Estimating model performance on external data sources from their summary statistics: a real-world benchmark

Tal El-Hay[1], Jenna M Reps[2], Chen Yanover[1]

[1]KI Research Institute, Kfar Malal, Israel; [2]Janssen Research and Development, Raritan, NJ, USA

Background

External validation of patient level prediction models is an essential step towards their implementation in clinical settings[1–7]. Often, such evaluation is costly or even infeasible as access to patient level data is typically limited; in some cases, however, external summary statistics may be available. We recently proposed a novel method that estimates model performance in external data sources from their limited statistical characteristics and analyzed its performance on synthetic and semi-synthetic data[8]. Here we test this method in real clinical settings using data from five US datasets and prediction models for various outcomes in individuals with major depression.

Methods

Datasets and clinical prediction tasks. We use five US observational healthcare databases mapped to the OMOP common data model:

- The IBM® MarketScan® Commercial Database (CCAE) includes health insurance claims across the continuum of care as well as enrollment data from large employers and health plans across the US who provide private healthcare coverage for employees, their spouses, and dependents.
- The IBM® MarketScan® Medicare Supplemental Database (MDCR) represents the health services of retirees in the US through employer-sponsored plans.
- The IBM® MarketScan® Multi-State Medicaid Database (MDCD) reflects the healthcare service use of individuals covered by Medicaid programs in numerous geographically dispersed states. The database contains the pooled healthcare experience of enrollees, covered under fee-for-service and managed care plans.
- Optum's Clinformatics® Data Mart (Optum CDM) is derived from a database of administrative health claims for members of large commercial and Medicare Advantage health plans. The database includes data over a 15-year period (1/2007 through 12/2021).
- Optum's longitudinal EHR (Optum EHR) repository is derived from dozens of healthcare provider organizations in the US, that include more than 57 contributing sources and 111K sites of care.

We focus on the prediction task "within patients who are pharmaceutically treated for major depressive disorder, what is the risk of developing <outcome> for the first time within 1 to 365 days after initial diagnosis of major depressive disorder". We investigate five outcomes: fracture, seizure, diarrhea, insomnia, and gastrointestinal bleed.

Estimation method. Given a patient level sample from an internal dataset and summary statistics from an external one, we first find weights that induce internal weighted statistics that are similar to the external ones. Next, we compute performance metrics such as area under the receiver operating characteristic curve (AUROC) and Brier (calibration) score using the internal weighted sample of labels and model predictions. To compute confidence intervals, we employ bootstrapping on the internal sample where for every bootstrap iteration the algorithm repeats both the reweighting and performance metric computation steps. The current implementation applies to binary outcome models and uses 'Table 1' statistics, e.g., prevalence of conditions in each outcomes group.

The proposed algorithm produces a weighted sample of the internal data that approximately emulates an out-of-distribution sample from the generating distribution of the external data. To provide a good approximant emulation and to allow reliable performance estimation, two assumptions should be met[8]: (1) the shared summary statistics are sufficiently detailed to capture most of the shift between the internal and external distributions; and (2) the internal dataset should have good coverage relative to the external one, i.e., the probability of features in the internal data source should be >0 whenever its external probability >0. The estimation package tests for various potential violations of this condition before employing reweighting. As such tests cannot detect every case of non-overlap, we also verify that the maximum standardized mean difference between attribute probability in the weighed internal and external samples does not exceed a predefined threshold (0.1).

***Evaluation setup*.** We evaluate the proposed external estimation methodology across the five prediction tasks and two model designs. Both models used a logistic regression with LASSO regularization[9], but with different feature sets. The first model (age/sex) only used age categories in 5-year buckets and sex as features and the second model (moderate-sized model) used age/sex plus 84 commonly used medical history features (e.g., history of hypertension).

Models were fit for each combination of prediction task, database and model design using the standard PatientLevelPrediction framework (75% train data with 3-fold cross validation to identify the optimal regularization and 25% test set for internal validation). For each combination of prediction task, development database, model design, and validation database we extract the internal performance of the model on the development test data; the true external validation of the model on the validation database; and the estimated external validation of the model on the validation database, computed using the estimation method and deriving weights based on the statistics of model features only.

To assess the benefit of the estimation method we compare the absolute difference between the internal and external validation performance to the difference between the estimated and external validation performance.

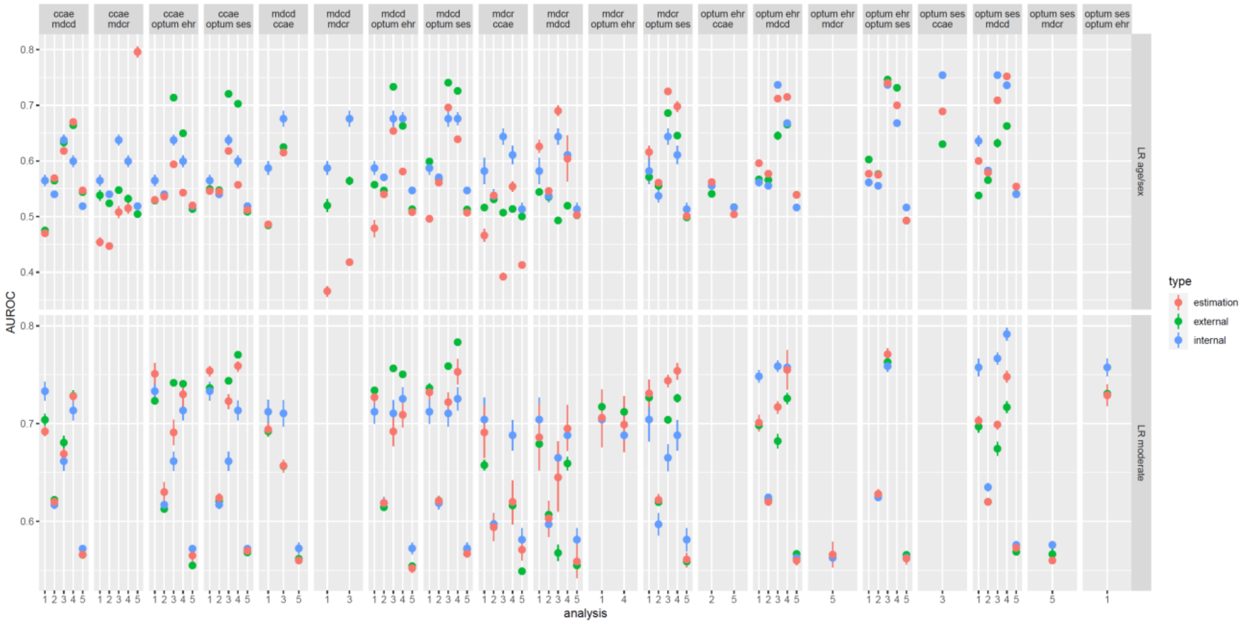
Results

Table 1 compares demographic and outcome statistics across databases. Note that CCAE and MDCD include mostly individuals under 65 years and a very small percentage of elderly individuals, while MDCR includes mostly older individuals and less than 3% of 20-64 years old ones.

Table 1. Characteristics of the five observational health databases.

	CCAE	MDCD	MDCR	Optum EHR	Optum CDM
N	2,365,324	660,158	205,789	3,309,284	1,678,579
female	68.6%	72.5%	67.1%	69.4%	67.5%
Age group (years)					
<20	12.4%	29.9%	0.0%	8.3%	8.0%
20-64	86.9%	67.1%	2.8%	71.1%	61.7%
65≤	0.7%	3.0%	97.2%	20.6%	30.3%
Outcome counts					
Seizure	9,058	6,515	1,778	18,597	9,341
Diarrhea	54,302	23,310	7,218	86,972	50,622
Fracture	9,772	4,407	4,281	20,655	16,618
GI bleed	8,172	5,700	3,304	21,291	12,775
Insomnia	77,754	30,201	6,950	114,422	64,778

Figures 1 and 2 show the estimation performance. For most moderate-sized models, the difference between estimation and external AUROC values is smaller than between internal and external ones (Figure 1, missing results indicate potential lack of overlap between the internal and external samples). In the age/sex model there are cases where the estimations are further away from then external AUROC than the internal one. Note that in some of these cases, the moderate-sized model tests do not provide estimations, suggesting that there is a large skew that could not be detected using only the age/sex features. Intriguingly, Brier score estimation are very accurate for both models (Figure 2).



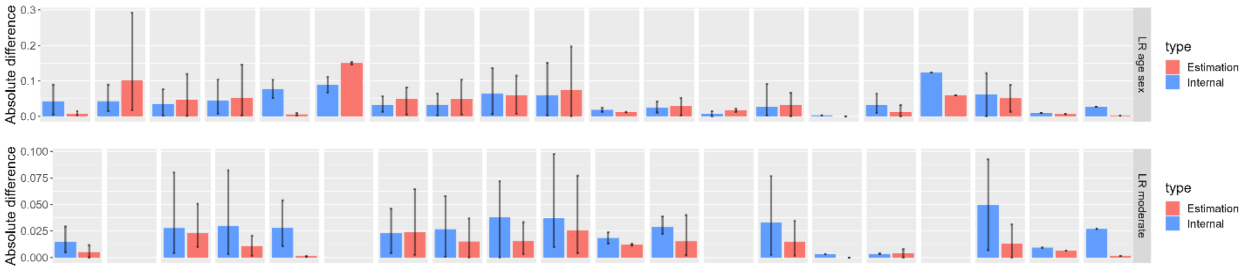


Figure 1. AUROC estimates. The top strips show detailed estimations. The bottom strips compare absolute difference between internal and external AUROC versus absolute difference between estimated and external ones; whiskers denote minimal and maximal absolute difference. In each strip, the top panel refers to the age/sex model and the bottom one – to the moderate-sized model.

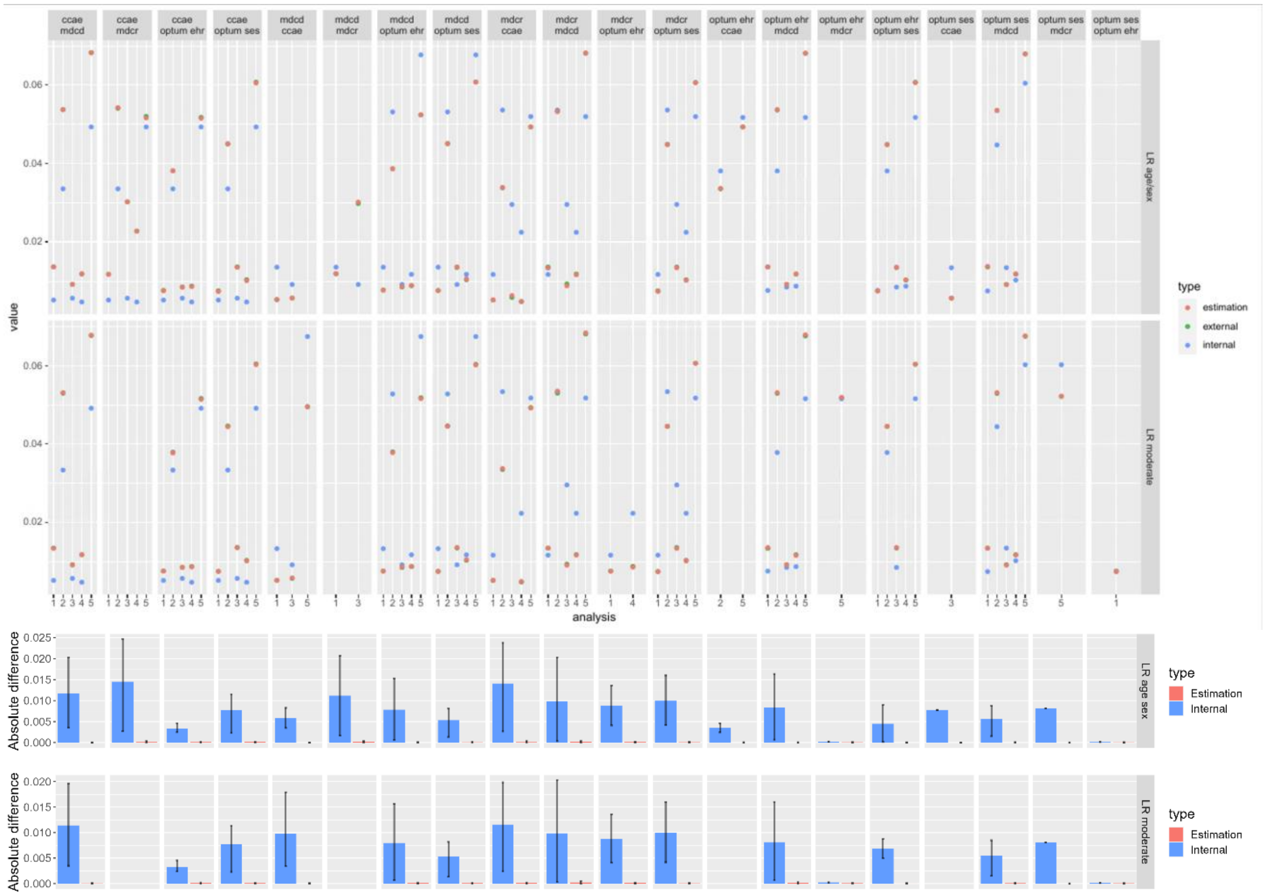


Figure 2. Brier score estimates. See Figure 1 for more details.

Conclusion

We tested an algorithm for estimating external performance on several claims and an EHR databases from the US and demonstrated good AUROC accuracy for moderate-sized models – involving dozens of features – and fair AUROC accuracy on age/gender models. Accuracy of Brier score estimation is excellent for both types of models.

While, eventually, an “actual” external validation is, likely, unavoidable, OHDSI researchers may apply this method to compare moderate-sized candidate models using solely external characterization studies, rule out poorly performing models and guide model refinement at an early stage. This method may also assist interrogation of data-shifts[10], e.g., to study the effect of population structure on external performance, it can be applied to external baseline statistics that are not split between outcome groups.

Future directions include exploring cases where AUROC estimation is fair while Brier estimation is accurate, testing the algorithm across geographies and along time, as well as testing with respect to specific data-shifts, e.g., population aging.

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. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2020;14(1):49-58. doi:10.1093/ckj/sfaa188

3. Reps JM, Williams RD, You SC, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. *BMC Med Res Methodol*. 2020;20(1):102. doi:10.1186/s12874-020-00991-3

4. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245-247. doi:10.1016/j.jclinepi.2015.04.005

5. Reps JM, Kim C, Williams RD, et al. Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study. *JMIR Med Inform*. 2021;9(4):e21547. doi:10.2196/21547

6. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. 2020;21(2):345-352. doi:10.1093/biostatistics/kxz041

7. Subbaswamy A, Adams R, Saria S. Evaluating Model Robustness and Stability to Dataset Shift. *ArXiv201015100 Cs Stat*. Published online March 15, 2021. Accessed April 4, 2021. http://arxiv.org/abs/2010.15100

8. El-Hay T, Yanover C. Estimating Model Performance on External Samples from Their Limited Statistical Characteristics. In: *Proceedings of the Conference on Health, Inference, and Learning*. PMLR; 2022:48-62. Accessed May 29, 2023. https://proceedings.mlr.press/v174/el-hay22a.html

9. The Book of OHDSI. https://ohdsi.github.io/TheBookOfOhdsi/

10. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med*. 2021;385(3):283-286. doi:10.1056/NEJMc2104626