# Understanding prediction model generalization through the lens of causality

**Su Li, Master's Student[1]**
**[1]Faculty of Social and Behavioural Sciences, Utrecht University, Netherlands**

## Introduction

Prediction models are widely used in medical domains and settings to help doctors detect whether the patients suffer from a particular disease (diagnosis) or to predict the occurrence of a specific outcome (prognosis).[1,2] Thus, robust model performance is essential for generating reliable and accurate decisions. However, no model is robust across different environments; models' performance may differ when evaluated in datasets that are not used to develop the model.[3] For example, university hospitals usually serve more severe cases than primary care physicians. So, a model based on university hospitals may perform worse in primary care physicians. In the healthcare area, this kind of change in patient distribution can be described as a change in 'case-mix'.[4]

Recently, the definition of 'case-mix' has been interpreted differently within the causal inference framework. In diagnostic prediction models, we typically predict the underlying diagnosis ($Y$) based on the symptoms ($X$). In this case, the underlying diagnosis serves as the cause, and symptoms are the effect, meaning the model predicts the cause from the effect. In this context, diagnostic prediction models follow an **anti-causal direction**. The case mix here is interpreted as a change in the marginal distribution of the diagnosis. In contrast, we typically predict the future outcome ($Y$) based on the current patient characteristics ($X$) in prognostic models. Patient characteristics can be seen as the cause, and the future outcome is the effect. Thus, the prediction direction here follows a **causal direction**. The case mix refers to the distribution of patient characteristics.

Calibration and discrimination are the two traditional predictive performance metrics.[3] Discrimination measures how well the model distinguishes the positive and negative cases, while calibration refers to the agreement between predicted and observed probability.[5] As demonstrated by Van Amsterdam,[6] calibration tends to remain stable in causal direction prediction (prognostic), while discrimination remains consistent in anti-causal direction predictions (diagnostic) under shifts in case-mix. This is because discrimination relies on the distribution of features conditioned on the outcome ($X \mid Y$), making it independent of changes in the marginal distribution of the outcome. Meanwhile, calibration depends on the outcome given the features ($Y \mid X$) and thus remains unchanged under the shifts of the marginal distribution of the features.

This study aims to use empirical analysis to validate that calibration remains stable for the causal-direction models. In contrast, discrimination remains stable for anti-causal models under shifts in case mix. If validated, this theory could help us understand the varying performances of different prediction models (prognostic and diagnostic) across diverse environments, guiding the selection of appropriate evaluation metrics. For prognostic models, changes in discrimination are expected when case-mix shifts across different settings, while calibration should remain stable. When evaluating diagnostic models' performance across various settings, changes in calibration are expected, but changes in discrimination are not. Meanwhile, depending on the aim of the task and the prioritization of either discrimination or calibration, model developers can choose to use entirely causal or entirely anti-causal variables to enhance the robustness of prediction models under shifts in case mix.

The eICU dataset is used to evaluate the hypothesis empirically across multiple environments. Several prognostic and diagnostic prediction models are developed using logistic regression, and their performances are assessed by discrimination and calibration across different training and test sets. The Results section presents descriptive statistics and the preliminary results, and the Discussion section summarizes the current results and outlines future plans.

## Methods

### *Data sources*

In this study, the eICU Collaborative Research Database (version 2.0)[7] will be used, which is a publicly available, multicenter database containing de-identified data on 200,859 ICU patients from 208 hospitals across the United States from 2014 to 2015[8]. This database includes detailed records on patient demographics, vital sign measurements, diagnosis information, treatment information, laboratory results, and outcomes such as mortality. Three variables can be used to split the dataset to create different clusters: `hospital_region`, `hospital_bed_size`, and `hospital_teaching_status`.

- `Hospital Region` clusters the 208 hospitals based on their geographical location (Midwest, Northeast, South, West, Unknown (those are missing in hospital region value)).
- `Hospital Bed Size` categorizes hospitals by their capacity ($<$ 100 beds, 100–249, 250–499, $\geq$ 500, Unknown).
- `Hospital Teaching Status` indicates whether the hospital is a teaching institution or non-teaching.

### *Analysis Methods*

Logistic regression (LR) is a widely used method to analyze the relationship between predictors (independent variables) and dichotomous outcomes (dependent variable). Since logistic regression is already well-calibrated and well-discriminated, we use it to verify that calibration remains invariant for prognostic prediction models while discrimination remains stable for diagnostic prediction models under shifts in case mix.

Calibration evaluates how well predicted probabilities agree with observed outcome proportion. A prediction model is said to be calibrated when the predicted risks are consistent with the observed proportions of the event. In this study, calibration error is measured with the smooth expected calibration error.[9] Discrimination refers to how well the model can separate between positive and negative cases. Discrimination is evaluated using the area under the curve (AUC). A score close to 1.0 indicates perfect discrimination, while near 0.5 suggests no discrimination.[10]

### *Research Design*

We will define several prediction tasks that can be classified either as prognosis (e.g. 30-day mortality) or diagnosis (e.g. underlying diagnosis ICU admission). Independent variables will be selected based on previous literature and expert recommendations, taking particular notice of whether the independent variables are causal of the outcome (in the case of prognosis), anti-causal (in the case of diagnosis), or neither. Observations with missing values in the outcome variables were removed from the dataset. For numerical variables, missing values will be imputed by their mean, while for categorical variables, missing values will be imputed by the value with the most frequency. We will split the dataset based on the variables `hospital_region`, `hospital_bed_size`, and `hospital_teaching_status`, selecting one subgroup as the training set, while the remaining subgroups will form the test sets. This process will repeat across all subgroups.

For each task and each testing hospital, we will record the difference in AUC ($\delta$-AUC) and calibration error ($\delta$-ECE) between the training fold ($\text{AUC}_0/\text{ECE}_0$) and the testing fold ($\text{AUC}_1/\text{ECE}_1$). To ensure comparability across subgroups and tasks, $\delta$-AUC and $\delta$-ECE will be standardized by their values in the training fold. According to the causal-case mix hypothesis, prognostic models are expected to maintain calibration across external validation under shifts in case-mix, while for diagnostic models, the discrimination remains the same. To evaluate these differences, we will use an F-test to compare whether the variances of $\delta$-AUC and $\delta$-ECE differ significantly between prognostic and diagnostic models. Specifically, we sum up the variances of $\delta$-AUC (or $\delta$-ECE) for prognostic and diagnostic models separately. Then, the F-test is used to test the null hypothesis that the variances of prognostic models are the same as those of diagnostic models.

$$\delta = \frac{(\mathrm{AUC}_1 - 0.5) - (\mathrm{AUC}_0 - 0.5)}{(\mathrm{AUC}_0 - 0.5)}$$

Taking discrimination as an example, diagnostic models are expected to have stable discrimination scores. So, the differences in discrimination between the training set and testing set for diagnostic models are expected to be 0. Thus, $\mathrm{VAR}(\delta_{\mathrm{diagnostic}})$ is also anticipated to be 0. While for prognostic models, discrimination scores are expected to vary between training and testing sets. Therefore, $\mathrm{VAR}(\delta_{\mathrm{prognostic}})$ is anticipated to be statistically significantly larger than the variances of diagnostic models.

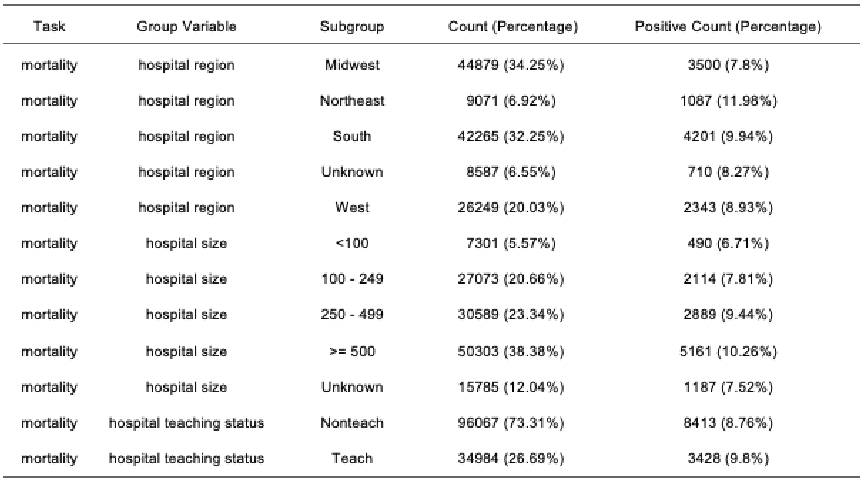
## Results

### *Descriptive Analysis*

So far, three tasks have been built: mortality (prognosis), age (diagnosis), and gender (diagnosis). The variables were selected according to the article by Tang et al.,[11] which selected variables based on several evaluation scores that are often used in ICU: SOFA, SAPs, APACHE II, and APACHE IV. The mortality model is a prognostic model that predicts whether the patient in the ICU survived (0) or died (1). Age and Gender are used as diagnostic models. Age is categorized by the median value (64 years), with age $\geq 64$ as 1 and age $< 64$ as 0. Gender is coded as male is 1 and female is 0.

Table  1 shows the sample size and positive case counts together with their percentages for mortality across different hospital characteristics: region, bed size, and teaching status. Using the mortality task as an example, it is evident from the table that the mortality rates vary significantly across different regions. The Northeast region has the lowest number of samples (9071, 6.92%) but the highest mortality rate (11.98%). However, the Midwest region has the largest sample size (44879, 34.25%) with the lowest mortality rate (7.8%). The South region has a relatively high sample size (42265, 32.25%) with a mortality rate of 9.98%. The West and Unknown have similar mortality rates of 8.93% and 8.27%, respectively. The rate of Unknown somewhat indicates the average mortality rate of all the groups. Subgroups categorized by `hospital_bed_size` show that both the sample size and mortality rate decrease as the number of beds reduces. Hospitals with more than 500 beds have the largest sample size (50303, 38.38%) and the highest mortality rate (10.26%), while those with smaller than 100 beds show the smallest sample size (7301, 5.57%) and the lowest mortality rate (6.7%). Teaching hospitals have a higher mortality rate (9.8%) compared to non-teach hospitals (8.76%).

**Table 1.** Distribution of Sample and Positive Case Counts by Hospital Characteristics.



| Task | Group Variable | Subgroup | Count (Percentage) | Positive Count (Percentage) |
|---|---|---|---|---|
| mortality | hospital region | Midwest | 44879 (34.25%) | 3500 (7.8%) |
| mortality | hospital region | Northeast | 9071 (6.92%) | 1087 (11.98%) |
| mortality | hospital region | South | 42265 (32.25%) | 4201 (9.94%) |
| mortality | hospital region | Unknown | 8587 (6.55%) | 710 (8.27%) |
| mortality | hospital region | West | 26249 (20.03%) | 2343 (8.93%) |
| mortality | hospital size | <100 | 7301 (5.57%) | 490 (6.71%) |
| mortality | hospital size | 100 - 249 | 27073 (20.66%) | 2114 (7.81%) |
| mortality | hospital size | 250 - 499 | 30589 (23.34%) | 2889 (9.44%) |
| mortality | hospital size | >= 500 | 50303 (38.38%) | 5161 (10.26%) |
| mortality | hospital size | Unknown | 15785 (12.04%) | 1187 (7.52%) |
| mortality | hospital teaching status | Nonteach | 96067 (73.31%) | 8413 (8.76%) |
| mortality | hospital teaching status | Teach | 34984 (26.69%) | 3428 (9.8%) |

### *Prediction Performance Across Hospital Characteristics*

The figures (Figure 1, Figure 2, and Figure 3) illustrate the calibration error (x-axis) and discrimination (AUC, y-axis) for the three tasks across subgroups grouped by `hospital_region`, `hospital_bed_size`, and `hospital_teaching_status`. The arrows start from the training set's performance to the test sets' performance. Different line styles correspond to different tasks, while symbols and colors represent subgroups and training environments, respectively. For the mortality task with the Midwest group as the training set, the arrows start at the training performance and spread to the test groups. The movements show a decline in performance, indicating increased calibration error and lower discrimination scores.

Across all groups, mortality prognosis models consistently show low and stable calibration errors, indicating that the prognostic model is well-calibrated. While AUC varies slightly (0.73-0.78), calibration performance remains stable across subgroups. In Figure 1, calibration errors stay near 0 for most training regions and show a slightly minimal increase when tested on others. Similarly, in Figure 2 and Figure 3, calibrations remain stable, confirming the reliability of these models.

In contrast, diagnostic models for age and gender result in low discrimination (AUC $\sim$ 0.5-0.65), suggesting weak discrimination ability. In Figure 2 and Figure 3, changes in calibration errors are minimal ($< 0.06$), but show clear variations compared to the values of prognostic models in the same group.

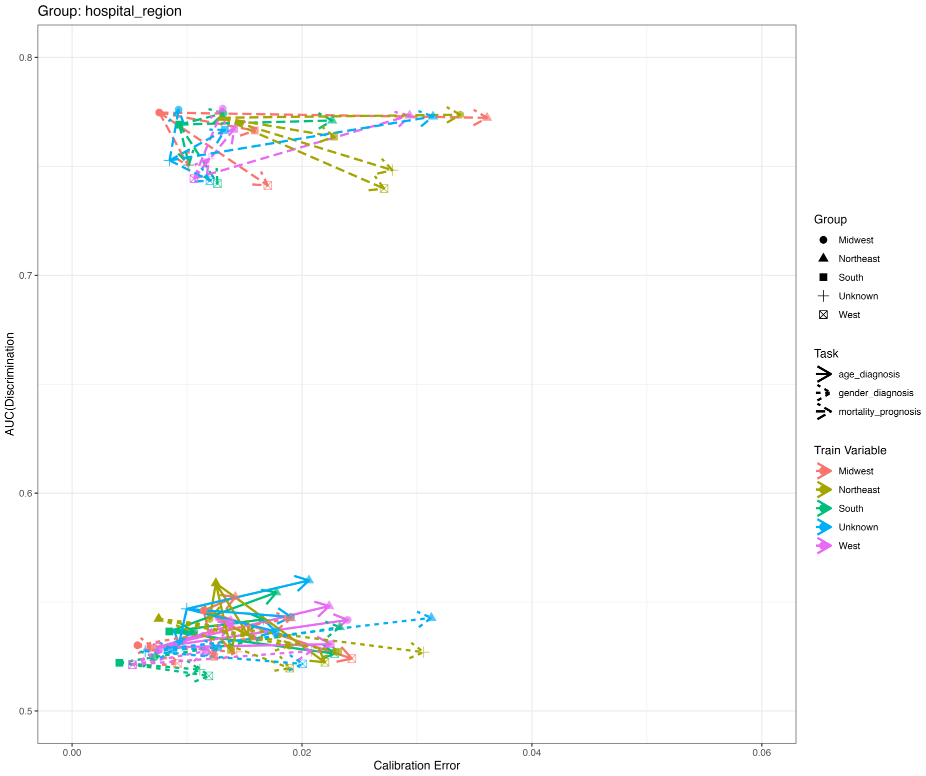


**Figure 1.** Discrimination vs. Calibration Error Across Hospital Regions and Tasks.

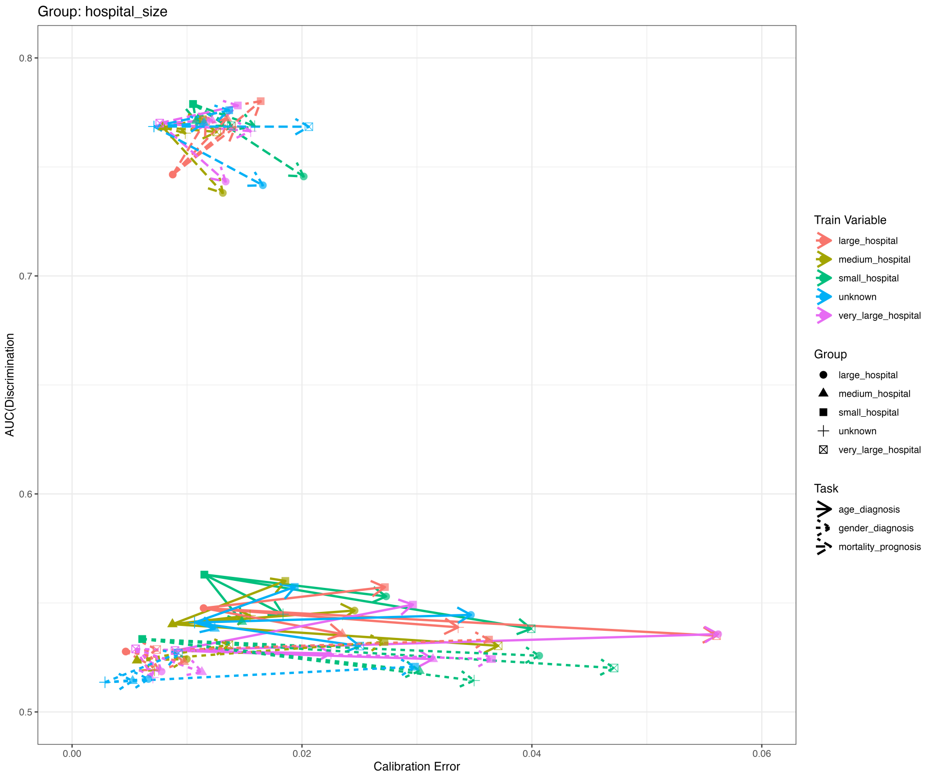


**Figure 2.** Discrimination vs. Calibration Error Across Hospital Sizes and Tasks.

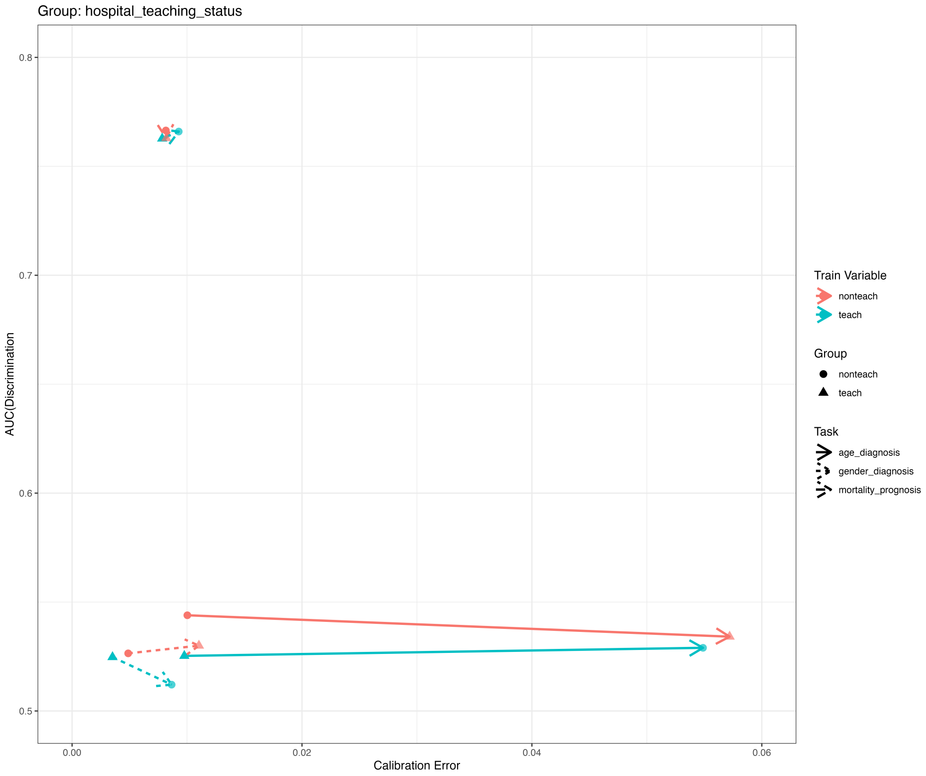


**Figure 3.** Discrimination vs. Calibration Error Across Hospital Teaching Status and Tasks.

## Discussion

The study aims to empirically validate that calibrations stay consistent for causal-direction (prognostic) models while discriminations remain stable for anti-causal (diagnostic) models under case-mix shift. Our findings partly support this hypothesis, as the calibration errors of the mortality prognostic model remain consistent across the training and test settings.

In Figures 1-3, the mortality prognostic model demonstrates robust and reliable performance across varying subgroups, with consistently high AUC values ranging from 0.75 to 0.8 and small calibration errors. The short arrow lengths and consistent directions indicate small variability between the training and testing sets, confirming the hypothesis that prognostic models maintain stability under external validation. The well-calibrated performances on both the training sets and the test sets are likely due to the predictors in the model covering comprehensive and multidimensional patients' physiological information that reduce the impact of confounding factors, thus ensuring robust performance under shift in case-mix. In contrast, Age diagnosis and gender diagnosis have low AUC values (0.5-0.6), reflecting unreliable results that lack strong comparative significance. The unsatisfactory performance may be due to the limited and unrepresentative predictors that fail to capture enough information for class distinction.

The current study includes only three tasks, and the low AUC value (0.5- 0.6) indicates the unreliability of the two diagnostic models, making them insufficient to support further analysis. Future work will focus on training more tasks and selecting appropriate predictors. Considering the results based on figures alone are not rigorous, we will use statistical tests to assess whether the variances of calibration and discrimination between the training fold and test fold differ significantly across prognostic and diagnostic models.

## References

1. Slieker R, van der Heijden A, Siddiqui M, et al. Performance of prediction models for nephropathy in people with type 2 diabetes: systematic review and external validation study. BMJ. 2021;374:n2134.

2. Shariat S, Karakiewicz P, Roehrborn C, Kattan M. An updated catalog of prostate cancer predictive tools. Cancer. 2008;113(11):3075-99.

3. Steyerberg E, Vickers A, Cook N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.

4. Thomas J. Risk Adjustment for Measuring Health Care Outcomes. vol. 16. 3rd ed. Int J Qual Health Care; 2004.

5. Steyerberg E. Clinical prediction models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer Science & Business Media; 2008.

6. Van Amsterdam J. A causal viewpoint on prediction model performance under changes in case-mix: Discrimination and calibration respond differently for prognosis and diagnosis predictions. arXiv. 2024.

7. Pollard T, Johnson A, Raffa J, Celi L, Mark R, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178.

8. Pollard T, Johnson A, Raffa J, Celi L, Badawi O, Mark R. eICU collaborative research database (version 2.0). PhysioNet. 2019.

9. Błasiok J, Nakkiran P. Smooth ECE: Principled reliability diagrams via kernel smoothing. arXiv. 2023.

10. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

11. Tang H, Jin Z, Deng J, et al. Development and validation of a deep learning model to predict the survival of patients in ICU. J Am Med Inform Assoc. 2022;29(9):1567-76.