npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01414-z

Extensive benchmarking of a method that estimates external model performance from limited statistical characteristics

Check for updates

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

Predictive model performance may deteriorate when applied to data sources that were not used for training, thus, external validation is a key step in successful model deployment. As access to patient-level external data sources is typically limited, we recently proposed a method that estimates external model performance using only external summary statistics. Here, we benchmark the proposed method on multiple tasks using five large heterogeneous US data sources, where each, in turn, plays the role of an internal source and the remaining—external. Results showed accurate estimations for all metrics: 95th error percentiles for the area under the receiver operating characteristics (discrimination), calibration-in-the-large (calibration), Brier and scaled Brier scores (overall accuracy) of 0.03, 0.08, 0.0002, and 0.07, respectively. These results demonstrate the feasibility of estimating the transportability of prediction models using an internal cohort and external statistics. It may become an important accelerator of model deployment.

Recent years have witnessed a sharp rise in the development of predictive machine learning models for healthcare applications; a striking example is the hundreds of prediction models for the diagnosis and prognosis of coronavirus disease 2019 (COVID-19), developed or validated between 2020-2022[1]. As such models are often trained on data from a limited number of "internal", fully accessible data sources, their application to "external" data—especially originating from different types of healthcare facilities, geography, and patient population—may result in inadequate performance[2,3]; e.g., such deterioration has been demonstrated for the widely implemented Epic Sepsis Model[4] or with various stroke risk scores in atrial fibrillation patients[5]. Thus, the task of verifying model transportability across different data sources, also known as external validation[6], gradually becomes a standard step in the life cycle of clinical prediction model development[3].

Practically, testing the performance of a predictive model on an external data source[6] entails identifying the model-relevant target units (e.g., patient cohort); extracting the underlying features (or independent variables) and outcome (dependent variable) for each unit; applying the model and calculating the predicted outcome values; and, finally, comparing the set of predicted and observed outcome values to obtain the performance measures of interest (e.g., area under the receiver operating characteristic, AUROC), in the entire cohort and, potentially, in key strata to assess model fairness[6].

Accurately redefining and extracting data elements (target units, features, outcome) in an external resource may be a daunting task. Harmonizing data to use standardized data structure, content, and semantics significantly reduces that burden, as definitions of model elements can be shared and readily applied across data sources. Still, even with harmonized data, external validation is an effortful task and, potentially, an iterative process in model development, aiming at selecting a well-performing model.

Previously, we developed a method that estimates the performance of a predictive model in external data sources, using only limited descriptive statistics[7] (Fig. 1). Briefly, the method seeks weights that induce internal weighted statistics that are similar to the external ones; then compute performance metrics using the labels and model predictions of the internal weighted units. These statistics may be task-specific and characterize the target population stratified by an outcome value or, more generally, describe the entire population (or predefined strata) within an external resource or in a geographical entity. Accordingly, these characteristics may be specifically extracted or could be obtained from previous studies, e.g., characterization studies (e.g., ref. 8) and national agencies. Therefore, the proposed method allows evaluation of model performance on external sources even when unit-level data is inaccessible but statistical characteristics are available. Moreover, once obtained, these statistics can be repeatedly used to estimate the external performance of multiple models, thus considerably reducing the overhead of external validation.

Here we assess the performance of the method in real-world clinical settings using data from five US datasets and prediction models for various outcomes. We leverage the infrastructure and tools developed by the Observational Health Data Sciences and Informatics (OHDSI, https://ohdsi.

[1]KI Research Institute, Kfar Malal, Israel. [2]Janssen Research and Development, Raritan, NJ, USA. ✉e-mail: talelh@kinstitute.org.il

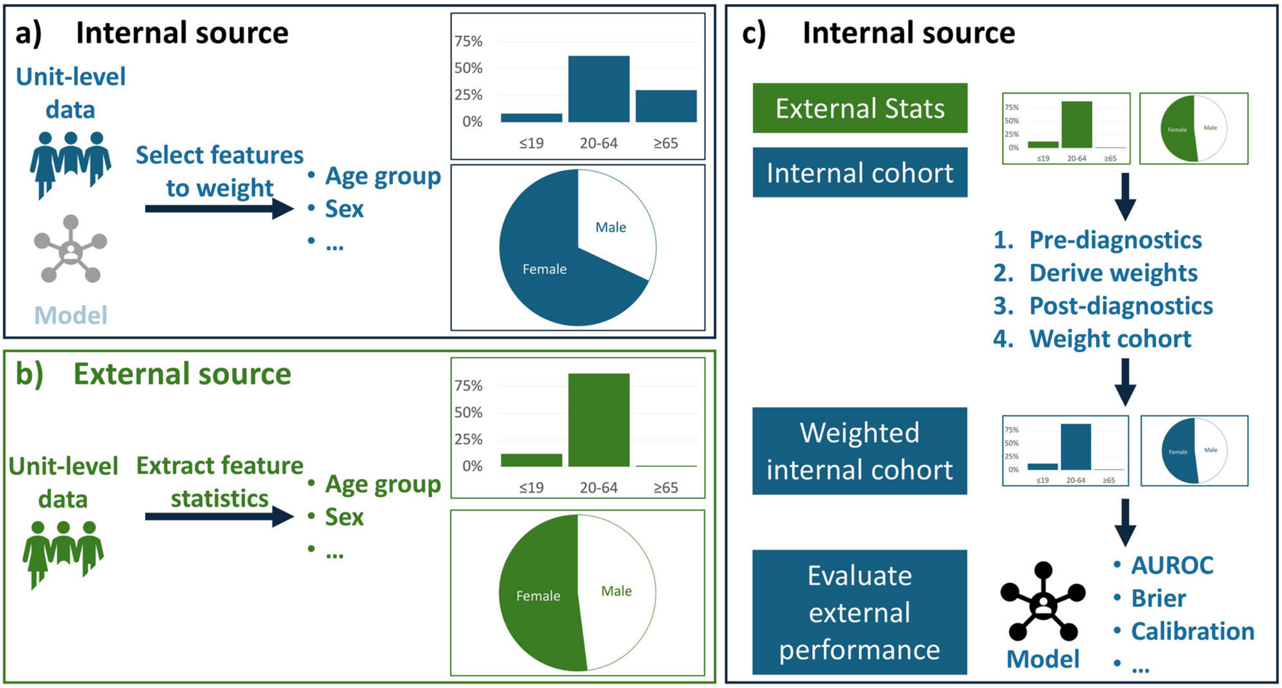
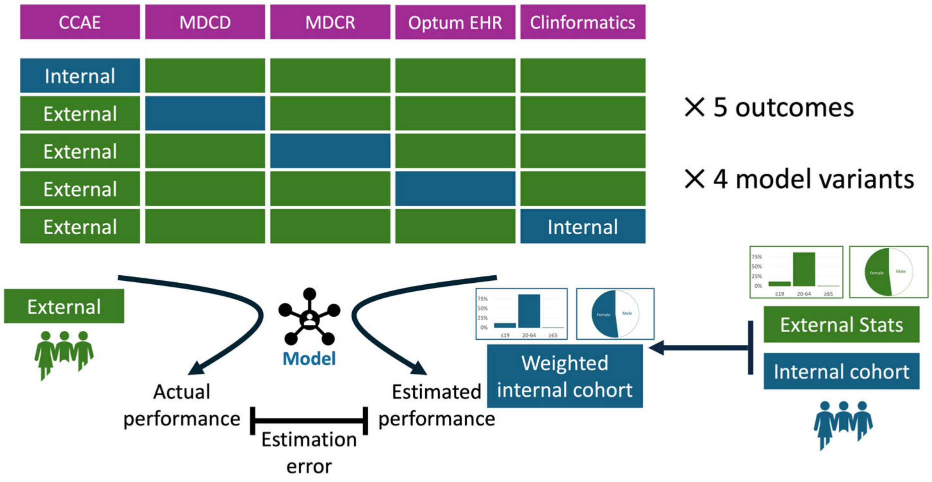


Fig. 1 | An illustration of the external performance estimation algorithm. The model evaluator selects a set of features, based, e.g., on the model's most important ones (**a**), and asks each external collaborator to provide their statistics over these features in their external cohort (**b**). The algorithm then attempts to weight the internal cohort to reproduce the external statistics and uses various diagnostic modules to test if the algorithm's underlying assumptions are met. Finally, to assess model performance, the algorithm applies the model to the internally weighted cohort (**c**).

Fig. 2 | The evaluation benchmark. In each tested configuration, we designated a data source as internal, and used its cohort to train an outcome prediction model. The performance of this model was computed directly on each of the four remaining external cohorts as well as estimated using their statistics with the proposed algorithm. The estimation error is defined as the absolute difference between the actual and estimated performance.



org/) community, a global, collaborative network of clinicians, researchers, and data scientists, whose mission is to improve the use of observational health data for research and healthcare decision-making.

Results

Benchmark overview

We set out to benchmark the accuracy of a method that estimates model performance in external data sources using only their limited statistical characteristics. Following Reps et al.[9], we defined, in five US data sources, a target cohort that included patients with pharmaceutically-treated depression; internally trained, in each given data source, models that predict patients' risk of developing diarrhea, fracture, gastrointestinal (GI) hemorrhage, insomnia, or seizure; extracted population-level statistics in the remaining four external cohorts and used these to estimate models' external performance (namely, predictive accuracy and calibration); then compared the estimated measures to the actual ones, as computed by testing the models in each external cohort (Fig. 2).

Table 1 presents the baseline characteristics of the target cohort in each data source as well as the prevalence of outcomes. Notably, age distribution varies significantly across data sources; e.g., elderly individuals (aged 65 years or more) amount to 0.7% in CCAE, 21–30% in the Optum® EHR and Clinformatics® data sources, and 97% in MDCR. Moreover, MDCR lacks children under 20 years whereas the other data sources have at least 8%.

Evaluation of the estimation method

In essence, the benchmarked method assigns weights to internal cohort units to reproduce a set of external statistics; as such, it could not be applied when certain statistics can not be represented as a weighted average of the

Table 1 | Baseline characteristics of target cohorts and outcome prevalence in sub-cohorts of patients with no recorded corresponding outcome prior to the index

	CCAE		MDCD		MDCR		Optum® EHR		Clinformatics®	
N	2,365,324		660,158		205,789		3,309,284		1,678,579	
Female	1,622,351	(68.6%)	478,848	(72.5%)	138,089	(67.1%)	2,297,254	(69.4%)	1,133,229	(67.5%)
Age ≤19	293,704	(12.4%)	197,335	(29.9%)	0	(0%)	273,708	(8.3%)	133,891	(8%)
Age 20–64	2,055,921	(86.9%)	442,793	(67.1%)	5760	(2.8%)	2,353,756	(71.1%)	1,035,538	(61.7%)
Age ≥65	15,699	(0.7%)	20,030	(3%)	200,029	(97.2%)	681,820	(20.6%)	509,150	(30.3%)
Seizure	9058	(0.5%)	6515	(1.4%)	1778	(1.2%)	18,597	(0.8%)	9341	(0.8%)
Diarrhea	54,302	(3.5%)	23,310	(5.7%)	7218	(5.7%)	86,972	(4%)	50,622	(4.7%)
Fracture	9772	(0.6%)	4407	(0.9%)	4281	(3.1%)	20,655	(0.9%)	16,618	(1.4%)
GI bleed	8172	(0.5%)	5700	(1.2%)	3304	(2.3%)	21,291	(0.9%)	12,775	(1%)
Insomnia	77,754	(5.2%)	30201	(7.3%)	6950	(5.5%)	114,422	(5.5%)	64,778	(6.5%)

internal cohort's features. For example, if statistics from MDCD include the proportion of subjects under 20 years, then there is no set of weights that yields this proportion in MDCR because indicator features of this group are zero in all MDCR units. Additionally, the optimization algorithm may fail to find appropriate weights because of higher-order dependencies between features.

Therefore, for a given pair of external and internal datasets, the success of the weighting algorithm depends on the set of provided statistics. Specifically, the more features used, the harder it is to find a solution. On the other hand, to provide accurate estimations, the weights should induce an accurate approximation of the joint distribution over the features and outcome, which affects model performance, suggesting that such features should be included. Consequentially, to balance between these two considerations, the benchmark used statistics of features with non-negligible model importance in each configuration (see "Methods" for details).

In the main analysis, we tested 400 configurations over combinations from five internal data sources; for each, the other four were marked as external; and four models were constructed for every outcome. Among these configurations, a few that used MDCR as the internal source failed to estimate external performance: XGBoost seizure model on external dataset MDCD, insomnia model on Optum® EHR, and diarrhea model on MDCD and Optum® EHR. Additionally, the medium-sized logistic regression seizure model failed on the external dataset CCAE.

Figure 3 compares the actual versus estimated external AUROC of internally trained outcome prediction models in each data source. In most cases, both the internal and external performance of the small linear models was inferior to that of the other, comparably performing ones (with the exception of insomnia, where the large logistic regression model dominated). Importantly, the estimated and actual external performance is similar, e.g., when training a linear diarrhea model with a medium-sized feature set on CCAE, the internal AUROC is 0.61, the actual external AUROC in MDCR is 0.587, and the estimated AUROC is 0.585.

To visualize the error distribution of the benchmarked algorithm, Fig. 4 compares the external performance estimation error versus the difference between internal and external performance as measured using AUROC, calibration, Brier score, and scaled Brier score. The upper quartile of AUROC estimation errors is usually below 0.02, whereas the values of internal-external AUROC difference are higher; for example, the median error in the internal resource MDCR benchmark is 0.011 (IQR 0.005–0.017), and the internal-external absolute difference is 0.027 (IQR 0.013–0.055). The other metrics accuracy differences are even more pronounced; for example in MDCR, calibration differences are 0.013 (0.003–0.050) versus 0.329 (0.167–0.836), Brier score differences are $3.2 \cdot 10^{-5}$ ($1.3 \cdot 10^{-5}$–$8.3 \cdot 10^{-5}$) versus 0.012 (0.0042–0.018), and scaled Brier score differences are 0.008 (0.001–0.022) versus 0.308 (0.167–0.440).

The effect of considered feature sets

To test the effect of the feature sets used for weighting, we applied the algorithm with the medium-sized feature set and used the small and large models (regardless of the model's features). The results of these tests suggest that using non-model-related features results in failure to obtain appropriate weights in some cases as well as less accurate results in others (Supplementary Notes 1–4 and supplementary Figs. 1–6). Additionally, we compared the estimation of linear model performance when using only important features (coefficient absolute value ≥0.1) versus using all features. The results suggest that using features with low coefficients leads to similar consequences of using unrelated features, giving inferior approximations relative to using important features.

Overall, the benchmarks that spanned various models and choices of feature-sets selected for weighting suggest that a good practice for executing the evaluation is using model specific feature sets and selecting them according to their model's importance.

The impact of sample size on estimation accuracy

Finally, using Clinformatics® as an internal data source, CCAE as an external one, and the logistic regression with a large feature set configuration, we tested the robustness of the estimation algorithm to internal and external sample sizes. We sub-sampled the internal cohort with sizes that range from 1000 to 250,000 units and used these sub-samples as input to the estimation algorithm, where the external statistics were computed from the entire external cohort. Similarly, we sub-sampled the external cohort and used the entire internal one to obtain the estimation. We applied this procedure for every outcome using two sub-sampling methods: simple uniform sampling and stratified sampling, where we preserve the original proportion of the outcome. For example, in stratified sampling, a 1000-unit sub-sample of Clinformatics® would have 8 seizure cases to maintain the prevalence of 0.8%.

Figure 5 shows the results of the size dependency benchmark. Both internal and external sub-sample sizes have an impact on the performance metrics, but the effect of the former is more pronounced: the algorithm fails to converge in most cases with 1000 units and in some of those with 2000 units; the variance and upper quartiles are larger; and error convergence is slower.

Discussion

We demonstrated the accuracy of a method that estimates the external performance of prediction models when external unit-level data is inaccessible and uses only internal data and limited external statistics. Specifically, AUROC approximation error 95th percentile was 0.03, calibration-in-the-large 0.08, Brier score 0.0002, and scaled Brier score 0.07. We recommend to use the proposed method with up to hundreds of features, as larger scale problems have not been tested.

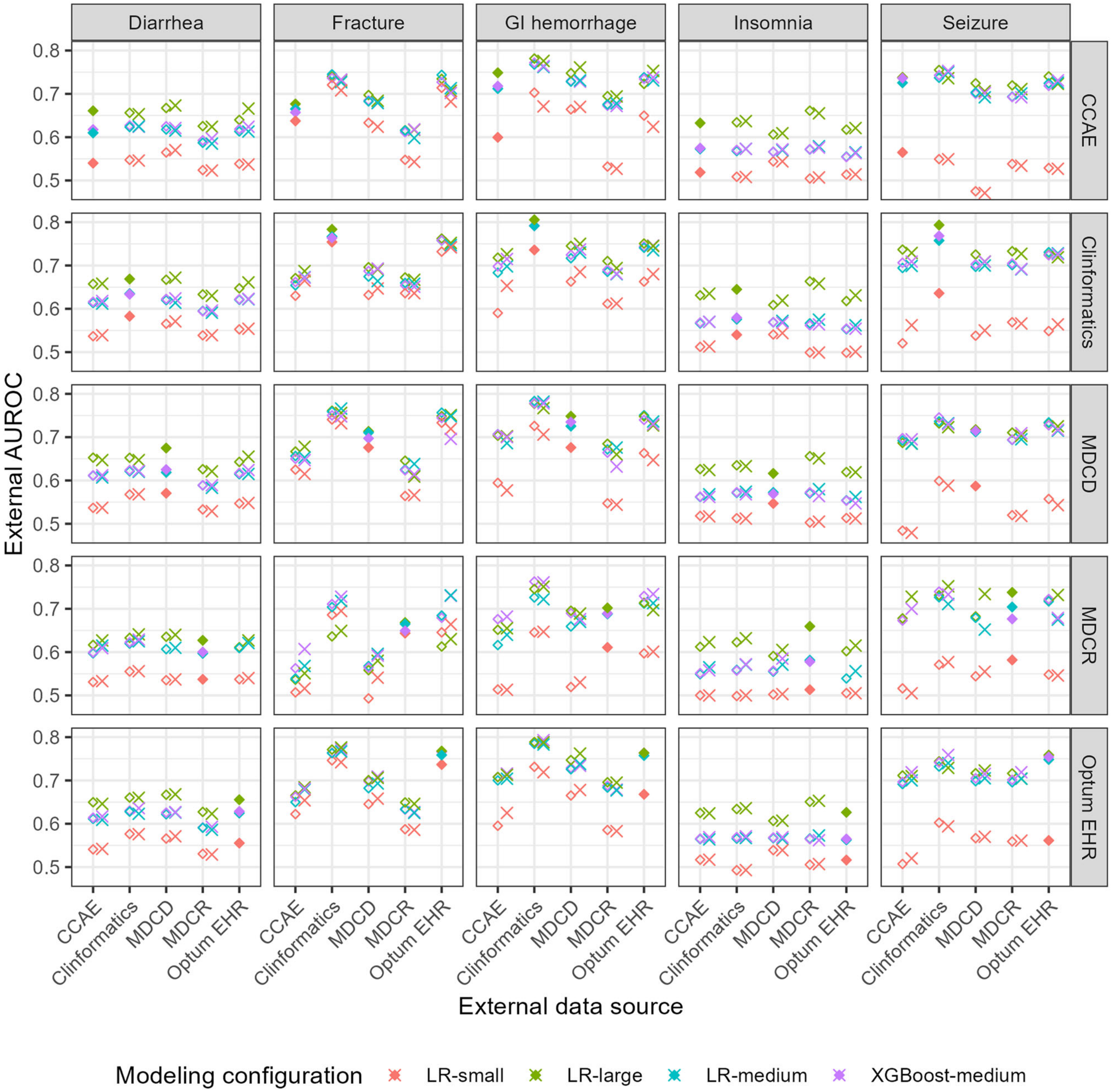


Fig. 3 | Estimating external model discrimination (AUROC). Filled and empty diamonds correspond to internal and external test performance, respectively; estimated AUROC values are marked by ×. Columns correspond to outcomes, and rows to internal data sources. LR logistic regresion.

Outcome prevalence in the tested cohorts is low (0.5–7.2%), potentially rendering some performance metrics (e.g., Brier) unreliable. Notably, estimations remain accurate even with these metrics. Moreover, we also computed calibration-in-the-large and scaled Brier scores, which are corrected for outcome prevalence and are, therefore, sensitive to small changes in the probability estimations of the outcomes; estimation errors of these metrics were an order of magnitude lower than the differences between their internal and external values.

Sampling benchmarks suggest that estimation accuracy depends on the sample size of the external cohort and, to a larger extent, on that of the internal cohort. These tests showed that good accuracy required more than 32,000 units in the internal sample. However, note that, due to low outcome prevalence, this accounts for around 150 up to 2000 cases. The size of the smaller group among cases and controls is likely the main determinant of accuracy.

The accuracy of the approximation algorithm depends on the diversity of the internal cohort as well as on the proper selection of features and transformations on which statistics are shared. Following the diversity requirement, we suggest inspecting the difference between internal and external statistics to ensure that there are no unrepresented external subgroups in the internal cohort, as well as using the tools provided in the package to assess overlap. Additionally, for reweighing, we recommend selecting statistics that involve features that have non-negligible predictive importance and the interactions of such features with the outcome. Unrelated features may result in noisy estimations.

The contribution of this work goes beyond the paper that introduced the tested framework[7]. First, experiments on real-world datasets give a broader view of the strengths and weaknesses of the framework, suggesting necessary conditions for accurate estimation. Second, these experiments highlighted practical challenges, which led to further refinements and improvements. These include guidelines to handle sparse and weakly informative features, diagnostic tools to address extreme distribution shifts, and a new customized optimization algorithm to handle large datasets, all implemented in an R package (https://github.com/KI-Research-Institute/LearningWithExternalStats).

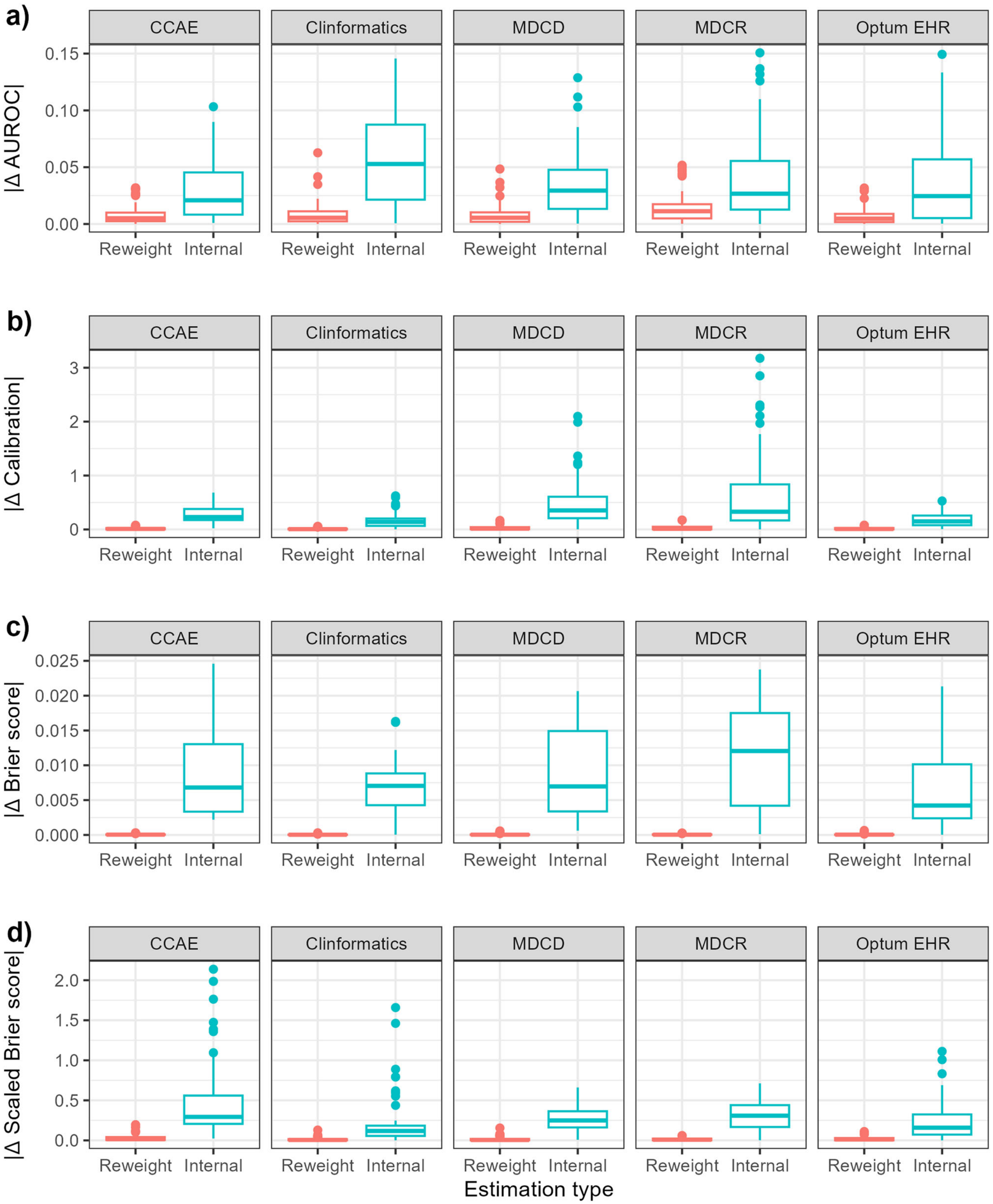


Fig. 4 | Summary of estimation performance by internal source. Absolute differences between estimated and actual external performance, as well as between internal and external performance (**a** AUROC, **b** calibration-in-the-large, **c** Brier score, **d** scaled Brier score). Each plot corresponds to an internal source, and boxes summarize performance overall external sources, outcomes, and models. The box center-line corresponds to the median and the box limits to the upper and lower quartiles. The whiskers extend to 1.5 times the inter-quartile range or to the most extreme value in the corresponding direction. Points correspond to outliers.

The problem of model evaluation under distribution shift has been addressed in several previous studies but with different data availability requirements. When no data is available, constrained sub-sampling of the internal data allows to estimate the worst-case performance under user-defined shifts in distributions of variable subsets[10]. In cases where an external *unlabeled* dataset is available, external performance can be evaluated using density ratio estimation, possibly augmented by user knowledge about the nature of the shift[11]. The method proposed here relies on a more common

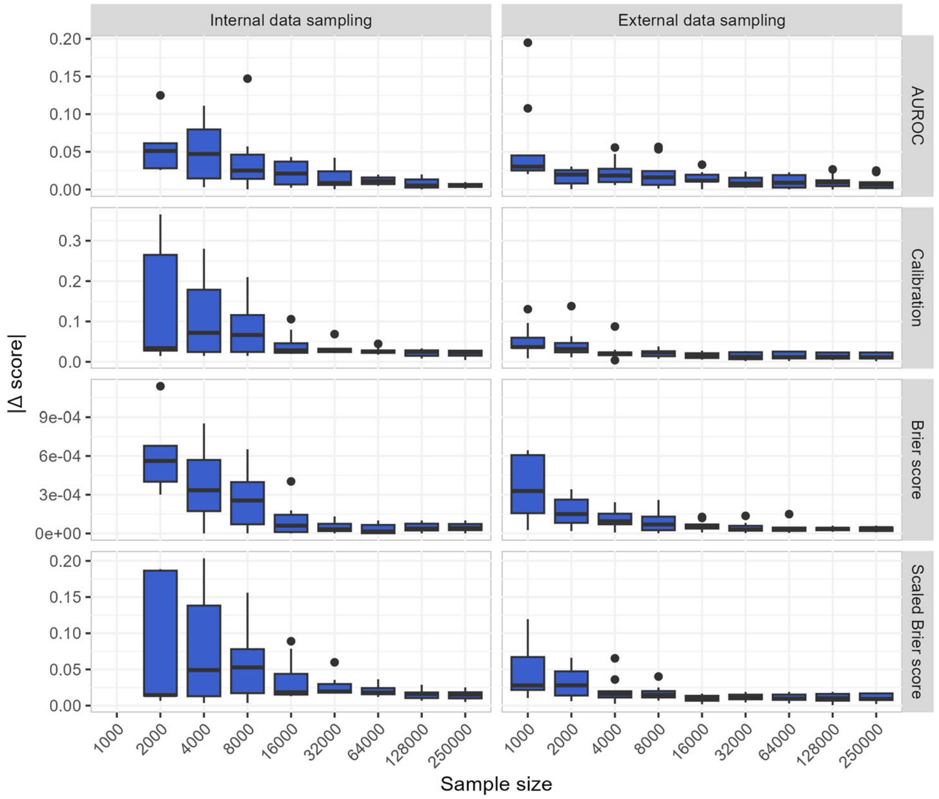


Fig. 5 | The impact of internal and external data size on evaluation accuracy. Each box corresponds to ten tests that span the five outcomes and both stratified and simple sampling. Only one test with an internal sub-sample of size 1000 converged; therefore, its result is not shown. Internal sampling of 2000 units yielded five results. An external sampling of 2000 yielded nine results. All other tests converged successfully.

scenario in which external statistics are available (or can be shared) but not full access to datasets. Therefore, it strikes a balance between the need to rely on specific information from external sources to give accurate estimations and the lack of access to detailed data.

This work has several limitations. First, the accuracy of the proposed method depends on the validity of the underlying assumptions of internal dataset diversity as well as on the proper selection of features and transformations. Therefore, this methodology should require thoughtful evaluation of the provided diagnostics and consideration of the included features. Second, the tests focus on US sources, thereby potentially ignoring geographic variability. Instead, we focused here on evaluation differences that stem from different natures of sources (i.e., claims versus providers' records), different health systems, and different populations. Third, it is currently limited to models that have up to hundreds of features. Fourth, we only used data sources mapped to OMOP format, thus avoiding the challenging task of standardizing cohort definition and feature extraction. However, while the OMOP format facilitates faster and more reliable cross-dataset research, data-shifts may still prevail, e.g., from differences in recording practices. It would be interesting to test the impact of the level of standardization on estimation accuracy in future benchmarks. Fifth, estimation accuracy depends on sample size. Specifically, the internal cohort requires hundreds to a few thousands of cases. However, this requirement is reasonable for models that are developed with external validity in mind. Finally, we assessed the method accuracy using AUROC, Brier score, scaled Brier score, and calibration-in-the-large, which capture the main dimensions of the model's performance. Other measures exist but have not been considered.

The proposed method is useful when only external summary statistics are given, but unit-level data is unavailable. In particular, in collaborative projects, the model evaluator could request summary statistics from external collaborators; then test multiple models and select an optimal one based on how well it is estimated to perform across external cohorts. This allows the evaluator to internally evaluate several candidate models before applying final evaluation, thereby expediting the development process. In this case, an optimal setting will include some external sources that provide the statistics and others that allow final actual validation after model selection.

We note that while this benchmark focused on EHR based models, the tested method can be applied to models that are based on other data types. For example, mobile and wearable devices can generate data that allows training useful models that are based on a few dozens of features (e.g., refs. 12,13).

Future work can improve the usability of the evaluation method in several ways. First, additional benchmarks on non-US datasets may shed light on their performance under more diverse sources of variability. Second, the ability to capture performance trends over time may also be an important use case. Third, while the current benchmarks assume a collaborative environment that allows the sharing of all statistics that are relevant to a specific model, it may be interesting to test its performance with limited or pre-computed statistics, for example, from national resources or database profiles.

In conclusion, the results of this benchmark study show that given sufficiently rich cohorts, a weighting approach gives an accurate estimation of the model's performance on external cohorts when only summary statistics are available. To facilitate this benchmark and general real-world cases, we adapted the underlying algorithm to handle large datasets. The tested approach may be useful both in preliminary assessments prior to deployment to new settings as well as a means to expedite collaborative model development.

Methods

Ethics

The use of Merative MarketScan® and Optum® databases were reviewed by the New England Institutional Review Board and were determined to be exempt from broad Institutional Review Board approval.

Study design

To investigate whether external summary statistics can be utilized to accurately estimate external performance we designed an experiment to compare the true external performance against the estimated performance and the internal validation performance. This was performed across five different observational healthcare data sources and five different prediction tasks.

Data sources

Five observational US healthcare database sources, including four insurance claims sources and one electronic healthcare record (EHR) source, were included in this study. All the data resources were mapped to a common format known as the Observational Medical Outcomes Partnership (OMOP) Common Data Model[14].

The Merative™ MarketScan® Commercial Database (CCAE) includes health insurance claims across the continuum of care (e.g., inpatient, outpatient, outpatient pharmacy, carve-out behavioral healthcare) as well as enrollment data from large employers and health plans across the United States that provide private healthcare coverage for employees, their spouses, and dependents. This administrative claims database includes a variety of fee-for-service, preferred provider organizations, and capitated health plans.

The Merative™ 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 Medicaid enrollees, covered under fee-for-service and managed care plans. It includes records of inpatient services, inpatient admissions, outpatient services, and prescription drug claims, as well as information on long-term care. Data on eligibility, as well as service and provider type, are also included. In addition to standard demographic variables, such as age and gender, the database includes variables such as federal aid category (income-based, disability, Temporary Assistance for Needy Families) and race.

The Merative™ MarketScan® Medicare Supplemental Database (MDCR) represents the health services of retirees in the United States with Medicare supplemental coverage through employer-sponsored plans. This database contains primarily fee-for-service plans and includes health insurance claims across the continuum of care (e.g., inpatient, outpatient, and outpatient pharmacy).

Optum®'s de-identified Clinformatics® Data Mart Database (Clinformatics®) is derived from a database of administrative health claims for members of large commercial and Medicare Advantage health plans. Clinformatics® is statistically de-identified under the Expert Determination method consistent with HIPAA and managed according to Optum® customer data use agreements. Administrative claims submitted for payment by providers and pharmacies are verified, adjudicated, and de-identified prior to inclusion. This data, including patient-level enrollment information, is derived from claims submitted for all medical and pharmacy healthcare services with information related to healthcare costs and resource utilization. The population is geographically diverse, spanning all 50 states.

Optum® de-identified Electronic Health Record dataset (Optum® EHR) is derived from dozens of healthcare provider organizations in the United States, that include more than 57 contributing sources and 111K sites of care. The data is certified as de-identified by an independent statistical expert following HIPAA statistical de-identification rules and managed according to Optum® customer data use agreements. Clinical, claims, and other medical administrative data is obtained from both inpatient and ambulatory EHRs, practice management systems, and numerous other internal systems. Information is processed, normalized, and standardized across the continuum of care from both acute inpatient stays and outpatient visits. Optum® EHR data elements include demographics, medications prescribed and administered, immunizations, allergies, lab results (including microbiology), vital signs and other observable measurements, clinical and inpatient stay administrative data, and coding diagnoses and procedures.

MDD study population

The target population consisted of patients with a diagnosis of major depressive disorder (MDD) for the first time and an antidepressant prescription recorded within 30 days of the initial MDD diagnosis. The prediction index was the date of the first MDD record per patient. Patients were excluded if they had less than 365 days of observation in the database prior to the index or had a history of mania, dementia, or psychosis. This definition has been used in prior OHDSI methodology papers[15].

Outcomes

Models were developed to predict the onset of five outcomes occurring within one year after index: seizure, diarrhea, fracture, gastrointestinal (GI) bleeding, and insomnia. All outcomes were defined based on corresponding diagnosis records; seizure and GI bleed also required that these diagnoses were given during an inpatient or emergency room visit[15]. For each outcome, patients were excluded if they had the outcome recorded prior to the index.

Prediction tasks

The prediction tasks of interest were: for patients in the MDD study cohort, predict the risk of *outcome* occurring for the first time within 1 day to 365 days after index.

Features

Three candidate feature sets were used to train prediction models:

- **Small**: only the patient's sex and age at index one-hot-encoded into 5-year buckets (0–4, 5–9, etc).
- **Medium**: patient's sex, age at index one-hot-encoded, and a constrained set of 64 phenotype predictors (see https://ohdsi.github.io/PatientLevelPrediction/articles/ConstrainedPredictors.html for more details).
- **Large**: patient's sex, age at index one-hot-encoded, and thousands of one-hot-encoded features representing whether the patient had a record of each medical condition and drug code recorded in the database prior to index. For example, diabetes is often recorded in OMOP databases via the SNOMED-CT code 73211009. If a patient had code 73211009 recorded prior to index, their feature value is 1 and 0 otherwise, for the feature "Had code 73211009 recorded prior to index".

Note that, by design, none of these features can be missing. Age and sex are required fields in OMOP data. The other features simply represent whether a code is recorded, and this information is never unknown.

Model development and validation

Prediction models (logistic regression with L1 regularization and gradient boosting machines) were developed within each data source per prediction task. For each data source and prediction task pair, labeled data were extracted consisting of the features and true outcome (whether the patient developed the outcome within 1-year of prediction index: class 1 represents those who did and class 0 represents those who did not) for each unit (i.e., patient in the target cohort). A model was developed using the labeled data by splitting the data into 75% training data and 25% testing data, then implementing 3-fold cross-validation using the training data to identify the optimal regularization hyper-parameter and then finally fitting a model using the optimal hyper-parameter and all the training data. Internal validation was performed by applying the model to the left-out testing data and comparing the predicted risk with the true outcome.

Each model was externally validated across the other data sources by applying the model to make predictions using the labeled data from the four other data sources and comparing the predicted risk with the true label. This provided the true external validation performance.

The proposed estimation of the external performance algorithm was also implemented for each model, using summary statistics of the model's important features from each external cohort. This provided the estimated external validation performances.

Performance metrics

We assessed model discrimination using the area under the receiver operating characteristic curve (AUROC); model calibration using calibration-in-the-large, i.e., the ratio of the mean predicted risk across the test study population to the true observed risk; and overall accuracy using the Brier score, corresponding to the sum of the squared differences between predicted risk and true label.

The tested cohorts have highly imbalanced outcome rates, which is often the case in medical cohorts. As the Brier score is sensitive to such imbalance, we also assessed the overall accuracy using the *scaled Brier score*[16]. This score scales Brier by its maximum value under a non-informative reference model that outputs a constant probability of the outcome, regardless of the features. Specifically,

$$\text{Brier}_{\text{scaled}} = 1 - \frac{\text{Brier}}{\text{Brier}_{\max}}, \tag{1}$$

where $\text{Brier}_{\max} = \bar{p}(1 - \bar{p})$ and $\bar{p}$ is the mean of model's predictions.

Overview of the external performance estimation method

The performance estimation algorithm is a scalable variant of the method presented in ref. 7, as illustrated in Fig. 1. Given a classifier, an internal (test) cohort, and summary statistics from an external cohort, this method aims to assign weights to units in the internal cohort that reproduce the statistical properties of the external one. Next, it computes performance metrics using the classifier predictions and true labels on the weighted internal cohort.

In the next section, we will describe the previous method for completeness. In the following, we will introduce more efficient algorithms to handle large cohorts.

Weighting an internal dataset to reproduce external statistics and estimating performance

Suppose we have an internal test cohort $\mathcal{D}_{\text{int}} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n_{\text{int}}}$ with n_{int} units, where $\boldsymbol{x}_i$ are feature vectors, and y_i are binary outcomes; as well as summary statistics, $\boldsymbol{\mu}_{\text{ext}}$, from an external cohort with n_{ext} units. The summary statistics are defined as empirical averages of a set of transformations on unit-level observations composed of features and an outcome label:

$$\boldsymbol{\mu}_{\text{ext}} \equiv \frac{1}{n_{\text{ext}}} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{ext}}} \boldsymbol{\phi}(\boldsymbol{x}_i, y_i), \tag{2}$$

where $\boldsymbol{\phi}(\boldsymbol{x}_i, y_i)$ is a vector valued function. For example,

$$\boldsymbol{\phi}(\boldsymbol{x}_i, y_i) = \{\boldsymbol{x}_i \cdot y_i,\ \boldsymbol{x}_i \cdot (1 - y_i),\ y_i\} \tag{3}$$

allows computation of feature means in subsets of units with and without the outcome (as often reported in a study's Table 1).

To produce a weighted cohort that has similar statistical properties as the external cohort, we search for a set of non-negative weights $\{w_i\}_{i=1}^{n_{\text{int}}}$ that sum to one, such that $\boldsymbol{\mu}_{\text{ext}} = \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{int}}} w_i \cdot \boldsymbol{\phi}(\boldsymbol{x}_i, y_i)$. We refer to $\{w_i, \boldsymbol{x}_i, y_i\}_{i=1}^{n_{\text{int}}}$ as a weighted cohort.

Our previous work cast the task of finding internal weights that reproduce the external statistical properties as an optimization problem[7]. We start by denoting the simplex of n_{int}-dimensional weights by $\Delta_{n_{\text{int}}}$, i.e.,

$$\Delta_{n_{\text{int}}} \equiv \left\{ \boldsymbol{w} \in \mathbb{R}^{n_{\text{int}}} : \sum w_i = 1,\ w_i \geq 0 \right\}. \tag{4}$$

As we assume that n_{int} is larger than the dimension of ϕ, there may be infinite number of possible weight vectors $\boldsymbol{w} \in \Delta_{n_{\text{int}}}$ such that $\boldsymbol{\Phi}_{\text{int}}^{\top} \boldsymbol{w} = \boldsymbol{\mu}_{\text{ext}}$, where $\boldsymbol{\Phi}$ is a matrix whose rows are $\boldsymbol{\phi}_i \equiv \boldsymbol{\phi}(\boldsymbol{x}_i, y_i)$. Therefore, we propose to search for weights that satisfy this equality and are as close to uniform as possible. We use the *Kulback-Leibler (KL) divergence* as a measure of proximity, where $KL(\boldsymbol{w} \parallel \boldsymbol{1}/n) \equiv \sum_{w_i} w_i \log \frac{w_i}{1/n}$. This approach requires solving the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{w}} && KL(\boldsymbol{w} \parallel \boldsymbol{1}/n) \\ &\text{such that} && \boldsymbol{\Phi}_{\text{int}}^{\top} \boldsymbol{w} = \boldsymbol{\mu}_{\text{ext}} \quad \text{and } \boldsymbol{w} \in \Delta_{n_{\text{int}}}. \end{aligned} \tag{5}$$

KL-divergence is a convex function, and so are the constraints. Therefore, generic convex optimization libraries allow minimizing this function while satisfying the feature average equalities on cohorts with thousands of units[7].

Given a weighted dataset $\{w_i, \boldsymbol{x}_i, y_i\}_{i=1}^{n_{\text{int}}}$ and probabilistic classifier outputs $\{p_i\}_{i=1}^{n_{\text{int}}}$, we compute different performance metrics, such as the Brier score and AUROC, using weighted versions of these metrics. For example, the Brier score measures the mean squared error of the probabilistic predictions, i.e., $\frac{1}{n_{\text{ext}}} \sum_{i=1}^{n_{\text{ext}}} (p_i - y_i)^2$. The weighted cohort allows to approximate the Brier score in the external cohort simply using the weighted score $\frac{1}{n_{\text{int}}} \sum_{i=1}^{n_{\text{int}}} w_i (p_i - y_i)^2$. To estimate the weighted AUROC, we use the `WeightedROC` **R** package.

Efficient approximate weighting

As in this benchmark we deal with larger scale problems than those tested in ref. 7, we developed a more efficient algorithm by approximating the problem represented in Equation (5). First, to satisfy equality constraints, we formulate the following optimization problem:

$$\min_{\boldsymbol{w} \in \Delta_{n_{\text{int}}}} f(\boldsymbol{w}) \equiv \left\| \boldsymbol{\Phi}_{\text{int}}^{\top} \boldsymbol{w} - \boldsymbol{\mu}_{\text{ext}} \right\|_2^2. \tag{6}$$

Second, to maintain weights that are as close to uniform as possible, we use an *exponentiated gradient* algorithm[17] and initialize the algorithm with uniform weights $w_i^0 = 1/n_{\text{int}}$. At every iteration t of the gradient descent algorithm, it attempts to decrease the objective while maintaining the normalization constraints using the following updates:

$$\bar{\boldsymbol{w}} = \boldsymbol{w}^t e^{-\alpha \nabla f(\boldsymbol{w})} \tag{7}$$

$$\boldsymbol{w}^{t+1} = \frac{\bar{\boldsymbol{w}}}{\sum_{i=1}^{n} \bar{w}_i}, \tag{8}$$

where $\nabla f(\boldsymbol{w})$ is the gradient of f, specifically, $\nabla f(\boldsymbol{w}) = \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\top} \boldsymbol{w}^t - \boldsymbol{\mu}_{\text{ext}})$, and α is a pre-specified optimization rate. As every iteration locally minimizes $f(\boldsymbol{w}^{t+1}) + \alpha \cdot KL(\boldsymbol{w}^{t+1} \| \boldsymbol{w}^t)$[17], it tends to reduce the objective while attempting to maintain proximity between the weights in consecutive steps. Therefore, we use the heuristic of initializing the search with uniform weights.

Assumptions and conditions for running the algorithm

We pose two assumptions about the underlying data distributions of the internal and external cohorts that should be satisfied to give accurate estimations. First, for all the features contributing to the prediction model, whenever the external joint probability of a given feature and outcome values is greater than zero, so is their internal one. We call this assumption *one-sided positivity*, as it is analogous to the positivity, also known as overlap, assumption in causal inference[18]. Second, we assume that the external distribution is relatively close to the internal one among the set of distributions that have expectations $\boldsymbol{\mu}_{\text{ext}}$. Intuitively, the plausibility of this assumption increases with the richness of the transformation ϕ of which we share external statistics. In other words, the assumption states that the statistics shared between the external and internal systems are sufficient to give a good approximation of the data shift between them.

Estimation pipeline

We implemented an estimation pipeline that combines the weighting algorithm and various tests to assess the feasibility of accurate estimations. Specifically, it includes the following steps:

1. Assess one-sided positivity for each element in the external statistics

2. If one-sided positivity holds: run the optimization algorithm to solve the problem represented in Equation (6) and obtain weights.
3. If external statistics are reproduced: estimate external performance using the weighted dataset.

Confidence intervals can be computed using repeated re-sampling and weighting of the internal cohort.

One-sided positivity tests

The following tests are performed before running the weighting algorithm, to assess one-sided positivity:

1. Verify that $\boldsymbol{\Phi}_{\text{int}}$ and $\boldsymbol{\mu}_{\text{ext}}$ contain the same features and feature-outcome transformations and do not contain missing values.
2. **Unary variables:** In case one of the columns of $\boldsymbol{\Phi}_{\text{int}}$ has a constant value across all units, make sure that the difference between this value and the corresponding entry in $\boldsymbol{\mu}_{\text{ext}}$ is less than a threshold (default 0.01).
3. **Binary variables:** First, if the external statistics of binary features are exactly the same as one of the internal binary values, we assume that the external cohort is composed of a sub-population with this value and use only the corresponding sub-population in the internal cohort for weighting. Second, we test if univariate weighting on this variable will not result in assigning most of the weights to a few internal units. Specifically, let n_0 and n_1 be the number of internal units with feature values 0 and 1. Let p_0 and p_1 be the proportion of such units in the external dataset. Then, we assume that when $\frac{p_0^2}{n_0}+\frac{p_1^2}{n_1}>(\frac{1}{40})^2$ most weights will be assigned to a few samples and the variance of any weighted estimator will be too large.
4. **Continuous variables:** Make sure that the external statistics is within the range of the internal cohort values. We allow a small slack (default 0.01).

Weighting algorithm pre-processing and parameter settings

Before running this algorithm, the features in the internal dataset are normalized by subtracting the mean and dividing by the standard deviations. The means and standard deviation are maintained and used later to apply the same transformation to the external statistics. This step maintains the relationships between the internal cohort and the external statistics while improving numerical stability and standardizing decision threshold. The convergence of the algorithm is examined by testing if the L_2 norm of the distance between the weighted internal and the external statistics is less than a threshold (default value 10^{-5}). The maximum number of iterations in this benchmark was set to 2000. If the weighting algorithm converges, we test that the maximum standardized mean difference between the weighted internal statistics and the external ones is less than a threshold (default 0.05).

Data availability

Data may be obtained from a third party and are not publicly available. The MarketScan CCAE, MDCD, and MDCR data that support the findings of this study are available from Merative (contact at: https://www.merative.com/documents/brief/marketscan-explainer-general) and the Optum EHR and Clinformatics datasets are available from Optum (contact at https://www.optum.com/en/business/life-sciences/real-world-data.html), but restrictions apply to the availability of these data, which were used under license for the current study.

Code availability

The benchmark code is available at https://github.com/ohdsi-studies/ExternalValidation. The code of the weighting-based estimation algorithm is available at https://github.com/KI-Research-Institute/LearningWithExternalStats.

Received: 13 June 2024; Accepted: 21 December 2024;
Published online: 27 January 2025

References

1. Wynants, L. et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
2. Collins, G. S. et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* **384**, e074819 (2024).
3. Balagopalan, A. et al. Machine learning for healthcare that matters: Reorienting from technical novelty to equitable impact. *PLoS Digit. Health* **3**, e0000474 (2024).
4. Wong, A. et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.* **181**, 1065–1070 (2021).
5. Reps, J. M. 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.* **20**, 102 (2020).
6. Riley, R. D. et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ* **384**, e074820 (2024).
7. El-Hay, T. & Yanover, C. Estimating model performance on external samples from their limited statistical characteristics. (eds Flores, G., Chen, G. H., Pollard, T., Ho, J. C. & Naumann, T.) *Proceedings of the Conference on Health, Inference, and Learning*, Vol. 174 of *Proceedings of Machine Learning Research*, 48–62 (PMLR, 2022). https://proceedings.mlr.press/v174/el-hay22a.html.
8. Kostka, K. et al. Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS. *Clin. Epidemiol.* **14**, 369–384 (2022).
9. Reps, J. M., Williams, R. D., Schuemie, M. J., Ryan, P. B. & Rijnbeek, P. R. Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. *BMC Med. Inform. Decis. Mak.* **22**, 142 (2022).
10. Subbaswamy, A., Adams, R. & Saria, S. Evaluating model robustness and stability to dataset shift. In: *International Conference on Artificial Intelligence and Statistics*, 2611–2619 (PMLR, 2021).
11. Chen, M. et al. Mandoline: Model evaluation under distribution shift. In: *International Conference on Machine Learning*, 1617–1629 (PMLR, 2021).
12. Gashi, S. et al. Modeling multiple sclerosis using mobile and wearable sensor data. *NPJ Digit. Med.* **7**, 64 (2024).
13. Dixon, W. G. et al. How the weather affects the pain of citizen scientists using a smartphone app. *NPJ Digit. Med.* **2**, 105 (2019).
14. Reisinger, S. J. et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J. Am. Med. Inform. Assoc.* **17**, 652–662 (2010).
15. Reps, J. M., Schuemie, M. J., Suchard, M. A., Ryan, P. B. & Rijnbeek, P. R. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J. Am. Med. Inform. Assoc.* **25**, 969–975 (2018).
16. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
17. Kivinen, J. & Warmuth, M. K. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.* **132**, 1–63 (1997).
18. Hernán, M. & Robins, J. *Causal Inference: What if* (CRC Press, 2020).

Acknowledgements

The authors would like to thank Nir Kalkstein for useful discussions regarding efficient optimization. The authors declare that no funding was granted for the study.

Author contributions

T.E.H., J.M.R., and C.Y. designed the study and wrote the manuscript. T.E.H. adapted the method to large datasets. J.M.R. performed the benchmark and analysis. All authors reviewed the results and the final manuscript.

Competing interests

J.M.R. is employed by Johnson & Johnson and owns shares in Johnson & Johnson. T.E.H. and C.Y. declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01414-z.

Correspondence and requests for materials should be addressed to Tal El-Hay.

Reprints and permissions information is available at http://www.nature.com/reprints