Estimating Model Performance on External Samples from Their Limited Statistical Characteristics

Tal El-Hay talelh@kinstitute.org.il
KI Research Institute, Kfar Malal, Israel

Chen Yanover chen@kinstitute.org.il
KI Research Institute, Kfar Malal, Israel

Abstract

Methods that address data shifts usually assume full access to multiple datasets. In the healthcare domain, however, privacy-preserving regulations as well as commercial interests limit data availability and, as a result, researchers can typically study only a small number of datasets. In contrast, limited statistical characteristics of specific patient samples are much easier to share and may be available from previously published literature or focused collaborative efforts.

Here, we propose a method that estimates model performance in external samples from their limited statistical characteristics. We search for weights that induce internal statistics that are similar to the external ones; and that are closest to uniform. We then use model performance on the weighted internal sample as an estimation for the external counterpart.

We evaluate the proposed algorithm on simulated data as well as electronic medical record data for two risk models, predicting complications in ulcerative colitis patients and stroke in women diagnosed with atrial fibrillation. In the vast majority of cases, the estimated external performance is much closer to the actual one than the internal performance. Our proposed method may be an important building block in training robust models and detecting potential model failures in external environments.

Data and Code Availability This paper uses the IQVIA Medical Research Data, primary care electronic medical records (EMRs) from the United Kingdom (IMRD-UK, version: 2019-03), incorporating data from THIN, A Cegedim Database (reference made to THIN is intended to be descriptive of the data asset licensed by IQVIA), and transformed to the Observational Medical Outcomes Partnership (OMOP) common data model (CDM; v5.1) (OHDSI, 2019). Definitions of cohorts, features, and outcomes are available through OHDSI demo ATLAS. Code is available at `https://github.com/KI-Research-Institute/external-evaluation`.

1. Introduction

Predictive models, such as disease risk scores, are typically trained on a single, or few, data sources but are often expected to work well in other environments, that may vary in their population characteristics, clinical settings, and policies (Steyerberg and Harrell, 2016). In many cases, model performance deteriorates significantly in these external environments, as demonstrated repeatedly (e.g., Ohnuma and Uchino (2017)), and most recently for the widely implemented proprietary Epic Sepsis Model (Wong et al., 2021) and for COVID-19 risk models (Reps et al., 2021).

Model robustness – that is, its ability to provide accurate prediction despite changes, e.g., in the characteristics of input covariates – can be demonstrated using external validation, the process of evaluating model performance on data sources that were not used for its derivation. However, full access to medical datasets is often limited due to privacy, regulatory and commercial factors. Therefore, we aim to estimate the performance of a given model on external sources using only their more commonly available statistical characteristics.

Here, we propose an algorithm which reweights individuals in an internal sample to match external statistics, potentially reported in preceding publications or characterization studies (e.g., Recalde et al. 2021); then estimates the performance on the external sample using the reweighted internal one. We focus on cases that are common in the healthcare domain, where the size of samples (that is, number of individuals) is much larger than the number of features. In such cases, infinite number of weight sets may recapitulate the external statistics, therefore the proposed algorithms searches for weights with a minimal divergence from a uniform distribution.

We first study the strengths and limitations of our suggested approach using simulated data; then split a sample from a primary care dataset into "internal" versus "external" subsets based on demographic information, and validate the approach using a prediction model for 3-year risk of complications in ulcerative colitis patients; and, finally, use the entire primary care data to estimate the performance of three stroke risk scores in seven external resources and compare it to the actual performance, as reported in a recent study (Reps et al., 2020).

2. Related Work

The task of evaluating model performance in external samples, often with (at least some) data shift (Finlayson et al., 2021), is tightly coupled with that of training robust models, as evaluation is a necessary step in model selection and optimization.

One line of work handling data shifts adopts ideas from causal inference. Specifically, causal models (Bareinboim and Pearl, 2016) can distinguish invariant relations between risk factors (e.g., biological or physiological) and outcomes from context- or environment-dependent mechanisms (Subbaswamy et al., 2019). Subbaswamy et al. (2021) developed a method for analyzing model robustness (or stability) that, given a model, a set of distribution-fixed (immutable) variables and a set of distribution-varying (mutable) variables, identifies the sub-population with the worst average loss; thus, enabling evaluation of model safety, with no external information.

Sample reweighting is commonly applied to adjust for confounders, either measured (Hainmueller, 2012) or unmeasured (Streeter et al., 2017), and to account for selection bias (Kalton and Flores-Cervantes, 2003), typically leveraging fully-accessible samples. Methodologically, the optimization problem we derive is similar to that studied for entropy balancing (Hainmueller, 2012), which attempts to reweight a sample (e.g., control group) so its prespecified set of moments exactly match that of another sample (e.g., treatment group), while maximizing the weight entropy (that is, keeping weights as close as possible to uniform). We note, however, that we explore a different use-case and, consequently, optimize over moments of an otherwise inaccessible sample (rather than samples from an accessible data source).

3. Estimation Algorithm

The goal of the proposed method is to estimate the performance of a prediction model, e.g., risk score, on an external sample, given some of its statistical properties, and using an internal, fully-accessible data. Briefly, we reweight an internal sample to obtain the external statistics, then compute model performance on the weighted sample as an estimate of the external performance.

3.1. Problem Formulation

Let $\boldsymbol{x}_i$ and y_i denote an observation (or feature) vector and a binary outcome[†], respectively, for an individual i. Suppose we have access to observations for n_{int} individuals in an *internal* sample:

$$\mathcal{D}_{\text{int}} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{n_{\text{int}}};$$

and summary statistics for an *external* sample (with n_{ext} individuals):

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

where $\boldsymbol{\phi}(\boldsymbol{x}_i, y_i)$ is a set of transformations on individual-level observations. For example:

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

allows computation of features mean in subsets of individuals with and without the outcome (as often reported in a study's Table 1).

We aim at estimating the performance of a model m on the external sample $\mathcal{D}_{\text{ext}}$, using $\boldsymbol{\mu}$ and observations from $\mathcal{D}_{\text{int}}$. To this end, we search for weights $\boldsymbol{w} \in [0, 1]^{n_{\text{int}}}, \sum_i w_i = 1$, such that the statistical properties of the weighted sample $\{\boldsymbol{x}_i, y_i, w_i\}_{i=1}^{n_{\text{int}}}$ approximate these of the external one. Let $\mathcal{W}(\boldsymbol{\mu}, \boldsymbol{Z})$ denote the space of such weight sets:

$$\mathcal{W}(\boldsymbol{\mu}, \boldsymbol{Z}) = \Big\{ \boldsymbol{w} \in \mathbb{R}^n : \boldsymbol{Z}^\top \boldsymbol{w} = \boldsymbol{\mu}, \ \sum w_i = 1, \\ w_i \geq 0, \ i = 1, \ldots, n_{\text{int}} \Big\}$$

[†] We focus here on binary outcomes, as these are commonly used – and reported – in healthcare applications; it is possible to extend the proposed approach to continuous outcomes, using an appropriate performance measure and statistical characteristics.

Where $\boldsymbol{Z}$ is a matrix whose rows are $\boldsymbol{z}_i \equiv \boldsymbol{\phi}(\boldsymbol{x}_i, y_i)$. As $\mathcal{W}(\boldsymbol{\mu}, \boldsymbol{Z})$ may be infinitely large, we propose to search for a set of weights that is also closest to uniform. This additional constraint is based on a *proximity assumption*, intuitively that the external distribution is relatively similar to the distribution in $\mathcal{W}(\boldsymbol{\mu}, \boldsymbol{Z})$ that is closest to the internal distribution.

Using the reweighted sample we can now estimate two types of performance measures:

- Measures that can be expressed as a pointwise loss function, $l(m(\boldsymbol{x}_i), y_i)$, for which we estimate the expected loss of the model in the external sample as:

$$\frac{1}{n_{\text{int}}} \sum_{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_{\text{int}}} w_i \cdot l(m(x_i), y_i).$$

  For example, for a model that computes the probability of an outcome y, we can estimate the expected negative log-likelihood by setting $l(m(\boldsymbol{x}_i), y_i) = -y_i \log(m(\boldsymbol{x}_i)) - (1 - y_i)\log(1 - m(\boldsymbol{x}_i))$.

- Non-decomposable measures that can be evaluated on weighted samples. For example, the area under the receiver operating characteristic curve (AUC).

Below we present a model independent scheme, which minimizes an *f-divergence* function (for example, maximizes the weights entropy); and in Appendix A, we derive a model (and loss) dependent scheme, which maximizes a weighted upper bound on the model loss and the regularized divergence function.

Model-independent optimization scheme. To find a weighted representation of an internal sample that replicates the external expectations, we solve the following optimization problem:

$$\min_{\boldsymbol{w} \in \mathcal{W}(\boldsymbol{\mu}, \boldsymbol{Z})} D_f(\boldsymbol{w} \| \mathbf{1}/n), \tag{1}$$

where $D_f(P\|Q)$, *f-divergence*, for discrete measures P and Q is:

$$D_f(P\|Q) = \sum_x f\left(\frac{P(x)}{Q(x)}\right) Q(x)$$

and $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function, with $f(1) = 0$. For example, when $f(t) = t \log t$, Optimization Problem (1) becomes:

$$\max_{\boldsymbol{w} \in \mathcal{W}(\boldsymbol{\mu}, \boldsymbol{Z})} \mathcal{H}(\boldsymbol{w}), \tag{2}$$

where $\mathcal{H}(\boldsymbol{w}) = -\sum_i w_i \log w_i$ is the entropy function.

We derive a dual formulation of Problem (2), similar to Hainmueller (2012), in Appendix B; and show that the optimal solution has the form:

$$w_i \propto e^{\boldsymbol{z}_i \cdot \boldsymbol{\nu}},$$

where $\boldsymbol{\nu} \in R^{|\phi|}$. In other words, the optimal weights are normalized exponents of a linear function of $\boldsymbol{Z}$. We note that, as the number of features is typically much smaller than the sample size, the solution to the dual problem is expected to be more numerically stable than the primal's.

In cases where $\mathcal{W}(\boldsymbol{\mu}, \boldsymbol{Z}) = \varnothing$, Problem (1) can be adjusted to trade-off, using hyper-parameter λ, the accuracy at which the weighted internal sample reproduce the external statistics and proximity and rewritten as:

$$\begin{aligned} &\min_{\boldsymbol{w}} \left(\left\| \boldsymbol{Z}^\top \boldsymbol{w} - \boldsymbol{\mu} \right\| + \lambda \cdot D_f(\boldsymbol{w} \| \mathbf{1}/n) \right) \\ &\quad \text{such that} \sum_i w_i = 1, w_i \geq 0, \end{aligned} \tag{3}$$

where the norm can be L_2 or L_1.

3.2. Detecting Estimation Failure

To estimate the performance of a model in an external sample $\mathcal{D}_{\text{ext}}$, with distribution $P_{\text{ext}}(\boldsymbol{z})$ of transformed features, we assume that $P_{\text{int}}(\boldsymbol{z}) > 0$ whenever $P_{\text{ext}}(\boldsymbol{z}) > 0$. This condition is analogous to the *positivity assumption* in causal inference, except that it is *one sided*. In other words, the support of $P_{\text{ext}}(\boldsymbol{z})$ can be a strict subset of the support of $P_{\text{int}}(\boldsymbol{z})$. Although this assumption cannot be verified, its violations can be detected when external expectations cannot be attained in the internal sample.

3.3. Implementation

We used R's CVXR (Fu et al., 2020) library to solve the optimization problem and WeightedROC library (Hocking, 2020) to compute weighted AUC. To deal with cases where $\mathcal{W}(\boldsymbol{\mu}, \boldsymbol{Z}) = \varnothing$ we used the relaxed Problem (3) with $\lambda = 10^{-6}$ and L_2 norm. To alleviate numerical issues, we set a minimum weight parameter (10^{-6}) and remove features with a small standard deviation ($< 10^{-4}$).

4. Empirical Evaluation

To evaluate the accuracy of the proposed algorithm, we estimated the external performance of various models using an internal sample and limited external statistical characteristics, in three scenarios: (a) simulating data using a *structural equation model* (Bareinboim and Pearl, 2016) for "internal" and "external" environments; training an outcome prediction model on the internal sample and evaluating its performance on the external one; (b) extracting a cohort of newly diagnosed ulcerative colitis individuals in IMRD-UK data; synthetically splitting this cohort into "internal" and "external" samples; training a complication risk model on the internal sample and evaluating its performance on the external one; and (c) extracting atrial fibrillation patient cohorts in IMRD-UK data as an internal sample; evaluating the performance of three stroke risk models in multiple inaccessible claim and EMR databases using their published statistical characteristics.

4.1. Synthetic Data

We simulated synthetic data using structural equation models that contain a hidden variable $H \in \mathbb{R}$, features $\boldsymbol{X} \in \mathbb{R}^p$, a binary outcome Y, and a deterministic binary variable A where $A = 0$ denotes an internal environment and $A = 1$ denotes an external one (Figure 1). This framework allows examining the strengths and limitations of the proposed algorithm subject to different types of data shifts.

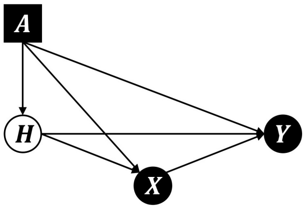


Figure 1: Graphical representation of the data-generating causal model. A is an environment variable (e.g., in a clinical setting, specific healthcare system), H is a hidden variable (encoding, for example, an individual's healthcare status), $\boldsymbol{X}$ is a set of observed features (e.g., prescribed medications or lab test results) and Y is a binary outcome (e.g., disease onset or progression).

We defined the simulations using the following structural equations model:

$$\begin{aligned} H &= \beta_{H,A} A + \epsilon_H \\ \boldsymbol{X} &= \boldsymbol{\beta}_{\boldsymbol{X},A} A + \boldsymbol{\beta}_{\boldsymbol{X},H} H + \boldsymbol{\beta}_{\boldsymbol{X},AH} AH + \boldsymbol{\epsilon}_{\boldsymbol{X}} \\ Y &\sim \text{Bernoulli}(\text{sigmoid}(f(\boldsymbol{X}, H, A))) \end{aligned}$$

where

$$f(\boldsymbol{X}, H, A) = \beta_{Y,A} A + \beta_{Y,H} H + \boldsymbol{\beta}_{Y,\boldsymbol{X}} \boldsymbol{X} + \boldsymbol{\beta}_{Y,A\boldsymbol{X}} A\boldsymbol{X} \,,$$

$\beta_{\boldsymbol{X},\cdot}$ and $\beta_{\cdot,\boldsymbol{X}} \in \mathbb{R}^p$ are coefficient vectors, the rest of the coefficients are scalars, $\epsilon_H \sim \mathcal{N}(0,1)$, $\epsilon_{\boldsymbol{X}} \sim \mathcal{N}(0, \mathbf{I}_p)$ are independent sources of variability and $\text{sigmoid}(z) = \frac{1}{1+e^{-z}}$.

This model is similar to the *anchor regression model* (Rothenhäusler et al., 2021), replacing the continuous outcome with a binary one. The dependency of $\boldsymbol{X}$ on the hidden variable H induces correlations between features, and the interaction term AH induces differences in the correlations structure between environments. Therefore, the coefficient $\boldsymbol{\beta}_{\boldsymbol{X},AH}$ controls the "strength" of the shift in correlations between features, depending on the environment; and the coefficient $\boldsymbol{\beta}_{Y,A\boldsymbol{X}}$ controls the shift in direct effect of $\boldsymbol{X}$ on Y.

4.1.1. Implementation

Here, we set the dimension of $\boldsymbol{X}$ to be $p = 10$ and sample coefficients $\beta_{H,A}$, $\beta_{Y,A} \sim \mathcal{N}(0, 0.2)$, $\boldsymbol{\beta}_{\boldsymbol{X},A} \sim \mathcal{N}(0, 0.2\mathbf{I}_p)$, $\beta_{X,H} \sim \mathcal{N}(0, \mathbf{I}_p)$, and $\beta_{Y,H} \sim \mathcal{N}(0,1)$. We let only X_1 and X_2 (but not X_3 to X_{10}) affect the outcome Y by setting $\beta_{Y,\boldsymbol{X}} = (1, 1, 0, \ldots, 0)$ and $\beta_{Y,A\boldsymbol{X}} = (-0.8, -0.2, 0, \ldots, 0)$.

As studies do not typically report correlations between features within each outcome class, we tested our algorithm in different scenarios of correlation shifts. Specifically, we used three configurations of $\boldsymbol{\beta}_{\boldsymbol{X},AH} \sim \mathcal{N}(0, \sigma_{\boldsymbol{X},AH})$, where $\sigma_{\boldsymbol{X},AH} = 0$, 0.5, or 1, emulating weak, medium, and strong correlation shifts, respectively.

Given a specific simulation model, we generated three data sets, namely internal training and tests sets and an external data set. We computed the mean and variance of every feature in $\boldsymbol{X}$, separately for individuals with $Y = 0$ and $Y = 1$, in the external set. Next, we trained an elastic net regularized logistic regression model on the internal training set and computed the AUC on the internal test and external sets. Finally, we applied the performance estimation algorithm on the internal test set, using external expectations, and compared the estimated AUC to the actual one.

Supplementary Figure 6 presents examples of generated samples with varying values of $\sigma_{\boldsymbol{X},AH}$. For each setting, we generated 200 models, and from each sampled data with varying sizes (n = 200, 500, 1000, 2000, 5000).

4.1.2. External Performance Estimation

The results of the proposed algorithm, in terms of divergence from uniform weights and AUC estimation accuracy, for different values of $\sigma_{\boldsymbol{X}|AH}$ and data size $n = 5000$ are shown in Table 1. As expected, weight divergence from uniform ($D_{KL}(\boldsymbol{w}\|\boldsymbol{1}/n)$) and estimation error grow with $\sigma_{X|AH}$.

Figure 2 presents the estimation error of the external AUC values, as a function of correlation shift strength and sample size, n. Estimation quality is lower for strong shifts in correlations which are not captured in the shared expectations, whereas milder differences result in good estimations. For comparison, the difference between internal and actual external AUC values is around 0.1 (Table 1).

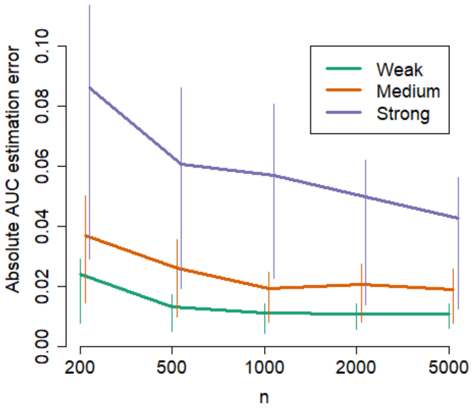


Figure 2: Estimation error (absolute value of the difference between actual and estimated external AUC values; y-axis) for weak, medium, and strong correlation shifts, as a function of sample size, n (x-axis). Whiskers correspond to 25-75 AUC percentiles, over 200 models.

Table 1: Algorithm performance in 5000-unit simulated datasets averaged on 200 sampling repetitions. The estimation error column presents the mean of the absolute values of errors. D_{KL}, Kullback Leibler divergence between the derived and uniform weights, and estimation error increase with stronger correlation shifts.

		Internal	External	Est.
$\sigma_{X\|AH}$	D_{KL}	AUC	AUC	Error
0.0 (Weak)	0.41	0.841	0.726	0.011
0.5 (Med.)	1.37	0.850	0.735	0.019
1.0 (Strong)	4.04	0.847	0.733	0.043

4.2. Synthetic Data Split: Complications of Ulcerative Colitis

Next, we studied the IMRD-UK primary care data and synthetically split it into ”internal” and ”external” sets based on various demographic criteria. Specifically, we trained a model on the internal sample to predict the 3-year risk of intestinal surgery (or death) in ulcerative colitis (UC) patients; estimated its performance on the external sample, using limited external statistics; and compared the estimated and observed performance.

4.2.1. Clinical Background

UC is a chronic inflammatory bowel condition with consistently increasing incidence rates in both newly industrialized and developed countries (Benchimol et al., 2014; Windsor and Kaplan, 2019; Kaplan and Windsor, 2021). The increase in its prevalence has a significant impact on healthcare financial burden due to chronically administered medications as well as hospitalizations and surgical procedures (Windsor and Kaplan, 2019).

UC pathogenesis is not well understood. Presumed risk factors for a more complicated disease include younger age at diagnosis, extensive disease, use of steroids and immunosupressive drugs, and being a non-smoker (Koliani-Pace and Siegel, 2019).

Table 2: Characteristics of internal and external samples, split by age.

	Internal: individuals > 34 years old, External: individuals ≤ 34 years old		Internal: individuals ≤ 64 years old, External: individuals > 64 years old	
	Internal	External	Internal	External
n	5577	1933	5616	1894
Townsend deprivation index				
Score	2.4 (±1.2)	2.6 (±1.3)	2.5 (±1.2)	2.4 (±1.2)
Available	4893 (87.7%)	1685 (87.2%)	4913 (87.5%)	1665 (87.9%)
Female	2752 (49.3%)	932 (48.2%)	2711 (48.3%)	973 (51.4%)
Smoking	1397 (25%)	362 (18.7%)	1393 (24.8%)	366 (19.3%)
Steroids	1597 (28.6%)	670 (34.7%)	1666 (29.7%)	601 (31.7%)
Body mass index (BMI)				
Underweight	105 (1.9%)	85 (4.4%)	140 (2.5%)	50 (2.6%)
Overweight	1535 (27.5%)	244 (12.6%)	1170 (20.8%)	609 (32.2%)
Perianal disease	66 (1.2%)	49 (2.5%)	97 (1.7%)	18 (1%)
Complications	900 (16.1%)	141 (7.3%)	457 (8.1%)	584 (30.8%)

4.2.2. Implementation

The UC onset cohort includes individuals with at least two diagnoses of inflammatory bowel disease (IBD) or with an IBD diagnosis and a prescription for an IBD medication; who have an ulcerative colitis diagnosis and no Crohn's disease diagnosis. We set index (or cohort entry) date to the first IBD diagnosis or medication prescription and required that individuals have a minimum observation of 365 days prior to index date. We excluded subjects with insufficient follow-up.

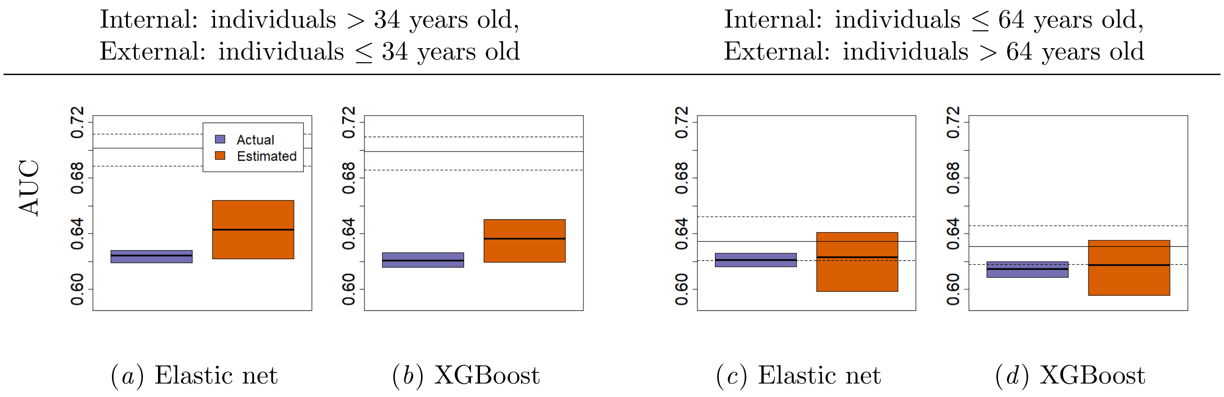
For each individual in the ulcerative colitis cohort we extracted a set of features, previously reported as associated with increased intestinal surgery risk (Koliani-Pace and Siegel, 2019). These include age (and age^2), sex, smoking, being underweight or overweight, presence of perianal disease, and use of steroids; and considered sets of predefined features (per OHDSI's Feature Extraction R library), e.g., drugs prescribed to a at least 1,000 subjects up to 90 days after index date. 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. Definition of all concept sets and cohorts are available at `https://atlas-demo.ohdsi.org/`.

We split the IMRD-UK data into internal and external sets based on individual age or country of living, as described below.

4.2.3. External Performance Estimation: Ulcerative Colitis, Split by Age

In the following experiments, we split the subset of individuals who live in England by their age. Specifically, in the first experiment, the internal set contained the 75% youngest subjects (≤ 64 years) and the external set – the 25% older ones; and in the second experiment, the internal set contains the 75% older individuals (> 34 years) and the internal set – the remaining older individuals. In each of these setups we randomly split the internal set to training (75%) and test (25%); we repeated the training-test random split 200 times. Next, we trained a linear model as well as a non-linear one, using XGBoost (Chen and Guestrin, 2016), computed model's AUC on the internal test and external sets, and estimated the external AUC using the internal set and the expectations of the external one. To maintain positivity and to emulate an environment dependent hidden factor, we excluded age from the feature set. The populations were different in several observed characteristics, notably, percentage of women, underweight and overweight; see Table 2 for details.

Overall, the external performance estimations, using either elastic net or XGBoost, are close to the actual ones (Figure 3), notably for external younger samples, where the difference between the internal and external performance is large (right panel).

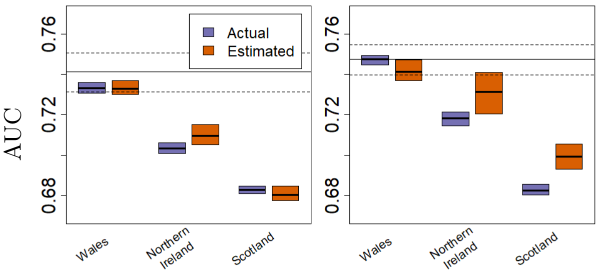

Internal: individuals > 34 years old,
External: individuals ≤ 34 years old

Internal: individuals ≤ 64 years old,
External: individuals > 64 years old

(*a*) Elastic net (*b*) XGBoost (*c*) Elastic net (*d*) XGBoost

Figure 3: Actual and estimated external performance in England UC cohorts, split by age. Boxes show the external median AUC and inter-quantile range (IQR, 25 and 75 percentiles) over 200 repetitions; solid line represents the median internal AUC and dashed lines represent the IQR.

4.2.4. External Performance Estimation: Ulcerative Colitis, Split by Country



(*a*) Elastic net (*b*) XGBoost

Figure 4: Actual and estimated performance in external, country-based UC samples. Boxes show the external median AUC and IQR over 200 repetitions; solid and dashed lines represent the internal median AUC and IQR, respectively.

Next, we split the UC cohort by country of residence and considered the sub-cohort of individuals living in England as the internal sample and those living in Scotland, Wales and Northern Ireland as three distinct external samples. Similarly to the age split analysis, we split the internal sample into training and test sets, repeatedly 200 times; trained a model on each training set; extracted expectations for the external samples; and evaluated model performance on the internal test and external sets.

The characteristics of different sub-populations are presented in Table 3; Figure 4 shows the external performance evaluation results, attesting to the (much) improved accuracy of the estimated AUC values, compared to internal performance.

4.3. Distinct Datasets: Stroke Risk Models

4.3.1. Clinical Background

Atrial fibrillation is a common cardiac rhythm disorder, associated with increased risk of stroke (Sagris et al., 2021). Risk factors associated with the occurrence of stroke include older age, various comorbidities (in particular, hypertension, diabetes, and renal disease) and smoking (Singer et al., 2013). To guide treatment, multiple risk scores have been devised and externally evaluated in several studies (van den Ham et al., 2015). Recently, Reps et al. (2020) replicated five previously published prognostic models that predict stroke in females newly diagnosed with atrial fibrillation; and externally validated their performance across nine observational healthcare datasets. Below, we use our proposed algorithm and the limited per-database statistical characteristics, as it appears in Reps et al. (2020), to estimate the external performance of these risk scores.

Table 3: Characteristics of ulcerative colitis, country-based sub-cohorts.

	England	Wales	Northern Ireland	Scotland
n	9469	1255	772	1772
Age (years)	48.7 (±18.9)	48.3 (±19.1)	46 (±18.2)	47 (±18.5)
Townsend deprivation index				
Score	2.5 (±1.2)	2.4 (±1.1)	2.9 (±1.3)	3 (±1.2)
Available	8265 (87.3%)	900 (71.7%)	634 (82.1%)	1541 (87%)
Female	4636 (49%)	602 (48%)	382 (49.5%)	909 (51.3%)
Smoking	2230 (23.6%)	313 (24.9%)	221 (28.6%)	484 (27.3%)
Steroids	2834 (29.9%)	408 (32.5%)	224 (29%)	668 (37.7%)
Body mass index (BMI)				
Underweight	248 (2.6%)	30 (2.4%)	24 (3.1%)	37 (2.1%)
Overweight	2276 (24%)	343 (27.3%)	200 (25.9%)	442 (24.9%)
Perianal disease	144 (1.5%)	16 (1.3%)	12 (1.6%)	11 (0.6%)
Complications	1315 (13.9%)	203 (16.2%)	123 (15.9%)	244 (13.8%)

4.3.2. Implementation

We downloaded Reps et al. (2020)'s analysis package and applied it to the IMRD-UK data, with the following modifications that adjust the study definitions to a primary care setting:

Target cohorts. We considered ECG-related procedures and conditions, in addition to measurements, within 30 days prior the atrial fibrillation diagnosis, as an optional inclusion criterion.

Outcome cohort. As stroke, typically not diagnosed in a primary care setting, may be poorly recorded for deceased individuals, we added death as an entry event to the stroke cohort.

Feature definitions. We extended the time window for extraction of model features to span the entire history of each individual until, and including, the date of the first atrial fibrillation event; included individuals with estimated glomerular filtration rate (eGFR) lower than 45 mL/min/1.73m^2 in the end stage renal disease cohort, as originally defined in the ATRIA risk model (Singer et al., 2013); and defined former smokers as individuals with an observation of smoker, as well as those diagnosed with tobacco dependence syndrome.

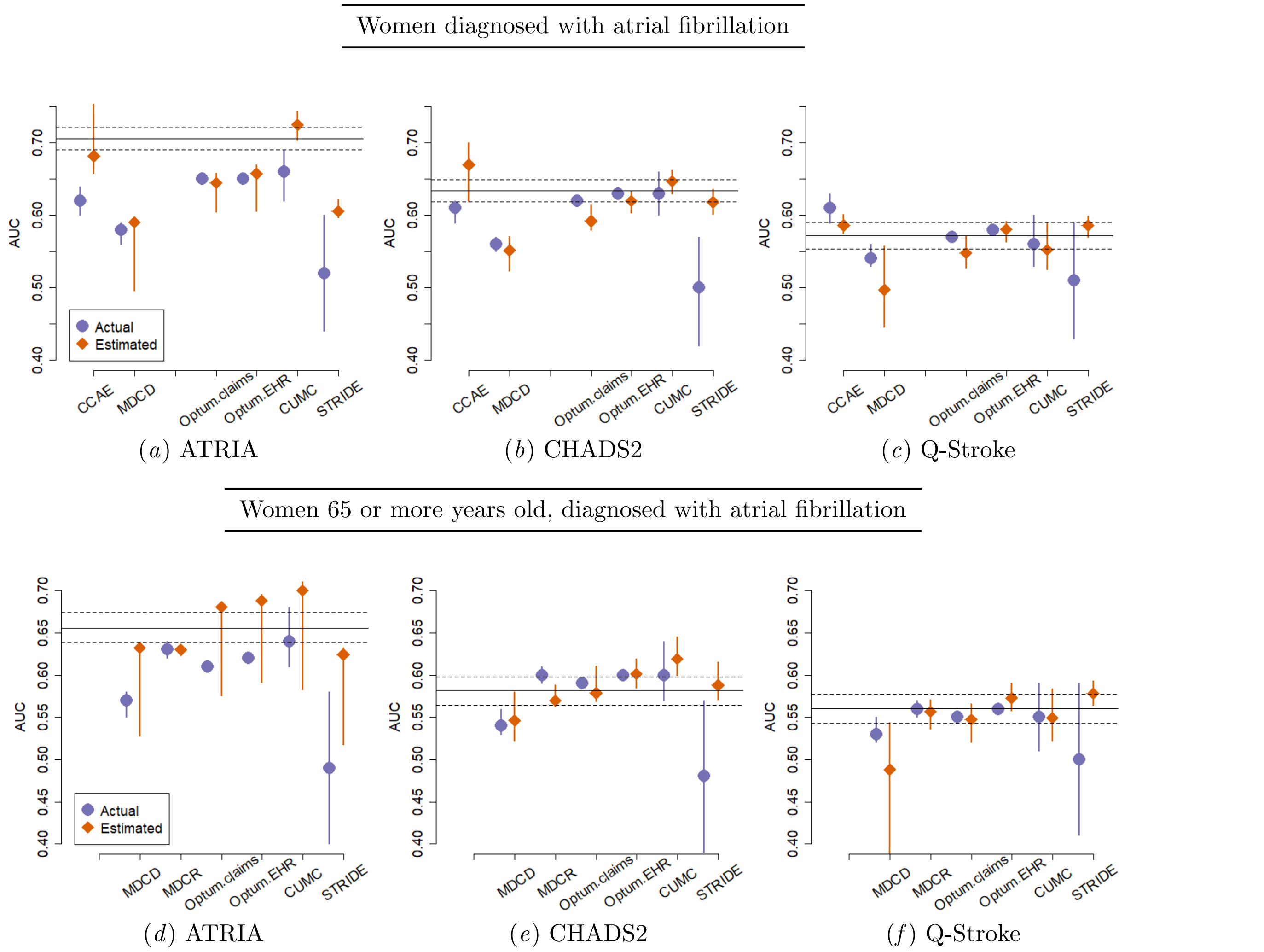
For each individual, the analysis package computed a stroke risk score given her set of features, as extracted from IMRD-UK; then, calculated score performance, vis-à-vis recorded stroke (and death) events. To estimate score performance in each external sample, we weighted individuals in the IMRD-UK data using the proposed algorithm to reproduce the sample's populations characteristics, as reported in Reps et al. (2020), and computed the score performance for the weighted individual cohort. We computed 95% confidence intervals using 1000 bootstrapping iterations.

Population attributes (Reps et al., 2020) include percentage of individuals in certain age groups (65-74 years, 75-85 years and above 85 years), comorbidities (hypertension, congestive heart failure, congestive cardiac failure, coronary heart disease, valvular heart disease, chronic and end stage renal disease, proteinuria, diabetes, and rheumatoid arthritis) and being a former smoker.

4.3.3. External Performance Estimation

A comparison between risk score performance, as reported by Reps et al. (2020), and the estimated performance is shown in Figure 5. For the full cohort (top panel), in three out of six datasets, the confidence interval of the ATRIA estimation overlaps the actual AUC (Figure 5a); in two other datasets, the estimation is better than the internal, IMRD-UK based performance. Qualitatively similar results are observed for the CHADS2 and Q-Stroke risk scores (Figure 5b and c, respectively); as well as for women 65 years or older (bottom panel).

We note that for two additional risk scores, Framingham and CHA$_2$DS$_2$VASc, and two datasets, Ajou University School Of Medicine (AUSOM) and Integrated Primary Care Information (IPCI), Reps



(a) ATRIA (b) CHADS2 (c) Q-Stroke

(d) ATRIA (e) CHADS2 (f) Q-Stroke

Figure 5: Performance estimation for three stroke risk scores across seven external datasets. Blue circles represent the actual AUC value as reported by Reps et al. (2020), red diamonds show the weighted estimations, and whiskers denote 95% confidence intervals. Solid lines represent the internal AUC, as computed in the IMRD-UK cohorts, with dashed lines denoting 95% confidence intervals.

et al. (2020) do not provide necessary statistical characteristics. Additionally, AUC values of IBM MarketScan® Medicare Supplemental Database (MDCR) were not provided for the full atrial fibrillation cohort and those of IBM MarketScan® Commercial Database (CCAE) are missing for the older female sub-cohort. Therefore, in all those cases, performance estimations are not reported.

5. Discussion

We presented an algorithm that estimates the performance of prediction models on external samples from their limited statistical characteristics; and demonstrated its utility using synthetic data, synthetic split of an ulcerative colitis cohort from a single database into age groups and by country of living, and a recent risk model benchmark of stroke risk models on multiple external samples. 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. It can thus direct development of robust models and accelerate deployment to external environments.

This study has several limitations. First, the algorithm relies on two assumptions: one-sided positivity and proximity. Both assumptions cannot be fully tested, but clear violations of the former one can be detected, for example, when the expected value of a feature is non-zero in the external distribution but all the individuals in the internal set have a zero value for that feature. Intuitively, proximity is more likely to be plausible when the statistical information becomes more detailed. Therefore, whereas our preliminary experiments involved only marginal statistics of features it may be informative to test the performance of the algorithm when more detailed statistics are available, for example interactions among features or information available in deep characterization studies (Burn et al., 2020). Second, when there are no weights that exactly reproduce the external statistics (Problem 1), we resort to a relaxed formulation (Problem 3), which requires tuning an additional hyper-parameter and may result in multiple, different solutions. Such a scenario may be common in high-dimensional models, where sample size is insufficient, compared to the number of constraints; we will investigate it further in future work. Lastly, for extremely large samples, optimizing over weights (Problem 3) may become prohibitively costly; large-scale optimization techniques may be used to overcome such challenges.

We believe that the proposed methodology can serve as a building block in network studies that aim to construct robust models across datasets when data sharing is limited, e.g., by regulatory constraints. Although federated learning methods may be a promising avenue for such scenarios, it would be interesting to explore in which cases the proposed algorithm can facilitate a one-shot federated learning scheme, that does not require deployment of federated algorithm clients in all network nodes.

In future work, we will combine the proposed algorithm with methods that aim to construct robust models such as those that leverage distributionally robust optimization (Bühlmann, 2020); study methods that exploit the relations between calibration and robustness (Wald et al., 2022); and look into decomposing AUC (Eban et al., 2017), so it can be optimized explicitly.

Institutional Review Board (IRB)

This study has been approved by IQVIA Scientific Review Committee (Reference numbers: 21SRC066, 22SRC002).

Acknowledgments

We thank Drs Roni Weisshof and Ramit Magen, Rambam Health Care Campus, Haifa, Israel, for their help in defining the ulcerative colitis predictive model; and Prof Seng Chan You, Yonsei University Health System, Seoul, Republic of Korea, for pointing us to the stroke external validation study. We are also grateful to KI Institute's researchers and, specifically, to Nir Kalkstein and Yonatan Bilu, for many fruitful discussions.

References

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, July 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1510507113. URL https://www.pnas.org/content/113/27/7345. Publisher: National Academy of Sciences Section: Colloquium Paper.

Eric I. Benchimol, Douglas G. Manuel, Astrid Guttmann, Geoffrey C. Nguyen, Nassim Mojaverian, Pauline Quach, and David R. Mack. Changing Age Demographics of Inflammatory Bowel Disease in Ontario, Canada: A Population-based Cohort Study of Epidemiology Trends. *Inflammatory Bowel Diseases*, 20(10):1761–1769, October 2014. ISSN 1078-0998. doi: 10.1097/MIB.0000000000000103. URL https://academic.oup.com/ibdjournal/article/20/10/1761-1769/4578853.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Edward Burn, Seng Chan You, Anthony G. Sena, Kristin Kostka, Hamed Abedtash, Maria Tereza F. Abrahão, Amanda Alberga, Heba Alghoul, Osaid Alser, Thamir M. Alshammari, Maria Aragon, Carlos Areia, Juan M. Banda, Jaehyeong Cho, Aedin C. Culhane, Alexander Davydov, Frank J. DeFalco, Talita Duarte-Salles, Scott DuVall, Thomas Falconer, Sergio Fernandez-Bertolin, Weihua Gao, Asieh Golozar, Jill Hardin, George Hripcsak, Vojtech Huser, Hokyun Jeon, Yonghua Jing, Chi Young Jung, Benjamin Skov Kaas-Hansen, Denys Kaduk, Seamus Kent, Yeesuk Kim, Spyros Kolovos, Jennifer C. E. Lane, Hyejin Lee, Kristine E. Lynch, Rupa Makadia, Michael E. Matheny, Paras P. Mehta, Daniel R. Morales, Karthik Natarajan, Fredrik Nyberg, Anna Ostropolets, Rae Woong Park, Jimyung Park, Jose D. Posada, Albert Prats-Uribe, Gowtham Rao, Christian Reich, Yeunsook Rho, Peter Rijnbeek, Lisa M. Schilling, Martijn Schuemie, Nigam H. Shah, Azza Shoaibi, Seokyoung Song, Matthew Spotnitz, Marc A. Suchard, Joel N. Swerdel, David Vizcaya, Salvatore Volpe, Haini Wen, Andrew E. Williams, Belay B. Yimer, Lin Zhang, Oleg Zhuk, Daniel Prieto-Alhambra, and Patrick Ryan. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. *Nature Communications*, 11(1), October 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18849-z. URL https://www.nature.com/articles/s41467-020-18849-z.

Peter Bühlmann. Invariance, Causality and Robustness. *Statistical Science*, 35(3):404–426, August 2020. ISSN 0883-4237, 2168-8745. doi: 10.1214/19-STS721. URL https://projecteuclid.org/journals/statistical-science/volume-35/issue-3/Invariance-Causality-and-Robustness/10.1214/19-STS721.full. Publisher: Institute of Mathematical Statistics.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Ryan Rifkin, and Gal Elidan. Scalable Learning of Non-Decomposable Objectives. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 832–840. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/eban17a.html.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*, 385(3):283–286, July 2021. ISSN 0028-4793. doi: 10.1056/NEJMc2104626. URL https://doi.org/10.1056/NEJMc2104626. Publisher: Massachusetts Medical Society.

Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R Package for Disciplined Convex Optimization. *Journal of Statistical Software*, 94(1):1–34, September 2020. ISSN 1548-7660. doi: 10.18637/jss.v094.i14. URL https://www.jstatsoft.org/index.php/jss/article/view/v094i14. Number: 1.

Jens Hainmueller. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1):25–46, 2012. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpr025. URL https://www.cambridge.org/core/product/identifier/S1047198700012997/type/journal_article.

Toby Dylan Hocking. WeightedROC: Fast, Weighted ROC Curves, February 2020. URL https://CRAN.R-project.org/package=WeightedROC.

Graham Kalton and Ismael Flores-Cervantes. Weighting methods. *Journal of official statistics*, 19(2), 2003.

Gilaad G. Kaplan and Joseph W. Windsor. The four epidemiological stages in the global evolution of inflammatory bowel disease. *Nature Reviews Gastroenterology & Hepatology*, 18(1):56–66, January 2021. ISSN 1759-5053. doi: 10.1038/s41575-020-00360-x. URL https://www.nature.com/articles/s41575-020-00360-x. Number: 1 Publisher: Nature Publishing Group.

Jenna L. Koliani-Pace and Corey A. Siegel. Prognosticating the Course of Inflammatory Bowel Disease. *Gastrointestinal Endoscopy Clinics of North America*, 29(3):395–404, July 2019. ISSN 1052-5157. doi: 10.1016/j.giec.2019.02.003. URL https://www.sciencedirect.com/science/article/pii/S1052515719300133.

OHDSI. *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI, 2019. ISBN 978-1-08-885519-5. URL https://books.google.co.il/books?id=JxpnzQEACAAJ.

Tetsu Ohnuma and Shigehiko Uchino. Prediction models and their external validation studies for mortality of patients with acute kidney injury: a systematic review. *PLoS One*, 12(1):e0169341, 2017.

Martina Recalde, Elena Roel, Andrea Pistillo, Anthony G. Sena, Albert Prats-Uribe, Waheed-Ul-Rahman Ahmed, Heba Alghoul, Thamir M. Alshammari, Osaid Alser, Carlos Areia, Edward Burn, Paula Casajust, Dalia Dawoud, Scott L. DuVall, Thomas Falconer, Sergio Fernández-Bertolín, Asieh Golozar, Mengchun Gong, Lana Yin Hui Lai, Jennifer C. E. Lane, Kristine E. Lynch, Michael E. Matheny, Paras P. Mehta, Daniel R. Morales, Karthik Natarjan, Fredrik Nyberg, Jose D. Posada, Christian G. Reich, Peter R. Rijnbeek, Lisa M. Schilling, Karishma Shah, Nigam H. Shah, Vignesh Subbian, Lin Zhang, Hong Zhu, Patrick Ryan, Daniel Prieto-Alhambra, Kristin Kostka, and Talita Duarte-Salles. Characteristics and outcomes of 627 044 COVID-19 patients living with and without obesity in the United States, Spain, and the United Kingdom. *International Journal of Obesity*, 45(11):2347–2357, November 2021. ISSN 1476-5497. doi: 10.1038/s41366-021-00893-4. URL https://www.nature.com/articles/s41366-021-00893-4.

Jenna M. Reps, Ross D. Williams, Seng Chan You, Thomas Falconer, Evan Minty, Alison Callahan, Patrick B. Ryan, Rae Woong Park, Hong-Seok Lim, and Peter Rijnbeek. 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 Medical Research Methodology*, 20(1):102, May 2020. ISSN 1471-2288. doi: 10.1186/s12874-020-00991-3. URL https://doi.org/10.1186/s12874-020-00991-3.

Jenna M. Reps, Chungsoo Kim, Ross D. Williams, Aniek F. Markus, Cynthia Yang, Talita Duarte-Salles, Thomas Falconer, Jitendra Jonnagaddala, Andrew Williams, Sergio Fernández-Bertolín, Scott L. DuVall, Kristin Kostka, Gowtham Rao, Azza Shoaibi, Anna Ostropolets, Matthew E. Spotnitz, Lin Zhang, Paula Casajust, Ewout W. Steyerberg, Fredrik Nyberg, Benjamin Skov Kaas-Hansen, Young Hwa Choi, Daniel Morales, Siaw-Teng Liaw, Maria Tereza Fernandes Abrahão, Carlos Areia, Michael E. Matheny, Kristine E. Lynch, María Aragón, Rae Woong Park, George Hripcsak, Christian G. Reich, Marc A. Suchard, Seng Chan You, Patrick B. Ryan, Daniel Prieto-Alhambra, and Peter R. Rijnbeek. Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study. *JMIR Medical Informatics*, 9(4):e21547, April 2021. doi: 10.2196/21547. URL https://medinform.jmir.org/2021/4/e21547.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):215–246, 2021. ISSN 1467-9868. doi: 10.1111/rssb.12398. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12398.

Marios Sagris, Emmanouil P. Vardas, Panagiotis Theofilis, Alexios S. Antonopoulos, Evangelos Oikonomou, and Dimitris Tousoulis. Atrial Fibrillation: Pathogenesis, Predisposing Factors, and Genetics. *International Journal of Molecular Sciences*, 23(1):6, December 2021. ISSN 1422-0067.

doi: 10.3390/ijms23010006. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8744894/.

Daniel E. Singer, Yuchiao Chang, Leila H. Borowsky, Margaret C. Fang, Niela K. Pomernacki, Natalia Udaltsova, Kristi Reynolds, and Alan S. Go. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. *Journal of the American Heart Association*, 2(3):e000250, June 2013. ISSN 2047-9980. doi: 10.1161/JAHA.113.000250.

Ewout W. Steyerberg and Frank E. Harrell. Prediction models need appropriate internal, internal-external, and external validation. *Journal of clinical epidemiology*, 69:245–247, 2016. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2015.04.005. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5578404/.

Adam J. Streeter, Nan Xuan Lin, Louise Crathorne, Marcela Haasova, Christopher Hyde, David Melzer, and William E. Henley. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *Journal of Clinical Epidemiology*, 87:23–34, 2017. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2017.04.022. URL https://www.sciencedirect.com/science/article/pii/S0895435616303341.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, April 2019. URL https://proceedings.mlr.press/v89/subbaswamy19a.html. ISSN: 2640-3498.

Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating Model Robustness and Stability to Dataset Shift. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2611–2619. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/subbaswamy21a.html.

Hendrika A. van den Ham, Olaf H. Klungel, Daniel E. Singer, Hubert G. M. Leufkens, and Tjeerd P. van Staa. Comparative Performance of ATRIA, CHADS2, and CHA2DS2-VASc Risk Scores Predicting Stroke in Patients With Atrial Fibrillation: Results From a National Primary Care Database. *Journal of the American College of Cardiology*, 66 (17):1851–1859, October 2015. ISSN 1558-3597. doi: 10.1016/j.jacc.2015.08.033.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On Calibration and Out-of-domain Generalization. *arXiv:2102.10395 [cs]*, January 2022. URL http://arxiv.org/abs/2102.10395. arXiv: 2102.10395.

Joseph W. Windsor and Gilaad G. Kaplan. Evolving Epidemiology of IBD. *Current Gastroenterology Reports*, 21(8):40, July 2019. ISSN 1534-312X. doi: 10.1007/s11894-019-0705-6.

Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, 181(8):1065–1070, August 2021. ISSN 2168-6106. doi: 10.1001/jamainternmed.2021.2626. URL https://doi.org/10.1001/jamainternmed.2021.2626.

Appendix A. Model-dependent optimization scheme

An upper bound of a model m's weighted loss l, up to a finite sample error, can be derived as follows:

$$\max_{\boldsymbol{w}\in\mathcal{W}(\boldsymbol{\mu},\boldsymbol{Z})} \sum_i w_i \cdot l(m(x_i), y_i).$$

The tightness of the bound may depend on the number of expectations we consider. Furthermore, as $\boldsymbol{z}$, and consequently $\boldsymbol{\mu}$, may not represent all inter-feature dependencies existing in the data, an additional constraint may yield improved estimations:

$$\max_{\boldsymbol{w}\in\mathcal{W}(\boldsymbol{\mu},\boldsymbol{Z})} \sum_i w_i \cdot l(m(x_i), y_i) - \lambda D_f(\boldsymbol{w}\|\mathbf{1}/n). \quad (4)$$

As we increase λ, the bound may become tighter but confidence may decrease.

Appendix B. Model-independent dual optimization problem

Recall that optimization Problem (2) is defined as follows:

$$\begin{aligned} \text{minimize}_w \quad & -\mathcal{H}(\boldsymbol{w}) \\ \text{such that} \quad & \boldsymbol{Z}^\top \boldsymbol{w} = \boldsymbol{\mu}, \boldsymbol{1}^\top \boldsymbol{w} = 1 \end{aligned} \tag{5}$$

where $\boldsymbol{w} \geq 0$. Denoting

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{Z}^\top \\ \boldsymbol{1}^\top \end{bmatrix}, \quad \boldsymbol{d} = \begin{bmatrix} \boldsymbol{\mu} \\ 1 \end{bmatrix}, \tag{6}$$

Problem (2) becomes:

$$\begin{aligned} \text{minimize}_w \quad & -\mathcal{H}(\boldsymbol{w}) \\ \text{such that} \quad & \boldsymbol{C}^\top \boldsymbol{w} = \boldsymbol{d} \end{aligned} \tag{7}$$

Following Equation 5.11 in Boyd et al. (2004) the dual function is:

$$g(\boldsymbol{\nu}) = -\boldsymbol{d}^\top \boldsymbol{\nu} - (-\mathcal{H})^*(-\boldsymbol{C}^\top \boldsymbol{\nu})$$

where $(-\mathcal{H})^*$ is the conjugate of the negative-entropy function (Boyd et al. (2004), p. 222):

$$(-\mathcal{H})^*(\boldsymbol{y}) = \sum_{i=1}^{n} e^{y_i - 1}$$

Therefore,

$$g(\boldsymbol{\nu}) = -(\boldsymbol{\mu}, 1)^\top \boldsymbol{\nu} - e^{-1} \sum_{i=1}^{n} e^{-(\boldsymbol{z}_i, 1)\boldsymbol{\nu}}$$

The Lagrangian of the primal problem is:

$$L(\boldsymbol{w}; \boldsymbol{\nu}) = \sum_i w_i \log w_i + \boldsymbol{\nu}^\top (\boldsymbol{C}\boldsymbol{w} - \boldsymbol{d}) \tag{8}$$

Let $\boldsymbol{\nu}^*$ be the optimal solution of $\max_{\boldsymbol{\nu}} g(\boldsymbol{\nu})$. Then, following Section 5.5.3 of Boyd et al. (2004), the solution of the primal problem minimizes the Lagrangian at $\boldsymbol{\nu}^*$:

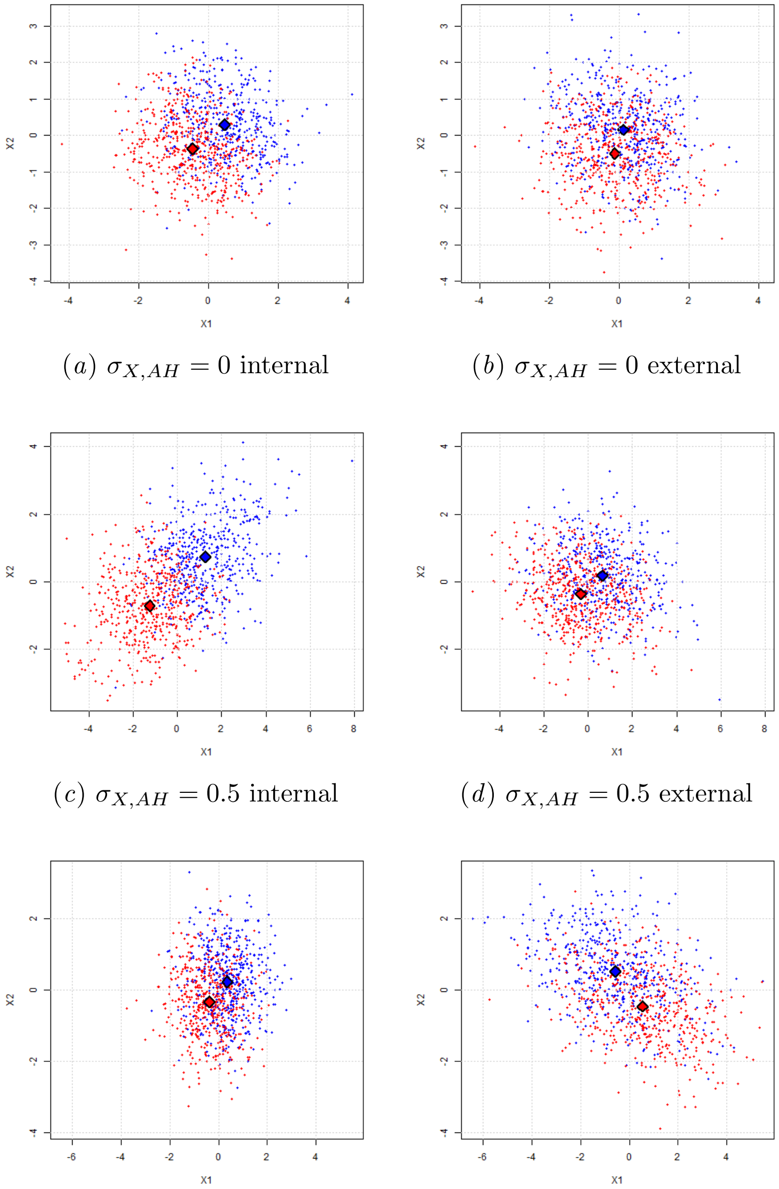
$$\frac{\partial L(\boldsymbol{w}; \boldsymbol{\nu}^*)}{\partial w_i} = \log w_i + 1 + (\boldsymbol{z}_i, 1)\ \boldsymbol{\nu}^* = 0$$

giving

$$w_i = e^{-1-(\boldsymbol{z}_i, 1)\boldsymbol{\nu}^*}.$$

This result shows that the optimal weights are normalized exponents of a linear function of the data points.

Appendix C. Supplementary Figures



(a) $\sigma_{X,AH} = 0$ internal

(b) $\sigma_{X,AH} = 0$ external

(c) $\sigma_{X,AH} = 0.5$ internal

(d) $\sigma_{X,AH} = 0.5$ external

(e) $\sigma_{X,AH} = 1$ internal

(f) $\sigma_{X,AH} = 1$ external

Figure 6: Simulation examples with varying values of $\sigma_{X,AH}$. Dot colors denote outcome class, diamonds represent class means. The shift in correlation between X_1 and X_2, given an outcome class, increases with $\sigma_{X,AH}$.