Evaluating confounding adjustment when sample size is small

Martijn Schuemie[1,2], Marc A. Suchard[2], Akihiko Nishimura[3] , Linying Zhang[4], George Hripcsak[4]

[1] Observational Health Data Analytics, Johnson & Johnson, [2] Department of Biostatistics, University of California, Los Angeles, [3] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, [4] Department of Biomedical Informatics, Columbia University Medical Center

Background

Observational studies estimating causal effects are vulnerable to confounding because groups receiving different treatments may differ in important aspects. OHDSI studies typically rely on large-scale propensity score (LSPS) models to adjust for these differences.[1] When treatment groups are sufficiently large, LSPS has proven to work well, both in terms of covariate balance and residual systematic error measured using negative controls.[2] However, little is known about LSPS's ability to adjust for confounding when treatment groups are small. To complicate matters, prior research shows that our ability to measure covariate balance — using the standardized difference of means (SDM) — degrades when sample size is limited.[3]

Methods

To measure performance of LSPS under small sample sizes, we take a large study population and randomly divide it into smaller partitions to simulate different data sites, as shown in **Figure 1**. After various adjustment strategies we pool the data again to compute a hazard ratio which we compare to the ground truth.

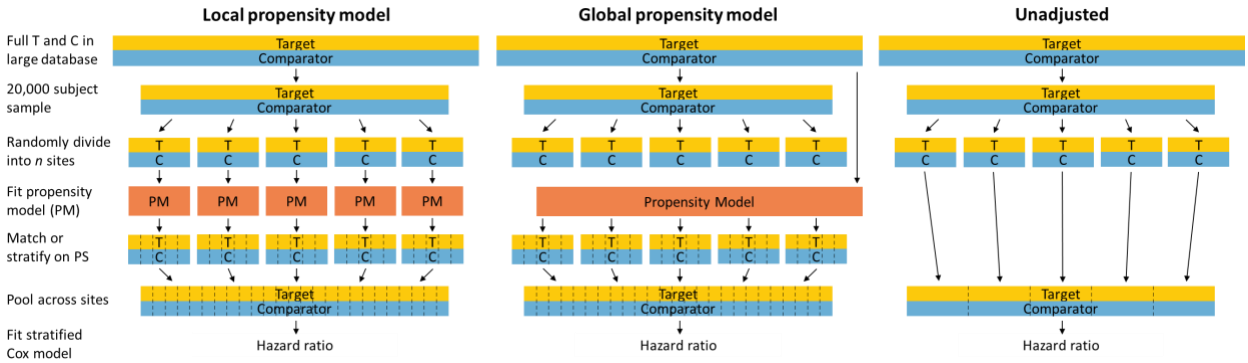

Figure 1. Simulating small data sites. We extract a target (T) and comparator (C) cohort from a large database and take a 20,000-person random sample. We then randomly divide these into *n* equally-sized sites. We evaluate propensity score adjustment using propensity models (PM) fitted at each simulated 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.

Ground truth

We specify four target-comparator treatment pairs, each with a set of negative control outcomes:

1. Lisinopril vs hydrochlorothiazide, with 76 negative controls (taken from LEGEND-HTN)
2. Lisinopril vs metoprolol, with 76 negative controls (taken from LEGEND-HTN)
3. Sitagliptin vs glimepiride, with 94 negative controls (taken from LEGEND-T2DM)
4. Sitagliptin vs liraglutide, with 94 negative controls (taken from LEGEND-T2DM)

For each negative control we generate three synthetic positive controls, with true effect size = 1.5, 2, or 4.

Data sources

We use the Merative MarketScan MDCD, MarketScan MDCR, and the Optum® de-identified Electronic Health Record dataset (Optum EHR).

Simulating smaller sites

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 *n* =

- 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
- 160 sites of 125 persons

Propensity score adjustments

We compare qualities of 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 the gold standard benchmark. 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.

Causal effects are estimated using Cox proportional hazards models, which are conditioned on the PS strata when performing PS stratification. A conditional Cox model does not include the stratum ID as predictor variable, but instead 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 (in the denominator). 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 a real-world setting, each site only shares a summary level data to be meta-analyzed. However, here we are primarily interested in assessing the small sample size performance of LSPS and therefore do not want to 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 *n* x 10 strata in the model when there are *n* sites.

Metrics

We use the following metrics to measure performance:

- Expected Absolute Systematic Error (EASE) is computed by first fitting a Gaussian distribution to the estimated negative control hazard ratios[5], and then taking the absolute expected value of that distribution.
- Geometric mean of the precision (1 / (standard error)^2) after empirical calibration,[4]
- 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.

Results

Figure 2 compares the EASE 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. In many scenarios the unadjusted analyses already produce low EASE scores, suggesting there is not much confounding to begin with. In situations where the unadjusted analyses do show significant systematic error, performance of the local PS adjustment does go down as sample size becomes smaller. For example, for sitagliptin vs liraglutide in the MDCR database we see an increase in EASE when sample size per site is equal to or smaller than 500.

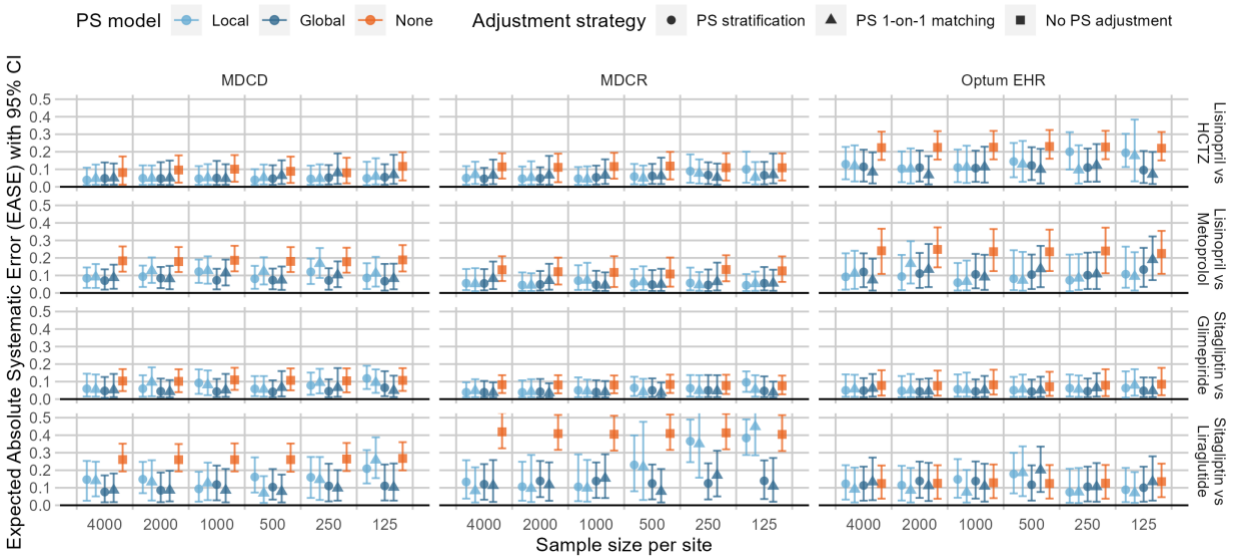

Figure 2. Expected Absolute Systematic Error (EASE) with 95% credible intervals per sample size.

Because it is difficult to compare performances when both precision and bias (as measured through EASE or coverage) vary at the same time, we additionally consider the precision of the calibrated confidence interval (CI). Because a calibrated CI by design achieves a common, pre-specified nominal coverage across different approaches, we can compare the approaches in the precision they produce (higher the better) **Figure 3** shows the precision of the calibration CI. Again, in general, we do not observe much difference between the local and global propensity scores, except for the sitagliptin vs liraglutide example in all three databases where in MDCR and Optum EHR we see a precision drop for the local model. This drop becomes larger as the per-site sample size goes down.

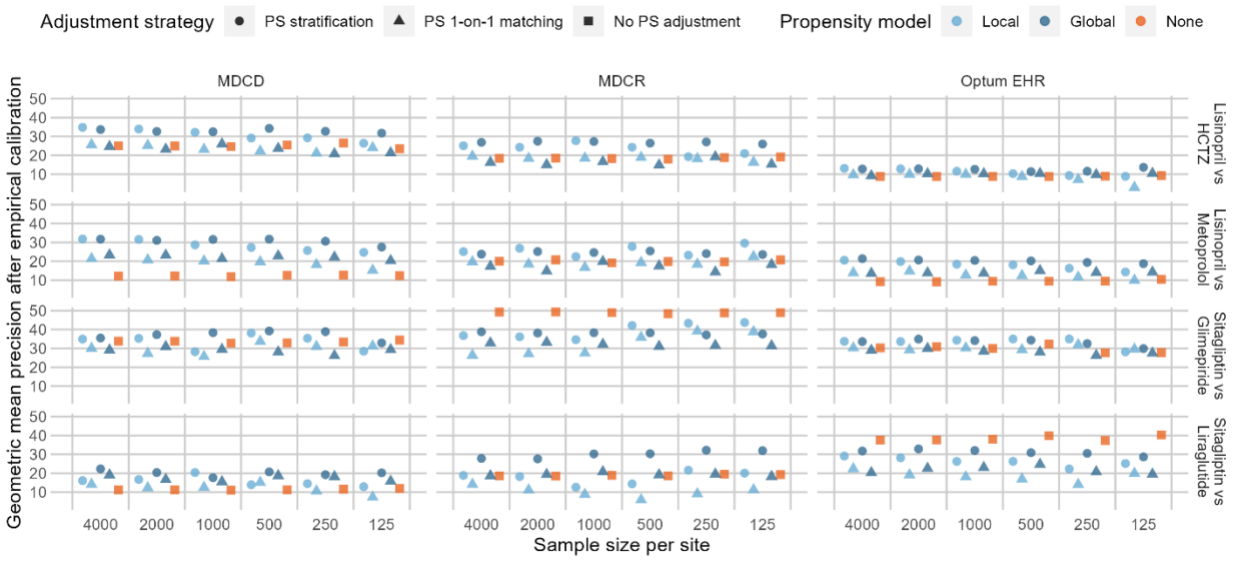

Figure 3. Geometric mean precision after empirical calibration based on both negative and positive control estimates.

If we consider our standard rule 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.

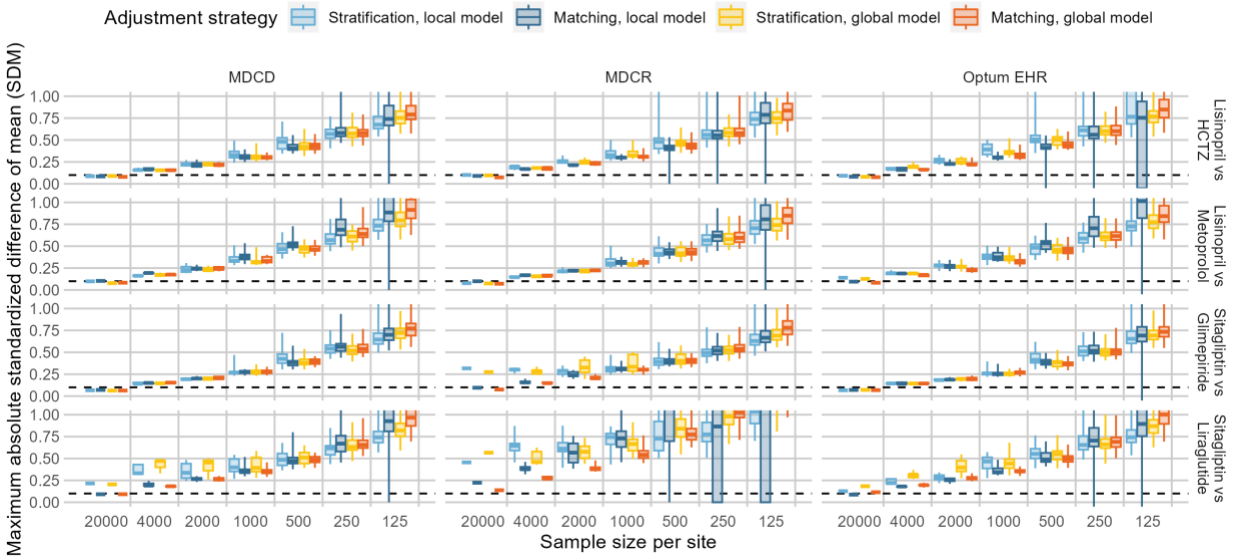

Figure 4. Maximum absolute standardized difference of mean (SDM) per sample size. Max SDM is computed at each site, resulting in a distribution characterized by box plots. A max SDM below 0.1 is considered to indicate balance.

Conclusion

In general we observe that LSPS remains capable of correcting for most confounding even when sites are small; difficulties still arise, however, in situations with known differences between treatments. New methods should be developed to address confounding in these situations, such as those having prior information about the confounders for which to adjust, dimensionality reduction before fitting propensity models, or cardinality matching. The framework described here can be used to evaluate these future methods.

As observed in our prior work,[3] our current balance metric of SDM almost always declares imbalance when sample size is small, even when there appears to be little residual confounding as measured using negative controls. A better balance metric, that at least can indicate when there is insufficient data to evaluate balance, should be developed.

As the OHDSI network grows, we will be able to study more and more rare exposures by combining data across many sites. Results here suggest that in many scenarios the LSPS method will be sufficient to address confounding in these studies. In some cases, however, new methodology is needed.

References

1. Zhang L, Wang Y, Schuemie MJ, Blei DM, Hripcsak G, Adjusting for Indirectly Measured Confounding Using Large-Scale Propensity Score, J Biomed Inform. 2022 Oct;134:104204. doi: 10.1016/j.jbi.2022.104204
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. Conover M, Shoaibi A, Ide J, Schuemie M, Evaluating the performance of Austin's standardized difference heuristic in observational cohort studies with varying sample size, OHDSI Symposium 2021
4. 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
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