# Team 6

# Predicting Discharge Dates for Hospitalized COVID-19 Patients: A Comparative Study of Machine Learning Models

**Likith Reddy Chintala , Nadia Siles, Sanjana Prasad, Shikhar Gupta**

## Introduction and Literature Review

In 2019, the outbreak of the novel respiratory virus SARS Coronavirus 2 (SARS-CoV-2) caused a global pandemic, leading to an unprecedented surge in hospitalizations, critical care admissions, and global excess mortality of 14.83 million deaths over 2020–2021 (Msemburi et al., 2023). This surge placed tremendous strain on healthcare systems worldwide, requiring the adoption of triage protocols to manage and prioritize patient care efficiently. Despite the United States ending the Federal COVID-19 Public Health Emergency on May 11, 2023, and lifting public health prevention and control interventions, the threat of future waves as herd immunity decreases remains a concern (Jacobsen et al., 2020). As new variants continue to emerge, there is an increasing risk of mutations that reduce the effectiveness of vaccines, potentially leading to a resurgence in cases that could overwhelm healthcare facilities (Gupta et al., 2023).

Accurate prediction of individual hospital patient trajectories is critical for better resource allocation and proactive interventions to improve patient outcomes during and after a pandemic. Various studies have explored this domain, employing different methodologies with varying degrees of success (Klein et al., 2023). IMPACC et al. (2022) introduced a novel approach, focusing on patients' trajectories based on their length of stay and oxygen consumption during hospitalization, identifying five distinct trajectories among 1,164 patients nationwide. Building on this work, Cole et al. replicated the trajectory grouping with a cohort of 21,000 patients within the Austin Ascension hospital network, demonstrating that electronic medical record (EMR) data with minimal cleaning and processing could be used to train a neural network model to predict mortality and severity in patients, aiding hospital resource planning and patient care strategies.

This paper aims to further advance the research initiated by Cole et al. by evaluating three machine learning models: linear regression (LR), feed-forward neural network (FFNN), and long short-term memory (LSTM). These models were used to predict patient discharge dates from hospitals. The study evaluated their performance using a dataset of 21,312 hospitalized COVID-19 patients from the Austin, Texas, Ascension Seton Hospital Network, focusing on predicting discharge times and mortality. This approach moves forward the efforts to apply machine learning to healthcare, particularly in managing infectious disease outbreaks and hospital logistics during ongoing and future health crises.

# Methods

### Electronic Medical Records (EMR) Data Preparation

Deidentified electronic medical records (EMR) data for 21,312 COVID-19 patients from March 2020 to September 2022 was obtained from the Ascension Seton Austin Hospital Network, including one children's hospital, three community hospitals, and one academic hospital. Patients were selected off having a confirmed COVID-19 diagnosis code (U07.1), with the first visit associated with this code being identified as the baseline admission (Day 0). All laboratory and vital data collected in at least 70% of patients on day 0 were included in the dataset. These became the set of 66 features for which daily data was used to train and test our model in addition to one hot encoded oxygen administration method and basic demographic information (sex, age, race, ethnicity, hospital admitted to).

### Data Preparation:

To validate our two hypotheses, we prepared two distinct dataset types - a time series dataset containing daily patient data and an aggregated dataset where features were summarized by taking the mean over the initial 48-hour period after admission.

### Time Series:

The time series dataset consisted of daily patient data in raw and processed formats. The raw version was composed of the initial data after integrating demographics, lab results, vital signs, and oxygen measurements for 20,000 patients, yielding 66 total features (13 non-time-series like age/weight and 53 time-series features). The processed version applied mean imputation to address missing values and scaling. For each patient, we constructed a 66x30 feature matrix covering 30 days, along with additional 66xN matrices to evaluate discharge prediction using N previous days' data. Since most discharges occurred within 30 days, the 66x30 matrices frequently contained trailing missing data imputed as -1.

### Aggregated:

For the aggregated dataset, we calculated the mean of all time-varying features (excluding categorical/demographic) over the initial 48-hour window post-admission. This condensed representation was then scaled analogously to the processed version of the time series data.

### Additional Preparation:

Several other preprocessing steps were undertaken. Deceased patients had a discharge date of 100, imputed as the maximal value. Min-max normalization was applied to non-time-series features in both datasets. We took the fourth root of these label values for prediction modeling purposes to mitigate the right skew from the discharge date distribution being concentrated in the first 10 days. These processed time series and aggregated datasets enabled the evaluation the relative predictive performance of granular versus summarized initial data representations.

**Exploratory Analysis**

Exploratory analysis involved examining the data for missing values, summarizing key numerical and categorical features, and visualizing distributions. The dataset included various demographic information, laboratory results, and vital signs recorded over time. Analysis revealed the following:

**Demographics:**

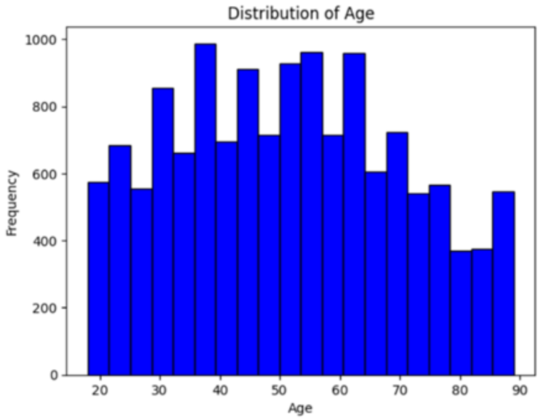
- A wide range of ages, with notable variability across different age groups.
- Ethnicity and race distributions were examined to understand demographic trends.

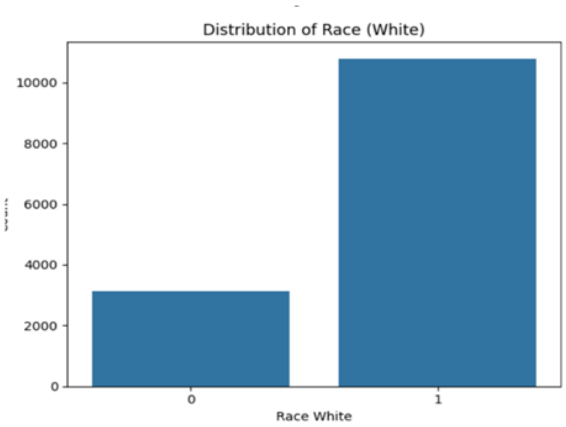
**Laboratory Data and Vital Signs:**

- The distribution of key lab results such as CRP, RBC, and Albumin was analyzed.
- Vital signs like mean arterial pressure and oxygen saturation were explored to understand their stability or variability over time.

Correlation analysis identified critical features associated with patient outcomes, including oxygen saturation, respiratory rate, hemoglobin count, and age. Based on this analysis, the hypothesis was that these features significantly impact discharge dates and can be used to train regression models for prediction.

**Observations:**





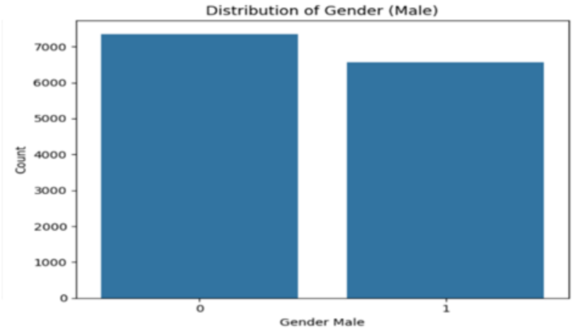


**Fig. 1.** Distribution of Demographics

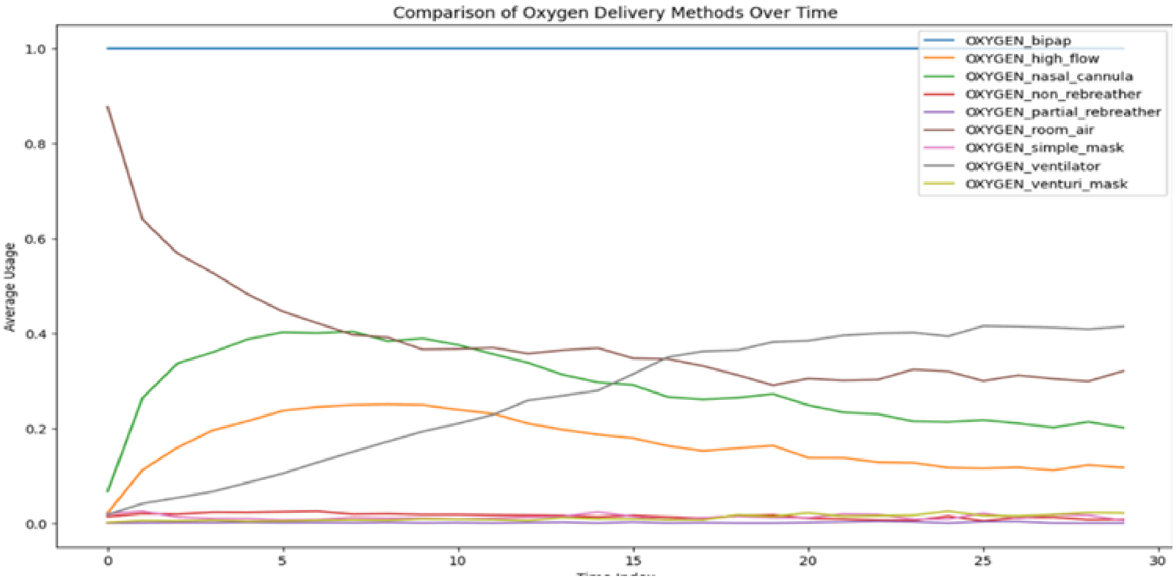


**Fig. 2.** The average usage of different oxygen delivery methods over time.

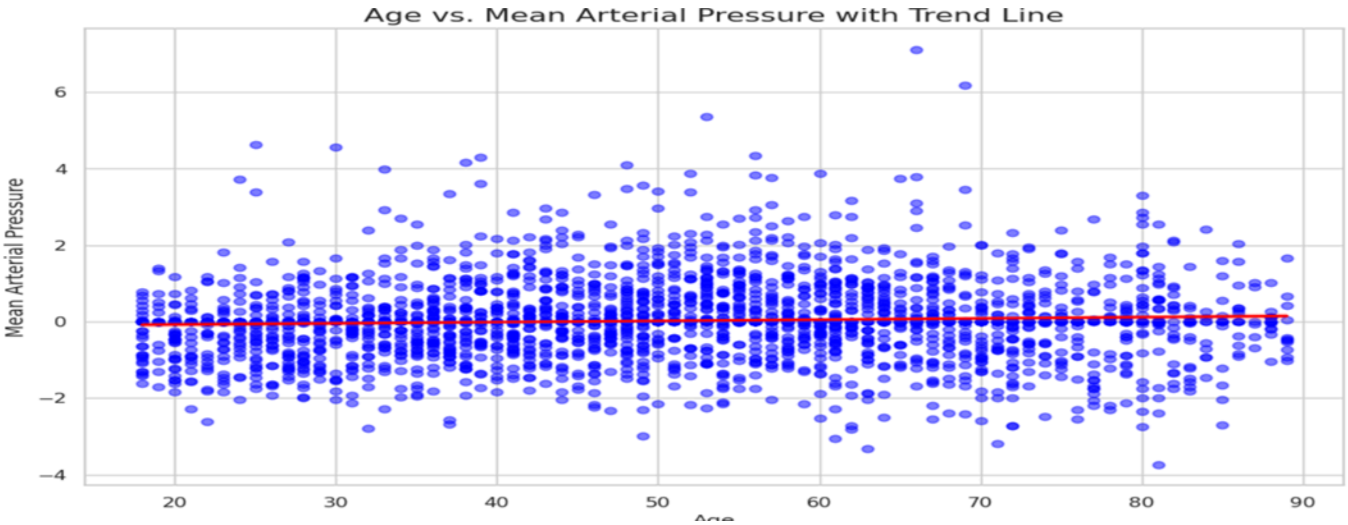


**Fig. 3.** Scatter plot of age and mean arterial pressure

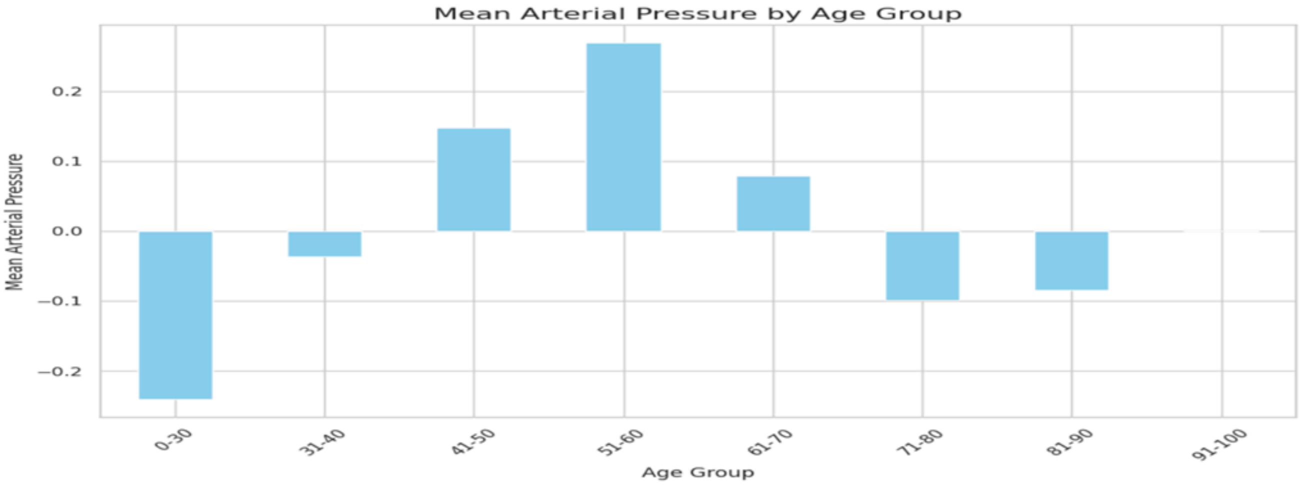


**Fig. 4.** Variation Of Mean Arterial Pressure By Age Group

By studying different features and their correlations, we found that some features, such as oxygen saturation, respiratory rate, hemoglobin count, race, age, and other methods of oxygen consumption, highly impacted the patient's survival outcomes. Figure 1 shows the distribution of age, sex, and race; these demographic features provide a profile of the typical hospitalized COVID-19 patient. Figure 2 clearly and comprehensively presents the trends and fluctuations in the usage of each method over time. The plot offers a detailed view of the average usage of various oxygen delivery methods across different time indices. Each line represents a different method, allowing you to quickly discern the differences in usage patterns and their evolution over time. Figure 3 shows the relationship between age and mean arterial pressure, with a relatively flat trend line. This visually confirms the weak correlation coefficient of 0.06, which is not significant, indicating a very slight increase in mean arterial pressure with age. Figure 4 helps us understand how the mean arterial pressure varies across age categories.

## Hypotheses:

The hypotheses underpinning this study were derived from existing research on leveraging electronic health record (EHR) data to prognosticate patient outcomes. Specifically, the work of Cole et al., which classified patients into broad categories but did not directly inform resource allocation decisions, prompted our primary hypothesis. We posited that predicting the precise discharge date for individual patients would enable healthcare institutions to manage resources more judiciously and dynamically in response to evolving needs. Furthermore, the study conducted by Laila Rasmy et al., which applied recurrent neural networks to EHR data for qualitatively assessing patient condition trajectories, inspired our objective to forecast a quantitative discharge day utilizing initial EHR inputs. Their approach demonstrated the feasibility of extracting clinically relevant insights from longitudinal EHR data.

HOSPITAL TO HOME
LEARN THE STEPS TO DISCHARGE HOME
Query on Real-Time Aggregate across pre-calculated and raw data
Materialized Table
Raw table
Oldest data
Completion Threshold
Latest data

**Fig. 5.** Depiction of the Hypotheses.

An ancillary hypothesis emerged concerning the relative predictive power of different data representations. While Rasmy et al. benefited from a large dataset of 247,960 patients' daily EHR records across 5 years, our study's data comprised the first 30 days of EHR data for a cohort of 20,000 patients and a separate aggregated summary of this cohort. This discrepancy in data volume and granularity prompted whether the higher resolution daily data sequence or the coarser aggregated records would better facilitate accurate discharge date prediction. Exploring this fundamental data representation issue held importance for our modeling approach and could inform best practices for future studies operating under data constraints.

## Modeling and Validation

To validate our hypotheses, we evaluated simple yet interpretable models as an initial step - linear regression and feedforward neural networks. Additionally, to assess the potential advantage of time-series modeling for the sequential daily data, we included long short-term memory (LSTM) neural networks. The dataset was split into an 80-20 training-test split, with labels based on the fourth root of discharge dates to account for right-skewed data. The models were trained using PyTorch, with the following architectures:

**Linear Regression (LN)**:
- All features were flattened before training.
- Mean squared error (MSE) was used as the loss function.

**Feed-Forward Neural Network (FFNN)**:
- Two fully-connected layers with 128 hidden size.
- ReLU activation function.
- Flattened features and used MSE loss.

**Long Short-Term Memory (LSTM)**:
- Torch LSTM layer with two layers, 128 hidden size, and 0.2 dropout.
- Fully-connected output layer.
- MSE loss.

Exploring these fundamental models was a rationale for first understanding how the EHR data informs predictions before potentially developing more complex approaches. If satisfactory performance could be achieved with these simple techniques, it could pave the way for deploying resource-efficient yet robust systems for optimizing hospital resource allocation. Conversely, poor performance would pinpoint areas requiring further investigation, either refining the data representations or adopting more sophisticated modeling strategies.

**Validation:**

Model performance was primarily evaluated using mean squared error (MSE) on the discharge date predictions, with visual analysis of prediction plots across models. To specifically assess our second hypothesis regarding optimal data quantity, we calculated an "extrapolation MSE" metric. This involved training models on data from only the first i days (i=1 to 30), then computing the MSE over only those patients with true discharge dates after day i. This enabled the determination of how many antecedent days of data were minimally required to achieve satisfactory forecasting accuracy for later discharges. The combined use of overall MSE and extrapolation MSE provided a comprehensive framework for validating hypotheses and model efficacy. Validation was performed by testing the models for the lowest MSE, using different days (1, 2, ..., 30) to assess model accuracy in predicting discharge dates.

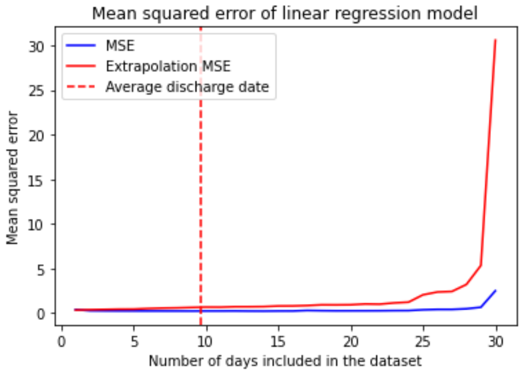
Additionally, we graphed the actual discharge date against the predicted discharge date. To do this, we iterated for true discharge date i, from 0 <= i <= 100. For all patients in the test set with true discharge date i, we ran the model on that batch of patients and plotted the average predicted discharge date across this batch.

## Results

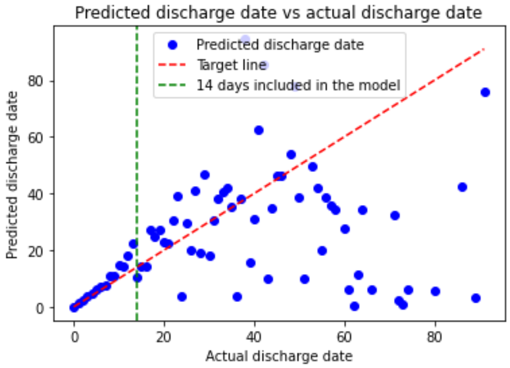
The results for the three models:

**Linear Regression (LN)**:

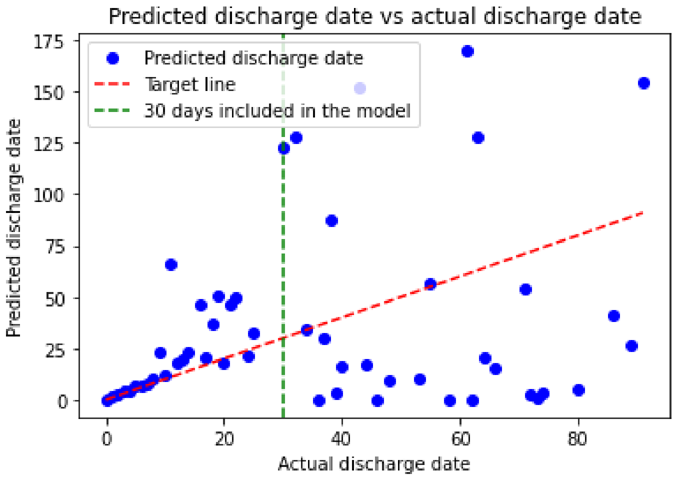
The best model included 14 days of data, with an MSE of **0.213** for raw data, **0.236** for mean-imputed, scaled data, and **0.524** for aggregated data. Interestingly, including more data did not necessarily result in a better result. From the graph of predicted vs actual discharge dates, shown for the 14 day model, the model interpolates data well, but then the predicted discharge dates start to fade after around 21 days. In the graph for the 30 day model, the fanning out begins right after day 30, showing that the extrapolation potential of this model is poor.



MSE vs No.of days of data included



14 days. MSE: **0.21**



30 days. MSE: **2.49**

**Fig. 6.** Performance of Time Series Dataset with Linear Regression.

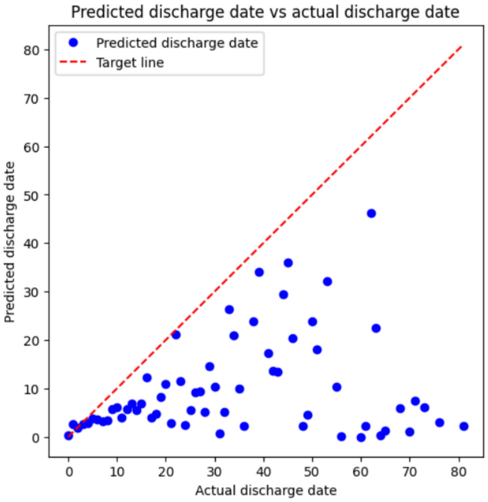
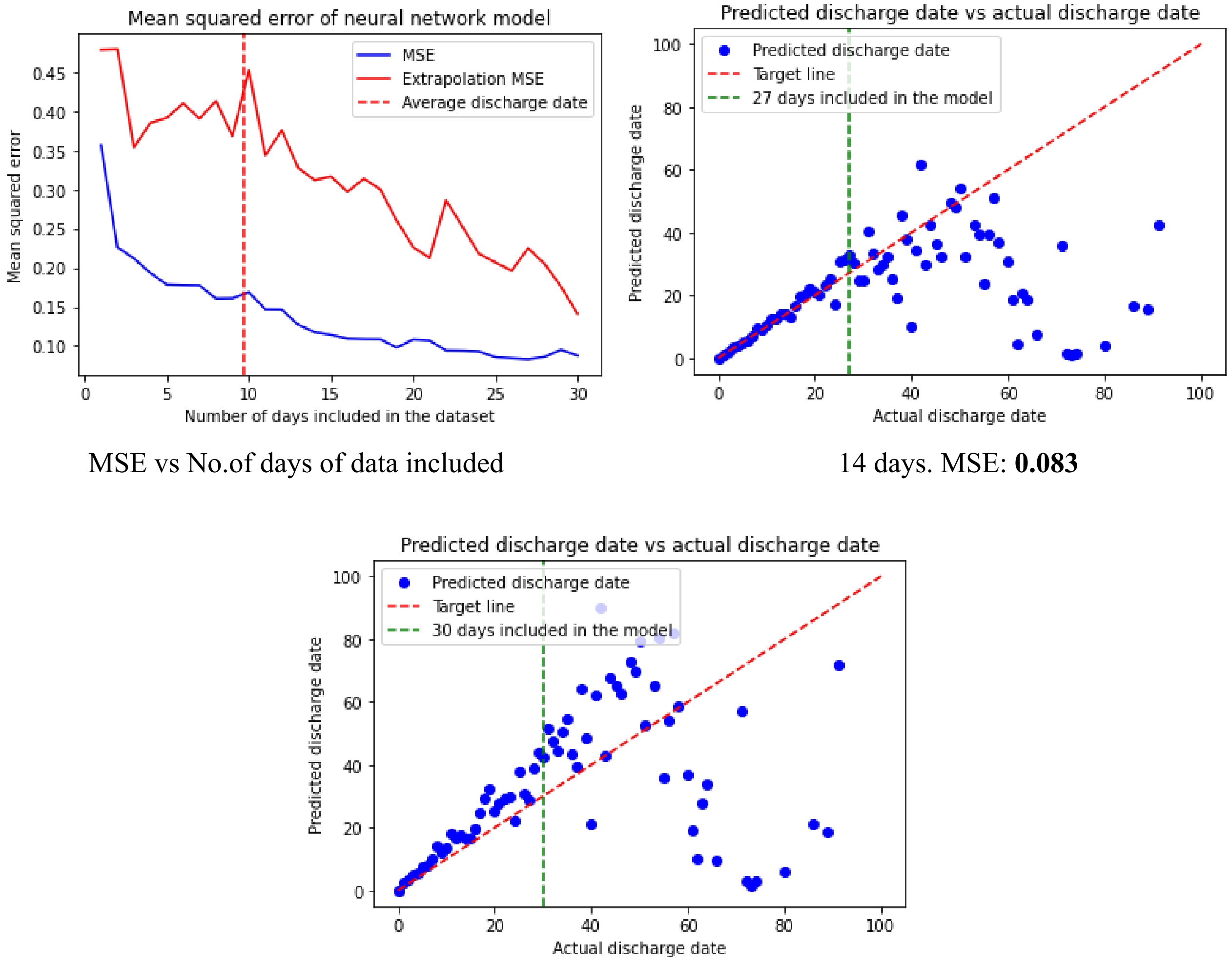


**Fig. 7.** Predicted Vs Actual for aggregated dataset for Linear Regression.

**Feed-Forward Neural Network (FFNN):**

The best model included 27 days of data, with an MSE of **0.083** for raw data, **0.197** for mean-imputed, scaled data, and **3.76** for aggregated data. As we'd expect, the model generally got better by including more data, and both the MSE and Extrapolation MSE reduced with including more data. From the graph of predicted vs actual discharge dates, shown for the 27 day model, the model interpolates data well but suffers some fan out in extrapolation. The graph for the 30 day data, though the overall MSE was slightly more on this model, exhibits better extrapolation loss and would probably be the best to use in practice.



MSE vs No.of days of data included

14 days. MSE: **0.083**

**Fig. 8.** Performance of Time Series Dataset with Feed Forward Neural Network.

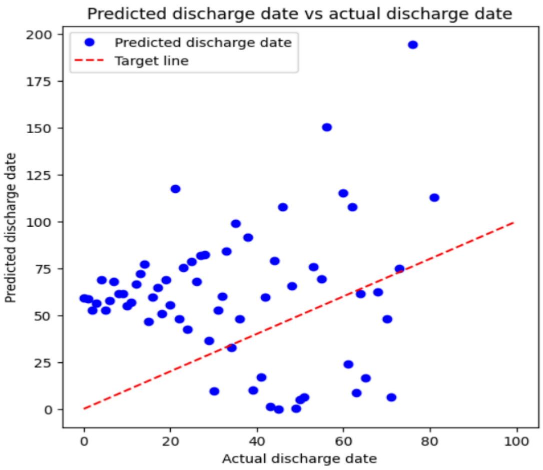


**Fig. 9.** Predicted Vs Actual for aggregated dataset for Feed Forward Neural Network.

**Long Short-Term Memory (LSTM)**:

The LSTM model performed poorly, predicting a discharge date of day 0 for all patients. Possible reasons include sparse raw data, lack of label diversity, and a need for more training examples. Due to the unavailability of RAM, we cannot even run LSTM for the aggregated dataset.

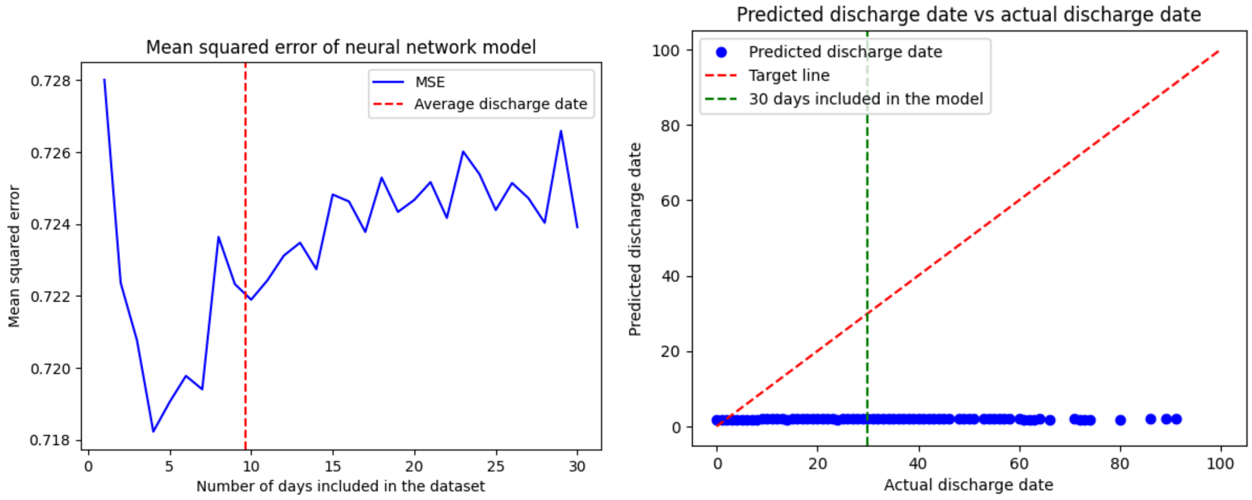


**Fig. 10.** Performance of Time Series Dataset with LSTM Neural Network.

## Discussion

Among the three models, the feed-forward neural network (FFNN) best predicted discharge dates. The LSTM model underperformed, likely due to the skewed nature of the data and the need for more training examples. Time-series data generally produced better results than aggregated data (i.e., the first 48 hours of EMR data), indicating that capturing temporal patterns is crucial for accurate predictions. Additionally, the FFNN model demonstrated better results with raw data than mean-imputed data, suggesting that imputation might introduce bias or noise in the dataset. This observation could have implications for future data preprocessing strategies. The time series data's performance is way better when comparing the model’s performance with the aggregated data. This might be due to the skewness of the data, which also requires fine-tuning the models specifically to the available dataset. However, after using whatever is available, we can infer that the time series data is the most useful format to predict a patient's discharge date.

## Conclusion

In conclusion, among the three models for predicting discharge dates, the feed-forward neural network (FFNN) outperformed the others, demonstrating that time-series data provides more accurate predictions than aggregated data. The linear regression model also showed reasonable performance, while the LSTM model struggled, possibly due to data sparsity and lack of label diversity. Future work should focus on refining the LSTM model, addressing data sparsity, and exploring alternative loss functions. Additionally, incorporating more training examples and exploring complex neural network architectures could improve the accuracy and reliability of discharge date predictions in hospital settings.

## Acknowledgments

**Nadia:** Literature Review, Problem Setup and Assistance in Validation. (Score :100)
**Sanjana:** Exploratory Data Analysis.(Score :100)
**Likith:** Hypotheses Generation, Assisting in data cleaning and Modeling.(Score :100)
**Shikhar:** Data Cleaning, Modeling & Validation.(Score :100)

## References

Cole Maguire, Elie Soloveichik, Netta Blinchevsky, Jaimie Miller, Robert Morrison, Johanna Busch, W. Michael Brode, Dennis Wylie, Justin Rousseau, Esther Melamed
medRxiv 2023.11.27.23297171; doi: https://doi.org/10.1101/2023.11.27.23297171

Naresh Doni Jayavelu Carly E. Milliren Carolyn S. Calfee Charles B. Cairns Monica Kraft Lindsey R. Baden Albert C. Shaw Al Ozonoff, Joanna Schaenman and et al. Phenotypes of disease severity in a cohort of hospitalized covid-19 patients: Results from the impact study. eBioMedicine, 83(62), 2022.

Msemburi, W., Karlinsky, A., Knutson, V. et al. The WHO estimates of excess mortality associated with the COVID-19 pandemic. Nature 613, 130–137 (2023). https://doi.org/10.1038/s41586-022-05522-2

Jacobsen GD, Jacobsen KH. Statewide COVID-19 Stay-at-Home Orders and Population Mobility in the United States. World Med Health Policy. 2020 Dec;12(4):347-356. doi: 10.1002/wmh3.350. Epub 2020 Jul 29. PMID: 32837774; PMCID: PMC7405141.

Gupta P, Gupta V, Singh CM, Singhal L. Emergence of COVID-19 Variants: An Update. Cureus. 2023 Jul 3;15(7):e41295. doi: 10.7759/cureus.41295. PMID: 37539393; PMCID: PMC10394493.

Klein B, Zenteno AC, Joseph D, Zahedi M, Hu M, Copenhaver MS, Kraemer MUG, Chinazzi M, Klompas M, Vespignani A, Scarpino SV, Salmasian H. Forecasting hospital-level COVID-19 admissions using real-time mobility data. Commun Med (Lond). 2023 Feb 14;3(1):25. doi: 10.1038/s43856-023-00253-5. PMID: 36788347; PMCID: PMC9927044.