Enriching Electronic-Health-Record Cohorts by Identifying Patients with Complete Data

Jeffrey G. Klann, PhD[1,2,3]; Darren Henderson[5]; Michele Morris[6]; Hossein Estiri, PhD[1,2,3]; Griffin Weber, MD, PhD[1,4]; Diane Keogh[1]; Shyam Visweswaran MD, PhD[6]; Shawn N. Murphy, MD, PhD[1,2,3]

[1]Harvard Medical School; [2]Mass General Brigham; [3]Massachusetts General Hospital; [4]Beth Israel Deaconess Medical Center; [5]University of Kentucky, Lexington, KY;[6]University of Pittsburgh, Pittsburgh, PA

Introduction

Research using electronic health records (EHRs) is hampered by data quality issues. Among these, missing data is among the most urgent and pervasive. Due to the complex regulatory environment of healthcare, it is difficult or impossible to link patient records from multiple institutions. Therefore, although a patient might have no record of diabetes in an EHR, this does not mean the patient does not, in fact, have diabetes. In fact, the patient might receive their diabetes care at another institution that is not included in the EHR. Such false negatives have the potential to significantly bias EHR research, by, for example, misrepresenting the prevalence of a disease or its treatment. For this reason, it is essential to ensure that patients included in a research cohort have a reasonable likelihood of complete data, also known as “low EHR discontinuity.” Because such patients are “loyal” to the healthcare system that is providing the EHRs, this is colloquially called a *loyalty cohort*.

Previous work by Lin et al. constructed a regression model predicting loyalty which used twenty EHR-derived variables to, using claims data to train and evaluate the model.[1] They found that higher loyalty scores correlated with a higher portion of encounters captured in the EHR compared to claims data. Although this approach was successfully evaluated in EHRs from two institutions, it does not provide a generalized solution that is adaptable to institution-specific practices, medical coding differences, and incompatible data models across healthcare IT implementations.

Here, we describe an approach for finding loyalty cohorts in a reusable way that can easily be implemented across multiple institutions, and then evaluate its performance at several diverse healthcare institutions.

Methods

We mapped the 20 concepts described by Lin et al. to specific codes used in the national ENACT network, which provides access to EHRs of more than 142 million patients. These concepts include healthcare utilization metrics such as multiple visits to the same provider, emergency department visits, multiple medication or diagnosis codes; and, specific measures indicating a patient’s primary care home, such as PSA tests, PAP tests, and mammograms.

We next translated the approach by Lin et al. into stored procedure programs that can run on Informatics for Integrating Biology and the Bedside (i2b2) databases, using Oracle or SQL Server. Our implementation is compatible with sites in the 57 site ACT network, making its application potentially immediate.

We evaluated the algorithm on two years of EHR data to predict the probability of a return to the healthcare system for care in year three. Three healthcare systems, Mass General Brigham, University of Pittsburgh, and University of Kentucky, comprising a total of 52 hospitals, participated in the evaluation. Each site extracted a cohort of patients with any inpatient or outpatient encounter between 01/01/2017 and 12/31/2018 who were over 18 as of 1/1/2017 and not deceased at any time after 1/1/2017. This study period was selected to avoid the changes in healthcare utilization during the peaks of the COVID-19 pandemic. We performed the evaluation using a two-year lookback period on which to compute the loyalty score and a one-year follow-up period of 1/1/2019-12/31/2019 to compute any return to the healthcare system.

We further enhanced the approach using the Machine Learns Health Outcomes (MLHO) machine learning toolkit to retrain the machine learning model on local data, optimizing it for cohort selection at individual sites. Retraining involved splitting the data 50%/50% into a training and holdout test set and then using a Generalized Linear Model (GLM) model (with a binomial family and a logit link) with 5-fold cross-validation, controlling for age and sex.

We computed the AUROC of both the original and retrained models when predicting return as well as the square root of the odds ratios for both methods. We also compared the loyalty score to the Charlson Comorbidity Index to ensure that the loyalty algorithm is not selecting only the extremely ill.

Each site performed its analysis using shared programs written in the R language, using standard packages as well as MLHO. All analysis was run locally at each site and only final results were shared outside of the home institution. This project was approved by the Institutional Review Boards of each individual institution.

Our approach outputs a loyalty score that can then be stored as a “derived patient fact” in the i2b2 database, allowing cohort selection in the graphical query tool with a user-selected level of loyalty.

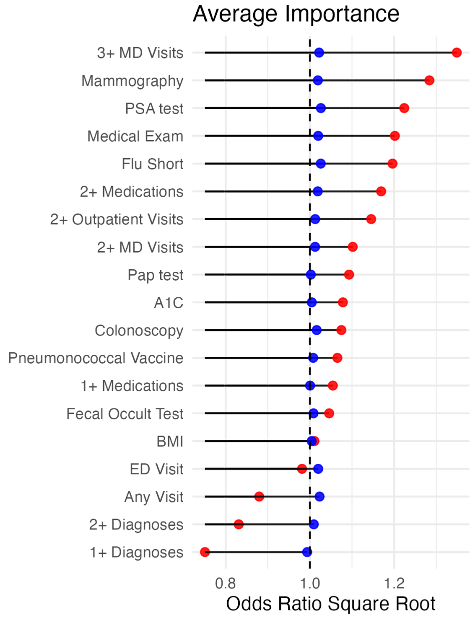

Figure 1: Contribution of each feature, on average across the three sites. Blue is the original regression equation; red is after retraining.

Results

In each of the three healthcare systems, we computed a loyalty score over a 2-year baseline period and used that score to predict a return to the healthcare system in a one-year follow-up period. After computing the loyalty score using the original equation, the average area under the ROC curve (AUROC) in predicting return was 0.776 across the three sites. After retraining the equation, the average AUROC increased to 0.807.

We found a low correlation of the loyalty score with the Charlson Comorbidity Index. Pearson’s correlation coefficient mean was 0.232, and we saw a similar distribution of comorbidity indices across all deciles of loyalty.

We finally examined the relative importance of each feature by plotting the square root of the odds ratios for each variable at each site. We found most sites had some common factors such as medication use, though the importance of specific screening measures (e.g., pap smear) varied across sites. The average odds ratios can be seen in Figure 1. Note that the blue dots, representing the original equation, are near 1.0 because a penalized estimation method was used. In contrast, we applied a generalized linear model because we were not concerned about overfitting as much as interpretability due to the small feature space.

Discussion

This loyalty cohort approach can immediately be utilized by any i2b2 site using the ACT network ontology, helping improve research cohorts by reducing the impact of missing data. With an average AUC above 0.8 after site-specific tuning, this tool reduces cohort size with high performance (in terms of cohort enhancement, not perfect prediction). Lin et al. found that just 60% completeness made machine learning classification acceptable.

One should be aware that this approach will shift the bias in the data toward patients with complete data, moving cohort demographics away from the mean. This could exclude patients with e.g., barriers to accessing healthcare.

We are presently enhancing the loyalty cohort approach with additional variables that have been shown to be good predictors of loyalty in other work by Weber. This other work has found the Dartmouth Atlas mapping from zip code to Hospital Referral Region (HRR) and the U.S. Department of Agriculture mapping from zip code to Rural-Urban Continuum Codes are highly correlated with loyalty.

We are implementing these approaches as an open-source package for i2b2. Sites will be able to install “derived fact scripts” that will create patient facts about e.g., loyalty score and visit-age interaction when data are refreshed. This will include a set of queryable terms that will allow these facts to be used in the graphical i2b2 query tool.

Our code is available open-source at https://github.com/i2b2plugins/loyalty_cohort. More information can be found here: [2]. This work was funded in part by ENACT (U24 TR004111).

References

1 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.

2 Klann JG, Henderson D, Morris M, Weber GM, Visweswaran S, Murphy SN. A Broadly Applicable Approach to Enrich Electronic-Health-Record Cohorts by Identifying Patients with Complete Data: A Multi-site Evaluation. *J Am Med Inform Assoc* doi:10.1093/jamia/ocad166