PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database

Harvineet Singh[1]*, **Vishwali Mhasawade**[2], **Rumi Chunara**[2,3]

1 New York University, Center for Data Science, **2** New York University, Tandon School of Engineering, **3** New York University, School of Global Public Health

* hs3673@nyu.edu

OPEN ACCESS

Citation: Singh H, Mhasawade V, Chunara R (2022) Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. PLOS Digit Health 1(4): e0000023. https://doi.org/10.1371/journal.pdig.0000023

Editor: Tom J. Pollard, MIT Laboratory for Computational Physiology, UNITED STATES

Received: August 3, 2021

Accepted: February 17, 2022

Published: April 5, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pdig.0000023

Data Availability Statement: All data used in the study is publicly accessible from PhysioNet website https://physionet.org/content/eicu-crd/2.0/

Abstract

Modern predictive models require large amounts of data for training and evaluation, absence of which may result in models that are specific to certain locations, populations in them and clinical practices. Yet, best practices for clinical risk prediction models have not yet considered such challenges to generalizability. Here we ask whether population- and group-level performance of mortality prediction models vary significantly when applied to hospitals or geographies different from the ones in which they are developed. Further, what characteristics of the datasets explain the performance variation? In this multi-center cross-sectional study, we analyzed electronic health records from 179 hospitals across the US with 70,126 hospitalizations from 2014 to 2015. Generalization gap, defined as difference between model performance metrics across hospitals, is computed for area under the receiver operating characteristic curve (AUC) and calibration slope. To assess model performance by the race variable, we report differences in false negative rates across groups. Data were also analyzed using a causal discovery algorithm "Fast Causal Inference" that infers paths of causal influence while identifying potential influences associated with unmeasured variables. When transferring models across hospitals, AUC at the test hospital ranged from 0.777 to 0.832 (1st-3rd quartile or IQR; median 0.801); calibration slope from 0.725 to 0.983 (IQR; median 0.853); and disparity in false negative rates from 0.046 to 0.168 (IQR; median 0.092). Distribution of all variable types (demography, vitals, and labs) differed significantly across hospitals and regions. The race variable also mediated differences in the relationship between clinical variables and mortality, by hospital/region. In conclusion, group-level performance should be assessed during generalizability checks to identify potential harms to the groups. Moreover, for developing methods to improve model performance in new environments, a better understanding and documentation of provenance of data and health processes are needed to identify and mitigate sources of variation.

upon completion of a training course and approval of data use request.

Funding: This study was funded by the National Science Foundation grant numbers 1845487 and 1922658. The funder had no role in design, conduct of the study or in preparing the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

With the growing use of predictive models in clinical care, it is imperative to assess failure modes of predictive models across regions and different populations. In this retrospective cross-sectional study based on a multi-center critical care database, we find that mortality risk prediction models developed in one hospital or geographic region exhibited lack of generalizability to different hospitals or regions. Moreover, distribution of clinical (vitals, labs and surgery) variables significantly varied across hospitals and regions. Based on a causal discovery analysis, we postulate that lack of generalizability results from dataset shifts in race and clinical variables across hospitals or regions. Further, we find that the race variable commonly mediated changes in clinical variable shifts. Findings demonstrate evidence that predictive models can exhibit disparities in performance across racial groups even while performing well in terms of average population-wide metrics. Therefore, assessment of sub-group-level performance should be recommended as part of model evaluation guidelines. Beyond algorithmic fairness metrics, an understanding of data generating processes for sub-groups is needed to identify and mitigate sources of variation, and to decide whether to use a risk prediction model in new environments.

Introduction

Validation of predictive models on intended populations is a critical prerequisite to their application in making individual-level care decisions since a miscalibrated or inaccurate model may lead to patient harm or waste limited care resources [1]. Models can be validated either on the same population as used in the development cohort, named internal validity, or on a different yet related population, named external validity or generalizability (or sometimes transportability) [2]. The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Statement strongly recommends assessing external validity of published predictive models in multiple ways including testing on data from a different geography, demography, time period, or practice setting [3]. However, the guidelines do not specify appropriate external validity parameters on any of the above factors. At the same time, recent studies in the computer science and biomedical informatics literature have indicated that sub-group performance of clinical risk prediction models by race or sex can vary dramatically [4,5], and clinical behavior can guide predictive performance [6]. Within the statistics literature are several methods for computing minimum sample size and other best practices for assessing external validity of clinical risk prediction models [7–9]. However, the assessment of sub-group-level performance and data-shifts are not explicitly considered in such guidance [10]. Moreover, recent analyses have shown that clinical prediction models are largely being developed in a limited set of geographies, bringing significant concern regarding generalizability of models to broader patient populations [11]. Amidst such rising challenges, an understanding of how differences among population and clinical data impact external generalizability of clinical risk prediction models is imperative.

When a model fails to generalize for specific patient groups (such as racial or gender identities), using it to guide clinical decisions can lead to disparate impact on such groups. This raises questions of equity and fairness in the use of clinical risk prediction models, for which performance on diverse groups has been repeatedly lacking [12–14]. Predictive discrimination quantifies how well a model can separate individuals with and without the outcome of interest (we study mortality prediction here). Calibration quantifies how well the predicted probabilities match with the observed outcomes. These measures can be used to check for aggregate

model performance across a study sample or within groups, but do not illuminate variation across groups. Hence, in assessment of generalizability, we add another set of measures to our analyses which we refer to as "fairness" metrics, following the algorithmic fairness literature [15–17]. Such performance checks are important, especially given the evidence on racial bias in medical decision-making tools [4,13,14].

The primary objective of this study is to evaluate the external validity of predictive models for clinical decision making across hospitals and geographies in terms of the metrics–predictive discrimination (area under the receiver operating characteristic curve), calibration (calibration slope) [18], and algorithmic fairness (disparity in false negative rates and disparity in calibration slopes). The secondary objective is to examine the possible reasons for performance changes via shifts in the distributions of different types of variables and their interactions. We focus on risk prediction models for in-hospital mortality in ICUs. Our choice of evaluation metrics are guided by the use of such models for making patient-level care decisions. We note that similar models (e.g., SAPS and APACHE scores) [19,20] are widely-used for other applications as well such as assessing quality-of-care, resource utilization, or risk-adjustment for estimating healthcare costs [19–21], which are not the focus of this study. Recently, prediction models for in-hospital mortality have been prospectively validated for potential use [22], or in case of sepsis, have even been integrated into the clinical workflow [23]. With access to large datasets through electronic health records, new risk prediction models leveraging machine learning approaches have been proposed, which provide considerable accuracy gains [24]. Being flexible, such approaches might overfit to the patterns in a particular dataset, thus, raising concerns for their generalization to newer environments [25]. We use the eICU dataset [26] as a test bed for our analyses. Past studies have employed the dataset for evaluating mortality prediction models [27,28]. As the dataset was collected from multiple hospitals across the US, it allows us, in a limited way, to test external validity across hospitals, diverse geographies, and populations.

Materials and methods

Analyses are based on data obtained from the publicly-available eICU Collaborative Research Database [26], designed to aid remote care of critically-ill patients in a telehealth ICU program. The database is composed of a stratified random sample of ICU stays from hospitals in the telehealth program where the sample is selected such that the distribution of the number of unique patient-stays across hospitals is maintained [26]. Data on 200,859 distinct ICU stays of 139,367 patients with multiple visits across 208 hospitals in the US between 2014 and 2015 are included. We followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline [29].

Data preprocessing

We follow the feature extraction and exclusion procedures including the exclusion criteria used by Johnson et al [27] which removes patient stays conforming to APACHE IV exclusion criteria [20] and removes all non-ICU stays. APACHE IV criteria excludes patients admitted for burns, in-hospital readmissions, patients without a recorded diagnosis after 24 hours of ICU admission, and some transplant patients [26]. Only patients 16 years or above are included. Age of patients above 89 years (which is obfuscated to adhere to HIPAA provisions) is coded as 90. After pre-processing, the dataset consists of 70,126 stays from 179 hospitals. For analyses, data is grouped at two levels–by individual hospitals and by U.S. geographic regions (Northeast, South, Midwest, West) [30]. Hospital-level analyses are restricted to the top 10 hospitals with the most stays, all of which have at least 1631 stays, to ensure enough examples

for model training and evaluation. Data is split into ten separate datasets using a hospital identifier for hospital-specific analyses and into four separate datasets using a region identifier for region-specific analyses. The outcome label is in-hospital mortality (binary). Mortality rates differed in the range of 3.9%-9.3% (1st-3rd quartile) across hospitals. Summary statistics by hospital and region are included in Table A and Table B in S1 Text.

Mortality prediction model

Features from the SAPS II risk scoring model [19] from the first 24 hours of the patient-stay starting from ICU admission were extracted and are summarized in Table C in S1 Text. These include 12 physiological measurements (vitals and labs), age, and an indicator for whether the stay was for an elective surgery. For features with multiple measurements, their worst values determined using the SAPS II scoring sheet (Table 3 in Le Gall et al [19]) are extracted. For example, for Glasgow Coma Score, we take the minimum value among the measurements. As previously employed for mortality prediction [27], we use logistic regression with ℓ_2 regularization using the implementation in scikit-learn v0.22.2 package with default hyperparameters [31]. Missing values in features are imputed with mean values computed across the corresponding columns of the full dataset. We experimented with other imputation methods as well, specifically imputation with mean or median across the train datasets and single imputation with a decision tree [32], however, the conclusions did not change. Features are then standardized to zero mean and unit variance using statistics from the train datasets. As sample size used in training models can affect generalizability, we control for this factor by fixing the number of samples used for training and testing. We use 1631 (or 5000) samples from each hospital (or region) while training and testing models. For model development, each dataset (for a hospital or region) is randomly split with the training set comprising 90% of samples and the validation set comprising the remaining 10%. The test set comprises all samples from the hospital (or region) different from the one included in the training set.

Statistical analysis

Performance metrics. Discrimination ability of the models is assessed using area under the receiver operating characteristic curve (AUC). For binary outcomes, calibration slope (CS) is computed as the slope of the regression fit between true outcomes and logits of the predicted mortality, with a logit link function. A perfectly calibrated model has a CS of 1. A value lower than 1 indicates that the risk estimates are extreme, i.e. overestimation for high risk patients and underestimation for low risk patients, and suggests overfitting of the model [33]. Hence, a value close to 1 is desirable. In addition to reporting AUC and CS computed on the test sets, we also report how much the metrics differ from their values computed on the validation set. This difference, known as the generalization gap, provides a quantitative measure of generalization performance (e.g. Jiang et al [34]). If this difference is high, i.e. the test metrics are worse than the train metrics, the model is said to lack generalizability. The allowable difference between test and train depends on the application context. Studies typically report confidence intervals around the difference and/or the percentage change relative to the train performance [35]. To measure fairness of model predictions, we use the racial/ethnic attributes to form two groups–African American, Hispanic, and Asian as one group and the rest as another. These will be referred to as *minority* and *majority* groups. Note that the racial/ethnic attributes are used only for the purposes of fairness analysis, these are not part of the model building process. We acknowledge that aggregating multiple groups does not represent an exposition of which groups are advantaged in the models. Based on the available dataset sizes, this approach serves to illustrate differences that would ideally be unpacked in detail in future work addressing

such issues. Disparity in false negative rates (DisparityFNR), and disparity in calibration slope (DisparityCS) are computed as the difference between the respective metric's value for the minority and the majority group. Differences in these two metrics have been employed in recent studies for bias analysis [16,17]. FNR quantifies the rate at which patients with the observed outcome of death were misclassified. Thus, a high FNR for the score may lead to an increase in undertreatment, and high DisparityFNR (in absolute value) highlights large differences in such undertreatment across groups. For the prediction threshold for FNR we use the mortality rate at the *test* hospital (assuming it is known beforehand). This threshold can be chosen in a more principled way, for example, based on decision-curve analysis [18], which will depend on the application context. We further acknowledge that there are myriad ways to define fairness that will depend on the context of the risk prediction's use and inputs from stakeholders [36].

Dataset differences. To address our secondary objective of studying external validity-specific performance changes, we test for dataset shifts across hospitals and geographies (i.e. whether the distributions of two datasets differ), and use causal graph discovery to explore the reasons for these differences. Dataset shifts are measured using squared maximum mean discrepancy (MMD^2) [37]. We perform the two-sample tests under the null hypothesis that the distributions are the same and threshold the resulting p-values at the significance level of 0.05. Details of the MMD^2 metric and the hypothesis test are included in Method A in S1 Text. To explain the shifts we leverage the recently introduced framework of Joint Causal Inference [38] which allows constructing a single graphical representation of how variables relate to each other, in the form of a causal graph. We use the Fast Causal Inference (FCI) algorithm [39] for constructing the causal graph as it is methodologically well-developed and requires fewer assumptions on the data generating process as it allows for the presence of unobserved variables affecting the observed variables in the data (i.e. unobserved confounders). We also include race as an indicator in the datasets, before running FCI, to study the causes of unfairness with respect to the race variable. Details of the causal modeling are included in Method B in S1 Text. Note that we use causal graphs as a compact representation of (conditional) independencies in the datasets. These are not meant to make statements about the causal effect of treatments on physiological variables, for which randomized controlled trials and other methods may be used to gather better evidence.

The metrics of interest–AUC, CS, DisparityFNR, DisparityCS, p-value, and MMD^2 –are averaged over 100 random subsamples of the datasets (i.e. resampling without replacement). While aggregating across hospitals (or regions), we report the median, 1st, and 3rd quartiles across all train-test set pairs which includes the 100 random subsamples in each pair.

Results

Fig 1 demonstrates the highly varied external validity of models across hospitals based on AUC, CS, generalization gap in AUC and in CS, DisparityFNR, and DisparityCS. Across all train-test hospital pairs, median AUC is 0.801 with 1st-3rd quartile range (IQR) as 0.778 to 0.832, and CS is 0.853 (IQR 0.725 to 0.983). AUCs are lower than the typical values for mortality risk prediction models of around 0.86 [19,22], although AUCs in the same range (around 0.8) have been observed in other studies (albeit in different populations) and were considered acceptable [40–42]. CS of around 0.8, as observed in our case, is considered to indicate overfitting [7]. Transferring a model trained on hospital ID 73, which is the hospital with the most samples, to other hospitals results in a median gap in AUC of -0.087 (IQR -0.134 to -0.046) and a median gap in CS of -0.312 (IQR -0.502 to -0.128). In aggregate, we observe a decline in the performance on the test hospitals relative to that on the train hospitals (Fig 1A). Across all

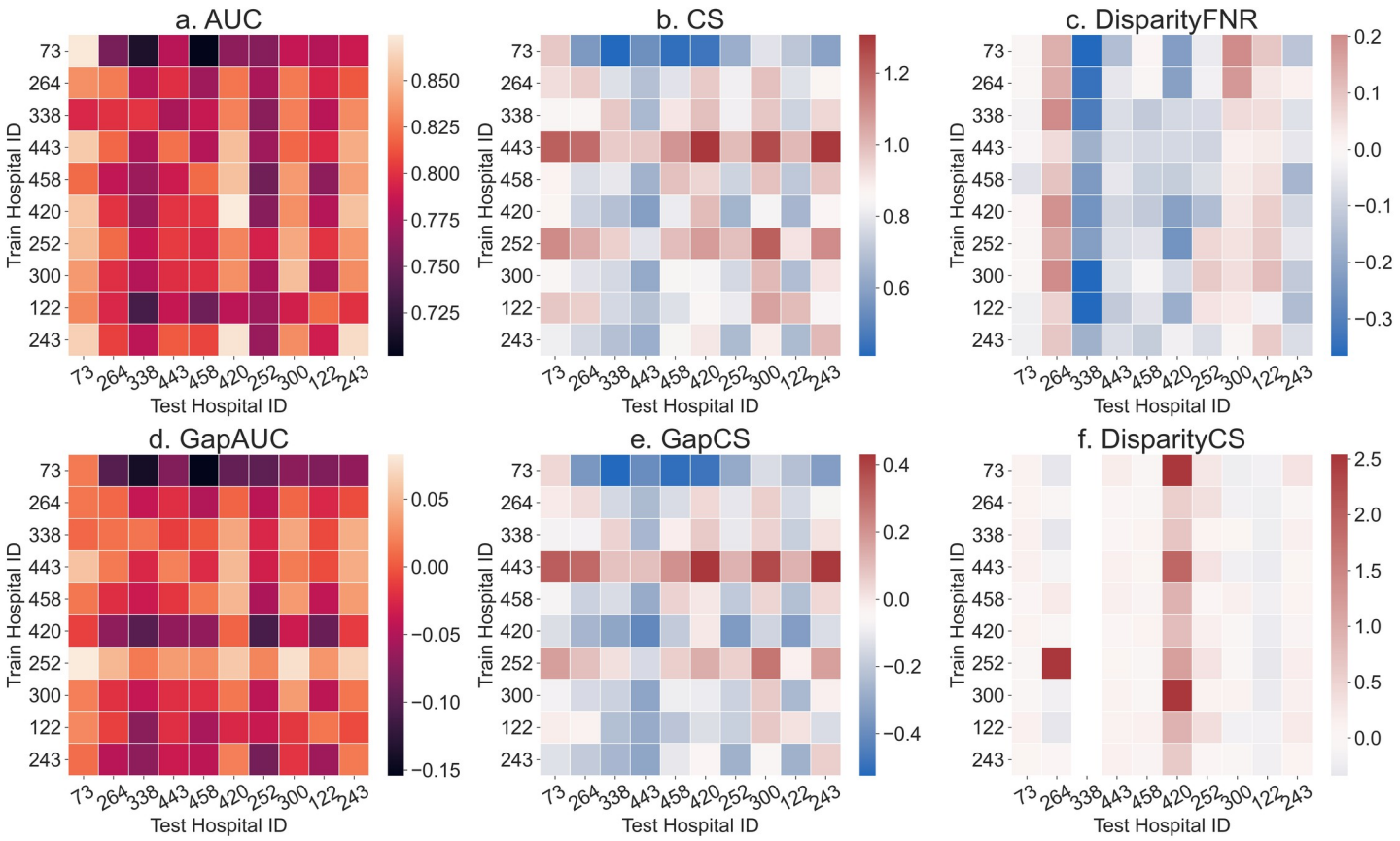

Fig 1. Generalization of performance metrics across individual hospitals. Results of transferring models across top 10 hospitals by number of stays. Models are trained and tested on a fixed number of samples (1631, the least in any of the 10 hospitals) from each hospital. Results are averaged over 100 random subsamples for each of the 10×10 train-test hospital pairs. All 6 metrics show large variability when transferring models across hospitals. Abbreviations: AUC, area under ROC curve; CS, calibration slope; FNR, false negative rate.

https://doi.org/10.1371/journal.pdig.0000023.g001

train-test hospital pairs, the median generalization gap in AUC is -0.018 (IQR -0.065 to 0.032) and the median generalization gap in CS is -0.074 (IQR -0.279 to 0.121). Fig 1B shows that the majority of models have CS of less than 1, indicating consistent miscalibration of mortality risk at test hospitals. This conforms with the typical observation of good discriminative power but poor calibration of SAPS II models [40,42–44]. The median values of AUC and CS are negative, indicating that both of them decrease in majority of the cases upon transfer. For comparison, the generalization gap in AUC for the SAPS II score in the original study by Le Gall et al [19] was -0.02 (AUC decreased to 0.86 in validation from 0.88 in training data), which is the same as the median gap here. Thus, for more than half of the hospital pairs the AUC drop is worse than the acceptable amount found in the original SAPS II study. Percentage changes in AUC and CS from train to test set, reported in Table D in S1 Text also indicate substantial drop in performance (in the range of -2.5% to -31.5% in AUC and -15.9% to -45.4% in CS). In some cases, for example for hospital ID 252, we observe an improvement in AUC (fourth row from bottom, Fig 1D). With regard to fairness metrics, DisparityFNR (absolute value) has median 0.093 (IQR 0.046 to 0.168), i.e. false negative rates across the racial groups differ by 4.6% to 16.8%. DisparityCS, i.e. the absolute value of difference in calibration across racial groups, is large as well (median 0.159; IQR 0.076 to 0.293). Considering that the ideal DisparityCS is 0, when CS is 1 for both the groups, the observed DisparityCS of 0.159 is large. There

are both positive and negative values in the disparity metrics (Fig 1C and 1F), i.e. models are unfair to the minority groups for some pairs and vice versa for others. Note that disparity metrics for hospital ID 338 are considerably different from others (third column in Fig 1C and 1F) due to the skewed race distribution with only 74 (3.2%) samples from the minority groups (Table A in S1 Text). We observe that the variation across hospitals in fairness metrics (DisparityFNR, and DisparityCS) is not captured by the variation in discrimination and calibration metrics (AUC and CS). Thus, fairness properties of the models are not elucidated by the standard metrics and should be audited separately.

Next, given concerns about the development of machine learning models in a limited set of geographies [11], we pool hospitals by geographic region, and validate models trained in one region and tested on another (Fig 2). Performance in terms of AUC and CS across regions improves as a result of pooling hospital data. Overall, AUC varies in a small range (median 0.804; IQR 0.795 to 0.813) as does CS (median 0.968; IQR 0.904 to 1.018). The same can be observed through generalization gaps in AUC and CS which are smaller–median generalization gap in AUC is -0.001 (IQR -0.017 to 0.016) and median generalization gap in CS is -0.008 (IQR -0.081 to 0.075). However, such pooling does not alleviate fairness metric disparities. DisparityFNR (absolute value) has a median value of 0.040 (IQR 0.018 to 0.074). This translates to, for example, a disparity between minority and majority groups of 6.36% (95% CI

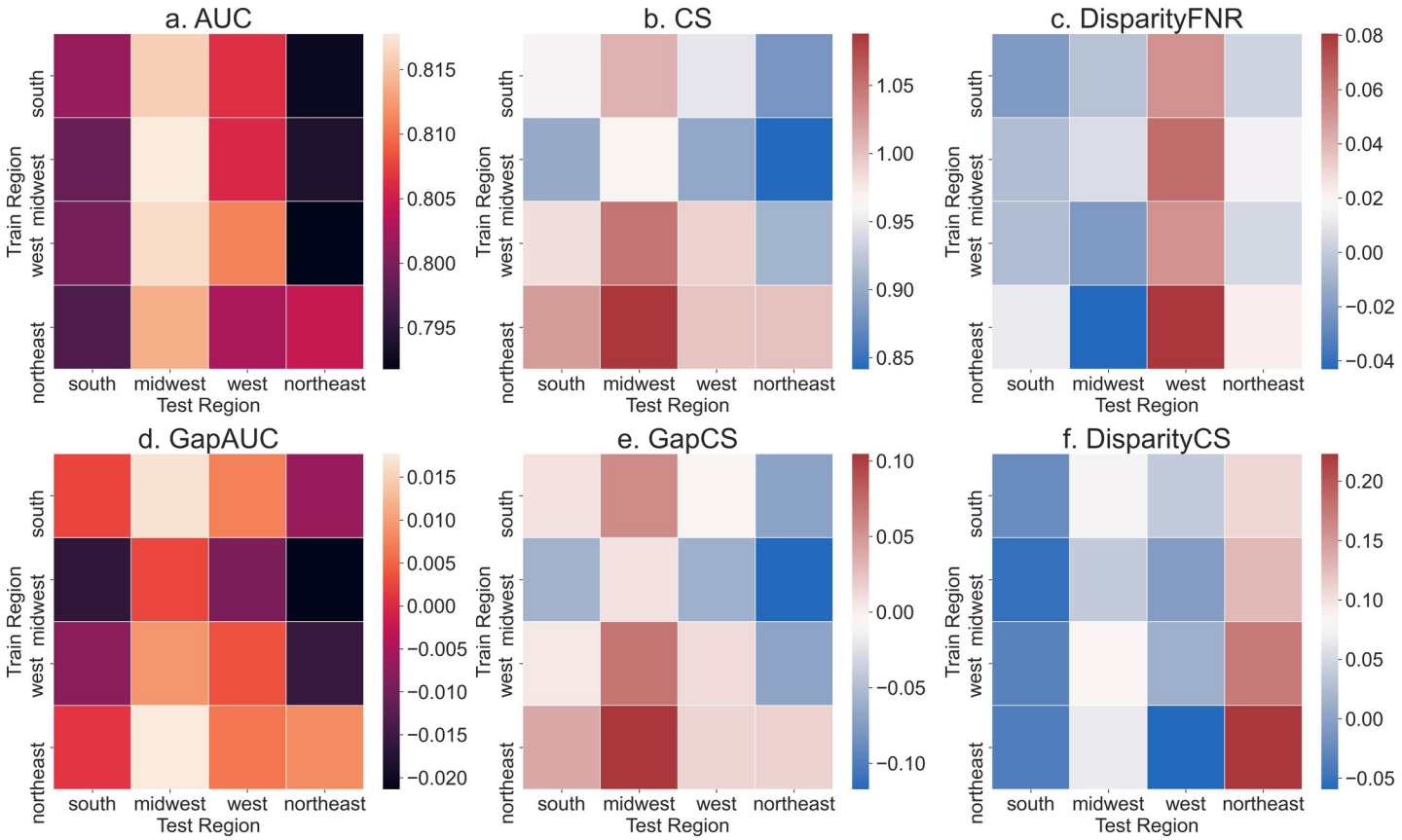

Fig 2. Generalization of performance metrics across US geographic regions. Results of transferring models after pooling hospitals into 4 regions (northeast, south, midwest, west). Models are trained and tested on 5000 samples from each region. Results are averaged over 100 random subsamples for each of the 4×4 train-test hospital pairs. DisparityFNR and DisparityCS show large variability when transferring models across regions. Abbreviations: AUC, area under ROC curve; CS, calibration slope; FNR, false negative rate.

https://doi.org/10.1371/journal.pdig.0000023.g002

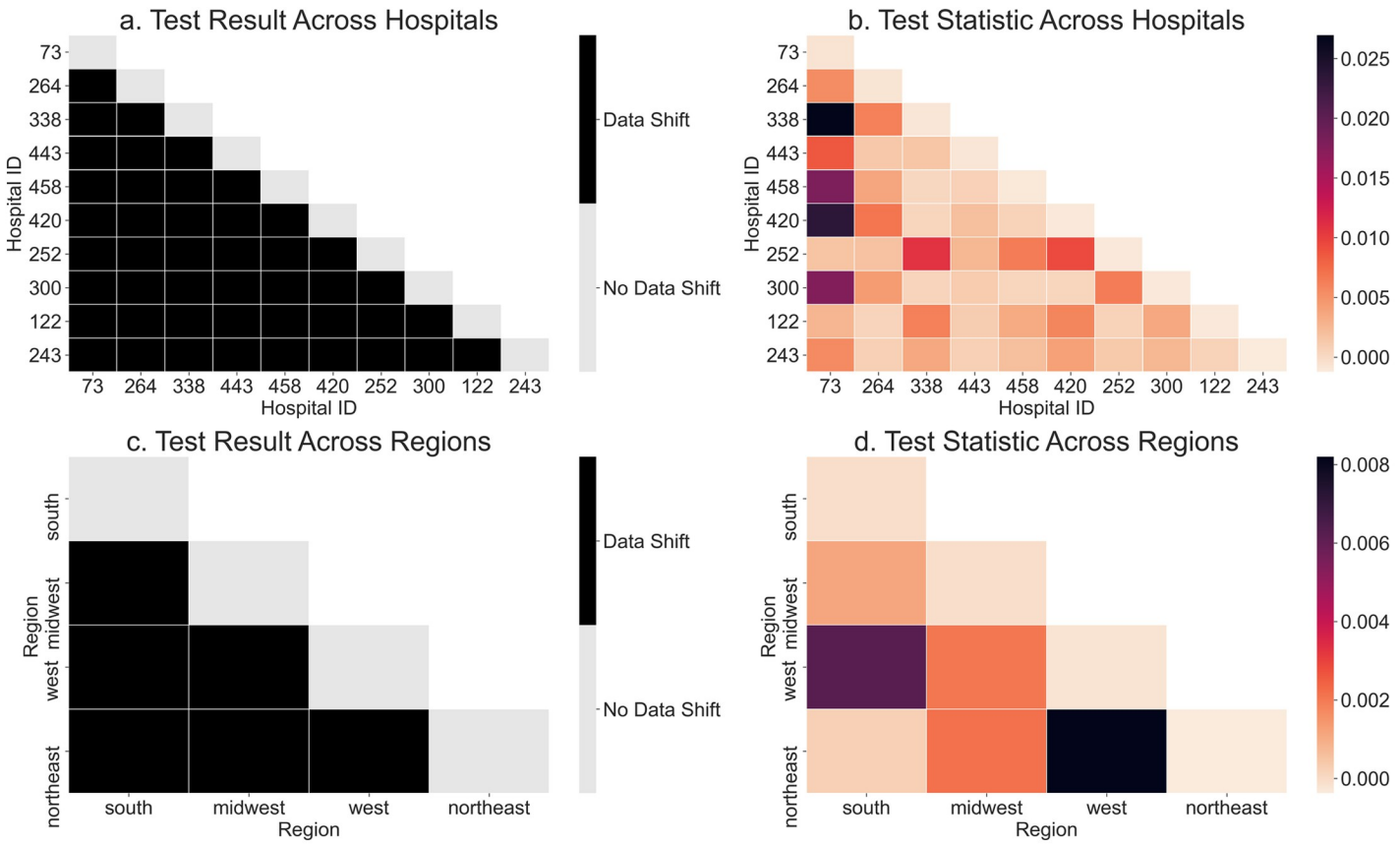

Fig 3. Statistical tests for dataset shifts. Results for two-sample tests with and without pooling of hospitals by region. Test results are plotted in (a,c) and test statistics are plotted in (b,d) to examine the test results in more detail. Since the order of hospitals considered in the two-sample test does not change the test statistic, we plot only the lower halves of the matrices. Results are averaged over 100 random subsamples. Feature distribution changes across all hospital and region pairs.

https://doi.org/10.1371/journal.pdig.0000023.g003

-7.66% to 17.42%) and 8.06% (95% CI -3.81% to 19.74%) when transferring models from Midwest to West and Northeast to West respectively (though CIs are large, both values are greater than 0%; one-sided one-sample t-test, $p = 10^{-5}$). DisparityCS (absolute value) is still high with a median value of 0.104 (IQR 0.050 to 0.167). For example, transferring models from South to Northeast (the region with the least minority population size) has a high DisparityCS (median 0.108; IQR -0.015 to 0.216). Percentage change in the test set metrics relative to the train set is reported in Table E in S1 Text which shows significant changes in DisparityFNR (ranging from -33% to 65%) and DisparityCS (ranging from -52% to 67%). Apart from geography, differences across hospitals can also be due to differences in patient load and available resources. In Figure B in S1 Text we include results for more fine-grained pooling of hospitals based on their number of beds, teaching status, and region where we again find consistent lack of generalizability in fairness metrics.

To investigate reasons for these performance differences across hospitals and regions, we first consider whether the corresponding datasets differ systematically. Fig 3 shows results from statistical tests for dataset shifts across hospitals and regions. Shifts across all pairs of hospitals are significant. Some hospitals are considerably different from others like hospital ID 73 in the first column of Fig 3B, which has significantly lower mortality rate than the other hospitals (Table A in S1 Text).

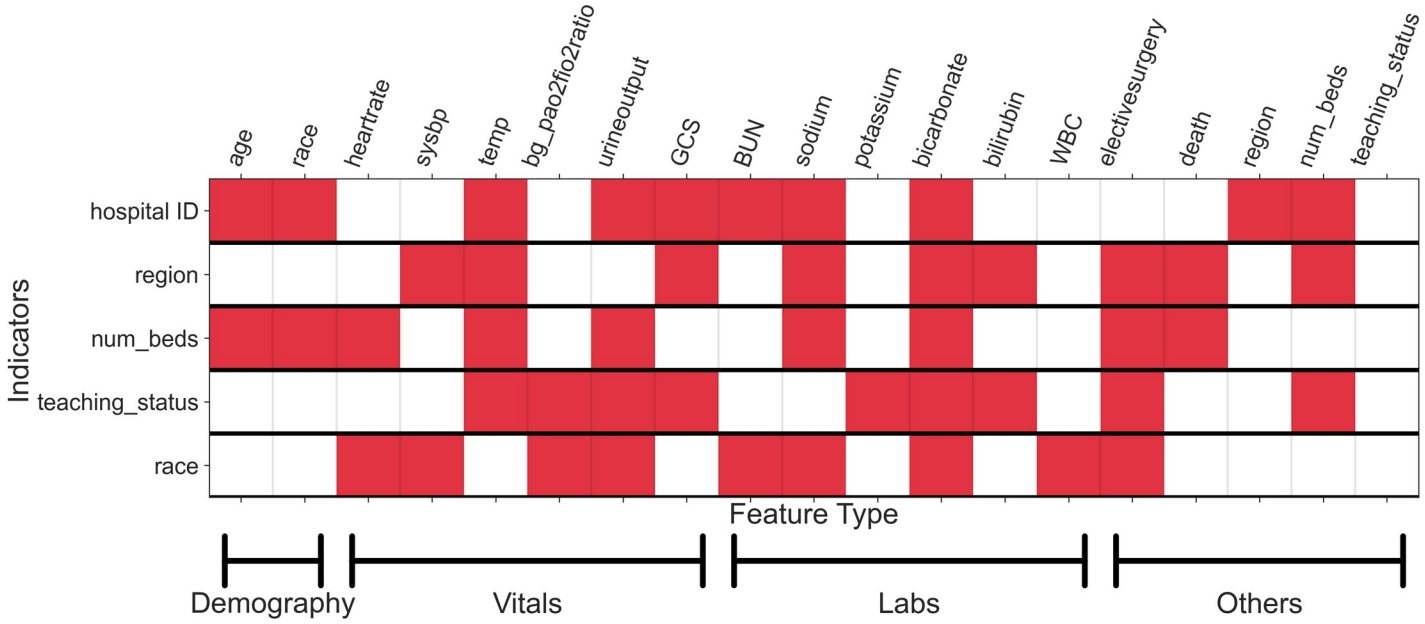

Fig 4. Shifts in variable distributions due to hospital, region, and other factors based on mortality causal graph. Each row represents (in red) the features which explain the shifts across each of the indicators labeling the row, i.e. the features with an edge from the indicator in the causal graph. For instance, shift across the hospital ID indicator (first row) is explained by shifts in the distributions of age, race, temp (or temperature), urine output, and so on. We observe that shifts are explained by changes in a few variables which are common across indicators. Full forms of the abbreviated feature names are added in Table C in S1 Text.

https://doi.org/10.1371/journal.pdig.0000023.g004

Finally, we study explanations for the observed shifts in Fig 3A via individual features. The discovered causal graph is included in Figure C in S1 Text which shows the estimated causal relationships among clinical variables, and which of these variables shift in distribution based on hospital, geography and other factors (i.e. which variables have a direct arrow from the indicators like hospital or region). Fig 4 summarizes the shifts from the causal graph. From Fig 4, we note that the distribution of all fourteen features and the outcome are affected either directly or indirectly by the hospital indicator. However, restricting to direct effects of hospital (first row in Fig 4), we observe that shifts are explained by few of the features–demography (age and race), vitals (3 out of 6), and labs (3 out of 6). Not all vitals and labs change directly as a result of a change in hospitals; changes in 3 of the vitals and 3 of the labs are mediated through changes in other features. We attempt to further understand these changes across hospitals by including hospital-level contextual information, namely, their region, size (number of beds), and teaching status. We observe that there exists common features that explain shifts across the three attributes (different rows in Fig 4). Thus, some of the variation across hospitals is explained through its region, size, and teaching status. But, notably, the three attributes do not explain all variation among hospitals and more contextual information is required. For the fairness analysis, we observe the direct effects of the race variable (last row in Fig 4). There are direct effects from the race variable to most vitals (4 out of 6), labs (4 out of 6), and indicator for elective surgery. These direct effects support past observations made on racial disparities, e.g. in access to specialized care (race → elective surgery) [45] and in blood pressure measurements (race → sysbp or systolic blood pressure) [46]. Out of the 9 features that are directly affected by the race variable, 4 also vary across the hospitals. This suggests that the distribution of clinical variables for the racial groups differs as we go from one hospital to another. As a result, the feature-outcome relationships learnt in one hospital will not be suitable in another,

leading to the observed differences in model performance (as quantified by the fairness metrics) upon model transfer.

Discussion

We retrospectively evaluate generalizability of mortality risk prediction models using data from 179 hospitals in the eICU dataset. In addition to commonly-used metrics for predictive accuracy and calibration, we assessed generalization in terms of algorithmic fairness metrics. To interpret results, we investigated shifts in the distribution of variables of different types across hospitals and geographic regions, leading to changes in the metrics. Findings highlight that recommended measures for checking generalizability are needed, and evaluation guidelines should explicitly call for the assessment of model performance by sub-groups.

Generalizability of a risk prediction model is an important criteria to establish reliable and safe use of the model even under care settings different from the development cohort. A number of studies have reported lack of generalizability of risk prediction models across settings such as different countries [47–50], hospitals [27,51–54], or time periods [55–57]. For instance, Austin et al [58] study the validity of mortality risk prediction models across geographies in terms of discrimination and calibration measures. They find moderate generalizability, however, the hospitals considered belonged to a single province in Canada. More importantly, these works did not investigate generalizability for different groups in the patient cohorts. Results in Fig 1 show that the fairness characteristics of the models can vary substantially across hospitals. Prior work investigating algorithmic fairness metrics in a clinical readmission task [4] did not investigate changes in the metric when models are transferred across care settings. A recent study of a mortality prediction model showed good performance across 3 hospitals (1 academic and 2 community-based) as well as good performance for subgroups *within* a hospital [22]. However, the change in performance for the subgroups *across* hospitals was not explored.

One strategy to tackle the lack of generalizability is to pool multiple hospital databases to potentially increase diversity of the data used in modeling [59]. However, available databases may not faithfully represent the intended populations for the models even after pooling. A recent study [11] found that US-based patient cohorts used to train machine learning models for image diagnosis were concentrated in only three states. However, the effect of using geographically-similar data on generalizability and fairness had not been studied previously. Results in Fig 2 suggest that pooling data from similar geographies may not help mitigate differences in model performance when transferred to other geographies. This finding adds more weight to the concerns raised about possible performance drop when transferring models from data-rich settings to low-resource settings [60].

Prior work has postulated multiple reasons for lack of generalizability [61] including population differences and ICU admission policy changes [47]. In Davis et al [55], reasons including case mix, event rate, and outcome-feature association are assessed. However, specific variables which shift across clinical datasets have not been examined. Through an analysis of the underlying causal graph summarized in Fig 4, we identify specific features that explain the changes in data distributions across hospitals. Demographics (age and race) differ across hospitals which is aligned with population difference being the common reason cited for lack of generalizability [47,55]. We find that vitals and labs differ as well but only a few change as a direct consequence of changes in the hospital setting. Often, the changes are mediated via a small number of specific vitals and labs. For understanding the root causes of the shifts, causal graph analysis helps to narrow down candidate features to analyse further. Significant differences found across ICUs in feature distributions and model performance calls for a systematic

approach to transferring models. Understanding the reasons for the lack of generalizability is the first step to deciding whether to transfer a model, re-training it for better transfer, or developing methods which can improve transfer of models across environments [62]. Factors affecting generalizability can potentially be due to variations in care practices or may represent spurious correlations learned by the model. These findings reinforce calls to better catalog clinical measurement practices [63] and continuously monitor models for possible generalizability challenges [61].

Beyond the descriptive analyses of dataset shifts with causal graphs, in Figure A in S1 Text, we examine whether the shifts can *predict* lack of generalizability. We plot the generalization gap in AUC and CS against the amount of shift, as measured in MMD^2, and report the correlation coefficient. Although we find only low correlation, this suggests a need to develop better methods of quantifying dataset shifts that can predict future model performance. One such metric derived from a model trained to discriminate between training and test samples was found to explain the test performance well (focusing on environments that primarily differ by case-mix, without attention to specific clinical or demographic shifts) [64]. We hope that current work motivates development and evaluation of such metrics on larger and more diverse populations and datasets. Recent work has also proposed methodological advances to ensure transferability of models across settings, for example, by pre-training on large datasets from related machine learning tasks [65,66] and with the help of causal knowledge about shifts [67–69]. Better metrics for dataset shift can help practitioners decide whether to transfer a model to a new setting based on how large the shift is between hospitals, for example.

Importantly, findings here showed that the race variable often mediated shifts in clinical variables. Reasons for this must be disentangled. As race is often a proxy variable for structural social processes such as racism which can manifest both through different health risk factors as well as different care (differences in health care received by patients' racial group are well-documented) [70–73], shifts across hospitals cannot be mitigated simply by population stratification or algorithmic fairness metrics alone. Indeed, better provenance of the process by which data is generated will be critical in order to disentangle the source of dataset differences (for example, if clinical practices or environmental and social factors are giving rise to different healthcare measures and outcomes). Following guidelines developed for documenting datasets [74] and models [75] in the machine learning community, similar guidelines should be established for models in healthcare as well [10]. An example is the proposal for reporting subgroup-level performances in MI-CLAIM checklist [76]. Further, the datasets have a skewed proportion of the minority population (as low as 3.2% for hospital ID 338, Table A in S1 Text). This may negatively impact model performance for minority groups since the number of training samples might be insufficient to learn group-specific outcome characteristics, which is a recognized source of bias termed as representation bias [77]. To reduce the impact of skewed data, algorithmic fairness literature proposes strategies such as reweighting data points for different groups based on their proportions or classification error [78–80]. Efficacy of such strategies to decrease the disparity metrics reported in our results can be investigated in future work. In sum, our findings demonstrate that data provenance, as described above, is needed in addition to applying algorithmic fairness metrics alone, to understand the source of differences in healthcare metrics and outcomes (e.g. clinical practice versus other health determinants), and assess potential generalizability of models.

Limitations

The main limitation of this study is that the results are reported for a collection of ICUs within the same electronic ICU program by a single provider. This collection captures only a part of

the diversity in care environments that a mortality prediction model might be deployed in. Further, our investigation is limited to models constructed using the SAPS II feature set containing 14 hand-crafted features for the mortality prediction task. Though we employ widely-used methods, our analysis is limited by the specific methods used for computing dataset differences, building predictive models and causal graphs. For analyzing reasons for dataset shifts, we could only investigate explanations based on the 14 features along with limited hospital characteristics (such as geographic region, number of beds, and teaching status). Multiple factors are left unrecorded in the eICU database, such as patient load, budget constraints, and socioeconomic environment of the hospital's target population, that may affect the care practices and outcomes recorded in the dataset. We do not investigate dataset shifts in time, which are common [81], as the eICU database includes patient records only for a year. Finally, we do not assess the effect of resampling data points on model performance to address issues of skewed mortality-class distribution or minority-majority group distribution.

Conclusion

External evaluation of predictive models is important to ensure their responsible deployment in different care settings. Recommended metrics for performing such evaluation focus primarily on assessing predictive performance of the models while ignoring their potential impact on health equity. Using a large, publicly-available dataset of ICU stays from multiple hospital centers across the US, we show that models vary considerably in terms of their discriminative accuracy and calibration when validated across hospitals. Fairness of models, quantified using their differential performance on racial groups, is found to be lacking as well. Furthermore, fairness metrics continue to be poor when validating models across US geographies and hospital types. Importantly, the pattern of out-of-sample variation in the fairness metrics is not the same as that in the accuracy and calibration metrics. Thus, the standard checks do not give a comprehensive view of model performance on external datasets. This motivates the need to include fairness checks during external evaluation. While examining reasons for the lack of generalizability, we find that population demographics and clinical variables differ in their distribution across hospitals, and the race variable mediates some variation in clinical variables. Documentation of how data is generated within a hospital where a model is developed specific to sub-groups, along with development of metrics for dataset shift will be critical to anticipate where prediction models can be transferred in a trustworthy manner.

Supporting information

S1 Text. Method A. Difference in Distributions and Statistical Tests. **Method B.** Causal Graph Discovery. **Method C.** Percentage Change. **Table A.** Summary statistics for data after grouping based on hospital ID. **Table B.** Summary statistics for data after grouping based on regions. **Table C.** List of features **Table D.** Percentage change in metrics from the train hospital to the rest. **Table E.** Percentage change in metrics from the train region to the rest. **Table F.** Description of edge types in causal graph. **Figure A.** Generalization gap in AUC and CS versus dataset shift. **Figure B.** Generalization of performance metrics across regions, number of beds, and teachingstatus. **Figure C.** Causal graph discovered with FCI algorithm.
(DOCX)

Author Contributions

Conceptualization: Harvineet Singh, Vishwali Mhasawade, Rumi Chunara.

Data curation: Harvineet Singh, Vishwali Mhasawade.

Formal analysis: Harvineet Singh, Vishwali Mhasawade.

Funding acquisition: Rumi Chunara.

Investigation: Harvineet Singh, Vishwali Mhasawade.

Methodology: Harvineet Singh, Vishwali Mhasawade, Rumi Chunara.

Project administration: Rumi Chunara.

Resources: Rumi Chunara.

Software: Harvineet Singh, Vishwali Mhasawade.

Supervision: Rumi Chunara.

Validation: Harvineet Singh, Vishwali Mhasawade.

Visualization: Harvineet Singh, Vishwali Mhasawade.

Writing – original draft: Harvineet Singh, Vishwali Mhasawade, Rumi Chunara.

Writing – review & editing: Harvineet Singh, Vishwali Mhasawade, Rumi Chunara.

References

1. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Mak. 2015; 35(2):162–9. https://doi.org/10.1177/0272989X14547233 PMID: 25155798
2. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999; 130(6):515–24. https://doi.org/10.7326/0003-4819-130-6-199903160-00016 PMID: 10075620
3. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015; 162(1):W1—W73. https://doi.org/10.7326/M14-0698 PMID: 25560730
4. Chen IY, Szolovits P, Ghassemi M. Can AI Help Reduce Disparities in General Medical and Mental Health Care? AMA J ethics. 2019; 21(2):167–79.
5. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. J Biomed Inform. 2021; 113:103621. https://doi.org/10.1016/j.jbi.2020.103621 PMID: 33220494
6. Beaulieu-Jones BK, Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ Digit Med. 2021; 4(1):1–6. https://doi.org/10.1038/s41746-020-00373-5 PMID: 33398041
7. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med [Internet]. n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9025 PMID: 34031906
8. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. Bmj. 2015;351. https://doi.org/10.1136/bmj.h3868 PMID: 26264962
9. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016; 69:245. https://doi.org/10.1016/j.jclinepi.2015.04.005 PMID: 25981519
10. Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Heal \& Care Informatics [Internet]. 2021; 28(1). Available from: https://informatics.bmj.com/content/28/1/e100289
11. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. JAMA [Internet]. 2020; 324(12):1212–3. Available from: https://doi.org/10.1001/jama.2020.12067 PMID: 32960230

12. Yadlowsky S, Hayward RA, Sussman JB, McClelland RL, Min Y-I, Basu S. Clinical implications of revised pooled cohort equations for estimating atherosclerotic cardiovascular disease risk. Ann Intern Med. 2018; 169(1):20–9. https://doi.org/10.7326/M17-3011 PMID: 29868850

13. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat Med. 2021; 27(1):136–40. https://doi.org/10.1038/s41591-020-01192-7 PMID: 33442014

14. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science (80-). 2019; 366(6464):447–53.

15. Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. Commun ACM. 2020; 63(5):82–9.

16. Seyyed-Kalantari L, Liu G, McDermott M, Chen I, Ghassemi M. Medical imaging algorithms exacerbate biases in underdiagnosis. 2021;

17. Barda N, Yona G, Rothblum GN, Greenland P, Leibowitz M, Balicer R, et al. Addressing bias in prediction models by improving subpopulation calibration. J Am Med Informatics Assoc [Internet]. 2020; Available from: https://doi.org/10.1093/jamia/ocaa283

18. Steyerberg EW, others. Clinical prediction models. Springer; 2019.

19. Le Gall J-R, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. Jama. 1993; 270(24):2957–63. https://doi.org/10.1001/jama.270.24.2957 PMID: 8254858

20. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006; 34(5):1297–310. https://doi.org/10.1097/01.CCM.0000215112.84523.F0 PMID: 16540951

21. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. Crit Care Med. 2006; 34(10):2517–29. https://doi.org/10.1097/01.CCM.0000240233.01711.D9 PMID: 16932234

22. Brajer N, Cozzi B, Gao M, Nichols M, Revoir M, Balu S, et al. Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality of Adults at Time of Admission. JAMA Netw Open [Internet]. 2020; 3(2):e1920733–e1920733. Available from: https://doi.org/10.1001/jamanetworkopen.2019.20733 PMID: 32031645

23. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. JMIR Med informatics [Internet]. 2020 Jul 15; 8(7):e15182–e15182. Available from: https://pubmed.ncbi.nlm.nih.gov/32673244 https://doi.org/10.2196/15182 PMID: 32673244

24. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. In: AMIA Annual Symposium Proceedings. 2017. p. 994. PMID: 29854167

25. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA. 2020;

26. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci data. 2018; 5:180178. https://doi.org/10.1038/sdata.2018.178 PMID: 30204154

27. Johnson AEW, Pollard TJ, Naumann T. Generalizability of predictive models for intensive care unit patients. arXiv Prepr arXiv181202275. 2018;

28. Cosgriff C V, Celi LA, Ko S, Sundaresan T, de la Hoz MÁA, Kaufman AR, et al. Developing well-calibrated illness severity scores for decision support in the critically ill. NPJ Digit Med. 2019; 2(1):1–8. https://doi.org/10.1038/s41746-019-0153-6 PMID: 31428687

29. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007; 147(8):573–7. https://doi.org/10.7326/0003-4819-147-8-200710160-00010 PMID: 17938396

30. Bureau USC. Census regions and divisions of the United States. US Census Bur website [Internet]. 2010; Available from: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in {P}ython. J Mach Learn Res. 2011; 12:2825–30.

32. Little RJA, Rubin DB. Statistical analysis with missing data. Third edit. Statistical Analysis with Missing Data. 2014. 1–381 p. (Wiley series in probability and statistics).

33. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019; 17(1):1–7. https://doi.org/10.1186/s12916-018-1207-3 PMID: 30651111

34. Jiang Y, Krishnan D, Mobahi H, Bengio S. Predicting the Generalization Gap in Deep Networks with Margin Distributions. In: International Conference on Learning Representations [Internet]. 2019. Available from: https://openreview.net/forum?id=HJlQfnCqKX

35. Wessler BS, Ruthazer R, Udelson JE, Gheorghiade M, Zannad F, Maggioni A, et al. Regional validation and recalibration of clinical predictive models for patients with acute heart failure. J Am Heart Assoc. 2017; 6(11):e006121. https://doi.org/10.1161/JAHA.117.006121 PMID: 29151026

36. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med. 2020; 3(1):1–8. https://doi.org/10.1038/s41746-020-0304-9 PMID: 32821854

37. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. J Mach Learn Res. 2012; 13(Mar):723–73.

38. Mooij JM, Magliacane S, Claassen T. Joint Causal Inference from Multiple Contexts. J Mach Learn Res [Internet]. 2020; 21(99):1–108. Available from: http://jmlr.org/papers/v21/17-123.html

39. Spirtes P, Glymour CN, Scheines R, Heckerman D. Causation, prediction, and search. MIT press; 2000.

40. Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G, et al. The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Intensive Care Med. 1996; 22(12):1368–78. https://doi.org/10.1007/BF01709553 PMID: 8986488

41. Katsounas A, Kamacharova I, Tyczynski B, Eggebrecht H, Erbel R, Canbay A, et al. The predictive performance of the SAPS II and SAPS 3 scoring systems: A retrospective analysis. J Crit Care. 2016; 33:180–5. https://doi.org/10.1016/j.jcrc.2016.01.013 PMID: 26883275

42. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med. 2006; 34(5):1378–88. https://doi.org/10.1097/01.CCM.0000216702.94014.75 PMID: 16557153

43. Beck DH, Smith GB, Pappachan J V, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med. 2003; 29 (2):249–56. https://doi.org/10.1007/s00134-002-1607-9 PMID: 12536271

44. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. Intensive Care Med. 2012; 38(1):40–6. https://doi.org/10.1007/s00134-011-2390-2 PMID: 22042520

45. Johnson A. Understanding Why Black Patients Have Worse Coronary Heart Disease Outcomes: Does the Answer Lie in Knowing Where Patients Seek Care? Am Heart Assoc; 2019. https://doi.org/10.1161/JAHA.119.014706 PMID: 31787054

46. Baldo MP, Cunha RS, Ribeiro ALP, Lotufo PA, Chor D, Barreto SM, et al. Racial differences in arterial stiffness are mainly determined by blood pressure levels: results from the ELSA-Brasil study. J Am Heart Assoc. 2017; 6(6):e005477. https://doi.org/10.1161/JAHA.117.005477 PMID: 28637779

47. Pappachan J V, Millar B, Bennett ED, Smith GB. Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system. Chest. 1999; 115(3):802–10. https://doi.org/10.1378/chest.115.3.802 PMID: 10084495

48. Wang X, Liang G, Zhang Y, Blanton H, Bessinger Z, Jacobs N. Inconsistent Performance of Deep Learning Models on Mammogram Classification. J Am Coll Radiol. 2020; https://doi.org/10.1016/j.jacr.2020.01.006 PMID: 32068005

49. Gola D, Erdmann J, Läll K, Mägi R, Müller-Myhsok B, Schunkert H, et al. Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease. Circ Genomic Precis Med. 2020; https://doi.org/10.1161/CIRCGEN.120.002932 PMID: 33170024

50. Reps JM, Kim C, Williams RD, Markus AF, Yang C, Duarte-Salles T, et al. Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study. JMIR Med Inf [Internet]. 2021 Apr; 9(4):e21547. Available from: http://www.ncbi.nlm.nih.gov/pubmed/33661754 https://doi.org/10.2196/21547 PMID: 33661754

51. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Informatics Assoc. 2014; 21(4):699–706. https://doi.org/10.1136/amiajnl-2013-002162 PMID: 24481703

52. Gong JJ, Sundt TM, Rawn JD, Guttag J V. Instance Weighting for Patient-Specific Risk Stratification Models. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: Association for Computing Machinery; 2015. p. 369–378. (KDD '15). Available from: https://doi.org/10.1145/2783258.2783397

53. Curth A, Thoral P, van den Wildenberg W, Bijlstra P, de Bruin D, Elbers P, et al. Transferring Clinical Prediction Models Across Hospitals and Electronic Health Record Systems. In: Cellier P, Driessens K,

editors. Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing; 2020. p. 605–21.

54. Desautels T, Calvert J, Hoffman J, Mao Q, Jay M, Fletcher G, et al. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. Biomed Inform Insights. 2017; 9:1178222617712994. https://doi.org/10.1177/1178222617712994 PMID: 28638239

55. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Informatics Assoc. 2017; 24(6):1052–61. https://doi.org/10.1093/jamia/ocx030 PMID: 28379439

56. Granholm A, Møller MH, Krag M, Perner A, Hjortrup PB. Predictive performance of the simplified acute physiology score (SAPS) II and the initial sequential organ failure assessment (SOFA) score in acutely ill intensive care patients: post-hoc analyses of the SUP-ICU inception cohort study. PLoS One. 2016; 11(12):e0168948. https://doi.org/10.1371/journal.pone.0168948 PMID: 28006826

57. Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks. In: Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, et al., editors. Proceedings of the 4th Machine Learning for Healthcare Conference [Internet]. PMLR; 2019. p. 381–405. (Proceedings of Machine Learning Research; vol. 106). Available from: https://proceedings.mlr.press/v106/nestor19a.html

58. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. J Clin Epidemiol. 2016; 79:76–85. https://doi.org/10.1016/j.jclinepi.2016.05.007 PMID: 27262237

59. Roth HR, Chang K, Singh P, Neumark N, Li W, Gupta V, et al. Federated Learning for Breast Density Classification: A Real-World Implementation. In: Albarqouni S, Bakas S, Kamnitsas K, Cardoso MJ, Landman B, Li W, et al., editors. Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning. Cham: Springer International Publishing; 2020. p. 181–91.

60. Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. JAMA Netw Open. 2018; 1 (5):e182658—e182658. https://doi.org/10.1001/jamanetworkopen.2018.2658 PMID: 30646173

61. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The Clinician and Dataset Shift in Artificial Intelligence. N Engl J Med [Internet]. 2021 Jul 14; 385(3):283–6. Available from: https://doi.org/10.1056/NEJMc2104626 PMID: 34260843

62. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Heal. 2020; 2(9):e489—e492. https://doi.org/10.1016/S2589-7500(20)30186-2 PMID: 32864600

63. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ [Internet]. 2018; 361. Available from: https://www.bmj.com/content/361/bmj.k1479 https://doi.org/10.1136/bmj.k1479 PMID: 29712648

64. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015; 68(3):279–89. https://doi.org/10.1016/j.jclinepi.2014.06.018 PMID: 25179855

65. Mustafa B, Loh A, Freyberg J, MacWilliams P, Wilson M, McKinney SM, et al. Supervised Transfer Learning at Scale for Medical Imaging. 2021.

66. Ke A, Ellsworth W, Banerjee O, Ng AY, Rajpurkar P. CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation. 2021.

67. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics [Internet]. 2019; 21(2):345–52. Available from: https://doi.org/10.1093/biostatistics/kxz041

68. Subbaswamy A, Saria S. I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models. 2020.

69. Singh H, Singh R, Mhasawade V, Chunara R. Fairness Violations and Mitigation under Covariate Shift. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency [Internet]. New York, NY, USA: Association for Computing Machinery; 2021. p. 3–13. (FAccT '21). Available from: https://doi.org/10.1145/3442188.3445865

70. Wenneker MB, Epstein AM. Racial Inequalities in the Use of Procedures for Patients With Ischemic Heart Disease in Massachusetts. JAMA [Internet]. 1989; 261(2):253–7. Available from: https://doi.org/10.1001/jama.1989.03420020107039 PMID: 2521191

71. Kjellstrand CM. Age, Sex, and Race Inequality in Renal Transplantation. Arch Intern Med [Internet]. 1988; 148(6):1305–9. Available from: https://doi.org/10.1001/archinte.1988.00380060069016 PMID: 3288159

72. Yergan J, Flood AB, LoGerfo JP, Diehr P. Relationship between patient race and the intensity of hospital services. Med Care. 1987;592–603. https://doi.org/10.1097/00005650-198707000-00003 PMID: 3695664

73. Blendon RJ, Aiken LH, Freeman HE, Corey CR. Access to Medical Care for Black and White Americans: A Matter of Continuing Concern. JAMA [Internet]. 1989; 261(2):278–81. Available from: https://doi.org/10.1001/jama.1989.03420020132045 PMID: 2909026

74. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, et al. Datasheets for Datasets. Commun ACM [Internet]. 2021 Nov; 64(12):86–92. Available from: https://doi.org/10.1145/3458723

75. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. In: FAT. 2019.

76. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020; 26 (9):1320–4. https://doi.org/10.1038/s41591-020-1041-y PMID: 32908275

77. Suresh H, Guttag J. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In: Equity and Access in Algorithms, Mechanisms, and Optimization [Internet]. New York, NY, USA: Association for Computing Machinery; 2021. (EAAMO '21). Available from: https://doi.org/10.1145/3465416.3483305

78. Chen RJ, Chen TY, Lipkova J, Wang JJ, Williamson DFK, Lu MY, et al. Algorithm Fairness in AI for Medicine and Healthcare. 2021. Available from: https://arxiv.org/abs/2110.00603

79. Krasanakis E, Spyromitros-Xioufis E, Papadopoulos S, Kompatsiaris Y. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-Aware Classification. In: Proceedings of the 2018 World Wide Web Conference [Internet]. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee; 2018. p. 853–862. (WWW '18). Available from: https://doi.org/10.3389/fendo.2018.00069 PMID: 29559954

80. Pfohl SR, Zhang H, Xu Y, Foryciarz A, Ghassemi M, Shah NH. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. Sci Rep [Internet]. 2022; 12 (1):3254. Available from: https://doi.org/10.1038/s41598-022-07167-7 PMID: 35228563

81. Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. Gigascience [Internet]. 2020; 9(8). Available from: https://doi.org/10.1093/gigascience/giaa079 PMID: 32729900