Predicting all-Cause 30-Day ICU Readmissions Using Robust Feature Selection

Arash Pakbin
a.pakbin@tamu.edu

Parvez Rafi
rafiparvez@tamu.edu

Sishir Subedi
sishirsubedi@tamu.edu

***Abstract*—Immature ICU discharge leads to early readmission. Most of the ICU early readmissions can be avoided and a significant amount of treatment costs can be saved. To be able to successfully predict readmission, prediction model should be interpretable so that the issues ignored while discharging a patient from ICU are properly identified. Complex machine learning models are not easily interpretable and consequently they are ineffective for clinicians. Therefore, our aim was to develop a prediction model which achieves a high discriminating characteristic and also is interpretable enough so it can be used to prevent the occurrence of early readmissions. In this project, we separated feature selection and modeling steps to build a robust predictive pipeline to predict 30-day all-cause ICU readmissions. Our model identified 24 highly predictive features which at best achieved prediction with 90% accuracy using simple logistic regression.**

Keywords—Intensive care units, ICUReadmissions, Machine Learning.

I. Introduction

A. Motivation

An Intensive Care Unit (ICU) is a special unit in hospitals where people with severe and life-threatening illnesses and injuries are transferred from patient wards to provide them critical and intensive treatment. After specific criteria are met, the patients are discharged from ICUs and returned to their wards. Intensive care is expensive[16, 4, 14, 1], and accounts for around 30% of total hospital costs and 1% of the US gross national product [16, 14]. Prudent decision making is required regarding admitting patients to and discharging from ICUs. If patients are discharged prematurely, it may result in inadequate levels of monitoring as well as early readmissions to ICUs [4, 2, 15, 12]. Internationally, 6-7% of people get readmitted to the ICU within 72 hours of being discharged[11]. Although unplanned ICU readmissions are uncommon, they have been linked to clinically adverse events, longer hospital stays and higher mortality [4, 8]. The Centers for Medicare and Medicaid Services (CMS) in the United States reported that 76% of hospital readmissions, which occur within 30-days, are potentially avoidable[3]. To reduce avoidable readmission rates, the Affordable Care Act of 2010 established the Hospital Readmissions Reduction Program, under which CMS could reduce payments to hospitals with high readmission rates [10]. Since then, hospitals and academics groups have been investing resources in identifying patients with high risk of early readmissions [5]. Therefore, predicting 30-day ICU readmission of patients would not only strengthen clinical decision making about whether a patient should be discharged from the ICU, but can also potentially reduce high costs associated with ICUs. Widespread adoption of Electronic Health Records (EHR) by hospital over last 10 years has provided great opportunity to develop clinical decision support systems by analyzing digital data of patient vitals, lab results, demographics and past diagnoses.

This work aims to provide an ensemble of different machine learning models trained from EHR data to provide predictions about whether patients are at risk of being readmitted to ICU within 30-days from discharge. We also provide an ensemble of different feature rankings models to generate a selection of most important predictors contributing to these predictions, making these predictions interpretable.

B. Related Work

In [7], Hoogendoorn et al. have compared two different approaches regarding mortality prediction which were predictive modeling and patient similarity approach. They have shown that predictive modeling outperforms the other one. They have used logistic regression on features picked by L2-regularization as their model to predict mortality. Consequently, we have used predictive modeling rather than patient similarity approach to build our model. In [6], Ghassemi et al. have shown that using latent topic features as well as structured features achieves the best performance in predicting the mortality. We have also included some latent topic features such as SAPS, SAPS II, SIRS, OASIS and LODS in our features to see if they show up in the top ranked features which are used in training our model.

C. Hypothesis

Our two layered approach of feature selection and predictive model will lead to the identification of an accurate predictive model with interpretable results. Selecting best features prior to the prediction will add a better understanding of how different risk factors contribute to the outcomes.

D. Model

In this project we will be using a two layer approach to build our predictive model. Our final model consists of a combination of a feature selection process and then prediction process. Different feature selection methods known to work best for this goal will be employed to produce multiple subsets of features. Additionally, these subsets will be analysed to understand their inherent relationships such as some specific features being repeated in all of the subsets which shows that they are more predictive. Once, we have good subsets of features, we will use them to train various predictive models and we will compare and analyze the results to reach the best possible combination for predicting ICU readmission within 30 days for MIMIC III dataset.

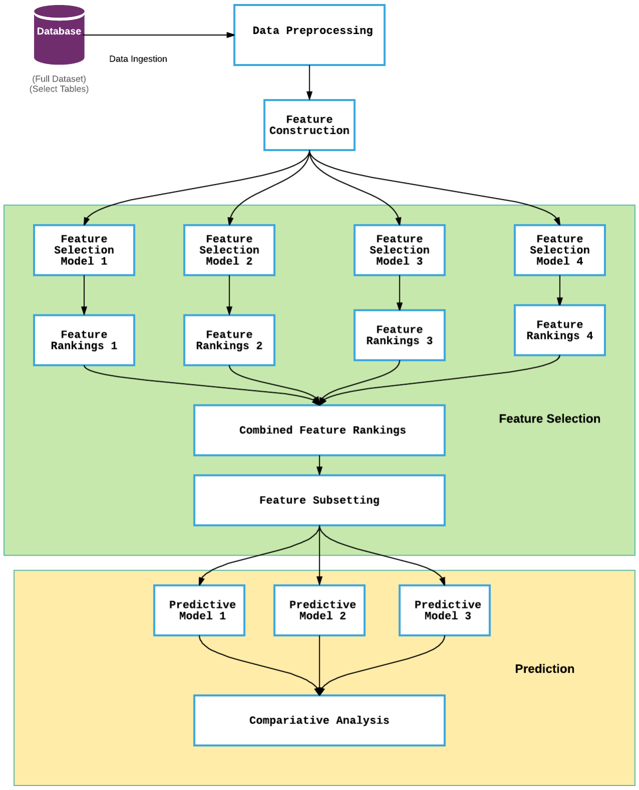


Fig. 1. Model Pipeline.

II. Data and Methods

The data analysis method aims to predict all cause 30 day ICU readmissions through robust feature selection and comparative modeling.

A. Dataset

The electronic health records collected between 2001 and 2012 at Beth Israel Deaconess Medical Center, Boston, MA, USA was used a data source for this analysis. It consists of de-identified 58,000 hospital admissions for 38,645 adults and 7,875 neonates. There were 34005 unique patients with ICU admissions. Among them, 5751 patients were readmitted two or more times. Since we were interested to predict ICU readmission within 30 days, we selected 1076 adult patients who had second ICU readmission record with less than 30 days. Figure 2 shows the readmission days distribution among the patients with two or more readmissions within 30 days. We analyzed demographics, laboratory events, chart events, and severity scores for potential predictor variables. All the available features for demographics and labortory events were used, while 20 features from charts events were selected based on previous work [13]. Features with more than 25% NAs were removed due to missing data. Table 1 describes a subset of 168 predictors collected from the database. Numerical variables were mean imputated while categorical variables were imputated with proportional distribution. Additionally, correlation analysis was conducted and 45 features with correlation 0.8 and higher were randomly removed resulting in a final modeling dataset with 123 features. Similarly,50:50 control-case sampling from the original dataset was conducted to randomly select 1076 patients with no readmissions.

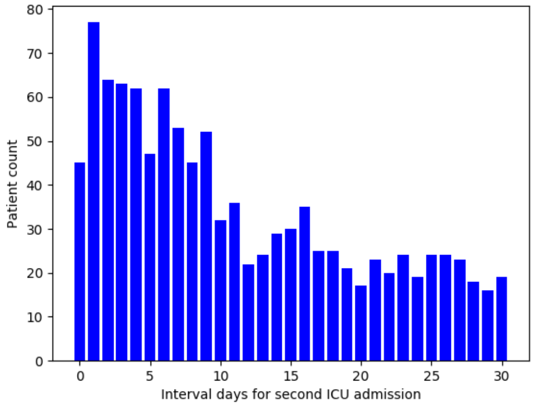


Fig. 2. Age distribution of patients with two or more ICU readmission.

Group	Predictors
Demographics	Age, Sex, Marital Status
Laboratory Events	Urea, Platelets, Magnesium, Albumin, Calcium, and others totaling 753 features
Charts Events	RespRate, Glucose, HR, SysBP, DiasBP, temp
Severity Scores	SAPS, SAP-II, SOFA

TABLE I. Subset of 168 Predictors used for Feature Selection.

B. Methods

The analysis pipeline consists of four major steps- feature selection, feature ranking, feature analysis, and comparative modeling.

1) Feature Ranking

To rank all the predictive features, we used a recursive feature elimination(RFE) algorithm [9]. RFE uses the coefficients of a linear model to select features by recursively removing the least important features from the model. We used three different models - Logistic Regression, Linear Support Vector Machine, and XGBoost classifier along with RFE to produce feature ranking sets. Additionally, we used a Least Absolute Deviations Basis Function and optimized the model using sum of squared to obtain a ranking for each feature based on its coefficients. KMeans clustering was used to determine the parameter(number of center) of the basis function (Figure 2). Finally, we had four sets of feature ranking using the above approaches. For each feature v in each model m was given equal weights and its rank across all models n were summed to obtained a final feature ranking f_v as following:-

$$f_v = \sum_{i=1}^{n} \frac{m_i(v)}{n}$$

The feature with the lowest value is ranked as the most important feature.

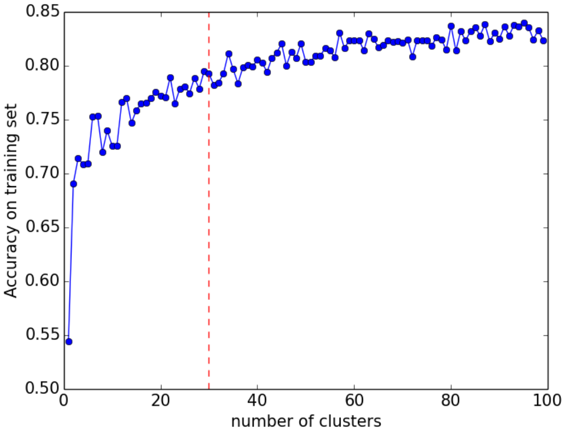


Fig. 3. KMeans clusering to optimize centers for feature selection using basis function.

2) Feature Selection

Ranked features obtained from the feature ranking analysis were used in a step wise Logistic Regression model to calculate the accuracy for adding a new featured in the model. Figure 4 shows the accuracy on validation set for the increasing number of features in the model. Top 24 features were selected as the best features based on the flattening of accuracy scores (Figure 3). Table II shows these features. Here, we note that many features derived from the same group were present, for example both minimum and maximum value for glucose were selected in the top 24 list.

Predictors
urea-n-max, glucose-min, sysbp-min, glucose-max, temp-max, Calcium-Total-max, hr-mean, urea-n-min, Chloride-var, sysbp-max, MCHC-min, PlateletCount-min, CalculatedTotalCO2-max, Chloride-min, PlateletCount-mean, Chloride-max, MCHC-mean, PlateletCount-var, Sodium-max, WhiteBloodCells-mean, Potassium-min, Glucose-min, Calcium-Total-var, MCHC-var

TABLE II. TOP 24 FEATURES SELECTED USING FEATURE SELECTION

The correlation analysis of the 24 features identified from the feature selection process showed that these features were highly uncorrelated with each other.

3) Feature Analysis

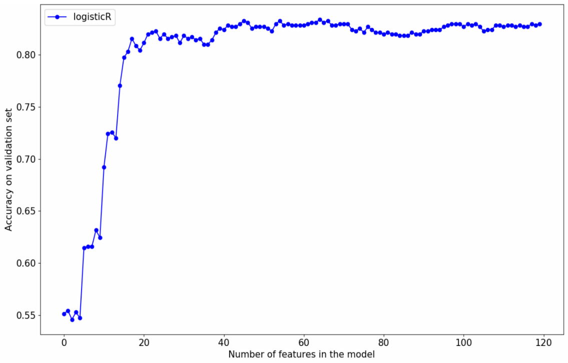


Fig. 4. Step wise feature selection from ranked features in the model.

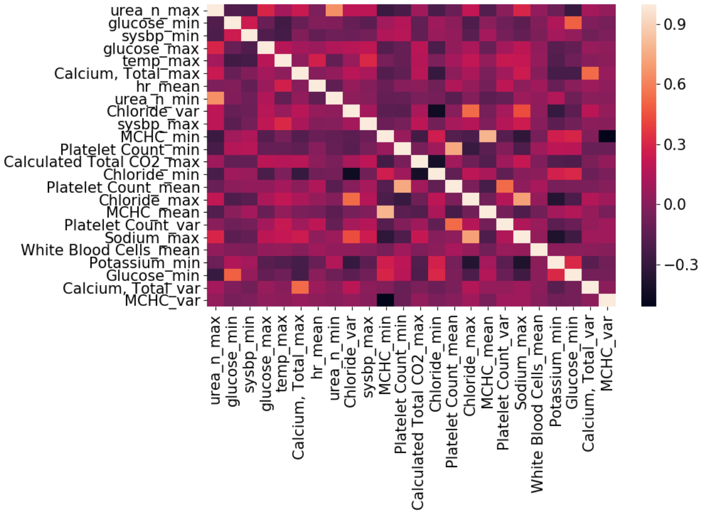


Fig. 5. Correlation plot of top 24 features identified.

We compared the top 24 features obtained from the feature selection procedure to the features selected by lasso (least absolute shrinkage and selection operator) regression. 7 out of 24 feature selected (alpha-0.033, accuracy-0.8) by lasso were present in the top 24 features list. This shows that about 30 percentage of features selected in our analysis were also selected independently by lasso analysis. This similarity highlights the accuracy of feature selection analysis and the importance of these common features in the prediction. These common features were - sysbp-min, Calcium-Total-max, Sodium-max, Chloride-min, Glucose-min, Potassium-min, MCHC-min.

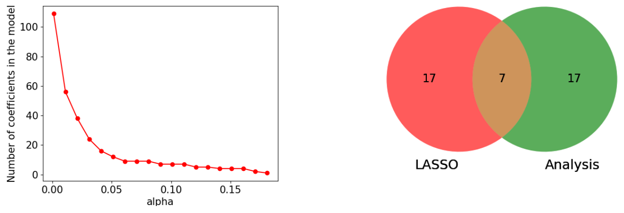
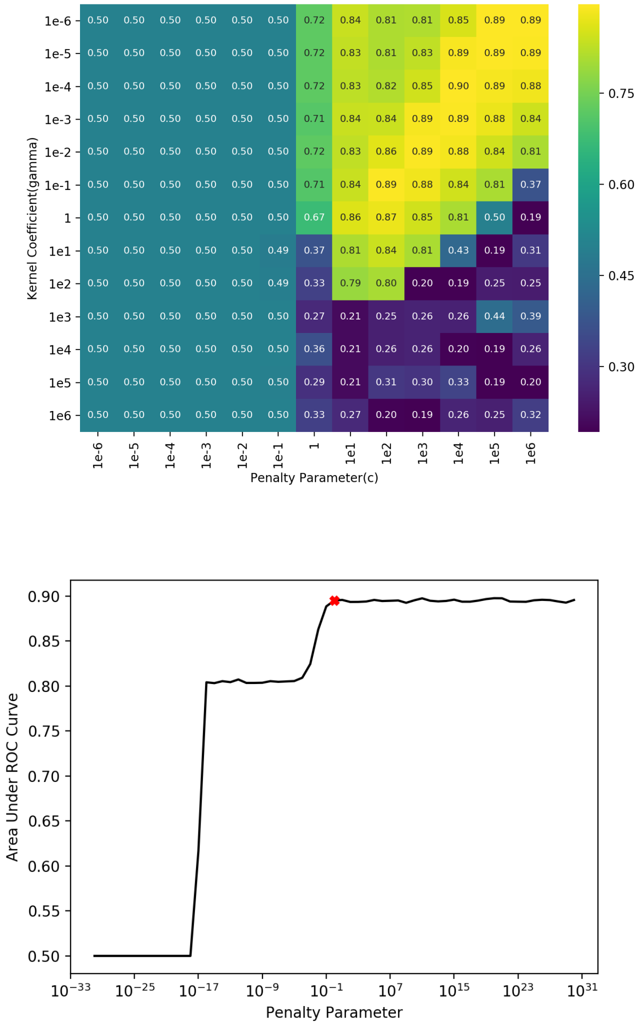


Fig. 6. Comparison between features obtained from LASSO and feature selection analysis.

4) *Comparative Modeling*

In this section, we used top 24 features to train three different models to predict re-admissions. We used a simple model as well as more complex and nonlinear models to carry out the prediction task. As our simple predictor, we used logistic regression and as our complex models we used XGBoost as well as SVM.
We used grid searching to find the optimum hyper-parameters of the models. It should be noted that the cross validation process of each specific value of the grid searching was carried out on the training set and the test set was set aside to measure the performance of the models. In Figure 7, we have shown the grid search results for different hyper-parameter values of SVM and Logistic Regression. The results of XGBoost grid searching could not be shown because of having 3 dimensions.



	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1	1e1	1e2	1e3	1e4	1e5	1e6
1e-6	0.50	0.50	0.50	0.50	0.50	0.50	0.72	0.84	0.81	0.81	0.85	0.89	0.89
1e-5	0.50	0.50	0.50	0.50	0.50	0.50	0.72	0.83	0.81	0.83	0.89	0.89	0.89
1e-4	0.50	0.50	0.50	0.50	0.50	0.50	0.72	0.83	0.82	0.85	0.90	0.89	0.88
1e-3	0.50	0.50	0.50	0.50	0.50	0.50	0.71	0.84	0.84	0.89	0.89	0.88	0.84
1e-2	0.50	0.50	0.50	0.50	0.50	0.50	0.72	0.83	0.86	0.89	0.88	0.84	0.81
1e-1	0.50	0.50	0.50	0.50	0.50	0.50	0.71	0.84	0.89	0.88	0.84	0.81	0.37
1	0.50	0.50	0.50	0.50	0.50	0.50	0.67	0.86	0.87	0.85	0.81	0.50	0.19
1e1	0.50	0.50	0.50	0.50	0.50	0.49	0.37	0.81	0.84	0.81	0.43	0.19	0.31
1e2	0.50	0.50	0.50	0.50	0.50	0.49	0.33	0.79	0.80	0.20	0.19	0.25	0.25
1e3	0.50	0.50	0.50	0.50	0.50	0.50	0.27	0.21	0.25	0.26	0.26	0.44	0.39
1e4	0.50	0.50	0.50	0.50	0.50	0.50	0.36	0.21	0.26	0.26	0.20	0.19	0.26
1e5	0.50	0.50	0.50	0.50	0.50	0.50	0.29	0.21	0.31	0.30	0.33	0.19	0.20
1e6	0.50	0.50	0.50	0.50	0.50	0.50	0.33	0.27	0.20	0.19	0.26	0.25	0.32

Fig. 7. Results of SVM and logistic regression grid searching to find the optimum hyperparameters.

Afterwards, we trained our model using the optimum hyper-parameters and whole training set. The models were then tested on the testing set and ROCs where computed and plotted as can be seen in Figure 8 .

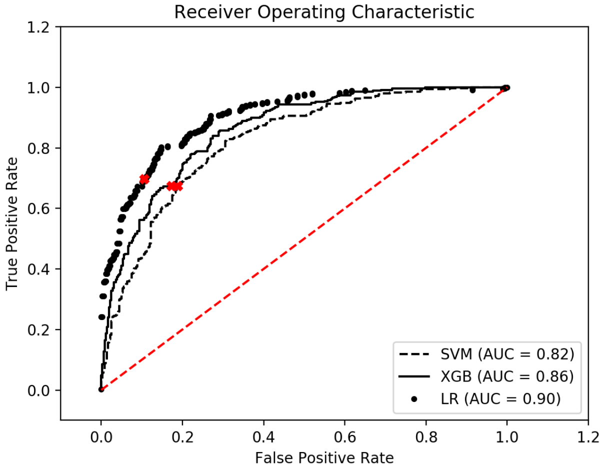


Fig. 8. ROC of XGBoost, SVM and Logistic Regression using 24 top features of the data.

It should also be noted that the red points on the ROC curves correspond to the best thresholds when maximizing with respect to fscore.

III. Conclusion

As can be seen, our model achieves a very high accuracy even using a simple model namely Logistic Regression. In addition to a high discriminating ability, our model is interpretable because of determining the top features before training our model. Figure 9 shows the distribution of values for each of top 24 features for readmission patients and non-readmission patients. The result shows that there is no significant difference between the marginal distributions of top 24 features in readmitted and non-readmitted patients. This result suggests that the difference lies in the conditional distribution of the features. Put another way, the difference between being readmitted and not being readmitted is due to the difference between the combinations of the values for different groups, given the marginal distributions are the same. This difference can simply be captured by a linear model such as Logistic Regression.

IV. Future Work

This study can be further improved by incorporating the following steps. First, we only included 15 important features from the charts event table due to the size limitation of our system. If we include all the features from the chart events table just like lab events in our model then it is likely that we will see improvement in our modeling. Second, the MIMIC database also consists of notes on each patient. This notes contain valuable information which can be used in the prediction. Thirdly, we can further expand this model into a time-series model where we can separately analyze data for

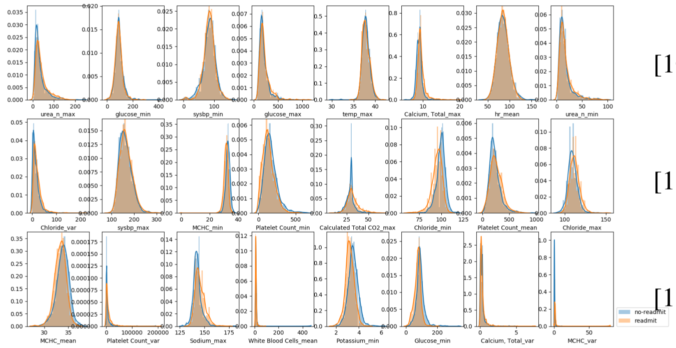


Fig. 9. Distribution of top 24 features for readmission and non-readmission patients.

each day and build a model which can predict readmission changes of a patient every day. Finally, we can also develop our predictive modeling into a real time modeling which keeps updating with the availability of new information and we can predict readmission probability of each patient at any time.

References

[1] Sydney ES Brown, Sarah J Ratcliffe, and Scott D Halpern. “An empirical derivation of the optimal time interval for defining ICU readmissions”. In: *Medical care* 51.8 (2013), p. 706.

[2] Carla A Chrusch et al. “High occupancy increases the risk of early death or readmission after transfer from intensive care”. In: *Critical care medicine* 37.10 (2009), pp. 2753–2758.

[3] Medicare Payment Advisory Commission et al. *Report to the Congress: promoting greater efficiency in Medicare*. Medicare Payment Advisory Commission (MedPAC), 2007.

[4] Thomas Desautels et al. “Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach”. In: *BMJ open* 7.9 (2017), e017199.

[5] Melanie Evans. “Healthcare’s’ moneyball’. Predictive modeling being tested in data-driven effort to strike out hospital readmissions.” In: *Modern healthcare* 41.41 (2011), pp. 28–30.

[6] Marzyeh Ghassemi et al. “Unfolding physiological state: Mortality modelling in intensive care units”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 75–84.

[7] Mark Hoogendoorn et al. “Prediction using patient comparison vs. modeling: A case study for mortality prediction”. In: *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE. 2016, pp. 2464–2467.

[8] F Shaun Hosein et al. “A systematic review of tools for predicting severe adverse events following patient discharge from intensive care units”. In: *Critical Care* 17.3 (2013), R102.

[9] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001–. URL: http://www.scipy.org/.

[10] Karen E Joynt and Ashish K Jha. “Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program”. In: *Jama* 309.4 (2013), pp. 342–343.

[11] Frida Kareliusson, Lina De Geer, and Anna Oscarsson Tibblin. “Risk prediction of ICU readmission in a mixed surgical and medical population”. In: *Journal of intensive care* 3.1 (2015), p. 30.

[12] Phillip D Levin et al. “Intensive care outflow limitation-frequency, etiology, and impact”. In: *Journal of critical care* 18.4 (2003), pp. 206–211.

[13] Oanh Kieu Nguyen et al. “Predicting all-cause readmissions using electronic health record data from the entire hospitalization: Model development and comparison”. In: *Journal of hospital medicine* 11.7 (2016), pp. 473–480.

[14] Andrew L Rosenberg and Charles Watts. “Patients readmitted to ICUs*: a systematic review of risk factors and outcomes”. In: *Chest Journal* 118.2 (2000), pp. 492–502.

[15] Norman Snow, Kathleen T Bergin, and Terrence P Horrigan. “Readmission of patients to the surgical intensive care unit: patient profiles and possibilities for prevention.” In: *Critical care medicine* 13.11 (1985), pp. 961–964.

[16] Evan G Wong et al. “Association of severity of illness and intensive care unit readmission: A systematic review”. In: *Heart & Lung: The Journal of Acute and Critical Care* 45.1 (2016), pp. 3–9.