Estimating Model Performance on External Datasets from Their Limited Statistical Characteristics: Application to 3-Year Surgery Risk in Ulcerative Colitis

Tal El Hay and Chen Yanover
KI Research Institute

The Israeli Institute for Applied Research in Computational Health

Background

External validation, that is performance evaluation of a model trained using an "internal" data source on other datasets, is increasingly recognized as an essential step in demonstrating model robustness[1]. A more proactive approach may seek a model that performs well on multiple datasets. A mechanism that estimates the performance of a given model on external sources from their statistical characteristics could support such a model optimization procedure.

Here, we propose an algorithm which reweighs samples in the internal dataset according to external dataset statistics, potentially reported in a preceding publication (as "Table 1") or a characterization study. The algorithm then uses the reweighted samples to estimate model performance on the external source. We validate our approach using a prediction model for 3-year risk of intestinal surgery in ulcerative colitis patients.

Methods

Data. We used primary care electronic medical records from the UK (IQVIA Medical Research Data incorporates data from THIN, A Cegedim Database; reference made to THIN is intended to be descriptive of the data asset licensed by IQVIA), which covers approximately 6% of the UK population, and is representative of the population in terms of demographics and major condition prevalence.

Ulcerative colitis use-case. For each patient in the ulcerative colitis cohort (identified based on diagnostic codes and prescriptions; see https://github.com/ohdsi-studies/IbdCharacterization for more details), we extracted a set of features, previously observed as associated with increased intestinal surgery risk[2]. The outcome considers procedure codes for colostomy, colectomy, ileostomy, small intestinal resection, stricturoplasty, balloon dilation, drainage of perianal abscess, drainage of intra-abdominal abscess, or death, within 3 years following index date (first diagnosis or related prescription). We excluded subjects with insufficient follow-up. All concept sets and cohorts have been defined as part of the IBD Characterization project.

Reweighing algorithm. Our goal is to obtain a (weighted) sample of the internal population with statistical properties that are similar to the ones available, e.g., as Table 1, for the external datasets. To avoid overfitting, we also require the set of weights to be close to uniform (i.e., maximal entropy). To derive an optimal set of weights, we define the following optimization problem:

$$\text{minimize}_{\,w} \left\| X_{internal}^{T} \cdot \vec{w} - \vec{\mu}_{external} \right\|_2 - \lambda \cdot \mathcal{H}(\vec{w}),$$
$$\text{such that } \sum_{i=1}^{n} w_i = 1, w_i \geq 0, \forall_i$$

where $X \in \mathbb{R}^{n \times p}$ is the feature matrix, $\vec{\mu} = \frac{1}{n}\sum_i \vec{x}_i$ is a p-dimensional vector of feature means, $\vec{w} \in \mathbb{R}^n$ are the inferred patient-specific weights, $\mathcal{H}(\vec{w})$ is the weight entropy and λ is a tunable parameter. The inferred weights induce a weighted sample $\{\vec{x}_i, y_i, w_i\}_{i=1}^{n}$ whose properties approximate the statistics of the external sample. In this study, we used this sample to approximate the area under the receiver operating characteristic curve (AUC) using R's *WeightedROC* library.

Algorithm evaluation. We trained, tuned, and validated the proposed approach on England-based patient cohorts and tested it on three external datasets, including patients residing in Scotland, Wales, and Northern Ireland.

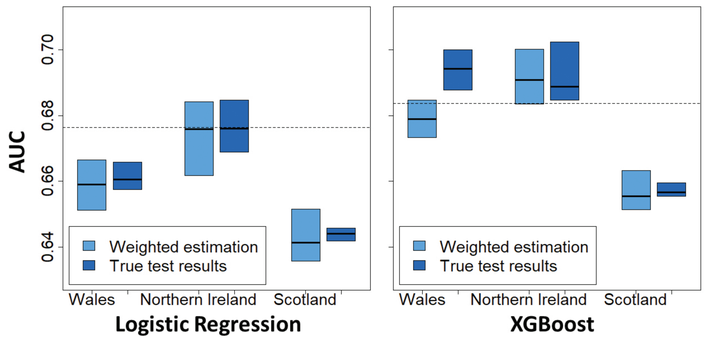
Contact: {talelh, chen}@kinstitute.org.il

Results

The following table summarizes the characteristics of the various location-based cohorts. Note the difference in various features, e.g., the younger age and greater smoking percentage in the Northern Ireland cohort; and the more prevalent use of steroids in Scotland.

	England train	England test	Wales	Northern Ireland	Scotland
No. of subjects	7559	1915	1255	772	1772
Age	48.9 ±18.9	48.0 ±18.9	48.3 ±19.1	46.0 ±18.2	47.0 ±18.5
Female	3700 (48.9%)	938 (49.0%)	602 (48.0%)	382 (49.5%)	909 (51.3%)
Smoking	1770 (23.4%)	461 (24.0%)	313 (24.9%)	221 (28.6%)	484 (27.3%)
Steroids use	2279 (30.1%)	555 (29.0%)	408 (32.5%)	224 (29.0%)	668 (37.7%)
Underweight	202 (2.7%)	46 (2.4%)	30 (2.4%)	24 (3.1%)	37 (2.1%)
Overweight	1839 (24.3%)	440 (23.0%)	343 (27.3%)	200 (25.9%)	442 (24.9%)
Perianal disease	126 (1.7%)	18 (0.9%)	16 (1.3%)	12 (1.6%)	11 (0.6%)
Gastrointestinal procedures	450 (6.0%)	104 (5.4%)	95 (7.6%)	48 (6.2%)	121 (6.8%)

The boxplots on the right show the distribution of actual (dark blue) and estimated performance (light blue) using regularized logistic regression and XGBoost. The dashed line shows the AUC for England's test set (as a baseline). In all instances (except Wales with XGBoost), AUC estimations are more similar to the actual values than the internal calculated performance (dashed line).



Conclusions

Our proposed algorithm can help in identifying models that perform well across multiple clinical settings and geographies, even when detailed test data from such settings is not available.

Notably, this is a work-in-progress, currently having several limitations. First, the divergence between internal and external validation sets in this study may be low and not representative of typical studies. Second, approximation quality depends on the level of detail of the statistical information. Third, the algorithm requires tuning a hyper parameter (λ) that may be crucial to avoid overfitting.

We believe that the proposed algorithm can serve as a building block in network studies that aim to construct robust models across datasets, using OHDSI's tools for extracting and sharing population-level statistics.

References

1. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. Published online June 21, 2021. doi:10.1001/jamainternmed.2021.2626
2. Koliani-Pace JL, Siegel CA. Prognosticating the Course of Inflammatory Bowel Disease. Gastrointest Endosc Clin N Am. 2019;29(3):395-404. doi:10.1016/j.giec.2019.02.003