**CPH 200C: Computational Precision Health Cornerstone** Wed Apr 17, 2024

# Problem Set 1: Risk Stratification

*Instructor: Prof Irene Chen* *Due date: Wed May 01, 2024*

Risk stratification allows clinicians to separate patients into high and low risk patients. The primary event of interest may include patient mortality, onset of a new disease, or hospital readmission. To analyze this problem, we typically use a supervised learning model to predict future events. The goal of this problem set is to develop your ability to conduct risk stratification from electronic health records. We will be examining a dataset of 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. Our goal is to predict which diabetic patients will be readmitted within 30 days of discharge.

**Instructions**: This is not a group project and students will be graded individually. This project has two deliverables:

- **A report summarizing your results**. The report should include point-by-point answers to the questions below. Please submit your report (in PDF format) via both the bcourses website for CPH200C and the Gradescope website for CPH200C (entry code: **DPWJ68**).
- **A zip file with your code**. Please submit both your code and report using the Gradescope. You will get feedback on both your report and code via Gradescope.

## 1 Diabetes Risk Stratification (23 pts)

We will be examining a dataset of 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. Our goal is to predict which diabetic patients will be readmitted within 30 days of discharge.

### 1.1 Data Exploration

Download the diabetes dataset and review the corresponding publication [SDG+14]. Graph the 30-day readmission rates by age, gender, and race. Which groups correlate with higher readmission rates?

### 1.2 Model Development

To build our risk stratification model, develop machine learning models in three model classes: linear, tree-based, and neural network. Across multiple train-validation-test splits, **we expect performance to be within or higher than 0.65-0.70.**

Describe your selected models and resulting performances, the hyperparameters searched over and chosen, and the best overall model. Graph your model performances in AUC with appropriate confidence intervals [CM04]. You may find the `confidenceinterval` Python package helpful.

Some factors to consider in your model development:

- Categorical features can be turned into one-hot encoding dummy variables. Beware multi-collinearity, which may require dropping one category.
- Some variables will have missing input data. How should you handle this missing data?
- According to [SDG+14], the interactions between variables may have strong predictive power for readmission rates.

### 1.3 Feature Importances

In order to better understand our models, we want to examine which features are most predictive. Using the feature importances of the linear model you chose, please graph the 10 most positively-predictive features and the 10 most negative-predictive features, along with their feature weights. Select 4-5 features of note and explain how these clinical factors might relate to readmission. You may find the variable descriptions in the supplementary materials of [SDG+14] helpful.

### 1.4 Subgroup Evaluation

Another area of interest is the performance of the best model across patient subgroups. Because comparing AUC across subgroups can be misleading due to varying thresholds [KZ19], we want to examine the performance for patient subgroups for the same fixed threshold.

Using the best predictive model, set the threshold for a fixed model accuracy (e.g., 95%) and report this threshold and model accuracy. Graph the subgroup accuracy across age, gender, and race subgroups with the appropriate confidence intervals. Which patient subgroups have higher or lower predictive accuracy? Similarly, graph the sensitivity and specificity for the same threshold and the same patient subgroups. Which patient subgroups have higher or lower sensitivity and specificity?

## 2 More About Diabetes (5 pts)

In Part 1, we predicted hospital readmission from emergency department visit information for diabetic patients to better understand diabetic risk stratification; however, this is only a small part of diabetes care for patients. Propose a different research project to leverage machine learning to better understand diabetes. Your response should include:

- The specific diabetes-related problem you are trying to solve
- The data modality/modalities you would use for this problem
- The machine learning model you would use
- How you would evaluate the model
- The intended health impact

While comprehensive answers can vary in length, we anticipate responses will be over 500 words.

## 3 MIMIC-IV Warm-up (1 pt)

Looking ahead, Problem Set 2 will look at distribution shift on the MIMIC-IV dataset. Let's do a little warm-up exercise now.

### 3.1 Getting Access

**Do not wait on this part because getting MIMIC-IV access can take 2-4 days to receive approval.**

To get access, please complete these steps:

- Create a Physionet account
- Complete the CITI training
- Submit evidence of your training
- Sign the Data Use Agreement

### 3.2 Patient Anchor Groups

For this portion, we recommend using AWS or Google BigQuery to access the dataset because downloading the CSV files can be time-consuming. Using your MIMIC-IV data access, create a table with the number of unique patients for each `anchor_year_group` from the `patients` table. What are anchor year groups and why are they necessary?

## 4 Feedback (1 pt)

How long did this problem set take you? The 1 pt will be given for any response to this question, and the value will only be used for calibration of future problem sets. Please separate out the MIMIC-IV data access from your response, i.e., report time on problem set outside of getting MIMIC-IV data access and report time spent getting MIMIC-IV access.

## References


[CM04] Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the roc curve. *Advances in neural information processing systems*, 17, 2004.

[KZ19] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019.

[SDG+14] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, John N Clore, et al. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.