Evaluating confounding adjustment when sample size is small

Fleur Vereijken[1], Jenna Reps[1,2], Marc A. Suchard[3], Akihiko Nishimura[4], Linying Zhang[5], George Hripcsak[6], Peter Rijnbeek[1], Ross D. Williams[1], Martijn Schuemie[2,3]

[1] Erasmus University Medical Centre, Rotterdam, The Netherlands, [2]Global Epidemiology, Johnson & Johnson, [3]Department of Biostatistics, University of California, Los Angeles, [4] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, [5]Data Science and Biostatistics, Washington University School of Medicine, St. Louis, [6]Department of Biomedical Informatics, Columbia University Medical Center

Background

Observational studies estimating causal effects are vulnerable to confounding. In Observational Health Data Sciences and Informatics (OHDSI), this is typically addressed using large-scale propensity score (LSPS) models[1], which—when applied to large datasets—have shown good covariate balance and low residual systematic error as measured by negative controls[2]. However, it is unclear how well these methods perform when sample sizes are limited. This is increasingly relevant as OHDSI uses federated networks of databases, where each participating node may have only limited local data. Recent work also shows that standardized mean differences (SMDs) can misleadingly indicate imbalance in small samples, leading to falsely rejecting the validity of a study[3]. In practice, small datasets are often set aside because they appear insufficient on their own, even though they may still contain valid and valuable information. Here, we emulate a federated network by splitting large real-world data into smaller sets and evaluate LSPS performance under these conditions.

Methods

To measure the performance of LSPS under small sample sizes, we take large study populations and divide them into smaller sets, where we fit propensity models. To maintain power, we then recombine the smaller sets, allowing us to compare effect estimates against a gold standard of real negative and synthetic positive controls where the true effect sizes are known.

Ground truth

We evaluate a large set of target-comparators to find a set where bias (measured through negative control outcomes) is large when not adjusting for any confounding. From this search we selected six target-comparators, varying in levels of bias, to examine the impact of confounding adjustment across different scenarios. The level of bias is expressed in the lower bound of the expected absolute systematic error (EASElb).

1. Losartan vs hydrochlorothiazide (EASElb = 0.26), with 76 negative controls (taken from LEGEND-Hypertension (HTN))
2. Quinapril vs propronalol (EASElb = 0.65), with 76 negative controls (taken from LEGEND-HTN)
3. Glimepiride vs saxagliptin (EASElb = 0.38), with 94 negative controls (taken from LEGEND-Type II diabetes (T2DM))
4. Sitagliptin vs dapagliflozin (EASElb = 0.06), with 94 negative controls (taken from LEGEND-T2DM)
5. Nortriptyline vs fluoxetine (EASElb = 0.51), with 52 negative controls (taken from LEGEND- Major depressive disorder (MDD))

6. Amitriptyline vs venlafaxine (EASEIb = 0.31), with 52 negative controls (taken from LEGEND-MDD)

To capture bias toward the null, we generate three synthetic positive controls per negative control. We synthesize positive controls, by adding simulated outcomes to real negative controls[4], achieving true effect sizes of 1.5, 2 and 4.

Simulating smaller sites

We then randomly divide the population into smaller partitions to simulate different data sites, as shown in Figure 1. From the full set of persons included at the start of the study (starting either treatment, having 365 days of observation prior, not being in both cohorts), we first randomly sample 20,000 patients. We then randomly divide (without replacement) the 20,000 patients into *k*=

- 5 sites of 4,000 persons
- 10 sites of 2,000 persons
- 20 sites of 1,000 persons
- 40 sites of 500 persons
- 80 sites of 250 persons

This approach allows us to assess how splitting data into smaller subsets impacts the ability of local propensity score models to account for confounding, while still maintaining overall power to evaluate bias

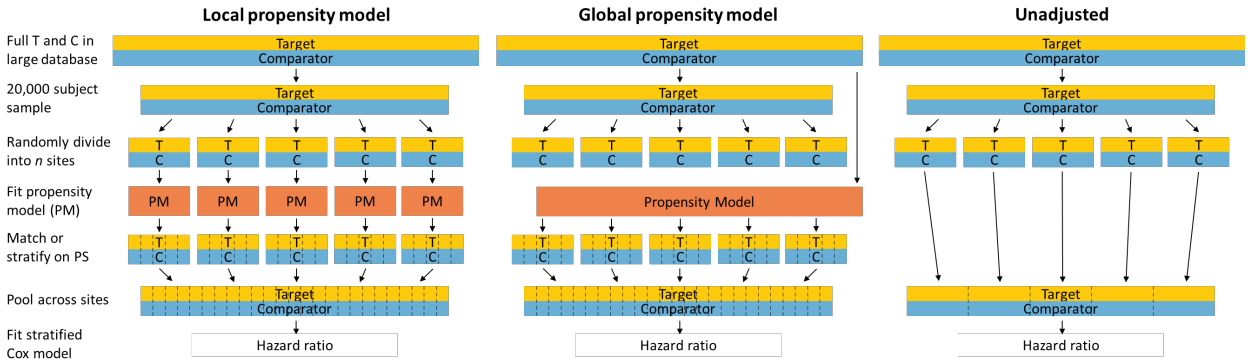

Figure 1. Simulating small data sites. We extract a target (T) and comparator (C) cohort from a large database and take a 20,000-person random sample. We then randomly divide these into *k* equally-sized sites. We evaluate propensity score adjustment using propensity models (PM) fitted at each simulated site (Local) or using a single PS model fitted on the original full data (Global), and compare this to no PS adjustment (Unadjusted). Data is pooled across simulated sites before fitting a stratified Cox model.

Data sources

We use the Merative Marketscan CCAE database. Later we will also include Optum EHR.

Propensity score adjustments

We compare treatment effect estimates under propensity scores (PS) computed in two different ways: using only the data at each site (`local`) and using the full population (`global`). The 'global' approach serves as a benchmark informing on any reduction in performance that is only due to the fact that as databases becomes smaller it becomes less likely an appropriate match for a patient exists. Subsequent 1-on-1 PS matching and PS stratification (into 10 equally-sized strata) are done locally at each site. Additionally, we also include an analysis without PS adjustment to assess the amount of confounding in a

study.

We estimate causal effects using Cox proportional hazards models, conditioning on the PS strata when performing PS stratification. A conditional Cox model limits the at-risk set in the likelihood denominator to only those subjects within the same stratum as the subject for which the likelihood is computed. It therefore allows for different baseline hazards within strata, while fitting a model across the entire population. When using 1-on-1 PS matching, we do not condition on matched sets.

Evidence synthesis

In many real-world settings, each site only shares summary-level results to be meta-analyzed. However, here we are primarily interested in assessing the small sample size performance of LSPS and therefore do not concern ourselves with how meta-analysis might also impact the quality of overall treatment effect estimate. We therefore forgo meta-analysis and pool person-level data from each site in fitting outcome models. These models do condition on the site. For PS stratification no PS site-strata are merged, resulting in a total of *k* x 10 strata in the model when there are *k* sites.

Metrics

We evaluated performance of treatment-effect estimates across four metrics, each reflecting a distinct aspect of confounding control:

- **Expected Absolute Systematic Error (EASE)** estimates the expected magnitude of residual bias by fitting a Gaussian distribution to the estimated negative control hazard ratios[5], and then taking the absolute expected value of that distribution. This quantifies how well confounding is controlled overall.
- **Geometric mean of the precision** (1 / (standard error)^2) after empirical calibration[6]. This reflects statistical precision – higher values indicate narrower confidence intervals for effect estimates.
- **Maximum standardized difference of mean (SDM)** is computed by dividing the difference between the mean in T and C by the standard deviation for each covariate and taking the maximum of the absolute value. This measures the largest imbalance in baseline covariates between treated and comparator groups.
- **SMD significance test** as proposed by Hripcsak et al. (2024) to assess balance under small-sample conditions[3]. A statistical test assessing whether standardized differences in covariates significantly exceed a prespecified threshold (e.g., 0.1), adjusting for sample size. It addresses the limitations of traditional SDM in small-sample settings by distinguishing meaningful imbalance from chance variation.

Results

Figure 2 shows a comparison of EASE scores when using the locally-fitted propensity model to when using the model fitted on the full data (both PS matching and stratification), and to no adjustment. Unadjusted analyses generally yielded high EASE scores, except for sitagliptin vs dapagliflozin, suggesting minimal confounding to begin with in this pair. In other target-comparator pairs, local PS adjustment consistently

lead to higher EASE scores compared to global PS adjustment, particularly as sample sizes decreased.

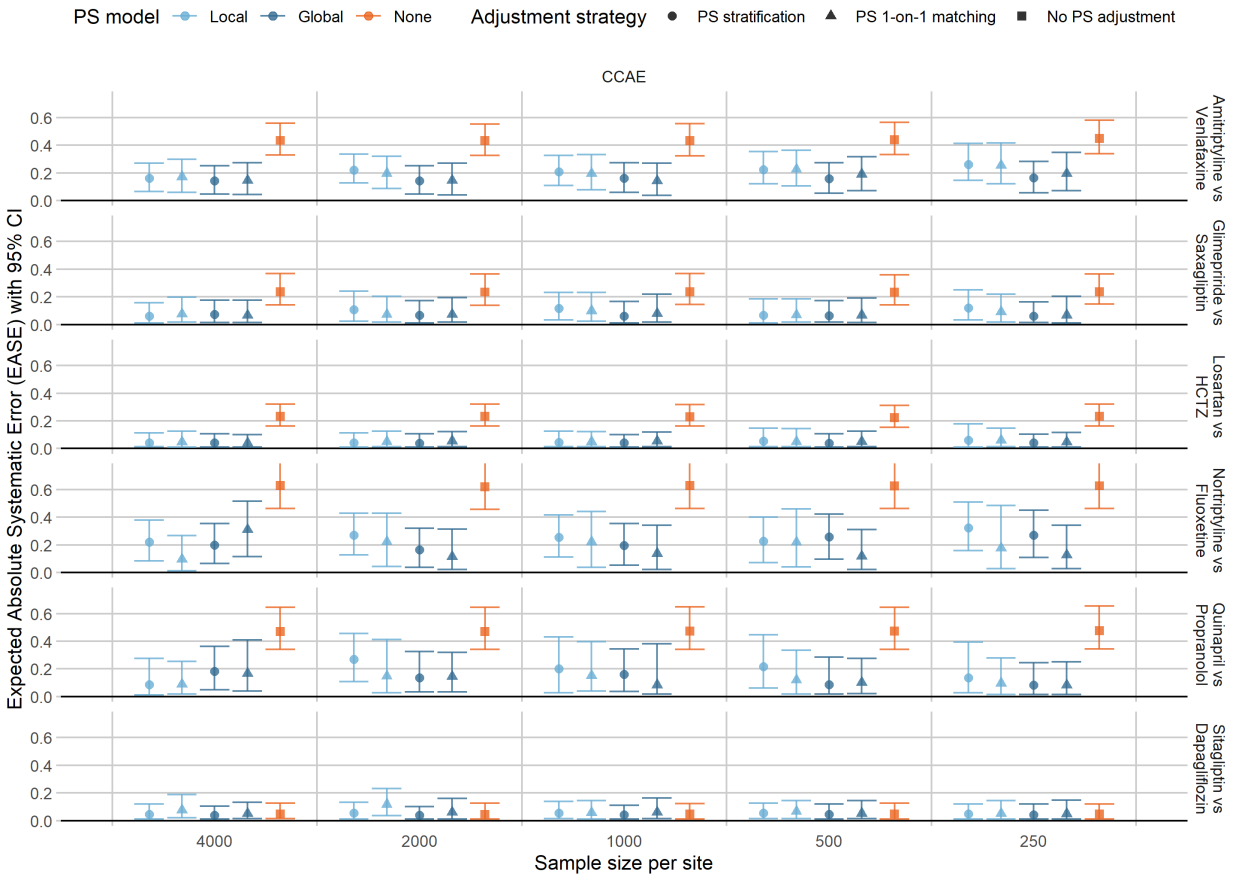

Figure 2. Expected Absolute Systematic Error (EASE) with 95% credible intervals per sample size.

To disentangle the effects of bias and precision, we examine the precision of the calibrated confidence interval (CI), which, by design, ensures a consistent nominal coverage across methods. This allows for comparing the approaches in the precision they produce (higher is better).

As shown in Figure 3, Across all target-comparator pairs, PS adjustment generally improves precision after empirical calibration. An exception is the sitagliptin vs. dapagliflozin comparison, where a slight reduction in precision is observed compared to no PS adjustment. The greatest gains in precision are seen with stratification, particularly when using the global propensity model. However, these gains do decrease when sample sizes decrease, specifically for the nortriptyline vs. fluoxetine and quinapril vs. propranolol comparisons

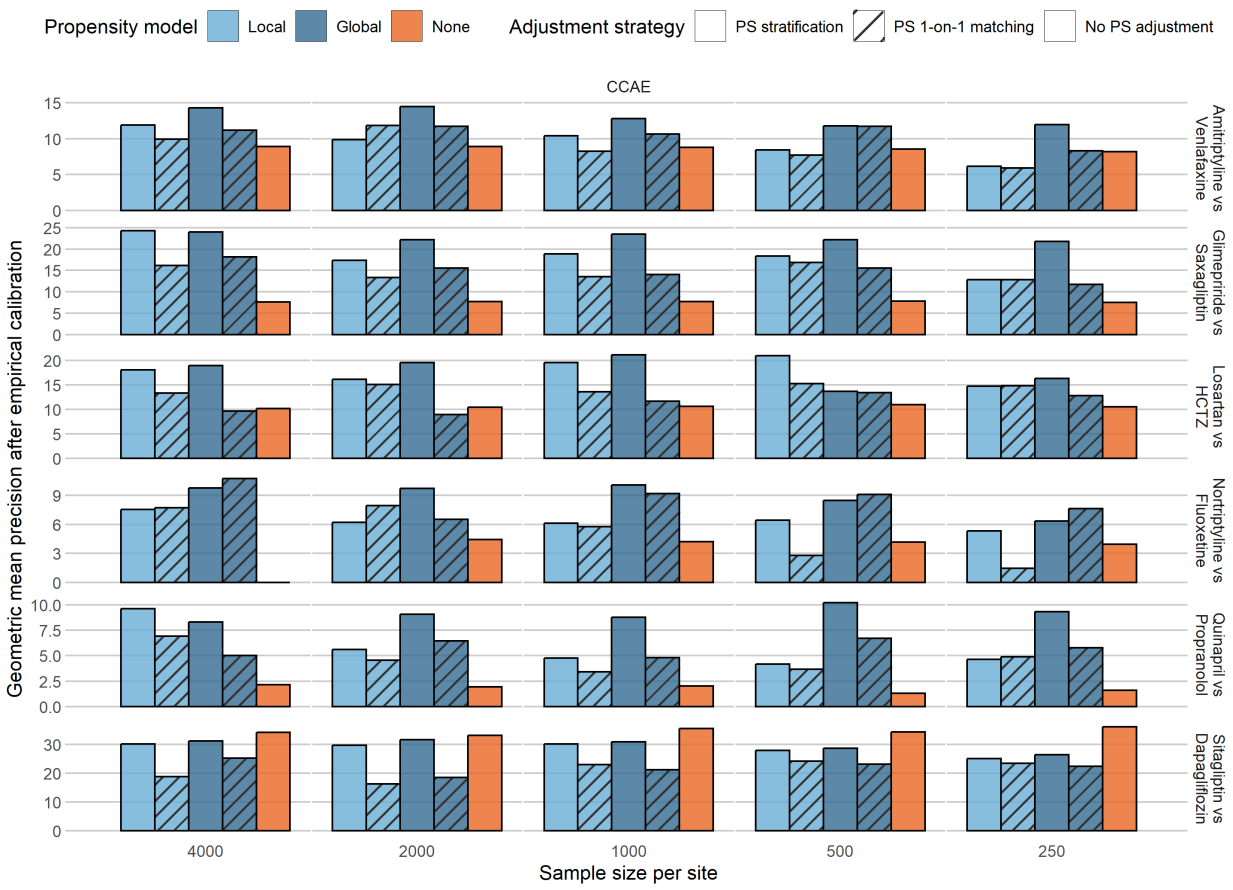

Figure 3. Geometric mean precision after empirical calibration based on both negative and positive control estimates.

If we consider a standard rule-of-thumb that the maximum absolute SDM must be no greater than 0.1 to declare balance between the two populations, we observe in Figure 4 that we always fail this diagnostic when sample size per site is <= 4,000.

Figure 4. Maximum absolute standardized difference of mean (SDM) per sample size. Max SDM is computed at each site, resulting in a distribution characterized by box plots. A max SDM below 0.1 is considered to indicate balance.

Following Hripcsak et al.[3], who showed that SMDs can be misleading in small samples, we apply their proposed SMD significance test. This method assesses whether observed SMDs significantly exceed a threshold (e.g., 0.1), accounting for sample size, and provides an overall count of significantly unbalanced covariates.

Figure 5 shows that for amitriptyline vs venlafaxine and losartan vs hydrochlorothiazide, covariate imbalance remains low until sample sizes drop below 2000. For other target-comparator pairs, matching (both local and global) consistently results in few significantly unbalanced covariates, regardless of sample size. In contrast, stratification approaches tend to yield more imbalanced covariates, even at larger sample sizes.

Figure 5. Significantly unbalanced covariates by sample size, based on the SMD significance test.

Conclusion

LSPS methods effectively reduce confounding, as measured by EASE, even in small sample settings. However, we observe declining performance in terms of precision after empirical calibration as sample size decreases. Some of this decline is also present when using the global propensity model, suggesting it partially may be attributed to lower power resulting from dividing the data in smaller and smaller strata.

Traditional SMD metrics often overstate imbalance in small datasets, while the SMD significance test—by adjusting for sample size—reduces false positives. Despite often having similar bias as measured by EASE, PS stratification tends to lead to many significantly imbalanced covariates. In contrast, for PS matching often no covariates are considered imbalanced even at the lowest sample sizes. This correction also limits statistical power, meaning true imbalances may go undetected when sample sizes are small. Hripcsak et al.[3] proposed that meta-analyzing balance diagnostics across sites (in network studies) or aggregating diagnostics alongside effect estimates helps mitigate the limitations caused by reduced power in individual sets. However, this has not been done in this study.

These findings suggest that LSPS still reduces confounding, even when sample sizes become small. However, for some of the target-comparator combinations confounder adjustment decreases when sample sizes become small. For these cases, LSPS with standard diagnostics, may be insufficient in low-sample settings. New strategies—such as incorporating prior knowledge of confounders, applying dimensionality reduction, or using cardinality matching—are needed to improve confounding control. The

framework presented here provides a basis for evaluating such future methods.

References

1. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. Int J Epidemiol. 2018 Dec 1;47(6):2005–2014. doi: 10.1093/ije/dyy120
2. Tian Y, Schuemie MJ, Suchard MA, Evaluating large-scale propensity score performance through real-world and synthetic data experiments, Int J Epidemiol. 2018 Dec 1;47(6):2005-2014. doi: 10.1093/ije/dyy120
3. Hripcsak G, Zhang L, Chen Y, Li K, Suchard MA, Ryan PB, Schuemie MJ, Assessing covariate balance with small sample sizes, medRxiv. 2024 Apr 24. doi: 10.1101/2024.04.23.24306230.
4. 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 Mar 13;115(11):2571–2577. doi: 10.1073/pnas.1708282114
5. 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 Jan 30;33(2):209-18. doi: 10.1002/sim.5925
6. Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Tian Y, Schuler A, Ryan PB, Madigan D, Hripcsak G, How Confident Are We about Observational Findings in Healthcare: A Benchmark Study, Harv Data Sci Rev. 2020;2(1):10.1162/99608f92.147cc28e. doi: 10.1162/99608f92.147cc28e