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

Background

External validation, that is the process of training a model on an "internal" data source and evaluating its performance on others[1], is increasingly recognized as an essential step in demonstrating model robustness[2]. 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 using only their statistical characteristics could support such a model optimization procedure. Here, we propose an algorithm which reweighs samples in the internal dataset to match external dataset statistics, potentially reported in a preceding publication (as "Table 1") or a characterization study; then estimate the performance on the external dataset using the reweighted internal set. We validate our approach using a prediction model for 3-year risk of intestinal surgery in ulcerative colitis patients; and discuss its future extensions and limitations.

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[3].

Ulcerative colitis use-case. For each patient in the ulcerative colitis cohort (identified based on diagnostic codes and prescriptions), we extracted a set of features, previously observed as associated with increased intestinal surgery risk[4]. These include age (and age^2), sex, smoking, being underweight or overweight, presence of perianal disease, and use of steroids or drugs prescribed to a at least 1,000 subjects up to 90 days after cohort entry date (index date; first diagnosis or related prescription). 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. 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. As there may exist multiple (often infinite) sets of "equally similar" weights, we propose to search for a set of weights that is also close to uniform (maximizes the weights entropy). More formally, to derive an optimal set of weights, we define the following optimization problem:

$$\text{minimize}_{\,w} \left\| X^T_{internal} \cdot \vec{w} - \vec{\mu}_{external} \right\|_2 - \lambda \cdot \mathcal{H}(\vec{w}) \qquad (1)$$

$$\text{such that} \sum_{i=1}^{n} w_i = 1, w_i \geq 0, \forall_i$$

where $X \in \mathbb{R}^{n \times p}$ is the feature matrix (p features for n patients), $\vec{\mu} = \frac{1}{n}\sum_i \vec{x}_i$ is a p-dimensional vector of feature means, $\vec{w} \in \mathbb{R}^n$ are the inferred patient-specific weights, $\mathcal{H}(\vec{w}) = \sum_i -w_i log w_i$ is the weight

entropy, λ is a tunable parameter, and *internal* or *external* refers to the data origin. We used the CVXR[5] R library to solve the optimization problem.

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, thus allowing estimation of performance measures of the external set. Specifically, in this study, we used this sample to estimate the area under the receiver operating characteristic curve (AUC) using R's *WeightedROC* library[6].

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. We trained both logistic regression and XGBoost[7] models and report the AUC of each model on the internal, internal-reweighted, and external test sets.

Results

We demonstrate the utility of the proposed algorithm using a risk prediction model, trained on England patient cohorts, and externally validated on datasets from Scotland, Wales, and Northern Ireland. Cohort characteristics are shown in Table *1*. Note the differences 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.

Table 1. Characteristics of ulcerative colitis patient cohorts in UK countries.

	England train		**England test**		**Wales**		**Northern Ireland**		**Scotland**	
No. of subjects	7559		1915		1255		772		1772	
Mean age (SD)	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%)

Figure 1 shows the distribution of actual (dark blue) and estimated (light blue) AUC values in "external" UK countries using regularized logistic regression and XGBoost. The dashed line shows the AUC for England's test set. In all instances (except Wales with XGBoost), AUC estimations are more similar to the actual values than the internally calculated performance (dashed line).

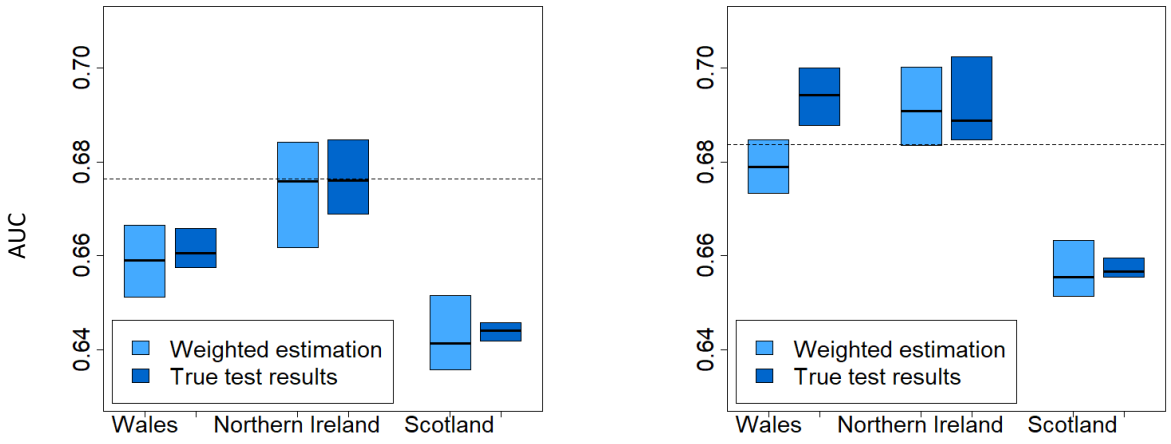

Figure 1. Actual versus estimated performance. For risk prediction models, trained on 80% of the England patient cohort (drawn multiple times), we present England's test performance (dashed line) as well as the median and inter quantile range of actual performance (AUC, dark blue) and estimated (light blue) performance on patient cohorts from the other UK countries.

Conclusion

We presented an algorithm that estimates the performance of an internally trained prediction model on external datasets from their limited statistical characteristics; and demonstrated its utility using an England-trained risk prediction model and datasets from Scotland, Wales, and Northern Ireland. 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, whereas our preliminary experiments involved only marginal statistics of features. Third, the algorithm requires tuning of a hyper parameter (λ) that may be crucial to avoid overfitting.

While we currently focus on population-level Table 1 like statistics, straightforward extensions can consider outcome-dependent (e.g., cases versus controls) feature statistics or more detailed statistical information available, for example, in deep phenotyping studies[8]. We will also explore other optimization schemes, such as entropy balancing for causal effects (EBAL)[9], and more efficient ways to handle high dimensional data. Finally, we plan to investigate the accuracy of our approach in more detailed empirical studies with external validation sets.

Importantly, our proposed algorithm can help identifying models that perform well across multiple clinical settings and geographies, even when detailed test data from such settings is not available. We believe that this 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/Citations

1. OHDSI. *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI; 2019. https://books.google.co.il/books?id=JxpnzQEACAAJ

2. 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

3. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Inform Prim Care*. 2011;19(4):251-255. doi:10.14236/jhi.v19i4.820

4. 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

5. Fu A, Narasimhan B, Boyd S. CVXR: An R Package for Disciplined Convex Optimization. *Journal of Statistical Software*. 2020;94(1):1-34. doi:10.18637/jss.v094.i14

6. Hocking TD. *WeightedROC: Fast, Weighted ROC Curves*.; 2020. Accessed August 22, 2021. https://CRAN.R-project.org/package=WeightedROC

7. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

8. Burn E, You SC, Sena A, et al. Deep phenotyping of 34,128 patients hospitalised with COVID-19 and a comparison with 81,596 influenza patients in America, Europe and Asia: an international network study. *medRxiv*. Published online June 28, 2020:2020.04.22.20074336. doi:10.1101/2020.04.22.20074336

9. Hainmueller J. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Polit anal*. 2012;20(1):25-46. doi:10.1093/pan/mpr025