Prediction of Short-term Hospitalization Outcomes utilizing Medical Notes using NLP Methods

Itay Kahane [1] **Roy Yatskan** [1]
github.com/ryatskan/MLHC-Project

1. Introduction

In modern healthcare, the ability to forecast patient outcomes is highly desirable, particularly within the high-stakes environment of the Intensive Care Unit (ICU) (Halpern & Pastores, 2010). With the advent of electronic health records (EHRs) and the exponential growth in healthcare data (Ramon et al., 2007), coupled with the increasing implementation of Machine Learning (ML) tools to enhance decision-making in various fields (Jordan & Mitchell, 2015), the application of similar for patient outcomes prediction problems is naturally on the rise as well (Nayyar et al., 2021). ICU hospitalization outcomes, such as mortality, length of stay, and readmission, are critical factors that can greatly influence both patient well-being and healthcare system efficiency (Shillan et al., 2019).

Despite the increase in available data, a large portion of it is implicitly stored within text notes taken by the medical crew (Johnson et al., 2016). For example, while the glucose level and weight upon admission require only minimal prepossessing, advanced Natural Language Processing (NLP) methods must be utilized in order to decipher on a large scale a text such as[1]:

```
"Pt c/o cough, non-productive, dry.
Tesselon and robitussin w/o effect.
Robitussin and codeine ordered snd
given w/ good effect.  R mid-to lower
base very diminished, MD aware.  4L NC,
tol well."
```

Fortunately, the growing capabilities of Natural Language Processing (NLP) tools allow reliable data extraction from such notes (Cambria & White, 2014). This study build upon such tools to enhance short-term ICU hospitalization predictions. We show that the utilization of data extracted from medical notes improve the performance of various common methods on three prediction tasks: Mortality, Prolonged Stay and Hospital Readmission.

2. Cohort Description

In this work we use a subset of MIMIC-III data set (Johnson et al., 2016). Within this subset, we have gathered data pertaining to a total of 40,156 hospital admissions, corresponding to 32,513 unique patients. Among these admissions, only 24,518 met the excluding criteria described in the follosing paragraph. In terms of patient outcomes, we have identified distinct categories. Specifically, there were 3,162 instances of patient mortality, 13,215 cases of extended hospitalization (defined as 7 days or more), and 1,116 occurrences of hospital readmission. Additionally, we have established a control group, which comprises 9,810 patients who did not experience any of these aforementioned outcomes. Given the non-mutually exclusive nature of our prediction targets, we have provided a comprehensive breakdown of the patient counts for each conceivable combination in Table 1

Table 1. Patient Counts in Each Outcome Combination: M denotes patient mortality, PS signifies prolonged hospitalization, and RA represents hospital readmission.

M	PS	RA	No. of Patients
0	0	0	9,810
0	0	1	10,593
0	1	0	1,117
0	1	1	1,882
1	0	0	324
1	0	1	629
1	1	0	52
1	1	1	111

our initial set of extracted features encompasses fundamental patient information, comprising vital signs and laboratory measurements. Additionally, we considered categorical attributes such as ethnicity, which we represented using a one-hot encoding scheme, as well as age, gender, and weight at the time of admission. Building upon this foundational feature set, we introduced three distinctive features into our analysis. Firstly, we incorporated the most recent nursing note recorded within the initial 48 hours of hospitalization, capturing potentially crucial contextual information. Sec-

[1] The cited text is only a fraction of the full note given for the considered patient

ondly, we included two medication-related features: one denoting the total number of medications administered to the patient during the 48-hour hospitalization period, and the other indicating the count of unique medications administered, offering insights into medication diversity and management within this critical time frame. These additional features enrich our dataset and enable a more comprehensive exploration of patient outcomes and healthcare patterns.

3. Methods

3.1. Inclusion and exclusion criteria

Our study primarily concentrated on patients who had been hospitalized for a minimum of 48 hours, with a specific emphasis on those patients who did not succumb during this initial time frame. The rationale behind this selection criterion is twofold: firstly, it ensures the availability of a substantial volume of relevant data, and secondly, it allows us to concentrate our efforts on patients for whom the medical team has a reasonable opportunity to provide assistance. Data derived from patients who passed away shortly after admission, within a matter of hours, may not be as pertinent to our research objectives. We only included in our study patients within the age range of 18-90 years old, to avoid outliers, and to focus our model on adults. Given our goal of developing a decision support tool for healthcare professionals based on available data, we implemented a 6-hour prediction gap, effectively excluding data from the final 6 hours within the initial 48-hour window (i.e., hours 42 to 48 since admission). Our predictive modeling efforts were then directed towards forecasting patient outcomes beyond this exclusionary period.

3.2. Data exploration and pre-processing

To refine the abundant available data on lab measurements, we generated three features per measurement, namely the minimum and the maximum value, as well as the mean of all measurements taken along the 48 hours period. Examining the patients age provided that many of them were not documented correctly and stated as 0, so our age exclusion criterion described above helped to avoid those outliers. We then created train and test sets per target, with a 80/20 division. Subsequently, we partitioned the data into training and testing sets, adhering to an 80/20 split ratio for each target variable. Addressing missing data, we adopted an imputation approach, where we replaced missing values with the mean value derived from all patients within the respective dataset. It is important to note that this imputation process was performed separately for each target variable to prevent any potential data leakage.

3.3. Models

In this study, we employ four primary models: Linear Regression, XGBOOST, Deep Averaging Network (DAN), and the BART transformer. However, our core evaluation methods focus on XGBOOST and Linear Regression.

3.4. ML Models for Tabular Data

Linear Regression: A ubiquitous ML model, Linear Regression establishes a linear relationship between labels and input features. It's known for its simplicity, good results with tabular data, and interpretability.

XGBoost: XGBoost is a decision-tree-based ensemble model. Along with similar models like CatBoost and AdaBoost, it's often considered as the SOTA for tabular data tasks (Chen & Guestrin, 2016).

For the inclusion of textual information, we apply feature extraction using TF-IDF:

TF-IDF: TF-IDF is a method that calculates a word's importance in a document compared to its occurrence in a set of documents. It highlights unique words and reduces the impact of frequent terms, such as stop-words. After extraction, we use the chi-squared test to select the top 400 word features.

3.5. NLP-specific Models

While our primary evaluation pipeline focuses on previously discussed models, in this section, we introduce two NLP-specific models. To our knowledge, these have not been previously applied to the MIMIC-3 dataset before.

Deep Averaging Network (DAN): DAN is a neural network model for text classification(Iyyer et al., 2015). It uses pre-trained embeddings to convert words into N-dimensional vectors, where semantically similar words have similar vectors. In its standard form, DAN embeds each word in the text, averages these embeddings, and then forwards them to a classification layer. Our adaptation modifies the network architecture to accommodate tabular data. Both the text and tabular data are processed separately through dedicated layers and are then concatenated for the final classification step.

BioBART: A pre-trained version of BART (Lewis et al., 2019), which is a transformer model inspired by the architectures of BERT and GPT models . **BioBART was trained on PUBMED articles and was not exposed to the MIMIC-3 dataset**(Yuan et al., 2022). A noteworthy attribute of BioBART, setting it apart from BERT, is its capability to handle a token size of 1024—twice that of BERT, which is especially useful for dealing with long clinical texts. In our approach, we perform supervised fine-tuning of BioBART on the notes, with the expectation that its specialized clinical understanding improves its ability to classify clinical patient notes.

3.6. Evaluation

We assess the models using a comprehensive set of metrics: F1 score, ROC and PR curves with boostrapping, the Shap interpretability test, and comparisons of calibration curves both pre and post-adjustment. The full evaluation is provided for only the XGBoost model, but more evaluation data is available inside the analysis notebook.

4. Results

As previously mentioned, our experimental approach involved the evaluation of three distinct models, namely XGBoost, Logistic Regression and Deep Averaging Network, across three target variables. This assessment was carried out under varying feature configurations, specifically utilizing: (1) exclusively tabular data, (2) exclusively textual data, and (3) a combination of both tabular and textual data. An F_1 comparison between the models is given in Table 2.

Table 2. Comparison of F_1 Scores on the Test Set for Logistic Regression and XGBoost Models, Employing Three Different Feature Sets: Exclusive Tabular Data (Tab), Exclusive Textual Data (Text), and Combined Data Sources (Both). Evaluation of Model Performance Across Three Target Variables: Patient Readmission (RA), Patient Mortality (M), and Patient Prolonged Stay (PR).

	Logistic Regression			XGBoost			DAN
Target	Tab	Text	Both	Tab	Text	Both	Both
RA	0.11	0.08	0.12	0.10	0.08	0.10	0.10
M	0.36	0.32	0.41	0.39	0.31	0.41	0.38
PR	0.68	0.72	0.73	0.70	0.72	0.73	0.72

In this context, the textual information encompasses all notes documented during the first 42 hours, falling under the categories of 'Nursing', 'Physician', or 'Other'. We have demonstrated that the integration of both tabular and textual information yields significant improvements in predictive accuracy for medical outcomes across three distinct objectives. Furthermore, our findings reveal small performance differences between the three models. While XGBoost exhibits a slightly superior predictive capability for patient readmissions compared to Logistic Regression, both models struggle in this particular task, with relatively poorer performance compared to mortality and prolonged hospitalization prediction, the latter being our least challenging task. However, it is noteworthy that in the last two objectives, patient mortality and prolonged stay prediction, XGBoost and Logistic Regression demonstrate similar predictive performance levels. However, our findings suggest that the Deep Averaging Network underperforms in comparison to traditional models like Logistic Regression and XGBoost. While the processing of averaged embeddings may be effective, the neural network's relatively poorer power for processing tabular data seems to negate these advantages. Focusing on the combined feature set, we provide ROC and Precision-Recall curves for both models in Figure 6 and Figure 1 respectively. Both curves reveal a strong similarity between Logistic Regression and XGBoost across all three targets and various operating point. The ROC curves also reveal the fact that the prediction of mortality and of prolonged stay are far easier task than predicting readmission. However, since the target classes in the case of mortality predicition is not well balanced (4,311 patients survived vs. 636 who perished), the order relation between the two latter tasks is not apparent, as it is indeed clear from the Precision-Recall curves.

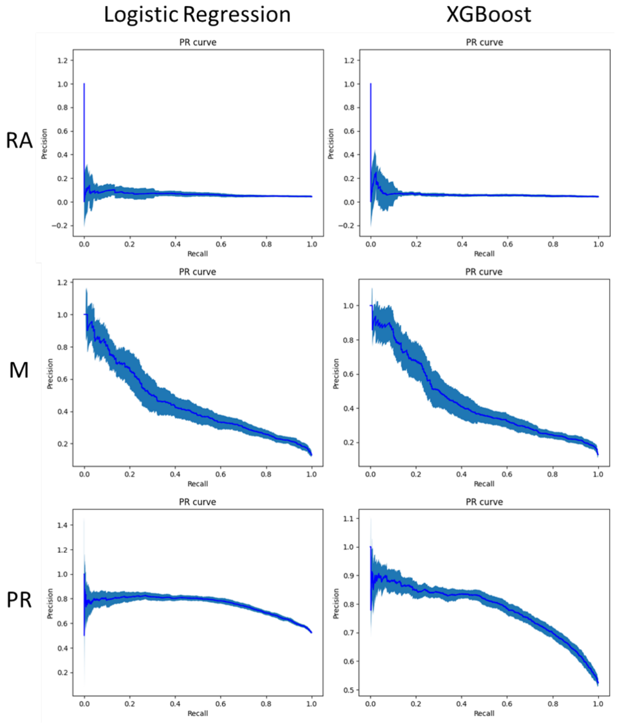

Figure 1. PR curves for Logistic Regression and XGBoost across Three Target Variables: Patient Readmission (RA), Patient Mortality (M), and Patient Prolonged Stay (PR).

We further choose to focus on the XGBoost algorithm. In Figures 3 through 5, we present SHAP (SHapley Additive exPlanations) Beeswarm plots for all three prediction targets. These graphical representations illuminate a rich diversity of features sourced from textual data that exert significant influence on the predictive capabilities of our models. Notably, the inclusion of a multitude of tabular laboratory test data in conjunction with textual information underscores the synergy achieved by harnessing both data sources, ultimately culminating in the formulation of our most robust predictive model. This integration stands as a testament to the inherent value of leveraging a multifaceted dataset in healthcare predictive modeling.

Finally, in Figure 2 we present calibration graphs that illustrate the performance before and after calibration for all three tasks. Our analysis reveals an interesting pattern in terms of calibration, wherein the predictive targets of readmission, mortality, and a future prolonged stay are arranged in order of increasing difficulty, with predicting readmission being the most challenging and forecasting a prolonged stay in the future being the most achievable among the three, similarly to the order in prediction abillity.

4.1. BART

To compare the state-of-the-art transformer architecture against the traditional NLP models, we fine-tune the BioBART model on the clinical notes. Due to BART's constraints with token size, we limited our input to only the most recent notes. For simplicity, we trained the model only for the prolonged stay target. The results showed that the fine-tuned BART lagged behind previous models, achieving an F1 score of 0.67. Given that BART typically offers better textual understanding than methods like TF-IDF and averaged embeddings, which don't preserve sentence order, we hypothesize that the limited text data length could be the primary factor in this result.

5. Discussion

In our study, we introduced four models designed to integrate both textual and tabular data, with the aim of improving predictive performance in the context of three critical healthcare objectives: patient readmission, mortality prediction, and prolonged stay estimation. Among these tasks, patient readmission emerged as the hardest task, while prolonged stay prediction posed the least complexity. We evaluated two model candidates, Logistic Regression and XGBoost, across various testing scenarios. Interestingly, our findings revealed a consistent trend: both models exhibited remarkably similar results across all tested configurations. This suggests that, within our specific experimental context, the choice between these two models may not significantly impact predictive outcomes, underscoring the robustness of our approach. In addition, our results indicate that the more recent architectures - the Deep Averaging Network and BART, do not surpass the performance of the conventional TF-IDF technique for the three targets.

References

Cambria, E. and White, B. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

Chen, T. and Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145%2F2939672.2939785.

Halpern, N. A. and Pastores, S. M. Critical care medicine in the united states 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical care medicine*, 38(1):65–71, 2010.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1681–1691, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1162. URL https://aclanthology.org/P15-1162.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Jordan, M. I. and Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

Nayyar, A., Gadhavi, L., and Zaman, N. Machine learning in healthcare: review, opportunities and challenges. *Machine Learning and the Internet of Medical Things in Healthcare*, pp. 23–45, 2021.

Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M., and Van Den Berghe, G. Mining data from intensive care patients. *Advanced Engineering Informatics*, 21(3):243–256, 2007.

Shillan, D., Sterne, J. A., Champneys, A., and Gibbison, B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical care*, 23:1–11, 2019.

Yuan, H., Yuan, Z., Gan, R., Zhang, J., Xie, Y., and Yu, S. Biobart: Pretraining and evaluation of a biomedical generative language model, 2022.

A. Calibration Curves

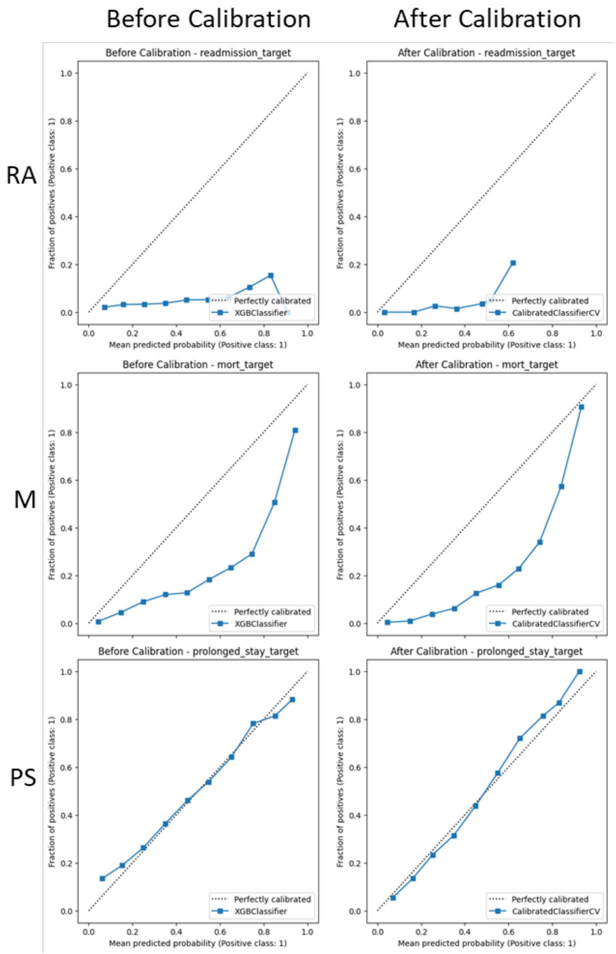

Figure 2. Calibration graphs of XGBoost, before (left) and after calibration on Patient Readmission (RA), Patient Mortality (M), and Patient Prolonged Stay (PR).

B. Shap values

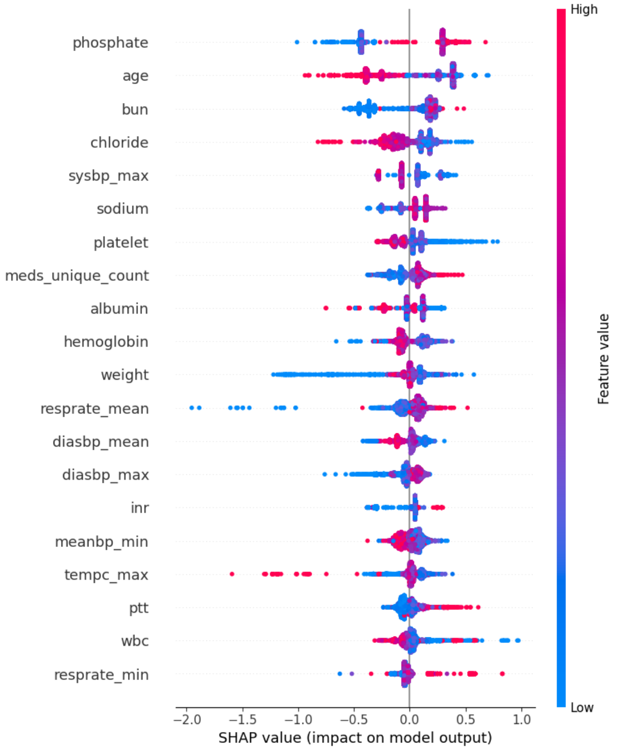

Figure 3. Shap interpretability results for the readmission target of the XGBoost model.

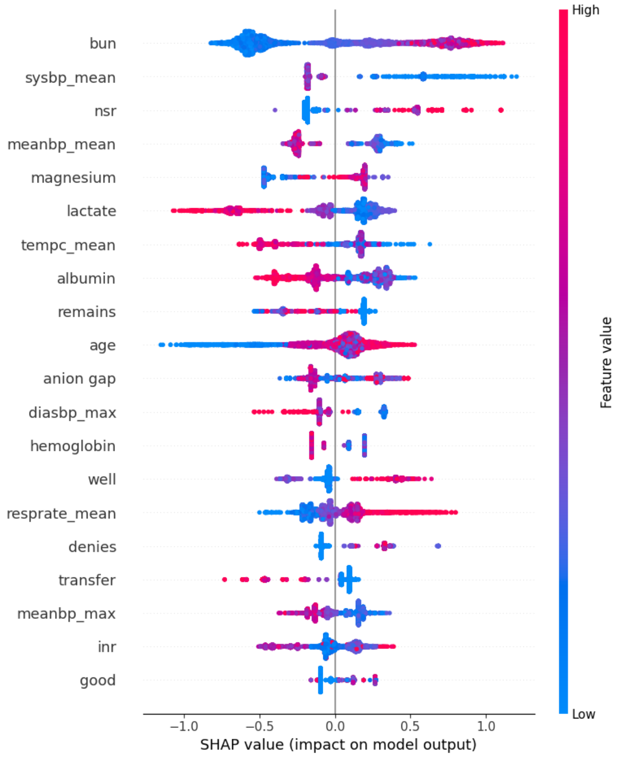

Figure 4. Shap interpretability results for the mortality target of the XGBoost model.

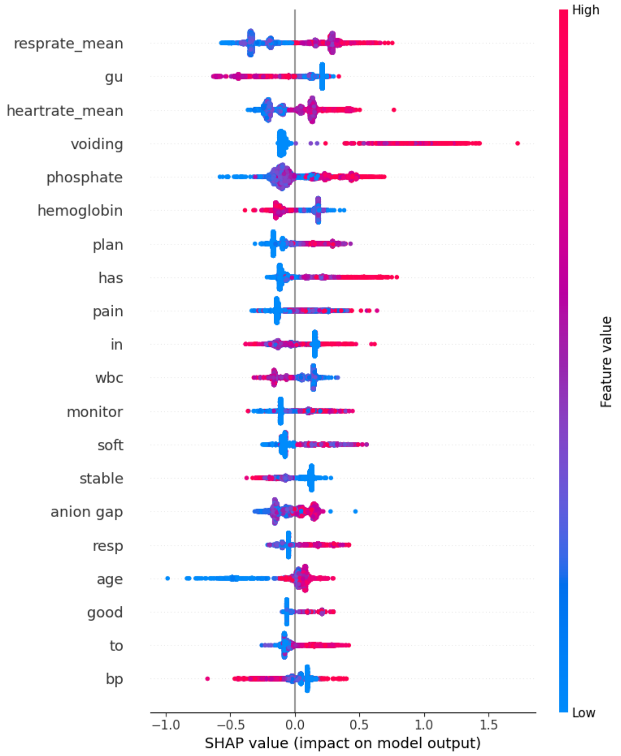

Figure 5. Shap interpretability results for the prolonged stay target of the XGBoost model.

C. ROC curve

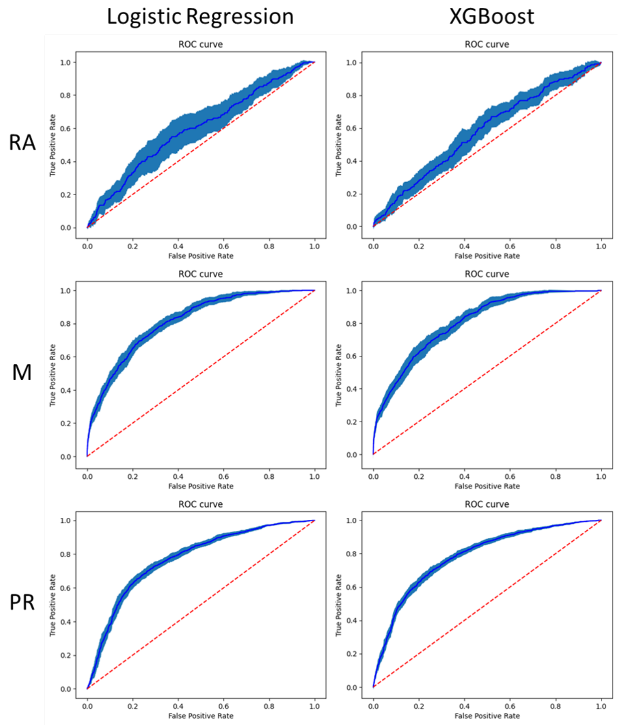

Figure 6. ROC curves for Logistic Regression and XGBoost across Three Target Variables: Patient Readmission (RA), Patient Mortality (M), and Patient Prolonged Stay (PR).