Using Machine Learning Models to Predict the Likelihood of Patient Readmission to ICU

Student Team: Jack Wright (Aims), Ryan Hanks (Approach), Yang Zhao (Significance, Innovation), Arman Koul (Approach), Jiaxin Lin (Significance, Innovation)

Clinical PIs: Dr. Nauder Faraday and Dr. Adam Sapirstein

Engineering PIs: Dr. Joseph Greenstein, Dr. Raimond Winslow, Ran Liu

Problem Statement

ICU readmission is a large problem in the healthcare system as preemptive or late discharge from the ICU can be financially costly and risky for the patient's well-being. The current techniques and metrics being used to evaluate patient recovery are insufficient to effectively capture chance of readmission, nor do they implement all of the data available in the data-rich ICU.

Project Aims

Intensive Care Units (ICUs) cater to individuals with severe injuries and illnesses. Therefore, the patients within suffer from acute anatomic and physiologic derangements requiring constant monitoring and more intense support from hospital staff. Their conditions necessitate rapid diagnosis and intervention of abnormalities to facilitate a period of recovery. A challenge arises, however, in the recognition of sufficient resolution of the pathophysiologic state such that the patient can be safely discharged to a lower intensity environment. The decision to discharge is currently based on the expertise of ICU clinicians, but there is currently no formal method to assist clinicians predict a patient's chance of success of readmission, so this process is imperfect.[1-2] As such, ICU readmission rates range from 2-20%.[3] Readmission rates depend on a variety of factors, including demographic characteristics, comorbidities, severity of illness score, duration of index ICU stay, type of ICU, discharge destination, etc. Regardless of what factors led to readmission, a major problem manifests in the rates of in-hospital death for those who are readmitted to the ICU. Compared to patients who are successfully discharged, those who are discharged but return to the ICU are 2-10 times more likely to die in hospital.[4-5] Unfortunately, current predictive models are insufficiently accurate. In general, these models are built on static parameters and don't take advantage of the large amount of complex data available in EHR, nor do they take time varying covariates into account.[6-10] So, generating an algorithm that is able to use all of the available data in order to accurately predict readmission to the ICU would have several important impacts. By improving physicians' ability to determine resolution of the pathophysiologic state and reduce premature discharge, we could expect morbidity and mortality rates to decrease as well as reduced healthcare costs for patients. Similarly, the information a predictive model provides would allow for better allocation of resources, by letting hospital staff know which ICU's have higher readmission rates and require more attention. Finally, the features examined during this project could potentially be generalized for use in future predictive models.

Aim 1: Process the complex data available in EHR into a form more suitable for analysis

We will start by writing codes to parse the data so that we can analyze which patients are readmitted, where they are readmitted, and what measurements we have available. This will be done in multiple steps. First, we will determine the number of unique hospital visits occurred (rather than searching for patient IDs). Then, we will determine which ICUs those visits were to, and which of those visits resulted in a readmission. Lastly, we will collect patient data from separate files into a single data structure to gain a more holistic view of what values are available for each patient.

Aim 2: Determine features useful for identifying which patients are at risk of readmission.

From here, we will select features based on the following criteria: features that clinicians value as important indicators of patient recovery, any additional features found to be statistically important, and features derived from the first two categories, e. g. a time rate of change for a particular measurement.

Aim 3: Develop a predictive model to quantify the likelihood of ICU readmission.

Once we have determined which covariates we will use, our next step is to generate a risk score based on this data. We will begin by using linear regression procedures on the processed clinical data to predict the readmission of patients. Following this, we will consider other methods such as a neural network or machine learning based approach to potentially improve prediction accuracy.

Significance and Innovation

Significance

Approximately 6% of patients discharged from ICUs in developed countries will be readmitted during the same hospital stay; however, this value could be underestimated as it is based on all admissions rather than only the number of patients who were discharged alive.[11] It has been reported that the mortality of those patients who are discharged before sufficient treatment of illness and readmitted to the ICU is 2~10-fold higher than ones who are discharged from the ICU and require no readmission.[4, 12] Readmission can also cause patients to incur drastically higher hospital costs.[12] The limited beds and resources needed for other critically ill patients may be occupied by individuals who are ready to be released but still remain in the ICU. Currently, no formal method is available to predict the readmission rate. The electronic health record (EHR) of the ICU patients is widely used by clinicians to determine the release of patients from ICU.[13-14] However, clinicians always have a hard time determining whether the patients require further intensive monitoring and intervention or can be discharged from the ICU to monitor rooms or the general wards safely due to the large scale and complexity of the EHR database. Therefore, developing mathematical models that accurately predict the probability of patient readmission to the ICU can help both the patients and ICU clinicians in numerous ways.

Accurately predicting the readmission rate of patients to the ICU also helps to manage the hospital resource allocation. The hospital will have a better sense of how many beds should be placed in each ICU since different ICUs may have different readmission rates. The readmission rate of the ICU can also be regarded as an indicator of the hospital care quality and used to institute the policy or management in order to lower the readmission rate. In addition, some features in the mathematical models for predicting readmission rate have the potential to be generalized to other situations, for example, they can be used to predict the likelihood of readmission into the hospital. Moreover, with the mathematical model, the readmission rates can be monitored to identify if they exceed or fall below historical standards, which has major implications for hospital throughput.

Innovation

The Acute Physiology and Chronic Health Evaluation (APACHE) IV model has been widely used to predict the ICU length of stay and mortality for critically ill patient groups using logistic regression procedure, though the accuracy and utility are limited.[15] Another mortality prediction scoring system, the Sequential Organ Failure Assessment (SOFA), determines the possibility of death based on the degree of dysfunction of six organ systems.[16-17] However, there is no validated and widely accepted model for predicting the probability of ICU readmission.[1-2] Both of these scoring systems are interested in testing stress response and its relationship to mortality. The stress response relates to the hormonal and physiological responses the body has to a stressor. We hypothesize that this stress response will have significant application to ICU readmission and therefore SOFA and APACHE are good parallels. But the primary objective of our project is to predict the likelihood of ICU readmission instead of the mortality of patients.

The readmission rates relate to several variables, including severity of the illness score, duration of the ICU stay for the first admission, type of ICU admitted, etc. The previous models which are used to predict ICU readmission focus on some static parameters, such as variables at admission and/or discharge from the ICU of the patients. However, none of these models consider time-dependent covariates (differences in the state of critical illness at admission and discharge from the ICU) as important parameters in the predictive models. Involving the time varying parameters may be helpful in improving the accuracy of the previous models designed to predict ICU readmission. Moreover, we hypothesize that ICU readmission is strongly related to the stress response of the patient. Therefore, identifying features that contribute to stress response will help in examining ICU readmission.

Approach

We plan to build a statistical model that uses time series data to predict a patient's chance of readmission to the ICU. We've broken down the construction of this model into three steps: understanding the data, selecting the features to use, and creating and testing models using different machine learning techniques.

Understanding the data is a critical step because it will provide a check to give the team confidence that our data is valid and will alert us to future difficulties we may have when building the model. It will also allow us to format the data in a way that is conducive to the later steps that we will do. Currently our dataset is not formatted for our intended use. The data set has numerous delimited text files that contain different types of data. This includes time series physiological data recorded in the ICU, categorical medical data, and qualitative clinician notes. Each file contains a particular type of data for the relevant Visit IDs and we need to aggregate and organize the data from these files so that unique Visit IDs are paired with their corresponding data. Our first step in this process has been determining which visits contain a discharge and readmission to the ICU within the first 48 hours and which don't. We've preliminarily selected 48 hours as a metric because we believe that readmissions within the first 48 hours are most indicative of an error in the decision to discharge a patient. Our second step in understanding the data involves stratifying the data by the ICU it was collected from and determining the readmission rate for each ICU in the Johns Hopkins Hospital. We will be ignoring the patients who die during their stay in the ICU as they do not have the opportunity to be readmitted. We must also consider and ignore patients that were readmitted to the ICU due to circumstances outside of the scope of their initial ailment. For instance, if a clinical error outside of the ICU caused a readmittance.

As we progress in organizing the data from each file, we will be calculating relevant statistics and confirming that the data we collected is reasonable. For example, for each text file, we will determine how many visits this data was recorded for, the mean value of the data, the standard deviation of the data, and the range of the data. These statistics will allow the feasibility of the data to be confirmed by Dr. Faraday and Dr. Sapirstein as well as by comparison to published statistics. It will allow us to filter out anomalies in our dataset. We have completed this process for the "Location.txt" file which tracks patient location over the duration of a hospital visit. Through this analysis, we have determined that the total number of visits is 10,285 and the number of readmissions is 647. From these figures, we can see that the readmission rate for our data set is 6.29% which is consistent with the national range of 2-20%3. The main goal of the "understanding the data" phase of the project is to successfully stratify our dataset to be localized in one spot and ensure that this data is usable and feasible.

The next step in our project will be selecting the features we believe to be pertinent in predicting ICU readmission. We are operating under the hypothesis that stress response and ICU readmission are correlated. Therefore, identifying features that are indicative of stress response is equivalent to identifying features that are indicative of ICU readmission. We will select these features in a number of ways. We will first investigate the features that are used by the current stress response scoring systems, SOFA and APACHE to help in our decision in which features to collect. We will also consult with Dr. Faraday and

Dr. Sapirstein to determine which features they deem significant through their clinical experience. After this, we will begin collecting derived features that will be obtained from our data set. Derived features will distinguish our model for readmission prediction from existing stress response models like SOFA and APACHE. Derived features will be calculated by creating formulas to consider multiple variables at one time. Because the ICU collects a significant amount of time series data, we predict that many of our derived features will involve a time component. By taking into account this temporal component, we will be able to more accurately describe the physiological state of the patient. An example of a derived feature that we are looking to implement is change in white blood cell count. The static white blood cell count itself is a useful feature, but many doctors look at the change in white blood cell count to be a good indication of ailment resolution. We will be looking to implement a number of derived features that involve percent change, absolute change, or delta of a certain metric over time. After collecting these features, we will use Lasso feature selection and thresholding to determine which of these features are statistically significant in predicting ICU readmission. We will also look at dimensionality reduction to potentially make our features feasible inputs into the models we are trying to build.

Once we select the features we believe are pertinent, we will begin building a Generalized Linear Model around the identified features. We are looking for our model to output a risk score ranging from 0-1. An output value higher than .5 will indicate a prediction of an unsuccessful discharge and a value lower than .5 will indicate a prediction of a successful discharge. This will give us a binary output and after testing our model on our testing data, we can use an ROC curve to get metrics as to how successful our model is at predicting ICU readmission. After testing our initial model, we will go back and look at the features we selected to see if there is an improvement that can be made. We will particularly look at the derived features as they are the most variable. Through a process of informed trial and error, we will refine our feature list through testing our model. We are also open to trying different regression techniques such as neural networks to try and build an optimally accurate model. Creating the derived features will be the most nuanced and different part of our approach and it will largely be figured out through informed trial and error. The success that we can potentially see from the use of derived features has far-reaching significance and it will be exciting to see what features we find to be significant in predicting a successful ICU discharge.

Milestones &Timeline

Research Aims		Checkpoint
AIM1	**Process the complex data available in EHR into a form more suitable for analysis**	**Nov. 30**
1.1	Pair the Visit IDs in the file with the corresponding data	Oct. 25
1.2	Determine the visits containing a discharge and a readmission to the ICU within the first 48 hours and others do not	Nov. 1
1.3	Stratify the data by the ICU (CCU, CVSICU, MICU etc.) and determine the readmission rate for each ICU in the Johns Hopkins Hospital	Nov. 15
1.4	Calculate the relevant statistics and verify statistics with clinicians and published values	Nov. 30
AIM2	**Determine features useful for identifying which patients are at risk of readmission.**	**Feb. 28**
2.1	Identify features that are indicatives of stress response	Dec. 14
2.2	Investigate the features that are used by the current scoring systems (SOFA and APACHE)	Feb. 5
2.3	Determine the clinical meanings of the features from 2.2	Feb. 20
2.4	Collect the derived features and determine the statistical significance of these features	Feb. 28
AIM3	**Develop a predictive model to quantify the likelihood of ICU readmission.**	**May. 15**
3.1	Build a GLM using the identified data/features	Mar. 10
3.2	Examine the risk score of the model	Mar. 21
3.3	Model testing	Apr. 2
3.4	Generate a ROC curve to see how successful the model is at predicting ICU readmission	Apr. 14
3.5	Check the selected features based on the testing results	Apr. 28
3.6	Use different techniques to build the model	May. 15

References:

(1) Jones, A. E., Trzeciak S., K. J. The Sequential Organ Failure Assessment Score for Predicting Outcome in Patients with Severe Sepsis and Evidence of Hypoperfusion at the Time of Emergency Department Presentation*. *Crit. Care*2010,*37*(5), 1649–1654.

(2) Jo, Y. S.; Lee, Y. J.; Park, J. S.; Yoon, H. Il; Lee, J. H.; Lee, C.; Cho, Y. Readmission to Medical Intensive Care Units : Risk Factors and Prediction. 2015, *56*(2), 543–549.

(3) Van Sluisveld, N.; Bakhshi-Raiez, F.; De Keizer, N.; Holman, R.; Wester, G.; Wollersheim, H.; Van der Hoeven, J. G.; Zegers, M. Variation in Rates of ICU Readmissions and Post-ICU in-Hospital Mortality and Their Association with ICU Discharge Practices. *BMC Health Serv. Res.*2017, *17*(1).

(4) De Araujo, T. G.; De Mello Rieder, M.; Kutchak, F. M.; Filho, J. W. F. Readmissions and Deaths Following ICU Discharge - A Challenge for Intensive Care. *Rev. Bras. Ter. Intensiva.*2013, *25*(1), 32–38.

(5) Rosenberg, A. L. Patients Readmitted to ICUs. A Systematic Review of Risk Factors and Outcomes. *CHEST J.*2000, *118*(2), 492.

(6) Piper, G. L.; Kaplan, L. J.; Maung, A. A.; Lui, F. Y.; Barre, K.; Davis, K. A. Using the Rothman Index to Predict Early Unplanned Surgical Intensive Care Unit Readmissions. *J. Trauma Acute Care Surg.*2014, *77*(1), 78–82.

(7) Lissauer, M. E.; Diaz, J. J.; Narayan, M.; Shah, P. K.; Hanna, N. N. Surgical Intensive Care Unit Admission Variables Predict Subsequent Readmission. *Am. Surg.*2013,*79*(6), 583–588.

(8) Cardoso, F. S.; Karvellas, C. J.; Kneteman, N. M.; Meeberg, G.; Fidalgo, P.; Bagshaw, S. M. Respiratory Rate at Intensive Care Unit Discharge after Liver Transplant Is an Independent Risk Factor for Intensive Care Unit Readmission within the Same Hospital Stay: A Nested Case-Control Study. *J. Crit. Care*2014, *29*(5), 791–796.

(9) Van Diepen, S.; Graham, M. M.; Nagendran, J.; Norris, C. M. Predicting Cardiovascular Intensive Care Unit Readmission after Cardiac Surgery: Derivation and Validation of the Alberta Provincial Project for Outcomes Assessment in Coronary Heart Disease (APPROACH) Cardiovascular Intensive Care Unit Clinical Prediction Model from a Registry Cohort of 10,799 Surgical Cases. *Crit. Care*2014, *18*(6).

(10) Ponzoni, C. R.; Corrêa, T. D.; Filho, R. R.; Neto, A. S.; Assunção, M. S. C.; Pardini, A.; Schettino, G. P. P. Readmission to the Intensive Care Unit: Incidence, Risk Factors, Resource Use, and Outcomes: A Retrospective Cohort Study. *Ann. Am. Thorac. Soc.*2017,*14*(8), 1312–1319.

(11) Kramer, A. A.; Higgins, T. L.; Zimmerman, J. E. The Association between ICU Readmission Rate and Patient Outcomes. *Crit. Care Med.*2013, *41*(1), 24–33.

(12) Dasta, J. F.; McLaughlin, T. P.; Mody, S. H.; Piech, C. T. Daily Cost of an Intensive Care Unit Day: The Contribution of Mechanical Ventilation. *Crit. Care Med.*2005, *33*(6), 1266–1271.

(13) Calvert, J.; Mao, Q. Q.; Hoffman, J. L.; Jay, M.; Desautels, T.; Mohamadlou, H.; Chettipally, U.; Das, R. Using Electronic Health Record Collected Clinical Variables to Predict Medical Intensive Care Unit Mortality. *Ann. Med. Surg.*2016, *11*, 52–57.

(14) Curtis, J. R.; Sathitratanacheewin, S.; Starks, H.; Lee, R. Y.; Kross, E. K.; Downey, L.; Sibley, J.; Lober, W.; Loggers, E. T.; Fausto, J. A.; et al. Using Electronic Health Records for Quality Measurement and Accountability in Care of the Seriously Ill: Opportunities and Challenges. *J. Palliat. Med.*2017, jpm.2017.0542.

(15) Zimmerman, J.; Kramer, A.; McNair, D.; Malila, F. Intensive Care Unit Length of Stay: Benchmarking Based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Crit Care Med*2006, *34*, 2517–2529.

(16) Matsumura, Y.; Nakada, T. A.; Abe, R.; Oshima, T.; Oda, S. Serum Procalcitonin Level and SOFA Score at Discharge from the Intensive Care Unit Predict Post-Intensive Care Unit Mortality: A Prospective Study. *PLoS One*2014, *9*(12).

(17) Raith, E. P.; Udy, A. A.; Bailey, M.; McGloughlin, S.; MacIsaac, C.; Bellomo, R.; Pilcher, D. V. Prognostic Accuracy of the SOFA Score, SIRS Criteria, and QSOFA Score for in-Hospital Mortality among Adults with Suspected Infection Admitted to the Intensive Care Unit. *JAMA - J. Am. Med. Assoc.*2017, *317*(3), 290–300.