Validating a clinical informatics consulting service using negative control reference sets

Michael L. Jackson[1], Saurabh Gombar[1], Raj Manickam[1], Robert Brown[1], Ramya Tekumalla[1], Yen Low[1]

[1] Atropos Health, New York, NY

Background

In the absence of evidence from randomized trials, clinical decisions can be supported using an "informatics consult" - a timely analysis of patient data collected in Electronic Health Records (EHRs). To provide results within a clinically relevant timeframe, an informatics consulting service must rely on automated statistical methods. This raises the question of whether systematic biases or uncontrolled confounding could cause effect estimates from the consulting service to differ from the underlying population parameters. We used negative control reference sets to assess the presence of bias in one such consulting service, the so-called "green button project."[1]

Methods

We used structured clinical data from Stanford Health Care (SHC, >3.2 million patients), Eversana (>120 million patients), and Healthjump (>120 million patients). Data from each source were converted to a time-centered common data model for access using the Advanced Cohort Engine. We used negative control reference sets defined Observational Health Data Sciences and Informatics (OHDSI) collaborative.[2] Each reference set consists of a treatment-comparator-outcome triad where no association is expected between outcome and treatment relative to the comparator. From the list of 200 reference sets we selected a subset of 46 based on expected sample sizes in each study arm. We used two versions of each reference set - with and without restriction to populations with the indication for treatment, for 92 total sets. For each reference set, we estimated the hazard ratio (HR) for the outcome between the treatment and comparator arms within each data source using adults with records between 2012-2022. Each estimate used a cohort study design, comparing two models: unadjusted, and adjusted via matching based on high-dimensional propensity score (HDPS).[3] For each data source and statistical method we computed summary statistics based on expected true hazard ratio of 1.0, including uncalibrated and calibrated false positive rates, expected absolute systematic error (EASE), and calibration of alpha values. Summary statistics were restricted to estimates with a minimum detectable relative risk (MDRR) < 5.

Results

The data sources included 1,767,333 (SHC), 25,696,709 (Healthjump), and 46,550,881 (Eversana) eligible patients. In unadjusted analyses, estimated HRs with MDRR <5 were possible in 37 (SHC), 71 (Healthjump), and 75 (Eversana) of the 92 reference sets. Unadjusted analyses showed high systematic error in the SHC data (EASE 0.19, and moderate systematic error in Healthjump and Eversana (EASE -0.06 in each) (Figure). False positive rates were substantially higher than the expected false positive rate of 0.05 (range, 0.23 - 0.28), although calibration reduced this rate (range, 0.04 - 0.13). The unadjusted models showed poor to moderate calibration of alpha values (expected area under the calibration curve, 0.5; actual range, 0.59-0.70).

PS matching was able to produce estimates with MDRR <5 for 30 (SHC), 64 (Eversana), and 60 (Healthjump) reference sets. PS matching effectively removed systematic bias within all three data sources (EASE range, -0.037 to 0.037). Uncalibrated false positive rates after PS matching ranged from 0.12 (Eversana/Healthjump) to 0.23 (SHC); calibration reduced these to 0.03 (Eversana), 0.05 (Healthjump), and 0.1 (SHC). PS matching showed moderate to good calibration (area under the calibration curve, range 0.54-0.62).

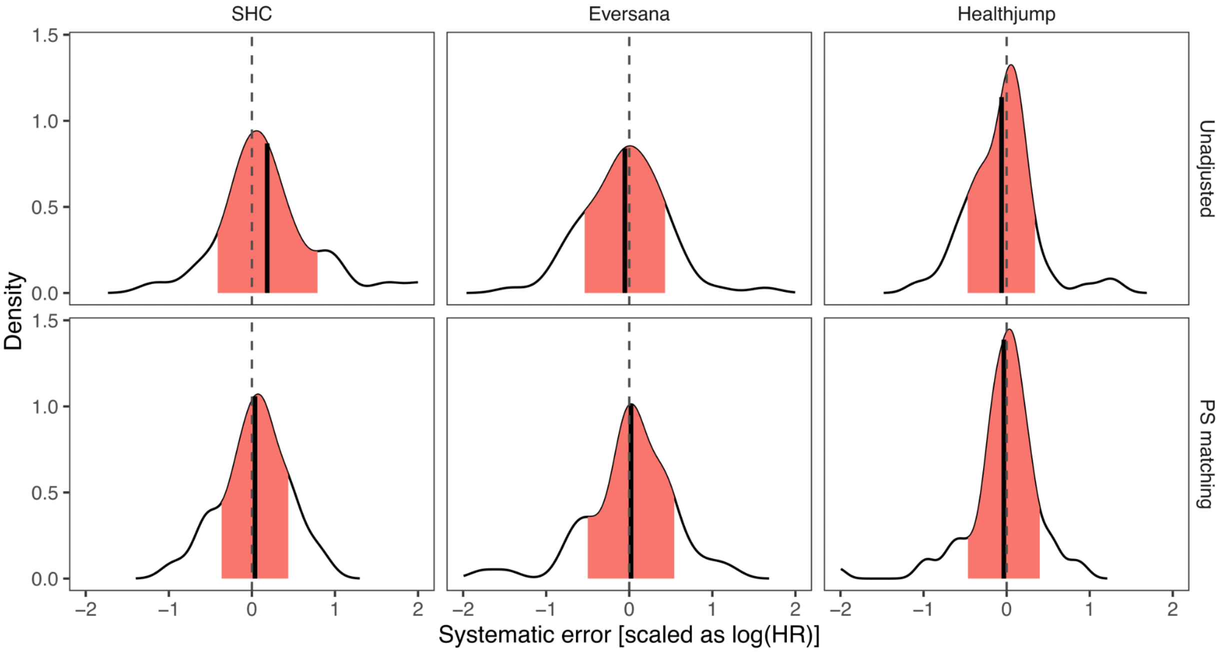

Figure: Systematic error of unadjusted and propensity score matching models against negative control reference sets across three data sources. Dashed lines represent zero error, solid lines indicate mean systematic error, and shaded regions indicate mean +/- one standard deviation

Conclusion

We found that propensity score matching produced estimates with good calibration and without systematic bias on a large group of negative control reference sets. This remained true even when the unadjusted data displayed systematic bias. These results suggest that automated, data-driven adjustment can mitigate bias when using real-world data for clinical informatics consulting.

References

1. Longhurst CA, Harrington RA, Shah NH. A “green button” for using aggregate patient data at the point of care. Health Aff (Millwood). 2014; 33:1229-35
2. Schuemie MJ, Cepeda MS, Suchard MA, et al. How confident are we about observational findings in healthcare: a benchmark study. Harvard Dat Sci Rev 2020; 2:1
3. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. Int J Epidem 2018; 47:2005-14