Using natural language processing and machine learning on caregiver notes to predict of ICU length of stay for emergency admissions

A Preprint

Anna Dupree
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90023
afdupree@ucla.edu

Carlotta Gherardi
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90023
cagherardi@gmail.com

Claire Hayashida
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90023
clairehayashida@ucla.edu

Neha Jonnalagadda
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90023
jvneha12@gmail.com

Allison Lynn
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90023
allisonlynn@g.ucla.edu

Olivia Motmans
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90023
olivia.motmans@gmail.com

Cassidy Sadowski
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90023
cassidysadowski@g.ucla.edu

March 13, 2025

Abstract

Assessing the severity of illness or injury is a critical challenge in healthcare decision-making. This study introduces a data-driven approach to evaluating patient acuity by using ICU length of stay (LOS) as a proxy for condition severity. We apply machine learning techniques, specifically regression models based on BioBERT and TF-IDF text processing, to predict ICU duration based on clinical caregiver notes. These notes capture both quantitative medical assessments and qualitative observations made by healthcare providers.

If successful, this approach could help with early identification of high-risk patients and bridge the gap between self-reported symptoms and clinical evaluations. Advanced NLP methodologies could be employed to analyze semantic similarity between patient-reported symptoms and caregiver notes. Using our model and this semantic bridge, we could provide a preliminary estimate of illness or injury severity. This method has the potential to improve triage efficiency, optimize resource allocation, and support decision making in emergency and critical care settings.

Keywords nlp · MIMIC · ICU LOS · machine learning · BERT · bioBERT · TF-IDF

1 Introduction

Suppose someone in rural Illinois is having chest pains and panicking due to said chest pains. The nearest hospital is an hour's drive away, and the cost of an emergency room visit is high. What can they do to determine whether or not it is worth it to make the trip?

1.1 Purpose

Our proposed model is a straightforward exploration of the relationship between clinical caregiver notes and ICU length of stay. However, the potential impact is much larger. Future studies could extend this model to bridge patient-reported symptoms with clinical medical notes, using advanced NLP methods. This semantic bridge, aligned with our model, could offer a preliminary estimate of illness severity.

In clinical settings, such a model could help medical providers with triage and patient prioritization in an emergency care setting. A sufficiently advanced model may even be helpful in resource allocation in resource-limited settings (e.g. assignment of ICU beds).

For patients, an extended model could support decision-making in seeking emergency care, which may reduce delays in treatment or prevent unnecessary hospital stays.

The progression of human knowledge is made with small steps. Past exploration of machine learning in medical applications has by no means been exhaustive. We aim to find one small but practical application of machine learning that could be integrated into a larger healthcare context, and, ideally, help make medical knowledge more accessible. For our imaginary Illinoisian, this tool has the potential to both reduce treatment delays and give a cost-effective approach to managing their healthcare needs.

1.2 Background

Numerous studies have been conducted on predicting ICU length of stay (LOS). Previous studies modeling LOS based on medical data have found up to a 94% accuracy in classification of LOS through machine learning methods 1, as well as determined primary predictive factors of LOS to be from both physiology and disease 2, and the day of data collection 3.

Use of natural language processing (NLP) in conjunction with machine learning methods have also shown promising results. One study predicted ICU 30-day readmission risk achieved with an AUROC curve of 0.74, using a nltk.word_tokenize function on discharge summaries from the MIMIC III dataset, and applying a support vector machine with radial basis function kernel (SVM-RBF) model 4. Another study found correlation between sentiment scores from nursing notes and 28-day in-hospital mortality among sepsis patients, using the TextBlob library 5.

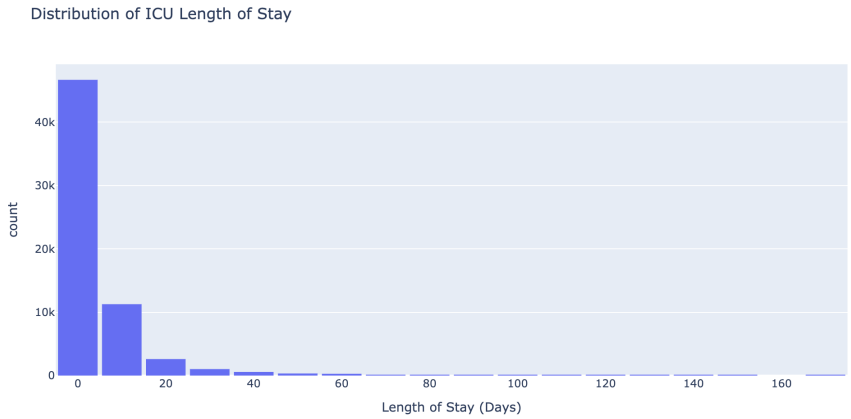
Despite inclusion of length of stay (LOS) as a feature in these predictive models, dedicated research on using NLP and sentiment analysis to predict LOS as a continuous variable remains limited. One study derived text variables to predict a binary outcome encompassing both an ICU LOS over 7 days or in-hospital death, where text data improved model performance; yet, mortality was the primary focus and LOS was not directly predicted 6.

Prediction of LOS can be valuable to both patient and provider. As ICU stays can be extremely expensive, providing patients and their families with a reliable estimate of stay can help them prepare both financially and emotionally. For providers, a reliable LOS prediction could be useful to improve patient flow and optimize resource utilization. Recognizing this gap, our study aims to apply NLP methods to model ICU LOS as a continuous outcome, thereby addressing an under explored area in critical care research.

2 Method

Our design steps for this study were: extract caregiver notes from the dataset; use BioBERT, Term Frequency-Inverse Document Frequency (TF-IDF), and Textblob for text processing; create a regression model with the processed text to predict ICU LOS.

Figure 1: Distribution of ICU LOS in the dataset.



2.1 Data

The MIMIC III database contains deidentified, health-related data of over 40,000 patients of the Beth Israel Deaconess Medical Center between 2001 and 2012 7, 8, 9. We used the following comma-separated value (.csv) files provided by MIMIC III v.1.4: ADMISSIONS.csv, ICUSTAYS.csv, NOTEEVENTS.csv, PATIENTS.csv.

The variables used in our model are listed below, and taken directly from the MIMIC III database.

- Hospital admission type
- ICU admission and discharge times
- Caregiver notes from discharge summaries

After extracting our target data, we cleaned caregiver note text by removing appropriate whitespace and placeholders used for protected health information (PHI).

Due to computational constraints, we limited our total combined observations for the training and testing sets to the first 10,000 observations out of over 60,000 that matched our target criteria. For sentiment analysis using Textblob, we used only 5,000 total observations for our training and testing sets.

2.2 Text Processing

In our study, we explored three different text processing methods: BioBERT, TF-IDF, and TextBlob.

2.2.1 BioBERT

BERT is an open source, deep learning language model introduced by Google in 2018. It improves natural language processing by considering both left and right context simultaneously. BioBERT is a BERT model pre-trained on a large corpus of biomedical literature 10. We used the BioBERT model to tokenize, vectorize, and transform our text data into embeddings to capture meaningful relationships between terms. Using BioBERT allows interpretations of clinical caregiver notes to account for nuances specific to the medical field.

2.2.2 TF-IDF

Term Frequency-Inverse Document Frequency or TF-IDF is a method to evaluate the importance of a word within a document, relative to a larger collection of documents. A term that is frequently seen across all documents would have a low TF-IDF score, and a term that is infrequent across all documents would have a higher TF-IDF score. The formula for calculating TF-IDF consists of two components:

- Term Frequency (TF)

$$TF(t,d) = \frac{\text{number of times term } t \text{ appears in document } d}{\text{total number of terms in document } d}$$

- Inverse Document Frequency (IDF)

$$IDF(t, D) = \log\left(\frac{\text{total number of documents in corpus } D}{1 + \text{number of documents in } d \text{ which contain term } t}\right)$$

This gives us a TF-IDF score for each word that appears in each document:

$$\text{TF-IDF} = TF(t, d) \times IDF(t, D)$$

2.2.3 Textblob

As an experimental exploration of our model, we also incorporated sentiment analysis using the package TextBlob. A sentiment score between -1 and 1 was generated for each caregiver note in the training set, and we attempted to predict this sentiment score through a TF-IDF and regression pipeline.

2.3 Regression Models

Our target variable is the ICU length of stay (LOS) of the patient. This variable will be mapped on to the caregiver notes, and we will train our regression model using the methods of Random Forest, Lasso regularization, gradient boosting, and a neural network.

Because of the complicated nature of medical information, we are chose models that can capture more complex patterns. Random Forest, which uses a decision tree approach with a subset of variables and bootstrapped data, was our main focus across all text processing methods. A variation of this model using a logarithmic transformation of ICU LOS as the target variable was also tested. We also explored gradient boosting, which further reduces variance by iteratively improving weak learners and weighting misclassified points more heavily. Lasso regression, a regularized linear approach designed to prevent overfitting by shrinking coefficients, was included to assess the predictive power of a more constrained model. An MLPRegressor neural network, which optimizes using stochastic gradient descent to learn hierarchical patterns from data, was also tested with TF-IDF features. Finally, we experimented with a Random Forest model using sentiment scores derived from TextBlob as input features.

3 Results

A table summarizing all model results can be seen below in Table 1.

3.1 Summary

The BioBERT text embeddings, when utilized with different predictive models, demonstrated varying degrees of effectiveness in estimating the duration of stay in the ICU based solely on textual descriptions of discharge summaries. The random forest model produced a Root Mean Square Error (RMSE) of 22.159, with an R^2 value of 0.5083, indicating that approximately 50.83% of the variation in ICU stay length could be explained by the text embeddings alone. Gradient boosting showed a slight improvement, achieving a lower RMSE of 21.787 and a higher R^2 of 0.5246, meaning it captured 52.46% of the variability in ICU stay duration from the discharge summary embeddings. Conversely, lasso regression resulted in a higher RMSE of 22.906 and

Table 1: Summary of Model Results. Note that in most cases, y = ICU LOS.

Text Processor	Regression Model	RMSE	R-squared
BioBERT	Gradient Boosting	21.787	0.5246
	Random Forest	22.159	0.5083
	Lasso	22.906	0.4745
TF-IDF	Random Forest	8.766	0.5782
	Lasso	8.848	0.5685
	Neural Network	8.860	0.5674
	Random Forest w/ log(y)	9.509	0.5017
	Random Forest w/ y = TextBlob Sentiment Score	0.0423	0.3236

a reduced R^2 of 0.474, reflecting a slightly weaker ability to explain the length of stay in the ICU through textual features.

The TF-IDF text processing approach exhibited notable differences in its ability to estimate ICU length of stay based on discharge summaries. The random forest model yielded a Root Mean Square Error (RMSE) of 8.766 with an R^2 value of 0.5782, indicating that 57.82% of the variance in ICU stay duration could be explained using TF-IDF-derived textual features. Lasso regression performed similarly, with a slightly higher RMSE of 8.848 and a marginally lower R^2 of 0.5685, suggesting a comparable predictive power. A neural network model using stochastic gradient descent resulted in an RMSE of 8.860 and an R^2 of 0.5674, offering nearly identical performance to lasso regression. Incorporating a logarithmic transformation of the target variable in the random forest model led to a slightly increased RMSE of 9.509 and a reduced R^2 of 0.5017, reflecting a decline in explanatory power. Additionally, when predicting ICU stay duration based on sentiment scores derived from TextBlob, the random forest model produced an RMSE of 0.0423 and a much lower R^2 of 0.3236, indicating a limited ability to capture overall variance in ICU length of stay. It should be noted that sentiment scores range from -1 to 1; as such, the lower RMSE may not be a relatively lower metric than RMSE in other models.

3.2 Interpretation

Our results indicate that the effectiveness of different text-processing techniques and regression models varies significantly when predicting ICU length of stay (LOS) based on discharge summaries. The performance of models using BioBERT embeddings was moderate, with the gradient boosting model performing the best ($R^2 = 0.5246$, RMSE = 21.787), followed by random forest ($R^2 = 0.5083$) and lasso regression, which performed the worst ($R^2 = 0.4745$). Lasso regression also underperformed with TF-IDF, indicating that feature selection and sparsity constraints may have reduced predictive power in this context.

TF-IDF-based models outperformed BioBERT-based models across all standard regression approaches, achieving the highest explanatory power with random forest ($R^2 = 0.5782$, RMSE = 8.766). We initially expected that pre-trained contextual embeddings from BioBERT would outperform a basic term-frequency-based representation. It seems that simple frequency-based representations may retain clinically relevant signals more effectively than dense embeddings in this specific predictive task.

The use of log-transformed ICU LOS in a random forest model resulted in a reduction in explanatory power ($R^2 = 0.5017$), suggesting that ICU LOS may not follow a simple log-normal distribution. Additionally, using TextBlob sentiment scores as a predictor yielded the lowest explanatory power ($R^2 = 0.3236$), indicating that sentiment scores alone do not capture enough relevant clinical information to predict ICU LOS accurately.

These results suggest that tree-based models, particularly those trained on TF-IDF representations, provide the most effective approach for predicting patient outcomes based on caregiver discharge notes. We can also see that caregiver notes have significant explanatory power in predicting ICU LOS, demonstrated by the relatively strong performance of models trained on this one feature alone.

4 Conclusion

4.1 Shortcomings

Currently, our models rely on text embeddings derived from discharge summaries, which inherently capture a retrospective view of the patient's condition after their hospital course has concluded. While these embeddings provide valuable insights into the full scope of a patient's ICU stay, they do not offer real-time predictions that could assist clinicians or patients in planning for ICU duration at the time of admission.

Additionally, our tree-based models and neural networks are often considered black-box models, making it difficult to understand which features contribute most to predictions. While Lasso regression offers some interpretability by selecting key features, it is not the best performing model.

ICU LOS is inherently a time-dependent problem, meaning that changes in patient status over time could provide additional predictive power. The study does not account for this aspect.

4.2 Future Implementation and Recommendations

The variation in performance across different text processing techniques suggests that further refinement in feature selection may help predictive accuracy. Topic modeling with Latent Dirichlet Allocation (LDA)

could extract meaningful thematic clusters of words that correspond to clinical diagnoses, helping the models capture more information from caregiver notes.

Our proposed use of this model is to provide a fast, clinically test-free method for assessing condition acuity. The purpose would be to make predictions for patients before arriving at a hospital, or even to doctors with a preliminary condition assessment before triage. Incorporating structured clinical data such as lab results or imaging reports would contradict the intended application of the model. However, integrating non-invasive patient demographics, such as age, marital status, and religious background, may improve the model's effectiveness. Some vital signs that could be taken at home with simple devices might also increase model effectiveness while still balancing our goal of layperson accessibility.

Additionally, applying feature selection techniques to admission summaries rather than discharge summaries could significantly enhance the utility of ICU stay predictions. While this may decrease the explanatory power of the text-based variable alone, combining admission notes with patient demographics may have an overall higher predictive power.

References

Merhan A. Abd-Elrazek, Ahmed A. Eltahawi, Mohamed H. Abd Elaziz, and Mohamed N. Abd-Elwhab. Predicting length of stay in hospitals intensive care unit using general admission features. *Ain Shams Engineering Journal*, 12(4):3691–3702, 2021.

William A. Knaus, Douglas P. Wagner, Jack E. Zimmerman, and Elizabeth A. Draper. Variations in mortality and length of stay in intensive care units. *Annals of Internal Medicine*, 118(10):753–761, 1993. PMID: 8470850.

Andrew A. Kramer and Jack E. Zimmerman. A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay. *BMC Medical Informatics and Decision Making*, 10(27):1472–6947, 2010.

Negar Orangi-Fard, Alireza Akhbardeh, and Hersh Sagreiya. Predictive model for icu readmission based on discharge summaries using machine learning and natural language processing. *Informatics*, 9(1), 2022.

Qiaoyan Gao, Dandan Wang, Pingping Sun, Xiaorong Luan, and Wenfeng Wang. Sentiment analysis based on the nursing notes on in-hospital 28-day mortality of sepsis patients utilizing the mimic-iii database. *Computational and Mathematical Methods in Medicine*, 2021(1):3440778, 2021.

Gary E. Weissman, Rebecca A. Hubbard, Lyle H. Ungar, Michael O. Harhay, Casey S. Greene, Blanca E. Himes, and Scott D. Halpern. Inclusion of unstructured clinical text improves early prediction of death or prolonged icu stay. *Critical Care Medicine*, 46(7):1125–1132, 2018.

Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-iii clinical database (version 1.4), 2016.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Marzyeh Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, C. K. Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. [Online].

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.