

Enriching Electronic-Health-Record Cohorts by Detecting Patients with Complete Data in Argos

Jeffrey Klann, PhD^{1,2}, Griffin M Weber, MD, PhD^{2,3}, Lina Sulieman⁴, Shawn N. Murphy, MD, PhD^{1,5}, Aaron Abend, MBA⁶, Diane Keogh^{2,7}

¹Mass General Brigham, Boston, MA; ²Harvard Medical School, Boston, MA; ³Beth Israel Deaconess Medical Center, Boston, MA; ⁴Vanderbilt University Medical Center, Nashville, Tennessee; ⁵University of Washington, Seattle, WA; ⁶Prognosis Data; ⁷i2b2 transSMART Foundation

Background

Research using Electronic Health Records (EHR) is hampered by missing data. A patient might have no record of diabetes in an EHR, not because they do not have diabetes, but because their diabetes care occurs at another institution not included in the EHR. This is a phenomenon called *EHR discontinuity*. Such false negatives could significantly bias EHR research, e.g., by misrepresenting disease prevalence and treatment patterns. For this reason, it is important to enrich patient cohorts by identifying patients with a reasonable likelihood of complete data. Because such patients are “loyal” to the healthcare system, this is colloquially called a *loyalty cohort*[1]. We present a validated machine learning approach to finding these patients, including an open-source tool - originally developed for i2b2 [2], and now adapted for the OMOP CDM.

Methods

Our approach builds on a previously validated machine learning model that estimates the likelihood that patients have complete data based on specific proxies for loyalty. The model was created using EHR data linked to administrative claims[3]. Features in this model include markers that indicate that (1) the patient’s preventative care is performed within the institution with at least two of the following “routine care” facts: body mass index measurement, colonoscopy, fecal occult blood test, general medical examination, having a1c ordered or value recorded, influenza vaccine, mammography, pap smear, pneumococcal vaccine, PSA test; and (2) the amount of healthcare utilization: any one or two diagnosis codes, any ED visit, any encounter, inpatient or outpatient, any one or two medications, any two outpatient encounters, any 2-3 visits with the same provider.

We mapped these concepts to standard OMOP vocabularies and implemented the scoring algorithm in R and SQL, supporting both SQL Server and Oracle. The tool generates a loyalty score that can be stored as a derived fact and used directly in cohort queries through our Argos tool. Sulieman et al. also applied a version of this approach to the *All of Us* Research Program using OMOP, confirming feasibility in a nationally representative dataset.[4]

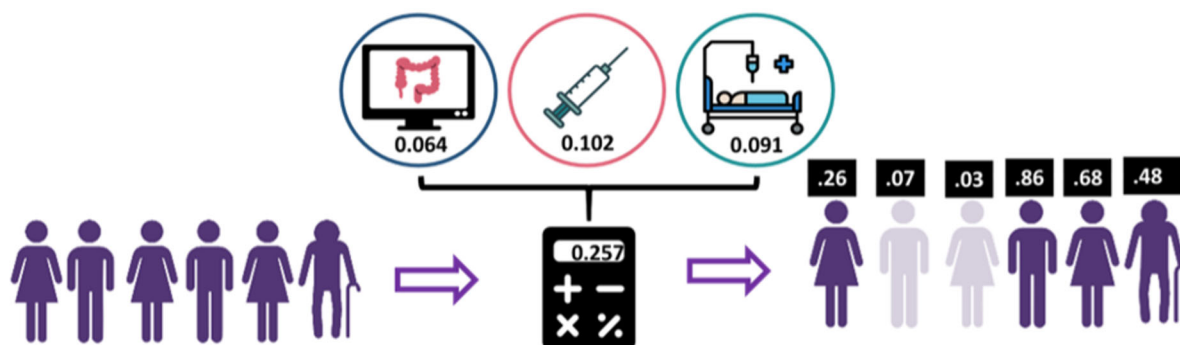


Figure 1. Patients are assigned a score based on 20 variables, comprising 124 medical concepts.

Results

We evaluated the tool at three different hospital systems by computing a loyalty score over a 2-year baseline period and using that score to predict a return to the healthcare system in a one-year follow-up period. After tuning the algorithm for individual sites, the area under the receiver operator curve (AUROC) was more than 0.80 at multiple sites.

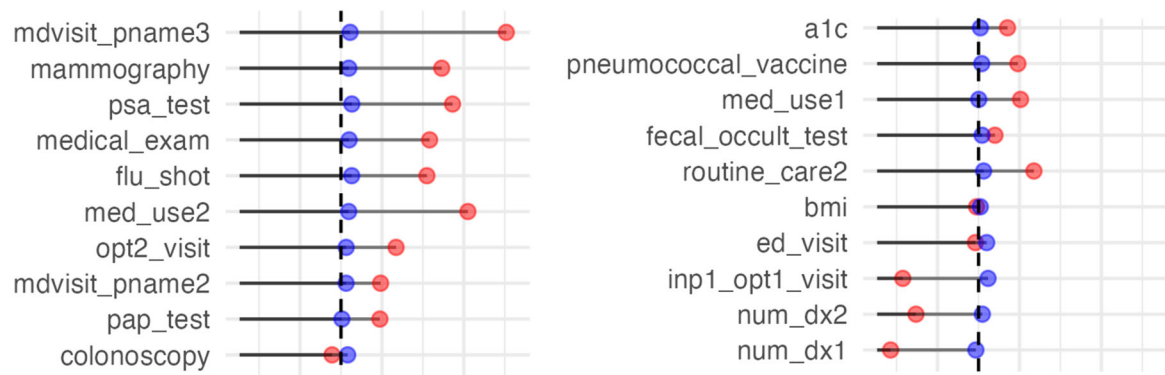


Figure 2. Square root of Odds Ratios at one site, before tuning (blue) and after tuning (red).

Conclusion

Enriching a cohort with loyal patients produces more complete data and can enhance the performance of machine learning models. We originally developed the software to run on the i2b2 data model, and we have now adapted it for OMOP and made it available to use in queries through our new Argos tool. The code is written in R and SQL and supports SQL Server and Oracle. It is available as open-source code at <https://github.com/i2b2/i2b2-digitaltwin>.

References

- 1 Weber GM, Adams WG, Bernstam EV, *et al*. Biases introduced by filtering electronic health records for patients with 'complete data'. *J Am Med Inform Assoc*. 2017;24:1134–41. doi: [10.1093/jamia/ocx071](https://doi.org/10.1093/jamia/ocx071)
- 2 Klann JG, Henderson DW, Morris M, *et al*. A broadly applicable approach to enrich electronic-health-record cohorts by identifying patients with complete data: a multisite evaluation. *J Am Med Inform Assoc*. 2023;30:1985–94. doi: [10.1093/jamia/ocad166](https://doi.org/10.1093/jamia/ocad166)
- 3 Lin KJ, Rosenthal GE, Murphy SN, *et al*. External Validation of an Algorithm to Identify Patients with High Data-Completeness in Electronic Health Records for Comparative Effectiveness Research. *Clin Epidemiol*. 2020;12:133–41. doi: [10.2147/CLEP.S232540](https://doi.org/10.2147/CLEP.S232540)
- 4 Sulieman L, Natarajan K. Using Cohort Diagnostics to Assess the Phenotypic Data Quality in All of Us Research Program. *OHDSI 2023 Collaborator Showcase*. OHDSI 2023 Global Symposium: https://www.ohdsi.org/wp-content/uploads/2023/10/CohortDiagnostics_AoU_OHDSI_2023_v2-Lina-Sulieman.pdf .