Group 12 Phase 1 Report

He Man
Lim Wen Sheng, Daniel
Luis Lee Zhi Rui
Siu Zhuo Bin

Literature Review

Hospital readmissions place a time and monetary burden on patients and society. Diabetes, a chronic metabolic disease that is highly associated with other diseases like heart and kidney disease, is worth studying as readmissions are common for diabetic patients. We have selected three studies for review – one model-building and validation study in Singapore (Soh et al., 2022), a large U.S. retrospective study that differentiates diabetes as a primary vs secondary diagnosis (Rubin et al,. 2023), and a machine-learning comparison on a database of over 100,000 records of diabetic patients (Shang et al., 2021). These models can flag high-risk patients with moderate discrimination where the most informative signals combine hospitalization history and care process variables, such as length of stay, discharge disposition, and medication history.

In Soh et al., a 30-day readmission model called LIPiD was developed and validated using data from 2,355 adults hospitalized for diabetes-related causes at National University Hospital in Singapore (2008-2015). LIPiD uses four predictors (length of stay, ischemic heart disease, peripheral vascular disease, and number of drugs) and achieves a C-statistic of 0.68 (0.66 - 0.70 with 10-fold cross-validation). Generalizability was further probed via three simulated “external” cohorts reflecting different readmission prevalences and comorbidity mixes (C-statistics of 0.64 – 0.68), for which reasonable calibration was demonstrated, particularly in lower predicted-risk deciles. LIPiD seems to be moderately robust in predicting readmission rates using these four predictors.

Rubin et al. analysed 8054 hospitalized adults with diabetes and separated patients by primary versus secondary discharge diagnosis of diabetes. Readmission was higher with primary diabetes diagnoses (22.2% vs 16.2%) and multivariable models were strongly discriminative in both strata (C-statistics of 0.836 vs 0.822). Inpatient diabetes consultation was also associated with lower odds of readmission among those with a primary diabetes diagnosis, which implies that quality of care during admission may be important in reducing readmission rates. Although the C-statistics of the models were promising, further evaluation and studies need to be done to validate their performance, such as by using randomized clinical trials.

Shang et al. used the Cerner Health Facts database which comprises 100,244 diabetes records from 1999-2008. They built models in KNIME to compare the Random Forest, Naïve Bayes, and tree ensemble models, and also addressed severe class imbalance via down and over-sampling. Compared to Soh et al. and Rubin et al., this study used a significantly larger database, but because 53.69% of patients had no readmission record and 11.22% were readmitted within 30 days, down and over-sampling was performed to compensate for this. Overall, RF seemed to be the best model by AUC (0.66) compared to the other two models.

Across all three papers, length of stay seemed to be a strong predictor of readmission rate. It was among LIPiD’s four variables in Soh et al. and also appeared in Shang and Rubin’s

models as an independent risk factor across strata. Shang et al. also quoted another study that a length of stay longer than 5 days was associated with a greater than 87% risk of readmission compared to a length of stay shorter than or equal to 2 days.

In terms of feature engineering, comorbidities were presented differently across the 3 papers. Comorbidities are strong predictors of readmission in Soh et al. and Shang et al. Rubin et al. used the Charlson Comorbidity Index to summarize a patient's overall burden of chronic illness into a single score. A higher score indicates higher risk of poor prognosis. Soh et al. picked individual conditions comprising ischemic heart disease and peripheral vascular disease, which are highly related to diabetes, as well as polypharmacy (number of drugs) for their LIPiD model. Soh et al reasoned that CCI places very high scores on diseases like AIDS and metastatic cancer, both of which are pathophysiologically unrelated to diabetes. This shows that in statistical analyses, subject matter knowledge is important in choosing the correct features for data modelling. In Shang et al., similar diagnoses for 16 types of diseases according to the ICD-9 coding were merged due to sparsity as the ICD-9 codes have several hundred categories. This was done to improve analyses and generate better statistical signals. Eventually, it turns out that secondary diagnoses are an important indicator of readmission rate in Shang et al.'s study.

In conclusion, these studies reveal some common themes. Length of stay and comorbidities seem to be strong predictors of readmission across both regression-based and machine learning approaches. Patient heterogeneity is important, as shown by Rubin et al., because the readmission risk differs between patients hospitalized for diabetes versus those for whom diabetes is only a comorbidity. Remaining gaps include the need for external validation of these models across diverse health systems and patient demographics, and better incorporation of other factors such as socioeconomic status and accessibility of healthcare into predictive models. Machine learning for outcome prediction in a healthcare setting requires machine learning models that are easily understood by healthcare practitioners and easy to use and interact with, such as via an interactive dashboard. Of course, these machine learning models should only be used with discretion as clinical knowledge is still key in healthcare, and the machine learning models must be evaluated properly before being used on a wide scale.

References

1. Rubin, D. J., Maliakkal, N., Zhao, H., & Miller, E. E. (2023). Hospital readmission risk and risk factors of people with a primary or secondary discharge diagnosis of diabetes. *Journal of Clinical Medicine, 12*(4), 1274. https://doi.org/10.3390/jcm12041274
2. Shang, Y., Jiang, K., Wang, L., Zhang, Z., Zhou, S., Liu, Y., Dong, J., & Wu, H. (2021). The 30-days hospital readmission risk in diabetic patients: Predictive modeling with machine learning classifiers. *BMC Medical Informatics and Decision Making, 21*(Suppl 2), 57. https://doi.org/10.1186/s12911-021-01423-y
3. Soh, J. G. S., Mukhopadhyay, A., Mohankumar, B., Quek, S. C., & Tai, B. C. (2022). Predicting and validating 30-day hospital readmission in adults with diabetes whose index admission is diabetes-related. *Journal of Clinical Endocrinology & Metabolism, 107*(10), 2865–2873. https://doi.org/10.1210/clinem/dgac380

EDA Report

The objective of this study is to predict diabetes patient readmission based on the Diabetes 130-US Hospitals for Years 1999-2008 database from the UCI Machine Learning Repository (ID: 296) comprising data collected from 130 US hospitals between 1999-2008.

```
Dataset Overview
Shape: (46861, 48)
Memory usage: 86.6 MB

Data Types
object    37
int64     11
Name: count, dtype: int64

Missing Values
                  Missing Count  Missing Percentage (%)
weight                    45073                    96.2
max_glu_serum             42246                    90.2
A1Cresult                 39553                    84.4
payer_code                31690                    67.6
medical_specialty         16030                    34.2
race                       1249                     2.7
diag_3                     1059                     2.3
diag_2                      258                     0.6
diag_1                       10                     0.0
readmitted                    1                     0.0
diabetesMed                   1                     0.0
```

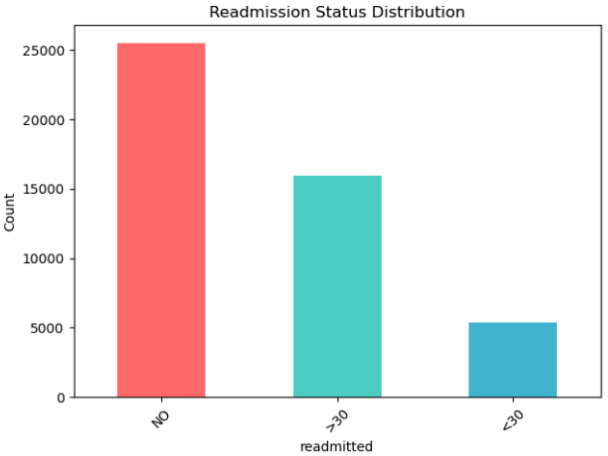

Table 1. Basic dataset information, and missing value count. **Fig. 1** Readmission status distribution

The raw dataset comprises 46861 records with 48 features each (**Table 1**). Table 1 also shows the number of missing values for each field in the dataset. The fields with large numbers of missing values should be removed as they do not contribute to the analysis. **Figure 1** shows the distribution of the readmissions in the three categories comprising NO (not readmitted), >30 and <30. The majority of patients are not readmitted, with the lowest being in the <30 category.

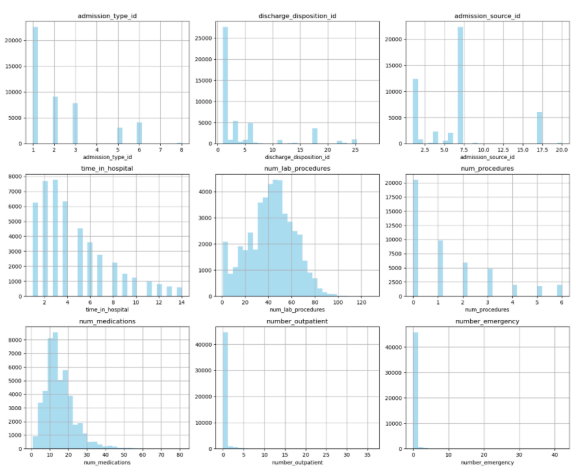

Fig. 2 Binned numerical feature distributions

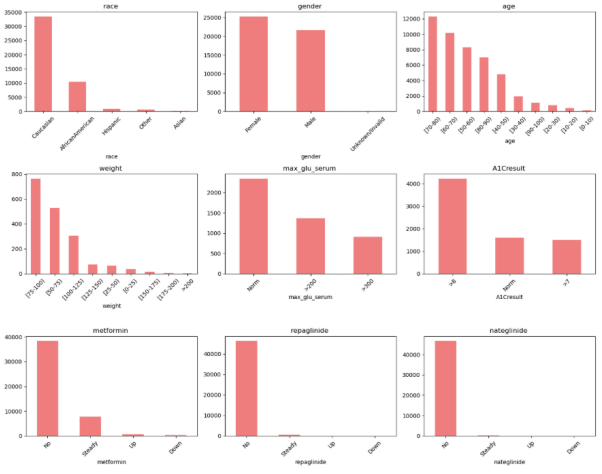

Fig. 3 Categorical feature distributions for low-cardinality features

Figures 2 and **3** show histogram plots for the binned patient number distributions for the numerical and low-cardinality categorial features (i.e., those with less than 10 categories), respectively. The mean, median, and standard deviations of the former and the unique value counts for all the categorical features are listed in detail in the

dashboard and EDA Jupyter notebook. From these two figures, it can be concluded that most patients in this dataset are older (60–80 years), highly comorbid (6–9 diagnoses), and managed with multiple medications, especially insulin and metformin.

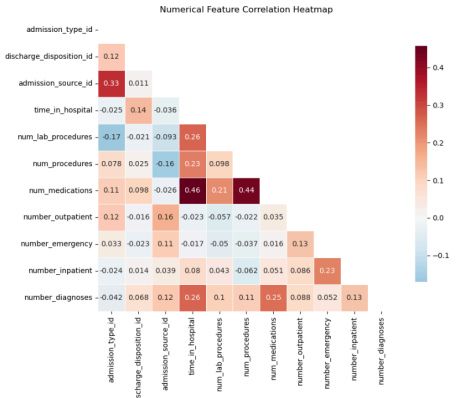

Fig. 4 Correlation heatmap between numerical features

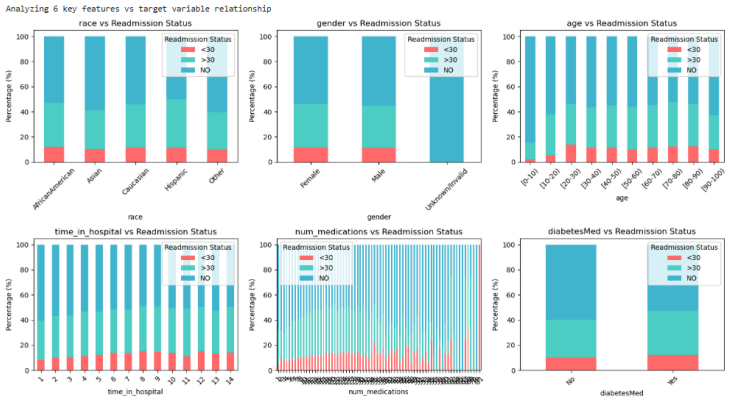

Fig. 5 Distribution of readmission status between different categories of six selected categorial features

Figure 4 shows the correlation between the numerical features. Highly correlated pairs of features include number of diagnoses with time in hospital, number of diagnoses with number of medications, number of medications with the number of procedures. **Figure 5** shows a percentage breakdown of the readmission status of the patients for each category for six selected categorical features. Overall, patients who are older and who spend more time in hospital have higher readmission rates.

Through this exploratory data analysis, we have gained the following key insights:
1. *Data Scale*: The dataset contains 101,766 samples and 50 features, with a moderate scale suitable for machine learning modeling
2. *Target Variable Distribution*: Readmission status shows an imbalanced distribution, requiring attention to class imbalance issues
3. *Data Quality*: There is a certain proportion of missing values that require appropriate preprocessing strategies
4. *Feature Diversity*: Contains both numerical and categorical features, requiring different encoding and standardization strategies

In the next phase, we will be focusing on
1. *Data Preprocessing*: Handle missing values, outlier detection and processing
2. *Feature Engineering*: Categorical variable encoding, numerical variable standardization, feature selection
3. *Model Selection*: Consider classification algorithms that handle class imbalance
4. *Model Evaluation*: Use appropriate evaluation metrics (such as F1-score, AUC, etc.)

Accuracy: 0.43

Classification Report:

	precision	recall	f1-score	support
0	0.64	0.48	0.54	16426
1	0.15	0.38	0.21	3422
2	0.42	0.38	0.40	10682
accuracy			0.43	30530
macro avg	0.40	0.41	0.39	30530
weighted avg	0.51	0.43	0.46	30530

Table 2. Classification results for preliminary logistic regression model

Accuracy: 0.54

Classification Report:

	precision	recall	f1-score	support
0	0.57	0.82	0.67	16426
1	0.20	0.02	0.04	3422
2	0.43	0.26	0.33	10682
accuracy			0.54	30530
macro avg	0.40	0.37	0.35	30530
weighted avg	0.48	0.54	0.48	30530

Table 3. Classification results for preliminary random forest model

We have performed preliminary analysis using two machine learning models comprising logistic regression and random forest, for which the classification results are shown in **Tables 2** and **3**, respectively. 70% of the data set was used to train the models with the remaining 30% used as test data. In future, the accuracy of both models will be improved with further data cleaning and pre-processing. Overall, the random forest demonstrated better precision and recall than the logistic regression model for predicting NO and >30 readmission categories. For the <30 category, Logistic Regression Model performed better in prediction.

Github repository and Dashboard

The raw .ipynb and .py files used for the EDA as well as the interactive dashboard are accessible at https://github.com/siuzhuobin/Team12_IT5006_Healthcare_Analytics_AY2526.

The interactive dashboard is accessible at https://team12it5006healthcareanalyticsay2526.streamlit.app/.