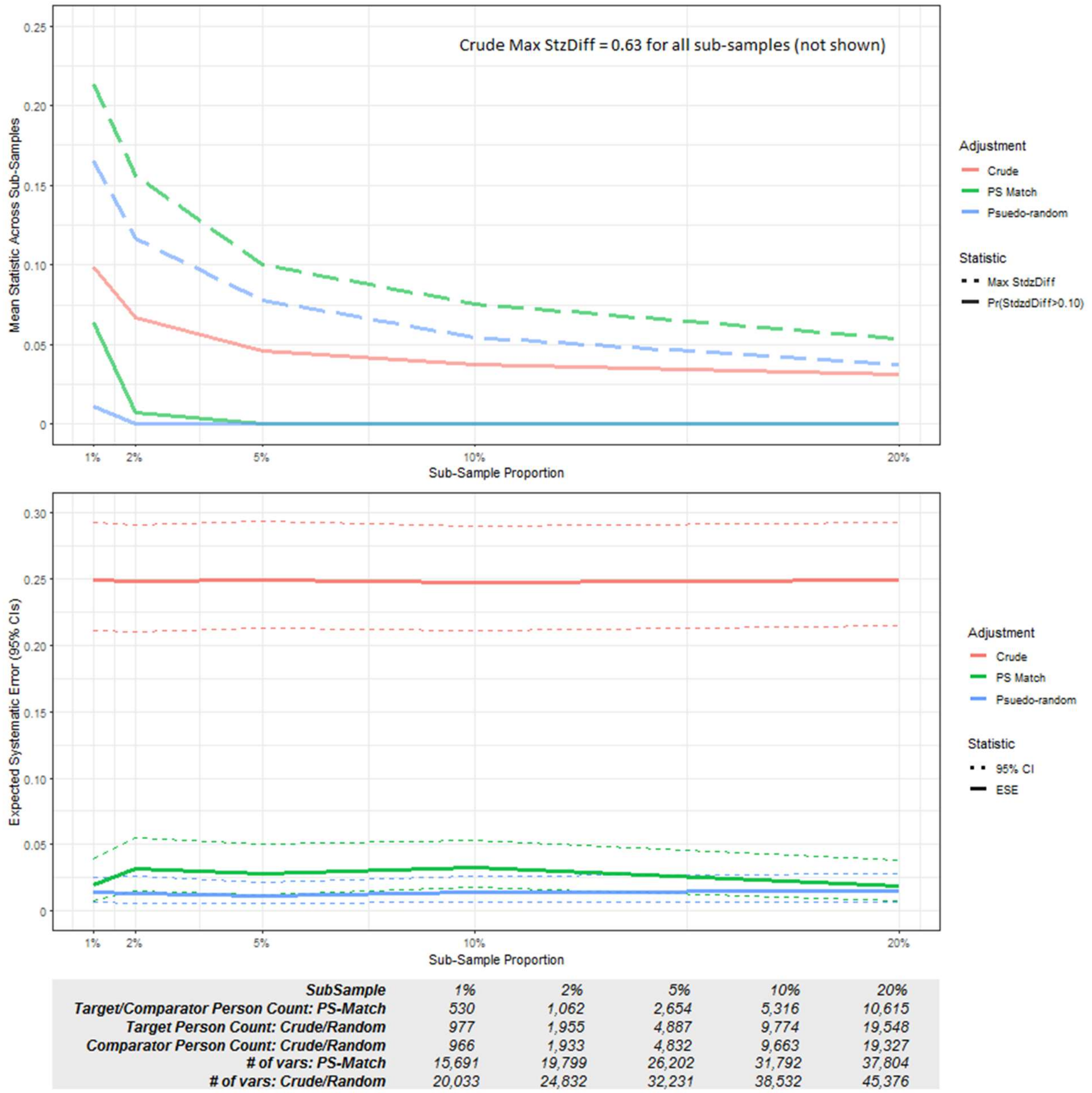
The three databases studied produced approximately equivalent findings, thus we chose to describe the findings for Optum DOD. Figures 1 and 2 show the results for the Optum DOD comparisons of ACEi-BB and ACEi-TZD, respectively. The crude approach failed the diagnostic (i.e. had a covariate with ASMD> 0.10) for both comparisons, in all databases, for all sub-samples considered. In Optum DOD, across all sub-samples, the average of the maximum standardized mean difference for the crude approach was approximately 0.63 and 0.77 for the ACEi-BB and ACEi-TZD comparisons, respectively. The PS-matched approach passed the diagnostic for larger samples but failed at sub-samples ≤5% for the ACEi-BB comparison and sub-samples ≤2% for the ACEi-TZD comparison. Of all covariates considered for both the crude and PS-matched approaches, the proportion with ASMD≥0.10 was relatively stable for sub-samples between 5% and 20%, increased slightly for 2% ($N_{\text{target-matched}}$=1,062) sub-samples and increased substantially for 1% sub-samples ($N_{\text{target-matched}}$= 530).

Despite changes in the ASMD diagnostic at lower sample sizes, the ESE statistic did not indicate any meaningful change in systematic error or the ability of the PS-matched approach to generate unconfounded estimates. For the crude ACEi-BB comparison, the ESE was approximately 0.25 across all sub-samples and approximately 0.01 for the pseudo-random approach. The ESE for the PS-matched approach ranged between 0.02 and 0.03, indicating a relatively unconfounded adjusted estimate nearly equivalent to the unconfounded pseudo-random approach. Findings were similar for the ACEi-TZD comparison, although confounding was less pronounced than in the ACEi-BB comparison. For both comparisons (ACEi-TZD and ACEi-BB), confidence bands around the PS-matched estimate for the smallest (1%) sub-sample contained the point-estimate for the (unconfounded) pseudo-random approach.

Conclusions

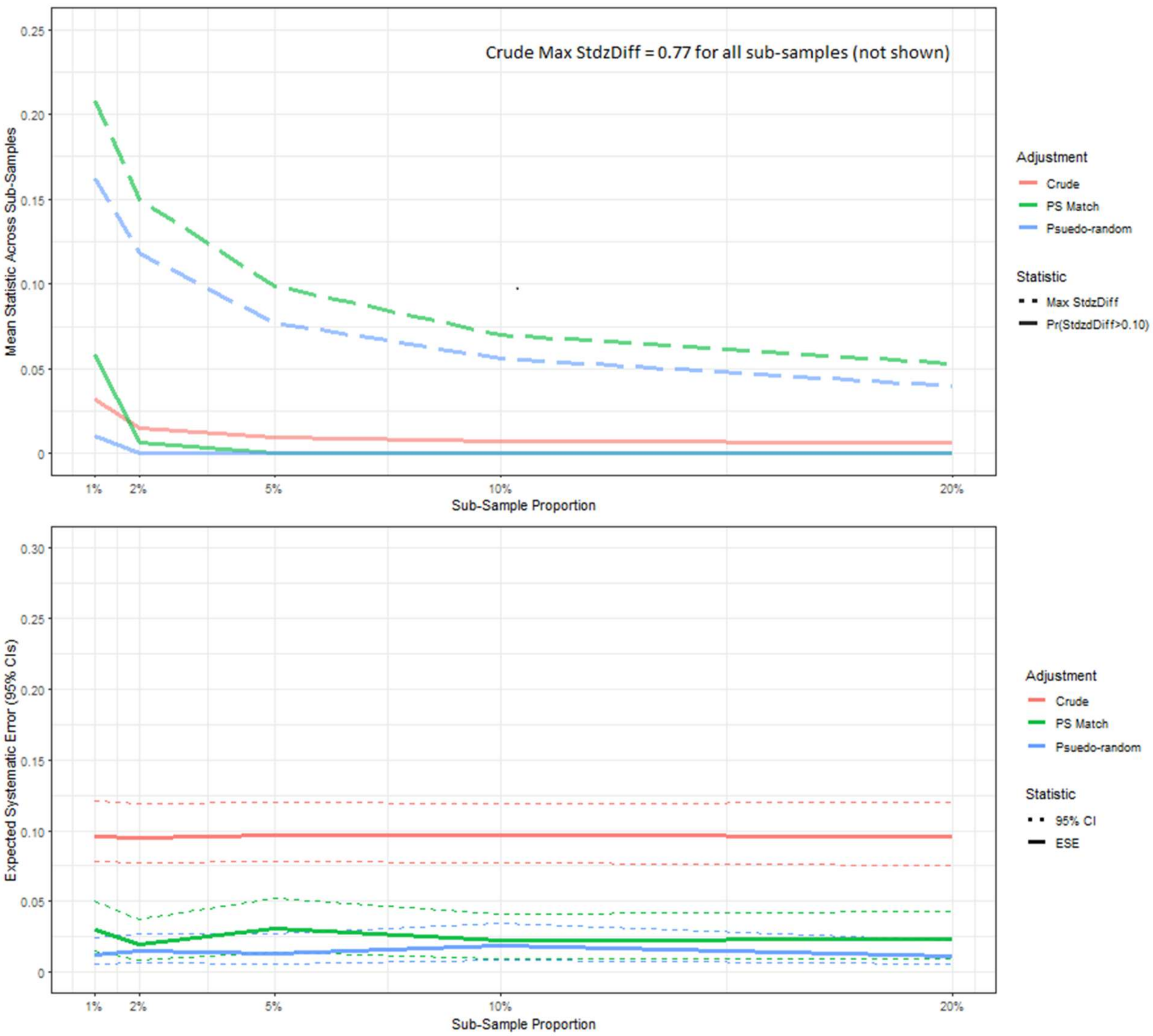
These results demonstrate a loss in specificity of Austin’s ASMD diagnostic for identifying problematic confounding at smaller sample-sizes (i.e., more likely to indicate problematic confounding even though none exists). This finding motivates future development of alternative strategies for understanding potential confounding in contexts where sample sizes are small and covariate sets are large.

Figure 1. Results for Optum DOD ACEi-BB comparison



SubSample	1%	2%	5%	10%	20%
Target/Comparator Person Count: PS-Match	530	1,062	2,654	5,316	10,615
Target Person Count: Crude/Random	977	1,955	4,887	9,774	19,548
Comparator Person Count: Crude/Random	966	1,933	4,832	9,663	19,327
# of vars: PS-Match	15,691	19,799	26,202	31,792	37,804
# of vars: Crude/Random	20,033	24,832	32,231	38,532	45,376

Figure 2. Results for Optum DOD ACEi-TZD comparison



SubSample	*1%*	*2%*	*5%*	*10%*	*20%*
Target/Comparator Person Count: PS-Match	*552*	*1,104*	*2,764*	*5,529*	*11,050*
Target Person Count: Crude/Random	*976*	*1,952*	*4,879*	*9,758*	*19,516*
Comparator Person Count: Crude/Random	*975*	*1,949*	*4,873*	*9,746*	*19,493*
# of vars: PS-Match	*14,952*	*18,835*	*24,946*	*30,288*	*36,171*
# of vars: Crude/Random	*18,588*	*23,076*	*29,950*	*35,871*	*42,234*

References

1. Austin, P. C. (2009). "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples." Stat Med 28(25): 3083-3107.
2. Suchard, M. A., et al. (2019). "Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis." Lancet 394(10211): 1816-1826.
3. Optum's de-identifed Clinformatics®Data Mart Database (2007-2020).