# Machine Learning in Maternal Healthcare: Predicting Post-Cesarean Hospital Stay

Group C

January 25, 2025

**Abstract**

This study addresses the need for generalizable models predicting hospital length of stay (LOS) for Cesarean sections (C-sections) using population-level data. Leveraging the Healthcare Cost and Utilization Project (HCUP) dataset, we employed various machine learning algorithms to establish LOS prediction models. Our research fills a significant gap in the literature by comparing predictive performances across diverse algorithms on a state-wide scale.

We conducted a comprehensive analysis of LOS prediction following C-sections using HCUP data from Maryland (2016-2019). After applying inclusion/exclusion criteria, 86,889 cases remained for analysis. Feature selection was implemented to verify the feature relevance to LOS prediction. Six machine learning models, including Linear Regression, Logistic Regression, Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP), were trained, tuned, and evaluated for LOS prediction. Results show that ICD-based predictors outperformed social determinants in C-section LOS prediction (with a 6% AUC improvement), which suggests the importance of incorperating physiological features or patient disease history for LOS prediction. Another important finding is that utilization of MLP outperforms other machine learning approaches for LOS prediction (with more than 10% AUC improvement). This comprehensive analysis offers insights into predicting C-section LOS, which is crucial for optimizing maternal healthcare and resource allocation. The code will be provided once this paper is published.

## 1 Introduction

The management of the peri-obstetric period is an essential aspect of maternal healthcare. C section complication has known to prolong the inpatient length of stay (LOS). Previous studies have shown that there has been a rise in rate in C section, and thus predicting the LOS in such context has become increasingly important. [1]

Predictive modeling in healthcare has evolved significantly, with numerous studies demonstrating the efficacy of machine learning algorithms in forecasting

LOS across various conditions.[2, 3, 4]. Several reviews have pointed out the lack of generalization ability of such models due to overly specific customization and data processing, limiting their use locally.[5, 6] Additionally, only a handful of studies focused on predicting the LOS of C-sections.[7, 8, 9], and these models never use non-clinical data.

Historically, research on LOS prediction for C-sections has been confined to studies involving one or a few institutions utilizing singular predictive models[7, 8, 9]. Indeed, few studies have used large-scale population data to predict outcomes from C-sections[10]. There have been a lack of comparative analyses across diverse predictive models and using comprehensive healthcare datasets. This gap in the literature signifies a missed opportunity for using advanced models to enhance C-section LOS predictions. Furthermore, explorations of LOS predictions using demographic and social determinants of health within large datasets—encompassing multiple hospitals across extensive regions—remain largely untapped. A solution addressing these gaps promises a comprehensive understanding of LOS determinants, transcending the limitations of single-center data to reflect a broader spectrum of patient demographics and care environments.

In this study, to address the lack of a generalizable C-section LOS model on a population level, we used the Healthcare Cost and Utilization Project (HCUP) databases. HCUP is one of the most comprehensive databases that provide information on patients' demographic backgrounds, emergency department encounters, ambulatory surgeries, and inpatient stays [11]. We conducted a novel examination of LOS predictions for C-sections, employing a variety of algorithms to compare their predictive performances in a state-wide population dataset.

# 2 Methodology

## 2.1 Dataset and Study Population

HCUP is collected by AHRQ (Agency for Healthcare Research and Quality) and is commonly used to analyze trends national trends in healthcare across different states. For our study, we utilized State Inpatient Database (SID) HCUP data from Maryland starting from 2016 to 2019. HCUP datasets include patient demographic features like 'AGE', 'RACE', 'FEMALE', and 'Homeless', alongside clinical variables such as 'LOS' (Length of Stay), 'ICD codes', and 'PCS codes'. We identified 86,889 cases where patients underwent Cesarean section (C-section) rather than abortion. To do this, ICD-10 Procedure Coding System codes are used to exclude patients with codes indicating Cesarean section for abortion ('10A00ZZ', '10A03ZZ', '10A04ZZ') and keep those with codes for Cesarean section for extraction of products of conception ('10D00Z0', '10D00Z1', '10D00Z2')[12].

## 2.2 Preprocessing and Feature coding

Our dataset contained both numerical and categorical features, necessitating specific preprocessing steps. For numerical attributes such as Age and the Number of ICD and PCS codes associated with each patient, we employed z-score normalization to standardize the data. For categorical variables such as demographic features and ICD codes, we employed one-hot encoding. This transformation technique expands categorical variables into binary columns, where each unique category is represented by a separate column. HCUP dataset offers comprehensive patient data but lacks crucial medical details like history, lab results, and in-hospital complications. To compensate this issue, we extracted clinically relevant insights from ICD and PCS codes.

Of paramount importance to obstetricians is the gestational age of patients, denoted in ICD codes (e.g., "Z3A32" signifies a gestational age of 32 weeks). Both low (<32 weeks) and high (>42 weeks) gestational ages correlate with elevated risks for maternal and neonatal complications, consequently prolonging LOS[13, 14].

To gauge the burden of comorbid conditions (e.g., hypertension, diabetes) on patients, we computed the Elixhauser comorbidity score[**?**]. This metric provides insights into predicted hospital resource utilization and mortality rates, with higher scores indicating greater medical complexity [15]. We leveraged the hcuppy Python library provided by HCUP, to derived the Elixhauser score for each patient.

The variability in how ICD and PCS codes are recorded poses a significant challenge for machine learning algorithms, particularly due to the generated sparse feature space when considered independently[16, 17]. To mitigate this issue, we utilized HCUP clinical category groupings to map codes to meaningful categories, thereby reducing feature dimensionality and enhancing model interpretability.

As per ACOG(American College of Obstetricians and Gynecologists), the hospital stay after a caeserian birth is typically 2 to 4 days[18]. Hence, for our labels, we grouped LOS into those greater than 4 days and those less than 4 days.

## 2.3 Feature Selection Method (mRMR)

To select which features might be relevant in our dataset to predict LOS, we employed Minimum Redundancy Maximum Relevance method, which is designed to optimize the informativeness of features while minimizing redundancy within them[19]. This provides an importance measurement for each independent variable, with higher scores suggesting more importance.

## 2.4 Models

We specified our LOS target on whether the individuals have a prolonged length of stay, which is determined by the mean and median length of stay in days in

the overall data. We performed the prediction task with six distinct machine-learning models on Maryland HCUP dataset. These models include Linear Regression, Logistic Regression, Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptron (MLP).

Linear Regression is a straightforward approach for prediction. In this study, it was adapted for classification purposes by manually setting a threshold. Similarly, Logistic Regression model was also introduced for prediction. Logistic Regression is a classic model for estimating the probability of binary outcomes. These two algorithms serve as baselines for our LOS prediction.

We also implemented more complicated models for the prediction task. Random Forest is an ensemble-leaning method that integrates multiple decision trees to enhance prediction accuracy and stability. In each decision tree, the inputs go through a series of decision nodes, and eventually culminating in a leaf node that represents the decision's outcome. In a Random Forest, each tree independently contributes to the decision-making process, with the final prediction determined by a majority vote [20]. Support Vector Machine (SVM) excels in high-dimensional spaces for identifying a hyperplane that best separates two classes. This model is particularly effective for classification tasks involving numerous variables [21]. Extreme Gradient Boosting (XGBoost) is an advanced iteration of the Gradient Boosting Decision Tree algorithm. It employs a sequence of decision trees arranged in a cascade. In the cascade, each tree refines the predictions of its predecessor, thereby enhancing the overall predictive accuracy [22]. Multilayer Perceptron is a basic form of neural network. It was chosen for its ability to discern intricate patterns within the data through its multiple layers and neurons [23].

For each model, we conducted hyperparameter tuning to optimize their performance by randomly search approach, which allows us to evaluate on different combinations of various parameters to look for the best outcome. Following the tuning process, we assessed the performance of each model using various evaluation metrics.

## 2.5 Metrics

We employed three primary scoring methods to evaluate the model performance: Accuracy (ACC), Area Under the Curve (AUC), and F1 Score. ‘ Accuracy serves as a straightforward metric, representing the overall correctness of the model in classification tasks. It reflects the proportion of true results (both true positives and true negatives) among the total number of cases examined. The Area Under the Curve (AUC) is a crucial metric in evaluating the performance of a binary classifier. A higher AUC value indicates better model performance, with a value of 1 representing a perfect classifier and a value of 0.5 suggesting no discriminative power. The F1 Score is a measure that combines precision (the true positive rate) and recall (sensitivity) into a single metric, it focuses on the balance of the two criteria while evaluating the accuracy of the model. A higher F1 Score signifies that the model has successfully achieved a balance between precision and recall, indicating both high precision and high recall. Consequently, a model

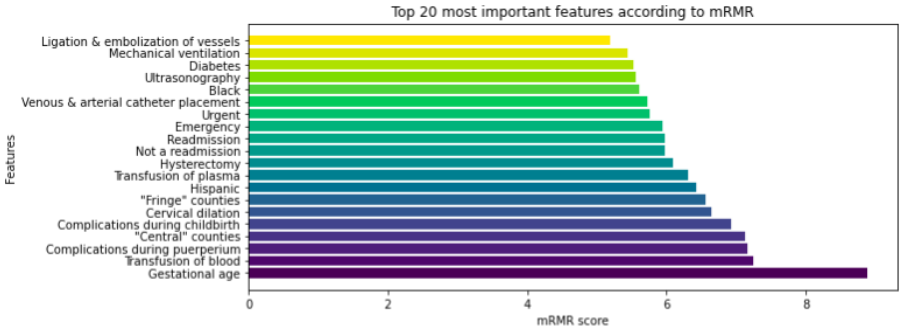


Figure 1: mRMR scores for top 20 important features

with a high F1 Score demonstrates overall superior performance.

# 3 Results

## 3.1 Participants

After selecting participants based on the inclusion and exclusion criteria previously mentioned, we included ***** from 2016 to 2020 in our final analysis. Their demographic and clinical characteristics are summarized in table 1.

## 3.2 Feature importance

mRMR identified gestational age as the most crucial factor in obstetric care prediction. Complications during childbirth and postpartum, like mental disorders and hypertension, are captured by PRG027 and PRG023 respectively. Interventions such as blood transfusions (ADM001, ADM002) and cervical dilation (FRS008) cater to various needs during C-section deliveries. Catheter placement (CAR024) and antibiotic administration (ADM015) were important in prediction. Mechanical ventilation (ESA003) and airway intubation (RES007) also prolonged the LOS. Contraceptive procedures (FRS013, FRS004) were considered in the prediction as well. These factors help predict post-C-section length of stay, showcasing the interplay between interventions, complications, and patient outcomes in obstetrics.

## 3.3 Classification Performance

In this part, experiment results for social determinant features, ICD-based features, and combined features are shown in Table 1, 2, and 3 respectively.

In the context of predicting cesarean section outcomes, it is evident that features derived from the ICD code offer superior performance over those based on social determinants. This is substantiated by the results presented in Table

Table 1: Classification performance on social determinant features.

| Model | ACC | AUC | F1 |
|---|---|---|---|
| Linear Model | 0.7323 (0.0200) | 0.5418 (0.0155) | 0.1898 (0.0479) |
| Logistic Regression | 0.7317 (0.0083) | 0.5463 (0.0133) | 0.2124 (0.0366) |
| Random Forest | 0.7306 (0.0042) | 0.5334 (0.0054) | 0.1587 (0.0159) |
| SVM | 0.7386 (0.0159) | 0.5255 (0.0273) | 0.0843 (0.0452) |
| XGBoost | 0.7328 (0.0069) | 0.5415 (0.0122) | 0.1890 (0.0380) |
| MLP | **0.7442 (0.0014)** | **0.6684 (0.0046)** | **0.3333 (0.0139)** |

Table 2: Classification performance on ICD-based features.

| Model | ACC | AUC | F1 |
|---|---|---|---|
| Linear Model | 0.5420 (0.0195) | 0.5820 (0.0095) | 0.3053 (0.0266) |
| Logistic Regression | 0.7594 (0.0062) | 0.5958 (0.0153) | 0.3402 (0.0418) |
| Random Forest | 0.7449 (0.0122) | 0.5538 (0.0342) | 0.2018 (0.1182) |
| SVM | 0.7610 (0.0152) | 0.5859 (0.0249) | 0.3152 (0.0457) |
| XGBoost | 0.7560 (0.0118) | 0.5889 (0.0346) | 0.3148 (0.1117) |
| MLP | **0.7762 (0.0017)** | **0.7287 (0.0084)** | **0.4697 (0.0167)** |

1 and Table 2, where, with the exception of the linear model, machine learning algorithms demonstrate enhanced predictive capabilities. Notably, MLP utilizing ICD-based features exhibits a significant enhancement, with a 6% increase in AUC and a 13% elevation in the F1-score. Furthermore, an integrative approach that combines both social determinant and ICD-based features does not yield any notable gains in predictive accuracy in comparison to the exclusive use of ICD-based features. This outcome further underscores the limited utility of social determinant features in LOS prediction, suggesting that they may not provide additional predictive value in such models.

In evaluating the efficacy of various predictive models, the MLP distinctly surpasses its counterparts, boasting an improvement exceeding 10% in both AUC and F1 score. This enhanced performance, however, comes at the expense of increased computational demand relative to more traditional methodologies. Acknowledging these computational considerations, we also advocate for ensemble models, specifically XGBoost and Random Forest. These models strike a balance by providing robust predictions for LOS while maintaining computational efficiency. Their integration serves to construct equitable and effective LOS forecasting systems, which are both practical and reliable for clinical application.

# 4 Discussion

clinical discussion In our research, we aimed to refine the prediction of hospital C-section LOS by incorporating data from the HCUP. Traditional methods, focusing on physiological data collected during the hospital stay, such as blood

Table 3: Classification performance on combined features.

| Model | ACC | AUC | F1 |
|---|---|---|---|
| Linear Model | 0.7323 (0.0200) | 0.5417 (0.0155) | 0.1898 (0.0478) |
| Logistic Regression | 0.7317 (0.0085) | 0.5465 (0.0138) | 0.2131 (0.0385) |
| Random Forest | 0.7295 (0.0028) | 0.5189 (0.0055) | 0.0882 (0.0214) |
| SVM | 0.7546 (0.0068) | 0.5739 (0.0086) | 0.3045 (0.0486) |
| XGBoost | 0.7326 (0.0059) | 0.5390 (0.0086) | 0.1790 (0.0256) |
| MLP | **0.7670 (0.0085)** | **0.7088 (0.0220)** | **0.4775 (0.0152)** |

pressure and lab values, offer insights but are limited by the need for real-time data collection, which may not be available in all healthcare settings. HCUP data, despite its less detailed nature, provides a broadly applicable dataset for predicting LOS across a diverse patient population. Because of HCUP's nationwide, population-level data, our model overcomes the limitations of accessing Electronic Health Record (EHR) data across various hospitals, enabling generalization across all states.

Our analysis focused on the geo-derived population-level patients' demographics and social determinants of health (DSD) characteristics and clinical data and the relative importance of various features in predicting LOS. Gestational age (GA) proved to be a pivotal factor, with premature or overdue births indicating potential complications and thereby extending hospital stays. This aligns with the current understanding that birth timing can reflect underlying health issues.

The predictive model based on ICD-10 features alone had the best performance, suggesting that a patient's medical history and overall health status prior to hospitalization are crucial determinants of LOS in C-section. Unexpectedly, the geo-derived social determinants of health did not significantly impact LOS predictions. This indicates that, despite the potential value of demographics and social factors, the population-level geo-derived DSDs did not have enough granularity to enhance our prediction. Our findings suggest that individual-level health determinants are critical and cannot be effectively replaced by aggregate, geographical data.

The establishment of appropriate prediction targets is crucial for assessing model performance. Our benchmarking efforts reveal significant challenges in formulating regression models aimed at LOS prediction, primarily attributed to the pronounced skewness in the data distribution. This skewness complicates the direct application of regression techniques. A noteworthy observation arises when converting the regression problem into a classification task by implementing a threshold criterion. The selection of this threshold profoundly influences the resultant performance metrics. For example, adjusting the threshold for defining a long-stay from 3 days (as per our parameterization) to 10 days markedly inflates the AUC values, potentially reaching as high as 80%. This inflation is primarily due to the reduced number of samples classified as long-stay. Consequently, it becomes imperative to consider the target for LOS predictions

carefully. The chosen metrics must be robust against the imbalances inherent in class distributions to ensure a genuine evaluation of the model's predictive capacity. Such diligence is essential for developing models that can faithfully represent and effectively predict LOS within clinical settings.

MLP has demonstrated superior performance over traditional machine learning algorithms in the present investigation. This superior performance is attributed to the model's overparameterization, allowing the MLP to capture the complexity of the dataset with a higher degree of granularity. Moreover, the process of hyperparameter optimization for conventional models presents substantial challenges, both in terms of complexity and computational time. For instance, the Support Vector Machine (SVM) optimization can extend beyond three days, whereas the MLP and XGBoost models complete the process in approximately an hour. Given the results, the MLP emerges as the preferred model among those evaluated, contingent upon the availability of computational resources. Nevertheless, this advantage comes at the expense of a significantly larger number of parameters, which may not be practical in scenarios with limited samples or computational constraints. Under such circumstances, ensemble-based methods are recommended, specifically Random Forest and XGBoost. These methods facilitate easier implementation and exhibit robustness to variations in hyperparameter configurations, making them well-suited for LOS prediction models.

# 5 Conclusion

In this study, we established LOS prediction model benchmarks based on the HCUP dataset. Our results show that utilizing disease-related features, encoding features in appropriate manners, and taking MLP or ensemble models as analytical tools are essential for successful LOS prediction. Gestational age is identified as the most important feature for LOS prediction and other features such as blood transfusion, complications during puerperium are also found. Challenges such as data skewness, training target setup and hyperparameter tuning for classifiers are also uncovered or explored in this study. In the future, it is considered to conduct more research comparing different states and employ advanced deep learning techniques to further improve the C-section LOS prediction models.

# References


[1] World Health Organization. Caesarean section rates continue to rise, amid growing inequalities in access, 2021. Accessed: 2024-03-22.

[2] Belal Alsinglawi, Osama Alshari, Mohammed Alorjani, Omar Mubin, Fady Alnajjar, Mauricio Novoa, and Omar Darwish. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific reports*, 12(1):607, 2022.

[3] Xin Ma, Yabin Si, Zifan Wang, and Youqing Wang. Length of stay prediction for icu patients using individualized single classification algorithm. *Computer methods and programs in biomedicine*, 186:105224, 2020.

[4] Tahani A Daghistani, Radwa Elshawi, Sherif Sakr, Amjad M Ahmed, Abdullah Al-Thwayee, and Mouaz H Al-Mallah. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. *International journal of cardiology*, 288:140–147, 2019.

[5] Kieran Stone, Reyer Zwiggelaar, Phil Jones, and Neil Mac Parthaláin. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, 1(4):e0000017, 2022.

[6] Vincent Lequertier, Tao Wang, Julien Fondrevelle, Vincent Augusto, and Antoine Duclos. Hospital length of stay prediction methods: a systematic review. *Medical care*, 59(10):929–938, 2021.

[7] Oona MR Campbell, Luca Cegolon, David Macleod, and Lenka Benova. Length of stay after childbirth in 92 countries and associated factors in 30 low-and middle-income countries: compilation of reported data and a cross-sectional analysis from nationally representative surveys. *PLoS medicine*, 13(3):e1001972, 2016.

[8] Emma Montella, Marta Rosaria Marino, Massimo Majolo, Eliana Raiola, Giuseppe Russo, Giuseppe Longo, Andrea Lombardi, Anna Borrelli, and Maria Triassi. Regression and classification methods for predicting the length of hospital stay after cesarean section: A bicentric study. In *Proceedings of the 6th International Conference on Medical and Health Informatics*, pages 135–140, 2022.

[9] Daniel Gabbai, Emmanuel Attali, Shai Ram, Uri Amikam, Eran Ashwal, Liran Hiersch, Ronni Gamzu, and Yariv Yogev. Prediction model for prolonged hospitalization following cesarean delivery. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 274:23–27, 2022.

[10] Jovanny Tsuala Fouogue, Aline Semaan, Tom Smekens, Louise-Tina Day, Veronique Filippi, Matsui Mitsuaki, Florent Ymele Fouelifack, Bruno Kenfack, Jeanne Hortence Fouedjio, Thérèse Delvaux, et al. Length of stay and determinants of early discharge after facility-based childbirth in cameroon: analysis of the 2018 demographic and health survey. *BMC pregnancy and childbirth*, 23(1):575, 2023.

[11] Healthcare Cost and Utilization Project (HCUP). https://hcup-us.ahrq.gov/, February 2024. Accessed: 2024-02-20.

[12] IQI 33 Primary Cesarean Delivery Rate. Accessed: February 15, 2024.

[13] Mohammed Galal, Ian Symonds, Henry Murray, Felice Petraglia, and Roger Smith. Postterm pregnancy. *Facts, views & vision in ObGyn*, 4(3):175, 2012.

[14] Seung Soo Lee, Hye Seong Kwon, and Hyung Min Choi. Evaluation of preterm delivery between 32+ 0-33+ 6 weeks of gestation. *Journal of Korean medical science*, 23(6):964–968, 2008.

[15] Brian J Moore, Susan White, Raynard Washington, Natalia Coenen, and Anne Elixhauser. Identifying increased risk of readmission and in-hospital mortality using hospital administrative data. *Medical care*, 55(7):698–705, 2017.

[16] Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality. HCUP Clinical Classifications Software Refined (CCSR) for ICD-10-PCS procedures, v2021.1. 2021. Accessed: February 15, 2024.

[17] Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality. HCUP Clinical Classifications Software Refined (CCSR) for ICD-10-CM diagnoses, v2021.2. 2021. Accessed: February 15, 2024.

[18] Cesarean birth. https://www.acog.org/womens-health/faqs/cesarean-birth. Accessed: February 15, 2024.

[19] Zhenyu Zhao, Radhika Anand, and Mallory Wang. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. 2019.

[20] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 10 2001.

[21] Nello Cristianini and Elisa Ricci. *Support Vector Machines*, pages 928–932. Springer US, Boston, MA, 2008.

[22] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[23] H. Taud and J.F. Mas. *Multilayer Perceptron (MLP)*, pages 451–455. Springer International Publishing, Cham, 2018.