Title: Generalizability Challenges of Mortality Risk Prediction Models: A Retrospective Analysis on a Multi-center Database

Short Title: Generalizability Challenges of Mortality Risk Prediction Models

Authors: Harvineet Singh[1], MTech; Vishwali Mhasawade[2], MS; Rumi Chunara[2,3], PhD

[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

Correspondence: Harvineet Singh

Center for Data Science, New York University

60 5th Ave, New York City, NY 10011

Email: hs3673@nyu.edu

Word count: Key points 198

Abstract 420

Body 4761

Keywords: Generalizability, in-hospital mortality prediction, fairness, eICU database, distribution shifts, machine learning, causal discovery

Key Points

Question: Does the sub-group level performance of mortality risk prediction models vary significantly when applied to hospitals or geographies different from the ones in which they are developed? What characteristics of the datasets explain the performance variation?

Findings: In this retrospective cross-sectional study based on a multi-center critical care database, mortality risk prediction models developed in one hospital or geographic region setting exhibited lack of generalizability to different hospitals/regions. Distribution of clinical (vitals, labs and surgery) variables significantly varied across hospitals and regions. Dataset shifts in race and clinical variables due to hospital or geography result in mortality prediction differences according to causal inference results, and the race variable commonly mediated changes in clinical variable shifts.

Meaning: Findings demonstrate evidence that such models can exhibit disparities in performance across racial groups even while performing well in terms of average population-wide metrics. Therefore, assessing subgroup performance differences should be included in model evaluation guidelines. Based on shifts in variables mediated by the race variable, understanding and provenance of data generating processes by population sub-group are needed to identify and mitigate sources of variation and can be used to decide whether to use a risk prediction model in new environments.

Abstract

Importance: Modern predictive models require large amounts of data for training and evaluation which can result in building models that are specific to certain locations, populations in them and clinical practices. Yet, best practices and guidelines for clinical risk prediction models have not yet considered such challenges to generalizability.

Objectives: To investigate changes in measures of predictive discrimination, calibration, and algorithmic fairness when transferring models for predicting in-hospital mortality across ICUs in different populations. Also, to study the reasons for the lack of generalizability in these measures.

Design, Setting, and Participants: In this multi-center cross-sectional study, electronic health records from 179 hospitals across the US with 70,126 hospitalizations were analyzed. Time of data collection ranged from 2014 to 2015.

Main Outcomes and Measures: The main outcome is in-hospital mortality. Generalization gap, defined as difference between model performance metrics across hospitals, is computed for discrimination and calibration metrics, namely area under the receiver operating characteristic curve (AUC) and calibration slope. To assess model performance by race variable, we report differences in false negative rates across groups. Data were also analyzed using a causal discovery algorithm "Fast Causal Inference" (FCI) that infers paths of causal influence while identifying potential influences associated with unmeasured variables.

Results: In-hospital mortality rates differed in the range of 3.9%-9.3% (1st-3rd quartile) across hospitals. When transferring models across hospitals, AUC at the test hospital ranged from 0.777 to 0.832 (1st to 3rd quartile; median 0.801); calibration slope from 0.725 to 0.983 (1st to 3rd quartile; median 0.853); and disparity in false negative rates from 0.046 to 0.168 (1st to 3rd quartile; median 0.092). When transferring models across geographies, AUC ranged from 0.795 to 0.813 (1st to 3rd quartile; median 0.804); calibration slope from 0.904 to 1.018 (1st to 3rd quartile; median 0.968); and disparity in false negative rates from 0.018 to 0.074 (1st to 3rd quartile; median 0.040). Distribution of all variable types (demography,

vitals, and labs) differed significantly across hospitals and regions. Shifts in the race variable distribution and some clinical (vitals, labs and surgery) variables by hospital or region. Race variable also mediates differences in the relationship between clinical variables and mortality, by hospital/region.

Conclusions and Relevance: Group-specific metrics should be assessed during generalizability checks to identify potential harms to the groups. In order to develop methods to improve and guarantee performance of prediction models in new environments for groups and individuals, better understanding and provenance of health processes as well as data generating processes by sub-group are needed to identify and mitigate sources of variation.

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] The statistics literature has developed methods for computing sample size and other best practices for assessing external validity of clinical risk prediction models[7–9] though sub-group and data-shift challenges 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 model 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 predicted probabilities match with 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 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 the study. Recently, prediction models for in-hospital mortality have been prospectively validated for potential use,[37] or in case of sepsis, have even been integrated into the clinical workflow.[75] 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.[22] Being flexible, such approaches might overfit to the patterns in a particular dataset, thus, raising concerns for their generalization to newer environments.[23] We use the eICU dataset[24] as a test bed for our analyses. Past studies have employed the dataset for evaluating mortality prediction models.[25,26] 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.

Methods

Analyses are based on data obtained from the publicly-available eICU Collaborative Research Database,[24] 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 number of unique patient stays across hospitals is maintained.[24] 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.[27]

Data Preprocessing

We follow the feature extraction and exclusion procedures including the exclusion criteria used by Johnson et al[25] 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.[24] 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).[28] 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 10 separate datasets using a hospital identifier for hospital-specific analyses and into 4 separate datasets using a region identifier for region-specific analyses. The outcome label is in-hospital mortality (binary). Summary statistics by hospital and region are included in eTable 1 and 2.

Mortality Prediction Model

Features from the SAPS II risk scoring model[19] from the first 24 hours of patient stay starting from ICU admission were extracted and are summarized in eTable 3. 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,[25] we use logistic regression with l_2 regularization using the implementation in scikit-learn v0.22.2 package with default hyperparameters.[29] Missing values in features are imputed with mean values across the corresponding columns and then standardized to zero mean and unit variance. 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

<u>Performance Metrics:</u> 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 an CS of 1. 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.[30] 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 generalization gap, provides a quantitative measure of generalization performance (e.g. Jiang et al[31]). 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.[32] 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 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.[33]

<u>Dataset Differences:</u> To address our secondary objective of studying external validity 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 reasons for

these differences. Dataset shifts are measured using squared maximum mean discrepancy (MMD^2).[34] 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 eMethod 1. To explain the shifts we leverage the recently introduced framework of Joint Causal Inference[35] 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[36] 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 race. Details of the causal modeling are included in eMethod 2. Note that we use causal graphs as a compact representation of (conditional) independencies in the datasets. These are not meant to make statements about causal effect of treatments on physiological variables, for which randomized controlled trials and other methods may be used for 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.

[Figure 1 here]

Results

Figure 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 typical values for

mortality risk prediction models of around 0.86,[19,37] although AUCs in the same range (around 0.8) have been observed in other studies (albeit in different populations) and were considered acceptable.[38–40] 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 median gap in CS is -0.312 (IQR -0.502 to -0.128). In aggregate, we observe a decline in the performance on test hospitals relative to that on the train hospitals (Figure 1a). Across all train-test hospital pairs, the median generalization gap in AUC is -0.018 (IQR -0.065 to 0.032) and median generalization gap in CS is -0.074 (IQR -0.279 to 0.121). Figure 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.[38,40–42] The median values of AUC and CS are negative, indicating that both 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 eTable 4 also indicate substantial drop in performance (in the range of -2.5% to -31.5%). In some cases, for example for hospital ID 252, we even observe an improvement in AUC (fourth row from bottom, Figure 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. 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 (Figures 1c, 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 Figures 1c, 1f) due to

the skewed race distribution with only 74 (3.2%) samples from minority groups (eTable 1). We observe that variation across hospitals in fairness metrics (DisparityFNR, and DisparityCS) is not captured by variation in the 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.

[Figure 2 here]

Next, given concerns about the development of machine learning models in a limited set of geographies, we pool hospitals by geographic region, and validate models trained in one region and tested on another (Figure 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) and CS in (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, pooling does not alleviate fairness metric disparities. DisparityFNR (absolute value) has median 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 -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 median 0.104; IQR 0.050 to 0.167. For example, transferring models from South to Northeast (region with the least minority population size) has 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 eTable 5 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 eFigure 2 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. Figure 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 Figure 3d, which has significantly lower mortality rate than the other hospitals (eTable 1).

[Figure 3 and 4 here]

Finally, we study explanations for the observed shifts in Figure 3a via individual features. The discovered causal graph is included in eFigure 3. Figure 4 summarizes the shifts in mortality causal graph variable distributions based on hospital, geography, and other factors (i.e. variables which have a direct arrow from the indicators like hospital or region). From the causal graph, we note that 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 Figure 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 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 Figure 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 Figure 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 high-quality care (race → elective surgery)[43] and in blood pressure measurements (race → sysbp).[44] Out of the 9 features that are directly affected by the race variable, 4 also vary across 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 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 should explicitly call for assessment of model performance for 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,[45–48] hospitals,[25,49–52] or time periods.[53–55] For instance, Austin et al[56] study 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 Figure 1 show that 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 shows good performance across 3 hospitals (1 academic and 2 community-based) as well as good performance for subgroups *within* a

hospital.[37] 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 data used in modeling.[57] 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 Figure 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.[58]

Prior work has postulated multiple reasons for lack of generalizability including population differences and ICU admission policy changes.[45] In Davis et al,[53] 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 analysis of the underlying causal graph summarized in Figure 4, we identify specific features that explain 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.[45,53] 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 reasons for lack of generalizability is the first step to deciding whether to transfer a model, re-train it for better transfer or developing methods which can improve transfer of models across environments.[59] Factors affecting generalizability can potentially be due to variations in care practices or represent

spurious correlations learned by the model. These findings reinforce calls to better catalog clinical measurement practices.[60]

Beyond descriptive analyses of dataset shifts with causal graphs, in eFigure 1, 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).[61] 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[62,63] and with the help of causal knowledge about shifts.[64–66] 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 (race differences in health care received by patients are well-documented),[67–70] 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 (i.e. if clinical practices or environmental and social factors are giving rise to different medical conditions). Following the guidelines developed for documenting datasets[71] and models[72] 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.[73] Our findings underscore the need to report the generalizability of such fairness metrics as well.

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,[74] as the eICU database includes patient records only for a year.

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 of 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.

Acknowledgments

Author Contributions: All authors were involved in conceptualizing and designing the study, and in preparing the manuscript. HS and VM performed statistical analysis. RC supervised the study and obtained funding.

Conflict of Interest Disclosures: Authors report no conflict of interest.

Funding/Support: This study was funded by the NSF grant number 1845487.

Role of the Funder/Sponsor: The funder had no role in design, conduct of the study or in preparing the manuscript.

Data/Code Availability: Scripts for data pre-processing are the same as that used by Johnson et al[25] and are available at https://github.com/alistairewj/icu-model-transfer. Scripts for model training and validation are available at https://github.com/ChunaraLab/medshifts.

References

1. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Mak*. 2015;35(2):162-169.
2. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515-524.
3. Moons KGM, Altman DG, Reitsma JB, 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.
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):167-179.
5. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform*. 2021;113:103621.
6. Beaulieu-Jones BK, Yuan W, Brat GA, 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.
7. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. n/a(n/a). doi:https://doi.org/10.1002/sim.9025
8. Seaman S, Pavlou M, Qu C, et al. Estimation of required sample size for external validation of risk models for binary outcomes. Published online 2021.
9. Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245.
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*. 2021;28(1). doi:10.1136/bmjhci-2020-100289
11. Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA*. 2020;324(12):1212-1213. doi:10.1001/jama.2020.12067

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):20-29.

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):136-140.

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):447-453.

15. Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. *Commun ACM*. 2020;63(5):82-89.

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

17. Barda N, Yona G, Rothblum GN, et al. Addressing bias in prediction models by improving subpopulation calibration. *J Am Med Informatics Assoc*. Published online 2020. doi: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):2957-2963.

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):1297-1310.

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):2517-2529.

22. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. In: *AMIA Annual Symposium Proceedings.* Vol 2017. ; 2017:994.

23. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. *JAMA*. Published online 2020.

24. 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.

25. Johnson AEW, Pollard TJ, Naumann T. Generalizability of predictive models for intensive care unit patients. *arXiv Prepr arXiv181202275*. Published online 2018.

26. Cosgriff C V, Celi LA, Ko S, et al. Developing well-calibrated illness severity scores for decision support in the critically ill. *NPJ Digit Med*. 2019;2(1):1-8.

27. 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):573-577.

28. Bureau USC. Census regions and divisions of the United States. *US Census Bur website*. Published online 2010. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in {P}ython. *J Mach Learn Res*. 2011;12:2825-2830.

30. 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.

31. Jiang Y, Krishnan D, Mobahi H, Bengio S. Predicting the Generalization Gap in Deep Networks with Margin Distributions. In: *International Conference on Learning Representations*. ; 2018.

32. Wessler BS, Ruthazer R, Udelson JE, et al. Regional validation and recalibration of clinical predictive models for patients with acute heart failure. *J Am Heart Assoc*. 2017;6(11):e006121.

33. 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.

34. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. *J Mach Learn Res*. 2012;13(Mar):723-773.

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

36. Spirtes P, Glymour CN, Scheines R, Heckerman D. *Causation, Prediction, and Search*. MIT press; 2000.

37. Brajer N, Cozzi B, Gao M, 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*. 2020;3(2):e1920733-e1920733. doi:10.1001/jamanetworkopen.2019.20733

38. Apolone G, Bertolini G, D'Amico R, 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):1368-1378.

39. Katsounas A, Kamacharova I, Tyczynski B, et al. The predictive performance of the SAPS II and SAPS 3 scoring systems: A retrospective analysis. *J Crit Care*. 2016;33:180-185.

40. 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):1378-1388.

41. 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):249-256.

42. 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):40-46.

43. Johnson A. Understanding Why Black Patients Have Worse Coronary Heart Disease Outcomes: Does the Answer Lie in Knowing Where Patients Seek Care? Published

online 2019.

44. Baldo MP, Cunha RS, Ribeiro ALP, 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.

45. 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):802-810.

46. 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*. Published online 2020.

47. Gola D, Erdmann J, Läll K, et al. Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease. *Circ Genomic Precis Med*. Published online 2020.

48. Reps JM, Kim C, Williams RD, et al. Can we trust the prediction model? Demonstrating the importance of external validation by investigating the COVID-19 Vulnerability (C-19) Index across an international network of observational healthcare datasets. *medRxiv*. Published online 2020. doi:10.1101/2020.06.15.20130328

49. 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.

50. 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*. KDD ’15. Association for Computing Machinery; 2015:369–378. doi:10.1145/2783258.2783397

51. Curth A, Thoral P, van den Wildenberg W, et al. Transferring Clinical Prediction Models Across Hospitals and Electronic Health Record Systems. In: Cellier P, Driessens K, eds. *Machine Learning and Knowledge Discovery in Databases.* Springer International Publishing; 2020:605-621.

52. Desautels T, Calvert J, Hoffman J, et al. Using transfer learning for improved

mortality prediction in a data-scarce hospital setting. *Biomed Inform Insights*. 2017;9:1178222617712994.

53. 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):1052-1061.

54. 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.

55. Nestor B, McDermott M, Boag W, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *arXiv Prepr arXiv190800690*. Published online 2019.

56. 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.

57. Roth HR, Chang K, Singh P, et al. Federated Learning for Breast Density Classification: A Real-World Implementation. In: Albarqouni S, Bakas S, Kamnitsas K, et al., eds. *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer International Publishing; 2020:181-191.

58. Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. *JAMA Netw Open*. 2018;1(5):e182658--e182658.

59. 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.

60. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;361. doi:10.1136/bmj.k1479

61. 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):279-289.

62. Mustafa B, Loh A, Freyberg J, et al. Supervised Transfer Learning at Scale for Medical Imaging. Published online 2021.

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

64. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. 2019;21(2):345-352. doi:10.1093/biostatistics/kxz041

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

66. 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*. FAccT '21. Association for Computing Machinery; 2021:3–13. doi:10.1145/3442188.3445865

67. Wenneker MB, Epstein AM. Racial Inequalities in the Use of Procedures for Patients With Ischemic Heart Disease in Massachusetts. *JAMA*. 1989;261(2):253-257. doi:10.1001/jama.1989.03420020107039

68. Kjellstrand CM. Age, Sex, and Race Inequality in Renal Transplantation. *Arch Intern Med*. 1988;148(6):1305-1309. doi:10.1001/archinte.1988.00380060069016

69. Yergan J, Flood AB, LoGerfo JP, Diehr P. Relationship between patient race and the intensity of hospital services. *Med Care*. Published online 1987:592-603.

70. Blendon RJ, Aiken LH, Freeman HE, Corey CR. Access to Medical Care for Black and White Americans: A Matter of Continuing Concern. *JAMA*. 1989;261(2):278-281. doi:10.1001/jama.1989.03420020132045

71. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for Datasets.

72. Mitchell M, Wu S, Zaldivar A, et al. Model Cards for Model Reporting. In: *FAT*. ;

2019.

73. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324.

74. Sáez C, Gutiérrez-Sacristán A, Kohane I, García-Gómez JM, Avillach P. EHRtemporalVariability: delineating temporal data-set shifts in electronic health records. *Gigascience*. 2020;9(8). doi:10.1093/gigascience/giaa079

75. Sendak, M. P., Ratliff, W., Sarro, D., et al. Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. JMIR medical informatics. 2020;8(7), e15182. https://doi.org/10.2196/15182

Figure titles

Figure 1. Generalization of performance metrics across individual hospitals.

Figure 2. Generalization of performance metrics across US geographic regions.

Figure 3. Statistical tests for dataset shifts.

Figure 4. Shifts in mortality causal graph variable distributions due to hospital, region, and other factors.

Figures

Figure 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.

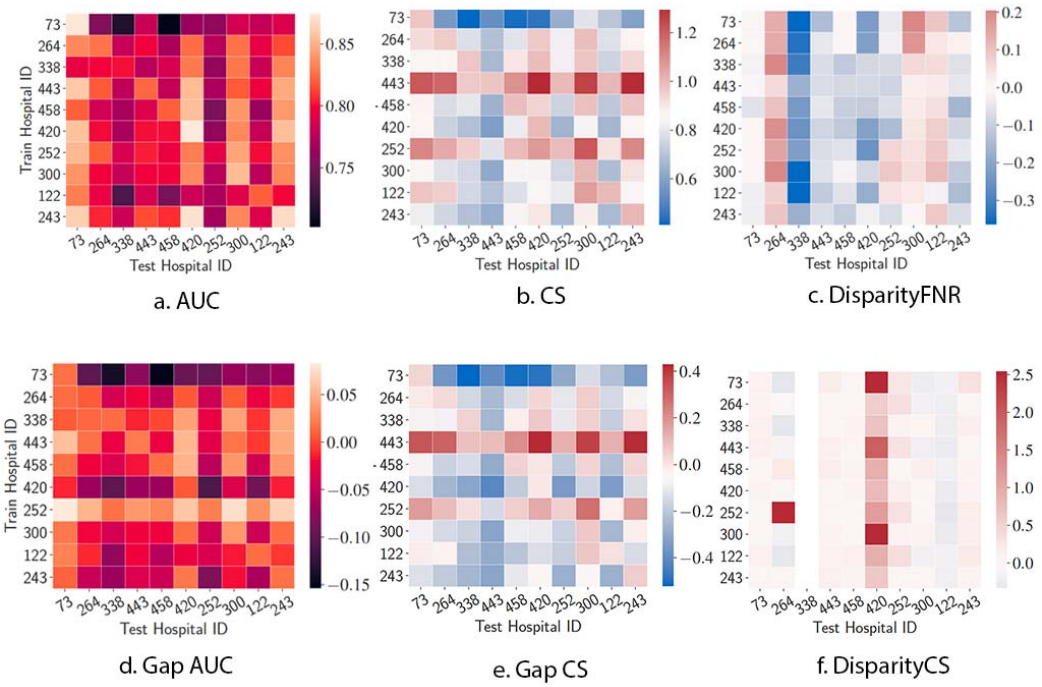

Figure 2. Generalization of performance metrics across US geographic regions. Results of transferring models after pooling hospitals into 4 regions (northeast, south, midwest, west). Models are trained and tested on 5000 samples from each region. Results are averaged over 100 random subsamples for each 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.

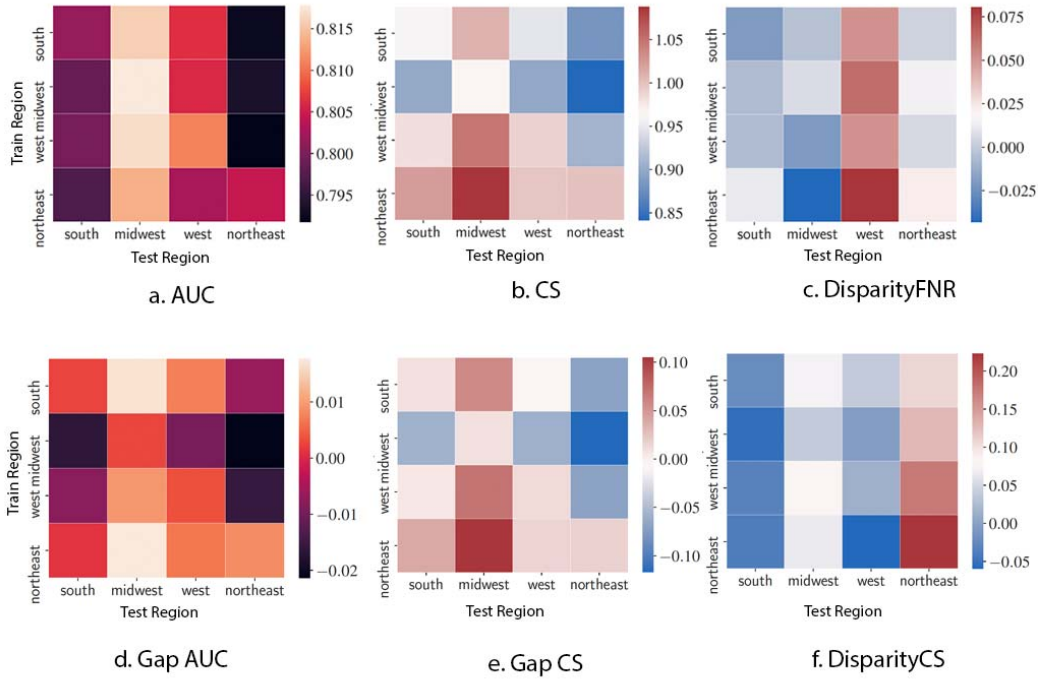

Figure 3. Statistical tests for dataset shifts. Results for two-sample tests with and without pooling of hospitals by region. Test results are plotted in (a,c) and test statistics are plotted in (b,d) to examine test results in more detail. Since the order of hospitals considered in the two-sample test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

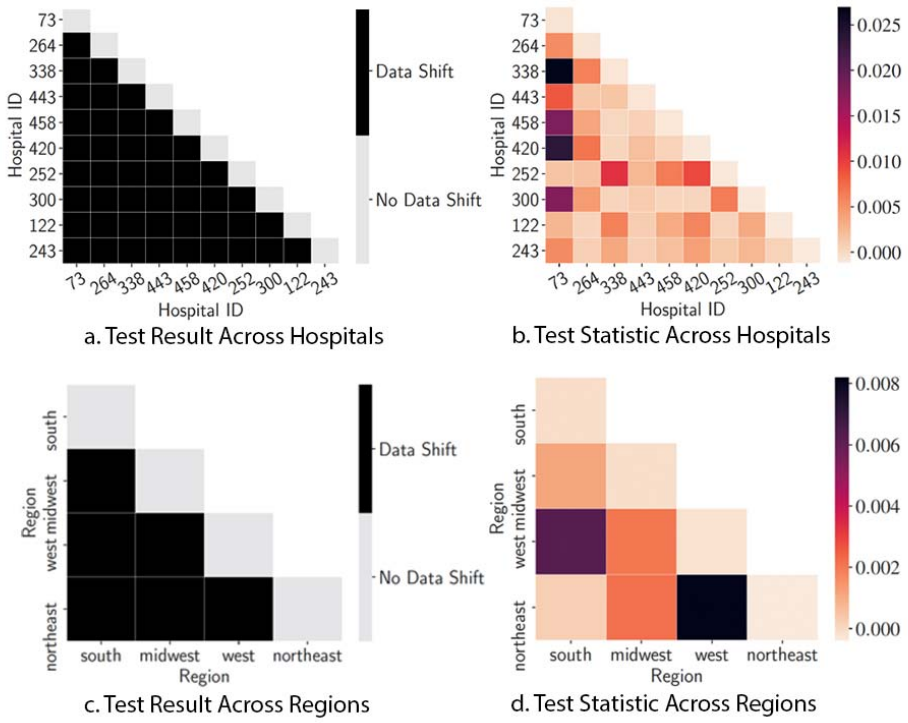
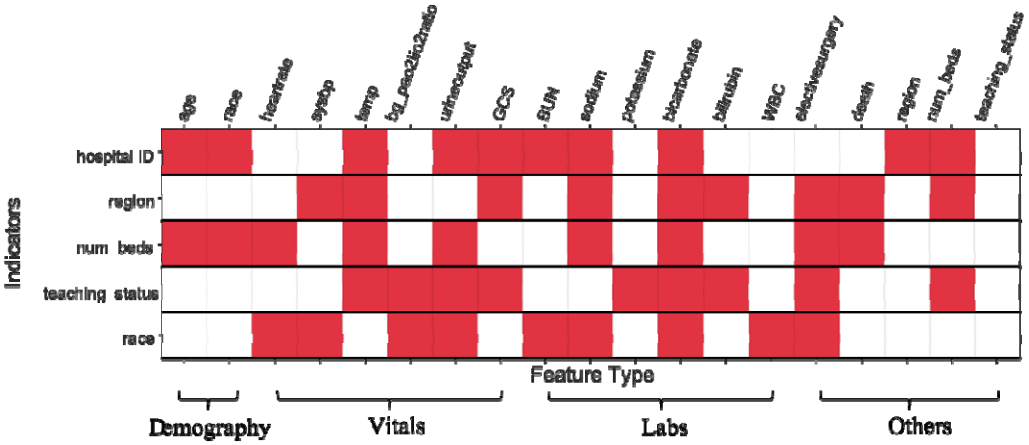
Figure 4. Shifts in mortality causal graph variable distributions due to hospital, region, and other factors. Each row represents (in red) the features which explain the shifts across each of the indicators given in column, i.e. the features with an edge from the indicator in the causal graph. For instance, shift across the hospital ID indicator (first row) is explained by shift in the distributions of age, race, temp, urineoutput, 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 eTable 3.