ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE E TECNOLOGIE DELLA SALUTE

Ciclo 36

Settore Concorsuale: 09/G2 - BIOINGEGNERIA

Settore Scientifico Disciplinare: ING-INF/06 - BIOINGEGNERIA ELETTRONICA E INFORMATICA

LEVERAGING THE POWER OF DATA TO PROVIDE BETTER HEALTHCARE SERVICES

Presentata da: Addisu Jember Zeleke

Coordinatore Dottorato

Igor Diemberger

Supervisore

Lorenzo Chiari

Co-supervisore

Rossella Miglio

Consuelo Basili

Esame finale anno 2024

ALMA MATER STUDIORUM
Università di Bologna

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi"

Academic discipline: ING-INF/06 Electronic and Informatics Bioengineering

PhD in Health and Technology

Leveraging the Power of Data to Provide Better Healthcare Services

Thesis submitted for the degree of PhD in Health and Technology – 36th cycle

Candidate: Addisu Jember Zeleke

Coordinator: Prof. Igor Diemberger

Supervisor: Prof. Lorenzo Chiari

Co-supervisors: Prof. Rossella Miglio, and

Dr. Consuelo Basili

Thesis defense: June 2024

Key words

Hospitalization

Prolonged Length of Stay (PLoS)

Emergency department

Childhood Cancer Survivor

Ovarian Failure

Nonsurgical Premature Menopause (NSPM)

COVID-19

Diseases Mapping

Performance Measure

Healthcare

Prediction

Classification and Regression

Machine Learning

Artificial Intelligence

Clustering

Spatiotemporal models

Global & Local regression

Count & Quantile regression

Survival models

Big-data

List of publications originated from PhD research outcomes

-**Peer-reviewed Journal contributions published**:

- **Addisu Jember Zeleke**, Pierpaolo Palumbo, Paolo Tubertini, Rossella Miglio, Lorenzo Chiari. Comparison of nine machine learning regression models in predicting hospital length of stay for patients admitted to a general medicine department. Informatics in Medicine Unlocked, Volume 47, 2024,101499, https://doi.org/10.1016/j.imu.2024.101499.
- **Zeleke AJ**, Palumbo P, Tubertini P, Miglio R, Chiari L. Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis. Front Artif Intell. 2023 Jul 28;6:1179226. PMID: 37588696; PMCID: PMC10426288, *ISI & Scopus Journal* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10426288/
- **Zeleke AJ**, Moscato S, Miglio R, Chiari L. Length of Stay Analysis of COVID-19 Hospitalizations Using a Count Regression Model and Quantile Regression: A Study in Bologna, Italy. Int J Environ Res Public Health. 2022 Feb 16;19(4):2224. PMID: 35206411; PMCID: PMC8871974, *ISI & Scopus Journal*. https://pubmed.ncbi.nlm.nih.gov/35206411/
- **Zeleke AJ**, Miglio R, Palumbo P, Tubertini P, Chiari L; Bologna MODELS4COVID Study Group of the University of Bologna and the National Institute for Nuclear Physics (INFN). Spatiotemporal heterogeneity of SARS-CoV-2 diffusion at the city level using geographically weighted Poisson regression model: The case of Bologna, Italy. Geospat Health. 2022 Dec 1;17(2). PMID: 36468589, *ISI & Scopus Journal.* https://pubmed.ncbi.nlm.nih.gov/36468589/

Journals contributions under review and preparation:

- **Zeleke AJ,** *et al* . Predicting the risk of nonsurgical premature menopause (NSPM) in survivors of childhood cancer in the presence of competing risks (under preparation)

Conference Proceedings

- **Zeleke AJ**, Miglio R, Palumbo P, Tubertini P, Chiari L. Prediction of Prolonged Hospitalization after Admission through the Emergency Department Using Machine Learning Techniques ; Society for Medical Decision Making (SMDM), 18th BIENNIAL EUROPEAN CONFERENCE, May 21-23,2023, Berlin, Germany [Conference proceedings-abstract]. https://www.smdmmeeting.org/18th-biennial-european-conference/
- **Zeleke AJ**, Miglio R, Palumbo P, Tubertini P, Chiari L. Machine learning regression models for predicting hospital stay for general medicine-specific patients, 44th Annual Conference of the International Society for Clinical Biostatistics (ISCB), Joint Conference with the Italian Region of the International Biometric Society, 27-31 August 2023, in Milan, ITALY. [Conference proceedings-abstract]. https://www.iscb2023.info/

Acknowledgement

በመጀመሪያ ክብር ምስጋና ለ ልዑል እግዚአብሔር ፣ ለናቴ ቅዱስ ባታ ፣ለናቴ ኪደን ምህረት ፣ለናቴ ቅድስት ማሪያም።

I would like to express my deepest gratitude to my advisor Prof. Lorenzo Chiari, for his invaluable guidance, patience, and support throughout the course of my PhD journey. His expertise, insightful feedback, and unwavering encouragement have been pivotal in shaping both this thesis and my academic growth.

I would also extend my appreciation to Pierpaolo Palumbo and Serena Moscato for their guidance and support during my PhD. My acknowledgment is also extended to my co-advisor Prof. Rossella Miglio for her support. I am also obliged and thankful to Prof. Yan Yuan for her guidance and advice during my period abroad at the University of Alberta, Canada. I am also thankful to all the current and former members of my lab (Biolab and Personal Health Systems lab) for their kind support and for providing a pleasant environment.

I would also give thanks to the Policlinico Sant'Orsola-Malpighi University hospital for funding my Ph.D. scholarship. I would also acknowledge the Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi" for their support.

I must also express my profound gratitude to my mother, my father, brothers, and sisters for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis.

Finally, I would like to express my appreciation to my Eden, for her support, encouragement in writing my thesis, and her positive attitude towards my life in general.

Thesis Abstract

Leveraging the power of data in healthcare can profoundly improve service delivery across multiple dimensions, from enhancing patient outcomes to optimizing operational efficiency. To engage patients and improve their health outcomes, it is essential to get a complete picture of their health journey, combining patient-generated information with official records. Modeling procedure plays a crucial role in medical response. They can contribute to obtain better and more accurate decision-making responses. To accurately predict disease outcomes, track disease progression, or identify risk factors—as health and wellbeing are affected by many factors —selecting the correct statistical and machine learning (ML) models is essential. As we are in the big data era, ML techniques stand out as the premier tools for extracting knowledge from data trends. Considering this scenario, this PhD project aims to develop and apply advanced statistical and machine learning models to analyze complex datasets, which are crucial for providing better healthcare services by leveraging the power of data on various health problems to inform decision-making and predict health-related outcomes.

Study 1 – part I (Chapter 2) aimed to estimate and compare different models to predict the Length of Stay (LoS) and the Prolonged Length of Stay (PLoS) of inpatients admitted through the emergency department (ED) in general patient settings. This aim is not only to promote any specific model but rather to suggest a decision-supporting tool (i.e., a prediction framework). We analyzed a dataset of patients admitted through the ED to the Sant'Orsola Malpighi University Hospital of Bologna, Italy, between January 1 and October 26, 2022. We considered six classification algorithms for predicting PLoS: Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GB), AdaBoost, K-Nearest Neighbors (KNN), and logistic regression (LoR). We evaluated the performance of these models with the Brier score, the area under the ROC curve (AUC), accuracy, sensitivity (recall), specificity, precision, and F1-score. We further estimated eight regression models for LoS prediction: Linear Regression (LR), including the penalized linear models Least Absolute Shrinkage and Selection Operator (LASSO), Ridge and Elastic-net regression, Support vector regression, RF regression, KNN, and eXtreme Gradient

Boosting (XGBoost) regression. The model performances were measured by their mean square error, mean absolute error, and mean relative error. A total of 12,858 eligible patients were included in our study, of whom 60.88% had a PloS. The Gradient Boosting classifier best predicted PloS (accuracy 75%, AUC 75.4%, Brier score 0.181), followed by the logistic regression classifier (accuracy 75%, AUC 75.2%, Brier score 0.182). These models were also shown to be adequately calibrated. Ridge and XGBoost regressions best predicted LoS, with the smallest total prediction error. The overall prediction error is between 6 and 7 days, meaning there is a 6–7 day mean difference between actual and predicted LoS. Our results demonstrate the potential of machine learning-based methods to predict LoS and provide valuable insights into the risks behind prolonged hospitalizations. In addition to physicians' clinical expertise, the results of these models can be utilized as input to make informed decisions, such as predicting hospitalizations and enhancing the overall performance of a public healthcare system.

Study 1 – part II (Chapter 3), as a specific focus within the previous study, specifically investigated a particular department/specialty within the hospital, namely the General Medicine (GM) department, which has the highest patient volume and heterogeneity. Closely examining hospitalization data is crucial because patients come with various conditions or traits. Therefore, we compared nine ML regression models in predicting hospital LoS for patients admitted to a GM department. The models' performance was assessed by means of root mean squared prediction error (RMSPE) and mean absolute prediction error (MAPE). Moreover, we used K-means clustering to group patients' medical and organizational criticalities (such as diseases, signs, symptoms, and administrative issues) into four clusters. Feature Importance plots and SHAP (SHapley Additive exPlanations) were employed to identify the top important features and enhance interpretability. The predictive performance of the different adopted models was between 11.00 and 16.16 days for RMSPE and between 7.52 and 10.78 days for MAPE. The eXtreme Gradient Boosting Regression (XGBR) model had the lowest prediction error, both in terms of RMSPE (11.00 days) and MAE (7.52 days). Sex, arrival via own vehicle/walk-in, ambulance arrival (118), light blue risk category, age 70 or older, and orange risk category are some of the top important features. The study's limitations include using a single cohort for model development and testing, hindering external validation crucial for

generalizability, and omitting vital signs and laboratory results, which are key for accurately predicting hospital stay lengths.

Study 2 (Chapter 4), conducted, when COVID-19 was at its peak, aimed to analyze the spatio-temporal patterns of the diffusion of SARS-CoV-2, the virus causing coronavirus 2019 (COVID-19). The study took place from February 1st, 2020 to November 20th, 2021 and accounted for space, sociodemographic characteristics and health conditions of the resident population. A second goal was to derive a model for the level of risk of being infected by SARS-CoV-2 and to identify and measure the place-specific factors associated with the disease and its determinants. Spatial heterogeneity was tested by comparing global Poisson regression (GPR) and local geographically weighted Poisson regression (GWPR) models. The key findings were that different city areas were impacted differently during the first three epidemic waves. The area-to-area influence was estimated to exert its effect over an area with a 4.7 km radius. Spatio-temporal heterogeneity patterns were independent of the sociodemographic and the clinical characteristics of the resident population. Significant individual risk factors for detected SARS-CoV-2 infection cases were old age, hypertension, diabetes, and comorbidities. More specifically, in the global model, the average SARS-CoV-2 infection rate decreased 0.93-fold in the 21–65 years age group compared to the >65 years age group, whereas hypertension, diabetes, and any other comorbidities (present vs absent), increased by 1.28-, 1.39- and 1.15-fold, respectively. The local GWPR model had a better fit than the GPR.

Study 3 (Chapter 5) aimed to identify and explore the hospital admission risk factors associated with the length of stay (LoS) by applying a relatively novel statistical method for count data using predictors among COVID-19 patients in Bologna, Italy. The second goal of this study was to model the LoS of COVID-19 patients to understand which covariates significantly influenced it and identify the potential risk factors associated with LoS in Bologna's hospitals from 1 February 2020 to 10 May 2021. The clinical settings we focused on were the Intensive Care Unit (ICU) and ordinary hospitalization, including low-intensity stays. We used Poisson, negative binomial (NB), Hurdle–Poisson, and Hurdle–NB regression models to model the LoS. The fitted models were compared using the Akaike information criterion (AIC), Vuong's test criteria, and Rootograms. We also used quantile regression to model the effects of covariates on

the quantile values of the response variable (LoS) using a Poisson distribution and to explore a range of conditional quantile functions, thereby exposing various forms of conditional heterogeneity and controlling for unobserved individual characteristics. Hurdle–NB provided the best fit based on the chosen performance criteria. As an output from the model, we found significant changes in average LoS for each predictor. Compared with ordinary hospitalization and low-intensity stays, the ICU setting increased the average LoS by 1.84-fold. Being hospitalized in long-term hospitals was another contributing factor for LoS, increasing the average LoS by 1.58 compared with regular hospitals. When compared with the age group [50, 60) chosen as the reference, the average LoS decreased in the age groups [0, 10), [30, 40), and [40, 50), and increased in the oldest age group [80, 102). Compared with the second wave, which was chosen as the reference, the third wave did not significantly affect the average LoS, whereas it increased by 1.11-fold during the first wave and decreased by 0.77-fold during out-wave periods. The quantile regression results showed that covariates related to the ICU setting, hospitals with longer hospitalization, the first wave, and the out-waves were statistically significant for all the modeled quantiles. Our study's results can help focus on the risk factors that lead to an increased LoS among COVID-19 patients and benchmark different models that can be adopted for these analyses.

Study 4 (Chapter 6) is an ongoing project that aims to adopt survival models and machine learning methods to predict the risk of nonsurgical premature menopause (NSPM) among childhood cancer survivors over time. Female Childhood cancer survivors face a heightened risk of NSPM due to treatment-induced toxicities. During treatment, patients are exposed to chemotherapy and radiation designed to kill cancer cells; however, these harsh exposures can also damage the surrounding healthy cells. The main task is to develop a landmark predictive model that will predict the probability of cancer patients developing NSPM based on the type and magnitude of exposure. Data was acquired for 7,891 female participants of the Childhood Cancer Survivor Study (CCSS), an ongoing multi-institutional North American cohort study of 5-year survivors of childhood cancer treated between 1970-1999. In CCSS, the follow-up concluded on 11/25/2016. The Outcome variables are the ovarian status and age at the event. The models we applied: Cause-specific hazard model, Fine-Gray regression model,

Time-specific Logistic Regression with Competing Risks , Random Survival Forest with Competing Risks (RSF-CR), and Neural Network with Competing Risks . Model performance and accuracy were measured using the time-specific area under the ROC curve (AUCt), the time-specific average positive predictive value (APt), and calibration curves on both the training and validation sets.

Table of Contents

CHAPTER ONE

General Introduction and Overview

1. Introduction

1.1. The power of data (or leverage data) in health care

Our modern digital world, especially in the healthcare sector, is immersed in an ocean of data, whether electronic health records, patient registries, or other structured and unstructured methods. This data, encompassing patient records, clinical studies, genomic sequences, and real-time health monitoring, forms the backbone of our understanding and advancement in healthcare. The use of this data in healthcare is expected to improve the quality of life, reduce operational costs, and support clinical decision-making, disease surveillance, and public health management [1]. Data drives information, information drives knowledge, and knowledge generates wisdom or value. Data will transform everything. It is essential, however, to use and apply data correctly and to ensure the quality of the data. As a result, the appropriate use and application of data in healthcare can improve patient outcomes, operational efficiency, and overall healthcare quality. It is crucial to use an appropriate methodology and sample size to generalize from the available data and infer meaningful insights or achieve better healthcare outcomes.

Leveraging the power of data in healthcare can deeply improve service delivery across multiple dimensions, from enhancing patient outcomes to optimizing operational efficiency. To engage patients and improve their health outcomes, it is important to get a complete picture of their health journey, combining patient-generated information with official records. "*There's no way you can move toward value-based care without incorporating patient-generated health data*" – Danny Sands [2]. Using real-world data, a data-driven approach can reduce health disparities and improve health equity, and hospitals and health systems can identify outcomes disparities caused by inequities and societal factors that influence health by using data.

By leveraging data and technology, the healthcare sector and its social care providers or partners can help improve the efficiency, effectiveness, and sustainability of efforts to address

health-related social needs as a regular component of healthcare delivery [3]. Advanced statistical methods and Artificial Intelligence (AI) have the potential to transform healthcare systems by enhancing efficiency and effectiveness, enabling universal health, and improving outcomes. To react to consumer needs faster, solve complex challenges, and uncover possibilities for better outcomes, hospitals and health systems must increase their efforts to manage, integrate, and analyze an unprecedented amount of data.

Modeling is a key part of public health response. It requires special attention to having better and more accurate decision-making responses. The adage 'Precision in modeling is the compass to health's true north' underscores this importance. To accurately predict disease outcomes, track disease progression, or identify risk factors—as health and wellbeing are affected by many factors —selecting the correct statistical and machine learning (ML) models is essential. As we are in the big data era, ML techniques stand out as the premier tools for extracting knowledge from data trends. In recent years, ML techniques, a subfield of artificial intelligence that allows machines to learn from past data and experiences, have become increasingly popular in predictive analytics due to their exceptional performance in managing large-scale datasets with consistent characteristics and noisy data [4]. According to observational research, ML is appropriate for building prediction models by extracting patterns from massive datasets [4]. ML methods will improve clinical decision-making in trials where multiple factors may influence outcomes. For example, if a treatment has a high rate of errors, it could lead to a number of problems or even death. As a result, researchers have begun to employ AI in medical treatment [5]. We employed high-dimensional datasets, settings with many categorical and mixed variables where traditional computer science techniques and algorithms fail to solve large and complex models but where other techniques like AI, ML, and Deep Learning (DL) allow researchers to produce reliable, repeatable decisions and results and uncover hidden insights through learning from relationships and trends in the data [6].

Overall, this PhD project aims to develop and apply advanced statistical and ML models to analyze complex datasets, which are crucial for providing better healthcare services by

leveraging the power of data on various health problems to improve informed decision-making and predict health-related outcomes.

1.2. Research questions

Data-driven decision-making processes can improve clinical outcomes and patient care by integrating them into healthcare services. It also aims to explore the potential of personalized data analysis for tailoring healthcare services based on individual patient needs, preferences, and health conditions. In order to explore how data can be leveraged to improve healthcare, we addressed the following research questions:

RQ1: Predictive Analytics and Machine Learning Models for Identifying Risk Factors in Prolonged Hospitalizations. How can predictive analytics and machine learning models, utilizing patient data, be employed to identify risk factors contributing to prolonged hospitalizations? Additionally, how can these models assist medical doctors in making informed decisions, enabling early interventions and better healthcare management? This question is specifically directed toward the University Hospital of Bologna Sant'Orsola-Malpighi (AOSP) and pertains to all departments or surgeries that admit patients through the emergency department.

RQ2: The General Medicine (GM) department has the highest patient volume and heterogeneity among other hospital specialties at ASOP. Examining hospitalization data closely is crucial because patients come with various conditions or traits. With this in mind, we sought to answer the following questions: What are the key determinants of patient length of stay (LoS) in the General Medicine (GM) department, considering factors such as medical background, demographic characteristics, arrival mode at the hospital, triage evaluation, type of diseases/signs/symptoms, and administration problems? Which specific machine learning regression models among the nine considered demonstrate superior performance in predicting hospital LoS for patients admitted to the General Medicine department?

RQ3 : Spatiotemporal Heterogeneity of SARS-CoV-2 Diffusion in Bologna. Health inequality? What was the degree of spatiotemporal heterogeneity of SARS-CoV-2 diffusion in Bologna? Were there distinct clusters or areas with varying levels of infection rates? How have SARS-CoV-2 transmission dynamics evolved over different periods within Bologna?

RQ4: Factors Associated with Longer Hospital Stays for COVID-19 Patients in Bologna. What are the key demographic, clinical, and contextual factors associated with longer hospital stays for COVID-19 patients in Bologna? How can socioeconomic factors, different time periods or waves, comorbidities, and disease severity be effectively integrated into regression models to predict length of stay (LoS)?

RQ5: Landmark Predictive Model for Nonsurgical Premature Menopause (NSPM) in Female Childhood Cancer Survivors. How can a landmark predictive model be developed to assess the probability of nonsurgical premature menopause (NSPM) in female childhood cancer survivors, considering treatment-induced toxicities such as chemotherapy and radiation? Additionally, how can this model aid in determining the necessity and timing of fertility procedures for these patients?

1.3. Thesis Outline

Following a brief explanation of the background concepts above, a detailed overview of the work presented in each chapter is provided below:

Chapter one: Describes the general introduction of the thesis, providing a comprehensive overview of the research's context, objectives, research questions, and thesis outline.

Chapter two: The study presented in this chapter aims to estimate and compare different models to predict the Length of Stay (LoS) and the Prolonged Length of Stay (PLoS) of inpatients admitted through the emergency department (ED) in general patient settings, offering insights that can contribute to informed decision-making in healthcare settings. This aim is not only to promote any specific model but rather to suggest a decision-supporting tool (i.e., a prediction framework).

Chapter three: This chapter discusses the Length of Stay (LoS) within a specific department/specialty, namely the General Medicine (GM) department, which has the highest patient volume and heterogeneity among other hospital specialties. This study, rooted in the significance of LoS as an efficiency metric, employs advanced ML regression models to predict and compare the actual number of days a patient stays in the hospital. By incorporating demographic and clinical information available at admission, the study aims to contribute

valuable insights into enhancing predictive accuracy for LoS in a high-volume and heterogeneous medical setting. Ultimately, the findings hold potential for optimizing resource allocation and improving overall efficiency in patient care within the General Medicine department.

Chapter four: Describes the spatiotemporal heterogeneity of SARS-CoV-2 diffusion in Bologna city using a geographically weighted Poisson regression model. This chapter explores the variation in infection rates across different clusters or areas and examines how the transmission dynamics of SARS-CoV-2 evolved over various periods or waves within Bologna.

Chapter five: This chapter introduces the technical, methodological, interpretational, and practical aspects of leveraging data for analyzing COVID-19 hospitalizations or length of stay (LoS) using a count regression model and quantile regression in Bologna, Italy. The results and findings suggest a framework for understanding the impact of various clinical factors and count models on healthcare service delivery.

Chapter six: The study presented in this chapter aims to predict the risk factors associated with NSPM before prespecified ages among female childhood cancer survivors, enabling early intervention and improved quality of life. As a late effect of cancer treatments, NSPM can occur at any point during a survivor's young adulthood, especially after reaching the 5-year survival milestone. The main objective is to facilitate deeper conversations among physicians, patients, and their families regarding the significance of fertility preservation. To achieve this, we used landmark predictive models that assess the likelihood of patients developing NSPM.

Chapter seven: This chapter presents the comprehensive conclusion drawn from the findings and discussions presented throughout the thesis. Additionally, recommendations based on the study's results are offered, suggesting potential avenues for further research or practical applications.

References :

1. Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. J Big Data. 2022;9(1):3. doi: 10.1186/s40537-021-00553-4. Epub 2022 Jan 6. PMID: 35013701; PMCID: PMC8733917.
2. Sands DZ, Halamka JD. PatientSite: Patient-centered communication, services, and access to information. In: Nelson R, Ball M, eds. Consumer Informatics: Applications and Strategies in Cyber Health Care: Springer 2004:20-32.
3. Robison LL, Armstrong GT, Boice JD, et al. The Childhood Cancer Survivor Study: a National Cancer Institute–supported resource for outcome and intervention research. Journal of clinical oncology 2009; 27(14): 2308.
4. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
5. Janiesch, C., Zschech, P. & Heinrich, K. Machine learning and deep learning. Electron Markets (2021).
6. Anjum, Nasreen; Asif, Amna; Kiran, Mehreen; Jabeen, Fouzia; Yang, Zhaohui; Huang, Chongwen; et al. (2021): Intelligent COVID-19 Forecasting, Diagnoses and Monitoring Systems: A Survey.

CHAPTER TWO

Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis

In this chapter, we draw heavily from our published work [∗], about prediction of prolonged length of stay for patients admitted through the emergency department to the Sant'Orsola Malpighi University Hospital of Bologna, Italy.

∗ *Zeleke AJ, Palumbo P, Tubertini P, Miglio R, Chiari L. Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis. Front Artif Intell. 2023 Jul 28;6:1179226. doi: 10.3389/frai.2023.1179226. PMID: 37588696; PMCID: PMC10426288. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10426288/*

Abstract

Objective: This study aims to develop and compare different models to predict the Length of Stay (LoS) and the Prolonged Length of Stay (PLoS) of inpatients admitted through the emergency department (ED) in general patient settings. This aim is not only to promote any specific model but rather to suggest a decision-supporting tool (i.e., a prediction framework).

Methods: We analyzed a dataset of patients admitted through the ED to the Sant"Orsola Malpighi University Hospital of Bologna, Italy, between January 1 and October 26, 2022. PLoS was defined as any hospitalization with LoS longer than 6 days. We deployed six classification algorithms for predicting PLoS: Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GB), AdaBoost, K-Nearest Neighbors (KNN), and logistic regression (LoR). We evaluated the performance of these models with the Brier score, the area under the ROC curve (AUC), accuracy, sensitivity (recall), specificity, precision, and F1-score. We further developed eight regression models for LoS prediction: Linear Regression (LR), including the penalized linear models Least Absolute Shrinkage and Selection Operator (LASSO), Ridge and Elastic-net regression, Support vector regression, RF regression, KNN, and eXtreme Gradient Boosting (XGBoost) regression. Themodel performances were measured by their mean square error, mean absolute error, and mean relative error. The dataset was randomly split into a training set (70%) and a validation set (30%).

Results: A total of 12,858 eligible patients were included in our study, of whom 60.88% had a PloS. The GB classifier best predicted PloS (accuracy 75%, AUC 75.4%, Brier score 0.181), followed by LoR classifier (accuracy 75%, AUC 75.2%, Brier score 0.182). These models also showed to be adequately calibrated. Ridge and XGBoost regressions best predicted LoS, with the smallest total prediction error. The overall prediction error is between 6 and 7 days, meaning there is a 6–7 day mean difference between actual and predicted LoS.

Conclusion: Our results demonstrate the potential of machine learning-based methods to predict LoS and provide valuable insights into the risks behind prolonged hospitalizations. In addition to physicians' clinical expertise, the results of these models can be utilized as input to make informed decisions, such as predicting hospitalizations and enhancing the overall performance of a public healthcare system.

Keywords: Emergency department, prolonged length of stay, machine learning, prediction, classification, regression

2.1. Introduction

2.1.1. Importance of addressing hospitalization LoS after an emergency department visit

The Length of Stay (LoS) measures the time a patient spends in a hospital, from admission to discharge. It is a key indicator of the quality of hospital services, including the speed and efficiency of patient treatment, the prevention of hospital-acquired infections, the ability to anticipate prolonged stays due to preexisting medical conditions, resource utilization, and the cost of inpatient care. LoS can also be used to evaluate the success of surgical procedures and patient outcomes. With an in-depth understanding of LoS and potential adverse events, hospitals can make informed decisions and improve patients' overall quality of care. Accurate LoS prediction enables the efficient use of medical resources, better clinical decision-making, and provision of useful prognostic information. In hospital management, LoS is critical in determining hospital costs and patient satisfaction. Furthermore, it is associated with disease severity and mortality (Paterson et al.,2006). During an ED visit, some predictors of hospital LoS were known before admission to the hospital. Prior studies have shown that patients in EDs have a longer LoS (Krochmal and Riley, 1994; Liew et al., 2003). It has been demonstrated that

extended hospital stays negatively affect clinical outcomes: according to Sud et al. (2017), long LoS is associated with increased mortality and readmission rates; the results of Bo et al. (2016) indicated that PLoS is associated with cognitive impairment, functional limitations, and higher burdens of comorbidity; the results of Emori and Gaynes (1993) also indicated that PLoS increased the risk of hospital acquired infections. Patients are prioritized based on their level of medical need in a triage plan to enhance healthcare and reduce mortality. Models that predict patient-related outcome measures and LoS are useful tools for maximizing healthcare utilization (Gellman, 1974). As a result, policymakers and clinicians could determine how to allocate resources among different approaches by comparing treatments across disciplines.

2.1.2. Methodological review/predictive modeling of PLoS

Machine learning (ML) provides innovative methods in data predictions that are widely used. Numerous studies have examined how different predictive models can predict LoS more accurately (Lu et al., 2015). A prediction model based on factors affecting LoS has been developed in previous studies using multiple supervised learning techniques. For categorical outcomes, including logistic regression (LoR), random forest (RF), Gradient Boosting (GB), K-nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), and artificial neural networks (ANN; Hachesu et al., 2013; LaFaro et al., 2015) were used to predict LoS. In a study by Chuang et al. (2018), LoR, SVM, RF, multivariate adaptive regression splines (MARS), classification and regression tree (CART), etc. were used to study the prediction of PLoS in patients undergoing general surgery. The RF classifier showed the highest performance. In another interesting study, for a continuous outcome, Caetano et al. (2015) used and compared regressors, including multiple regression (MR), RF regression, decision tree (DT), neural network (NN), and support vector machine (SVM) regression. The RF regression showed the highest performance. The performance of ensemble learning models (like RF, GB, AdaBoost) is usually better than that of single learning models (Han et al.,2019). An alternative, data-driven approach to predictive analytics in emergency care is available through preprocessing, data mining, and machine learning techniques applied to big data stored in electronic health records (EHRs; Yu et al., 2018). In other clinical data from inpatients with lower limb fractures, Colella et al. (2021) employed similar ML techniques to predict PLoS, by dividing the outcome variable

into two classes. Kirchebner et al. (2020) conducted an exploratory study on hospitalized schizophrenic patients to predict PLoS. This study selected the most significant features using a forward selection procedure. Then various machine learning classification algorithms were used for binary outcomes: with and without prolonged LoS. Overall in the literature, SVM, GB, LoR, NN, and RF are the most common and widely used supervised ML classifier algorithms used to estimate LoS (Jiang et al., 2010; Morton et al., 2014). **Table 1** provides a brief overview of ML models, prediction outcomes, and the target groups for which LoS was predicted.

Table 1. Brief review of ML models and patients groups for predicting hospital patients' LoS.

Ref.	**Outcome: Prediction type**	**ML models**	**Target group**	**Results**
Mekhaldi et al. (2020)	Regression	RF Regressor, Gradient Boosting Regressor;	General patients	GB performed better than RF; performance were checked by MSE, the R-squared and the Adjusted R-squared.
Daghistani et al. (2019)	Classification	RF, ANN,SVM, BN	Cardiac patients	RF model outperformed all other models : sensitivity (0.80), accuracy (0.80), and AUROC (0.94).
Tsai et al. (2016)	Regression	LR, ANN	Cardiac patients	LR model performed slightly better than ANN models, with a MAE value of 3.76 and 3.87
Symum and Zayas-Castro (2020)	Classification	DT C5.0, linear SVM, KNN, RF, and multi-layered artificial neural net	Chronic disease(congestive heart failure, acute myocardial infarction, COPD, pneumonia, type 2 diabetes.	For all patient groups, LSVM (Lagrangian SVM) with wrapper feature selection performed well.
Tanuja et al. (2011)	Classification	Naive Bayes; KNN; DT classifiers ; Multi-layer backpropagation	General patients	MLP and NB models had the best classification accuracy of around 85%, while KNN performed poorly with only 63.6% accuracy
Combes et al. (2014)	Regression & Classification	Two based models: *Classifier*: RF, LMT(Logistical model tree), MP, DT(C4.5-J48), NBTree, REPTree, and SVM. *Regression*: LR, SV regression, MLP,	Pediatric	Using 10-fold cross-validation, obtained the best performances in using logistic regression, and in continuous outcome SVM Regression showed a lower

		IRM(Isotonic regression model), M5P, PRLM(Pace regression linear models)		prediction error.
Etu et al. (2022)	Classification	LoR, GB, DT, and RF	COVID-19 Patients	The GB model outperformed the baseline classifier (LoR) and tree-based classifiers (DT and RF) with an accuracy of 85% and F1-score of 0.88 for predicting ED LoS
Alsinglawi et al. (2020a)	Regression	RF Regressor; GB Regressor; Stacking Regressor; DNN	Cardiovascular patients in the ICU	GB regressor outweighed the other proposed models, and showed a higher R-squared.
Kirchebner et al. (2020)	Classification	BT; KNN; SVM	Schizophrenic patients	Two factors have been identified as particularly influential for a prolonged forensic LoS, namely (attempted) homicide and the extent of the victim's injuries.
Thongpeth et al. (2021)	Regression	LR with three penalized linear (ridge, lasso, elastic net), and 4 ML model types: SVR, NN, RF and XGBoost	Chronic disease	The RF model had the best predictive performance with the smallest prediction errors, while linear ridge regression had the poorest prediction performance with the largest prediction errors.

LoR, logistic regression; LR, linear regression; RF, random forest; NB, Naive Bayes; ANN, artificial neural network; SVM, support vector machine; MLP, Multi-layer backpropagation; DT, decision tree; GB, Gradient Boosting; XGBoost, eXtreme Gradient Boosting; KNN, K-nearest neighbors; BN, Bayesian network; COPD, chronic obstructive pulmonary disease; MSE, mean square error; ICU, intensive care unit.

2.1.3. Related works

Previous research has investigated various methods of predicting LoS with varying scopes and settings. LoS can be predicted for all patients admitted to the hospital based on nonmedical factors such as type of admission, gender, race, insurance status, place of residence, and the cost of hospitalization, as well as medical characteristics like risk/severity measures, primary condition groups, emergency degree, and prior admissions. It is also possible to predict LoS for specific diseases or surgical procedures. The most frequently reported factors that affect the ED LoS are patient age, gender, triage category, mode of arrival, the requirement for an interpreter, admission, diagnostic complexity necessitating extra testing, and the availability of resources, including staff and beds (Asaro et al., 2007; Biber et al., 2013; Rahman et al., 2020). Patient characteristics influencing LoS, such as demographics and comorbidities, are often available at triage and admission (Tsai et al., 2016). Several studies in the literature have examined the LoS

trends in general patients (Tanuja et al., 2011; Mekhaldi et al., 2020), or in particular patient populations, focusing, for instance, on a certain age group (Ackroyd-Stolarz et al., 2011; Launay et al., 2018; Marfil-Garza et al., 2018; Sir et al., 2019) or specific health conditions (e.g., cardiology; García-González et al., 2014; Tsai et al., 2016; Chuang et al., 2018; Daghistani et al., 2019), peritoneal dialysis (Wu et al., 2020), schizophrenia (Kirchebner et al., 2020), knee arthroplasty (Song et al., 2020), COVID-19 (Vekaria et al., 2021; Etu et al., 2022; Zeleke et al., 2022), abdominal pain (Dadeh and Phunyanantakorn, 2020), mental health (Wolff et al., 2015), cardiovascular diseases (Almashrafi et al., 2016; Alsinglawi et al., 2020a), or in specific discipline areas or specialties such as spine surgery (Basil and Wang, 2019) and cancer surgeries (Laky et al., 2010; Gohil et al., 2014; Jo et al., 2021). However, most of these studies have had limited sample sizes and have not considered a wide range of clinical factors. In-hospital adverse events are known to increase the risk of prolonged Length of Stay (LoS) in older patients (Ackroyd-Stolarz et al., 2011).

A study of Length of Stay (LoS) in the emergency department of a tertiary care center (van der Veen et al., 2018) found a significant association between multiple chief complaints, including headaches and chest pain, laboratory/radiology testing, and consultation with prolonged hospitalization in the ED. Another population based study conducted in Osaka, Japan (Katayama et al., 2021) showed that factors such as old age, traffic accidents, lack of a permanent address, need for nursing care, and being solitary were associated with prolonged hospitalization for patients transported by ambulance. Another retrospective study of prolonged LoS in a tertiary healthcare center in Mexico (Marfil-Garza et al., 2018) showed that demographic and disease-specific differences, such as younger age, male gender, lower physician-to-patient ratio, emergency and weekend admissions, surgery, number of comorbidities, and lower socioeconomic status, were associated with a prolonged LoS. Diseases with the greatest risk for prolonged LoS included complex conditions like bone marrow transplant, systemic mycoses, parasitosis, and complex abdominal diseases like intestinal fistulas.

2.1.4. Aims

This study used various supervised machine learning algorithms to predict the length of stay for patients admitted through the emergency department in general patient settings. The outcome was analyzed as both a dichotomous (PLoS) and continuous (LoS) variable. Data was gathered from routine triage and ED admission processes and recorded in the hospital's electronic medical records. The best-performing model was selected to make predictions and gain meaningful insights for future patients.

2.2. Materials and methods

2.2.1. Study design and population

We screened for eligibility for all admissions to the hospital through the ED of the public University Hospital of Bologna Sant'Orsola-Malpighi (AOSP), Bologna, Italy, between January 1, 2022, and October 26, 2022. AOSP is a 1,500-bed tertiary care teaching hospital in Central-Northern Italy with 70,000 emergency department visits per year, this is one of the largest hospitals in the country (Fridman et al., 2022). All the necessary steps of the clinical pathway: ED triage, medical examination, hospital admission, and hospital discharge, are shown in Figure 1. We included all patients who visited the ED, were admitted to the hospital, and stayed until they got formal permission to discharge. Any patients who left the ED, were transferred to another hospital, refused the hospitalization, died, went away after the medical examination, left without being seen, or left without notice (detail as shown in Figure 2) were excluded from the analysis.

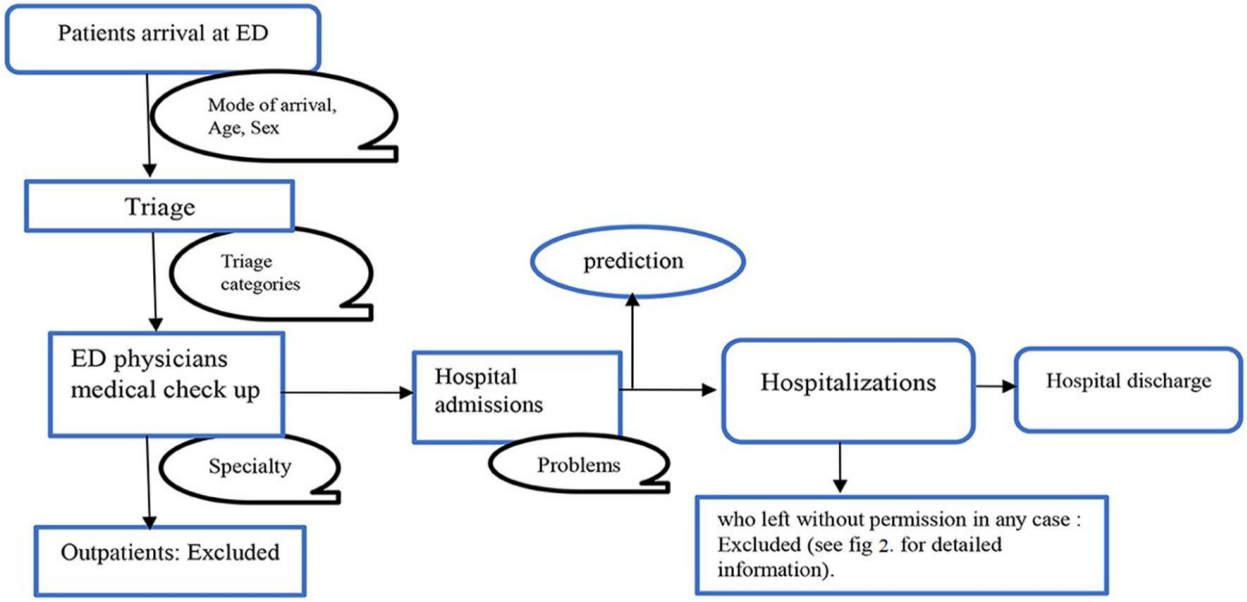


Figure 1: Clinical pathway.

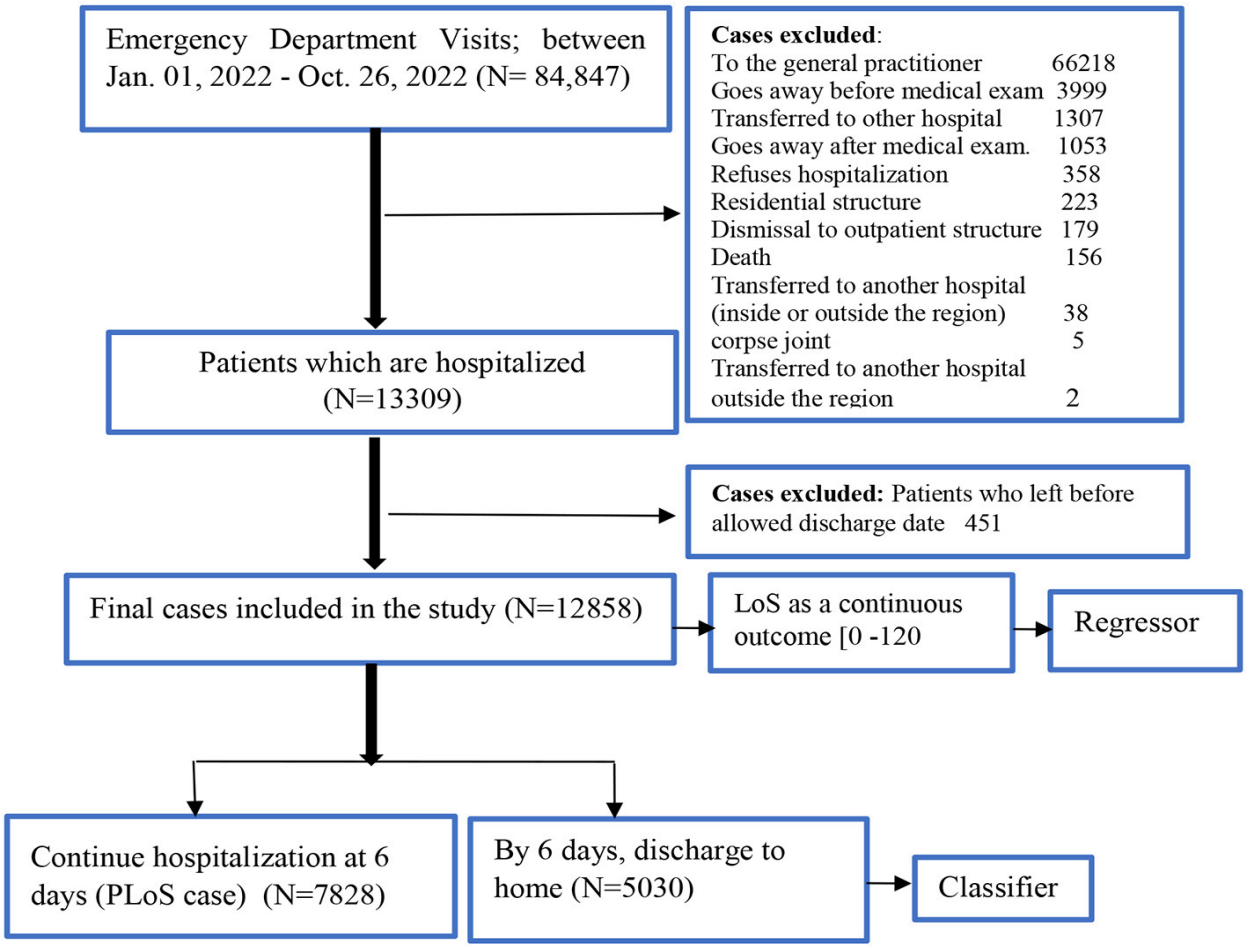


Figure 2: Flowchart of patients selection

2.2.2. Outcome variable

The primary outcome of this study was hospital length of stay (LoS) and prolonged length of stay (PloS). LoS is calculated as the number of days between admission and discharge. We defined PLoS threshold as any LoS that is longer than the reported average LoS (i.e., 6 days; Zoller et al., 2014; Song et al., 2020;Wuet al., 2020). The LoS was reclassified as binary (i.e., either "without PLoS < 6 'days' or with PLoS" $\geq$ 6 "days") for classification analysis, and LoS as a continuous outcome for regression analysis.

2.2.3. Independent variables

Any information collected at triage and available from ED admissions was considered as a predictor of LoS or PLoS. These include demographic factors (such as gender and age), mode of arrival/source of admission, risk categories as determined by triage at the entrance, and current problems or chief complaints. A detailed description of each independent feature, measure category, and outcome is presented in Supplementary Table 1.

2.2.4. Model development

2.2.4.1. Predictive models fitting and evaluation: binary outcome

The diagram in Figure 3 shows the data analysis framework we followed for developing and evaluating our predictive model. The main objective is to predict the categorical class labels of new data points or instances based on past observations. Based on the literature, six common classification algorithms were selected for comparison: GradientBoosting (GB), random forests (RF), support vector machine (SVM), K-Nearest Neighbors (KNN), AdaBoost, and logistic regression (LoR). The model with the highest prediction performance was used to identify predictive factors contributing to the outcome. We randomly divided the data into training (70%) and testing or validation (30%) sets. The analyses were performed in Scikit-learn in Python (Jupyter notebook version; Pedregosa et al., 2011).

Hastie et al. (2009) provide detailed explanations, but here we provide a brief overview of ML techniques, and hyperparameters tuning setting.

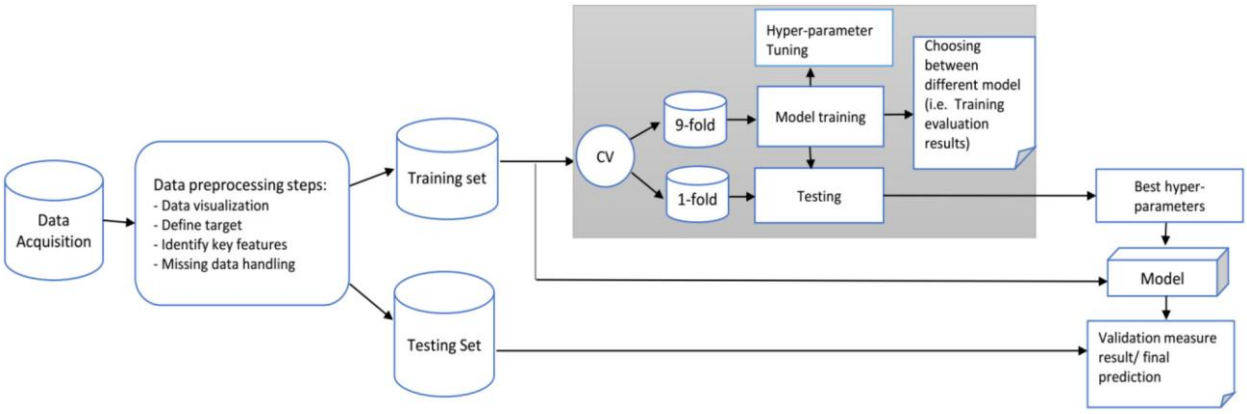


Figure 3: Proposed framework for our prediction model.

i. Random forest

In statistical applications, Random Forests (RF) are a commonly used type of supervised machine learning that can be utilized for both classification and regression tasks (Breiman, 2001; Genuer et al., 2010). RF predicts outcome labels for a group of samples by building several decision trees using a random set of covariates. The weak classifier can be transformed into a strong one by taking the majority of votes for classification and averaging in regression. To enhance the classification accuracy, multiple decision trees are combined in RF to form an ensemble classification algorithm. Each tree is grown using a bootstrapped sample from the original data (Qi, 2012). An ensemble ML method combines a series of underperforming classifiers to produce an improved classifier. The mechanism for this combination differs between ensemble algorithms. In this study, the RF model was created using the *sklearn.RandomForestClassifier* package in Python (Pedregosa et al., 2011).

ii. Gradient Boosting (GB)

Gradient Boosting is an ensemble learning model that employs decision trees as its base classifier, without bootstrap sampling (Luo et al., 2020). GB aims to create a robust predictive model by combining weak learning models, considering the bias of all previous decision trees in the model. Furthermore, unlike randomization in other methods, GB focuses on fixing the target outcomes in order to minimize errors. In this study, the GB model was constructed using the *sklearn.GradientBoostingClassifier* package in Python (Pedregosa et al., 2011).

iii. Support vector machines (SVMs)

In SVMs, the data is separated using a large gap or hyperplane to deal with linearly non-separable problems. It works by finding an optimal separating hyperplane in the feature space for classification. The Python sklearn.SVC package was used to build the SVM model for this study (Pedregosa et al., 2011).

iv. AdaBoost classifier

Similar to GB, AdaBoost classifier is also a boosting algorithm, converting a set of weak learners into a single strong learner. However, they differ on how they create weak learners during the iterative process. In GB, as mentioned, it is to minimize the cumulative predicted errors. Still, in AdaBoost it focuses on training the prior miscalculated observations and alters the data distribution to improve sample weight values. The Python sklearn.AdaBoostingClassifier package was used to build the AdaBoost model for this study (Pedregosa et al., 2011).

v. K-Nearest Neighbors (KNN)

KNN is an instance-based algorithm, which labels the test record based on its distance from similar data during training (i.e., which analyzes the similarities between the new data and the existing data and adds the new data into the category that is highly similar to the available categories). The only step in building the model is storing the training dataset. Then, the algorithm finds the closest data points in the training dataset, or its “nearest neighbors” to predict a new data point (Keller et al., 1985). Python sklearn. KNeighborsClassifier package was used to build the AdaBoost model for this study (Pedregosa et al., 2011).

vi. Logistic regression (LoR)

The LoR model is widely used in binary classification problems. The parameter of interest is estimated using maximum likelihood estimation. Similarly, Python *sklearn.LogisticRegression* package was used for this classifier.

2.2.4.1.1. Hyperparameters settings

Every machine learning (ML) technique requires the optimization of hyperparameters to enhance its performance. To develop a well-performing generalized model, it is crucial to carefully select the hyperparameters. Different algorithms will have distinct sets of hyperparameters. The hyperparameter tuning summary for each type of classifier and their descriptions used for this analysis are shown in Table 2.

Model classifiers	Hyperparameter tuning description
RF	# of _estimators = 200; longest path between root node and leaf node, max_ depth=15; class_ weight='balanced'; Number of maximum features for each tree , max_ features = sqrt; min_ samples_ split=2; min_ samples_ leaf=1; random_ state=42)
GB	# of estimators=200, max_depth=4, and loss=ls
KNN	Number of neighbors = 10; algorithms = "auto"; leaf_ size = 1; p = 1; weights='uniform'
AdaBoost	Similar to RF, define the Decision tree (Dt) classifier first in the same setting and then boost the Dt fit by AdaBoostClassifier.
SVM	Kernel=linear; degree of similarity, gamma=0.01; regularization, C = 10
LoR	No critical hyperparameters need to be tuned.

Table 2. Hyperparameter tuning summary

2.2.4.1.2. Performance measure

In building a prediction model, evaluating its performance and accuracy is important. Various metrics were used to assess the model's accuracy, including the Brier score, AUC, accuracy, sensitivity, specificity, precision, and F1-measure (Steyerberg et al., 2010). Calibration curve plots were also employed to visualize the calibration power of each model and ensure that the model fitted the data optimally. By carefully evaluating the predictive power of the model, we can ensure that the results produced by the model are reliable and can be trusted for decision-making purposes in the healthcare system.

Brier score is an overall performance measure, a measure of the accuracy of a predicted probability score (i.e., mean squared error of probability estimate). A low Brier score suggests an excellent overall performance (Steyerberg et al., 2010).

$$BS = \sum_{i=1}(\hat{p}(y_i) - y_i)^2/n$$

An evaluation metric like accuracy calculates the proportion of correct predictions (both positive and negative) out of all the predictions made by the model. Achieving the highest accuracy level is important. Sensitivity or recall reflects the number of positive predictions that were accurately identified, while specificity measures the same for negative predictions. A higher recall indicates that more true values were correctly predicted. The F1-score balances precision and recall by taking the harmonic mean of both values. The overall predictive accuracy of the model was evaluated by determining the area under the receiver operating characteristic curve (AUC). Calibration is crucial in developing and validating clinical prediction models, which refers to the match between predicted and observed risks (Steyerberg, 2019). In the case of binary outcomes, calibration measures the agreement between estimated and observed probabilities of occurrence. Calibration curves were used to assess calibration. A perfect model's calibration curve would be diagonal, meaning that the predicted probabilities align with the observed probabilities.

2.2.4.1.3. Variables importance

The most effective prediction model was utilized to determine the importance of variables. Identifying key factors in machine learning predictions is crucial. The metric used to evaluate this is the mean decrease in impurity, which calculates the average change in the impurity of nodes across all trees in the ensemble, taking into account the proportion of samples that reach each node. A higher value generally means that the feature is more significant. With high-dimensional datasets, it is crucial to properly select and rank covariates for both prediction and interpretation purposes.

2.2.4.2. Predictive models fitting and evaluation: continuous outcome

In order to minimize information loss in a classification task, we also explored it as a continuous outcome and employed regression models. Our study employed eight different learning algorithms, including linear regression (LR) and its penalized versions (Lasso, Ridge, and Elastic Net regression), as well as Support Vector Regression, Random Forest Regression, K-Nearest Neighbors (KNN), and XGBoost Regression.

i. Linear regression (LR)

This method involves fitting a linear equation to the data to establish a relationship between the independent variables and the dependent or outcome variable. The equation can then be used to make predictions based on the input data. The linear regression model is typically expressed in the following form:

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}$$

where y_i is the continuous outcome value of subject i, β_0 is intercept, β_j is the coefficient of feature j, and x_{ij} is feature j of subject i.

It is possible to estimate the regression parameter of a linear regression model using the least square method by minimizing the error term in the unknown β_j.

$$\hat{\beta} = argmin_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right\}$$

ii. Ridge regression

It works by finding the coefficients that minimize the sum of error squares by applying a penalty to those coefficients (Tibshirani, 1996).

$$\hat{\beta} = argmin_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 \right\}$$

λ is the regularization parameter that we are going to optimize.

iii. Lasso regression

The same task but uses the sum of absolute values of the weights for the penalty (Tibshirani, 1996).

$$\hat{\beta} = argmin_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

iv. Elastic-Net

A combination of lasso and ridge regression that reduces bias, better than lasso or ridge regressions (Friedman et al., 2009).

$$\hat{\beta} = argmin_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1\sum_{j=1}^{p}{\beta_j}^2 + \lambda_2\sum_{j=1}^{p}|\beta_j|\right\}$$

In contrast to prediction models, regression models focus on estimating the relationship between a set of independent variables and a continuous outcome variable. Instead of categorizing the outcome into specific classes, the regression models aim to predict the continuous value of the outcome based on the given set of independent variables. The performance measure used in regression models is typically the mean squared error, or the root mean squared error, which represents the average deviation between the predicted and actual values of the outcome variable. Regression models aim to minimize these errors, thereby providing a more accurate prediction of the continuous outcome.

Using a loss function helps us evaluate the performance of a prediction model by quantifying the difference between the predicted and the actual values. Mean square error (MSE), mean absolute error (MAE), and mean relative error (MRE) were calculated to measure the prediction performance of each model. MSE is the most widely used loss function for continuous outcomes. Still, we also considered MAE and MRE to get a more comprehensive understanding of the performance. The lower the value of the loss function, the better the model's prediction performance.

$$MSE = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}\ ;\ MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}\ ;\ \text{and}\quad MRE = \frac{\sum_{i=1}^{n}(|\hat{y}_i - y_i|/y_i)}{n}$$

where $\hat{y}_i$ and y_i are the predicted LoS and actual LoS for the ith test data.

2.3. Results

Patient selection;

Figure 1 illustrates a flowchart of patients' eligibility for analysis in the emergency department of triaging system. A total of 84,847 patient visits were recorded at the ED between January 1 and October 26, 2022. After filtering for exclusion criteria, 12,858 patients were available for analysis.

Patients characteristics summary;

Of the 12,858 eligible patients included in the study, 60.88% had a prolonged length of stay (LoS). The median age of the patients was 72 years, and the elderly age groups (50–69 and 70+) had longer LoS than the other age groups. The male patients comprised 52.6% (6,757/12,858) of the total population. 51.7% of the patients arrived at the hospital via ambulance and had a longer stay compared to those who arrived by car or on foot. In the triage categories, patients with red codes, which indicate an higher severity at the ED admission, had a longer LoS, while green and white codes showed shorter stays. Light blue codes were also associated with prolonged LoS.

The most common problems among the patients were dyspnea (15.2%), abdominal pain (9.9%), and fever/hyperpyrexia/hyperthermia (8.5%). The majority of patients were seen by specialists in general medicine (29.2%), geriatrics (12.6%), astanteria or casualty department (10.7%), obstetrics and gynecology (9.0%), and pediatrics (4.7%). A detailed breakdown of patient characteristics can be found in Table 3.

TABLE 3 Presenting characteristics of patients who visited the ED of ASOP, Bologna, Italy, 2022 (n = 12,858).

Factor	Total, n (%) (n = 12,858)	PLoS (i.e., ≥6 days)		Proportion difference (%) (with PLoS—without PLoS)
		With PLoS (n = 7,828) 60.88%	Without PLoS (n = 5,030) 39.12%	
Age, median	72	-	-	-
Age categories, n (%)				
(0–17)	1,170 (9.1)	329 (4.2)	841 (16.7)	−12.5
(18–29)	679 (5.3)	158 (2.1)	521 (10.4)	−8.3
(30–49)	1,772 (13.8)	616 (7.9)	1,176 (23.4)	−15.5
(50–69)	2,364 (18.4)	1,554 (19.9)	810 (16.1)	3.8
70 or older	6,873 (53.5)	5,171 (66.1)	1,702 (33.9)	32.2
Gender, n (%)				
Male	6,101 (47.4)	3,928 (50.2)	2,173 (43.2)	7.0
Female	6,757 (52.6)	3,900 (49.8)	2,857 (56.8)	−7.0
Mode of arrival, n (%)				
Ambulance–118	6,645 (51.7)	4,624 (59.1)	2,021 (40.2)	18.9
Own vehicle/walk-in	4,769 (37.2)	2,204 (28.2)	2,565 (51.0)	−22.8
Others[a]	1,444 (11.2)	1,000 (12.8)	444 (8.8)	4.0
Triage category				
Red	807 (6.3)	539 (6.9)	268 (5.3)	1.6
Orange	4,360 (33.9)	2,367 (30.2)	1,993 (39.6)	−9.4
Light blue	4,253 (33.1)	3,065 (39.2)	1,188 (23.6)	15.6
Green	3,224 (25.1)	1,784 (22.8)	1,440 (28.6)	−5.8
White	214 (1.7)	73 (0.9)	141 (2.8)	−1.9
Specialty, n (%)				
General medicine	3,757 (29.2)	2,995 (38.3)	762 (15.1)	23.2
Geriatric	1,624 (12.6)	1,252 (16.0)	372 (7.4)	8.6
Astanteria/casualty department	1,450 (10.7)	809 (10.3)	641 (12.7)	−2.4
Obstetrics and gynecology	1,159 (9.0)	114 (1.5)	1,045 (20.8)	−19.3
Pediatrics	609 (4.7)	193 (2.5)	416 (8.3)	−5.8
General surgery	571 (4.4)	276 (3.5)	295 (5.9)	−2.4
Infectious and tropical diseases	533 (4.1)	372 (4.8)	161 (3.2)	1.6
Orthopedics and traumatology	481 (3.7)	378 (4.8)	103 (2.0)	2.8
Urology	405 (3.2)	99 (1.3)	306 (6.1)	−4.8
Coronary unit	377 (2.9)	283 (3.6)	94 (1.9)	1.7
Pediatric surgery	376 (2.9)	77 (1.0)	299 (5.9)	−4.9
Gastroenterology	308 (2.4)	237 (3.0)	71 (1.4)	1.6
Cardiology	150 (1.2)	96 (1.2)	54 (1.1)	0.1
Intensive care	141 (1.1)	113 (1.4)	28 (0.6)	0.8
Pneumology	135 (1.1)	111 (1.4)	24 (0.5)	0.9
Nephrology	105 (0.8)	91 (1.2)	14 (0.3)	0.9
Oncology	93 (0.7)	67 (0.9)	26 (0.5)	0.4
Vascular surgery	89 (0.7)	65 (0.8)	24 (0.5)	0.3
Missing values	76 (0.6)	30 (0.4)	46 (0.9)	−0.5

(Continued)

TABLE 3 (Continued)

Factor	Total, *n* (%) (*n* = 12,858)	PLoS (i.e., ≥6 days)		
		With PLoS (*n* = 7,828) 60.88%	Without PLoS (*n* = 5,030) 39.12%	Proportion difference (%) (with PLoS—without PLoS)
Others[b]	419 (3.3)	170 (2.2)	249 (5.0)	−2.8
Problems, *n* (%)				
Dyspnea	1,954 (15.2)	1,446 (18.5)	508 (10.1)	8.4
Abdominal pain	1,268 (9.9)	739 (9.4)	529 (10.5)	−1.1
Fever/hyperpyrexia/hyperthermia	1,090 (8.5)	761 (9.7)	329 (6.5)	3.2
Problems in pregnancy > 20th week	944 (7.3)	70 (0.9)	847 (16.8)	−15.9
Non-specific minor disorders	579 (4.5)	395 (5.0)	184 (3.7)	1.4
Chest pain of suspected cardiovascular cause	524 (4.1)	329 (4.2)	195 (3.9)	0.3
Sincope/pre-sincope	344 (2.7)	220 (2.8)	114 (2.3)	0.5
Generalized asthenia	325 (2.5)	257 (3.3)	68 (1.4)	1.9
Politrauma—contusive	301 (2.3)	198 (2.5)	103 (2.0)	0.5
Pain at the side	278 (2.2)	100 (1.3)	178 (3.5)	−2.3
Nausea and/or vomiting repeated	269 (2.1)	150 (1.9)	119 (2.4)	−0.5
Heart palm/irregular wrist	251 (2.0)	156 (2.0)	95 (1.9)	0.1
Altered level of consciousness	234 (1.8)	165 (2.1)	69 (1.4)	0.7
State of confusion	213 (1.7)	162 (2.1)	51 (1.0)	1.1
Hematochezia/rectorrage/melena	194 (1.5)	136 (1.7)	58 (1.2)	0.6
Lower limbs injury	187 (1.5)	157 (2.0)	30 (0.6)	1.4
Cough/congestion	181 (1.4)	105 (1.3)	76 (1.5)	−0.2
Lower limbs pain	160 (1.2)	137 (1.8)	23 (0.5)	1.3
Chest pain not suspected due to cardiovascular cause	158 (1.2)	92 (1.2)	66 (1.3)	−0.1
Pallor/anemia	137 (1.1)	108 (1.4)	29 (0.6)	0.8
Request for urgent specialist advice	135 (1.0)	94 (1.2)	41 (0.8)	0.4
Macro-hematuria	130 (1.0)	70 (0.9)	60 (1.2)	−0.3
Diarrhea	121 (0.9)	85 (1.1)	36 (0.7)	0.4
Request for prescription or performance	120 (0.9)	75 (1.0)	45 (0.9)	0.1
Swollen/edematous leg	119 (0.9)	104 (1.3)	15 (0.3)	1.0
Weakness of extremities/symptoms associated with cerebrovascular disease	118 (0.9)	89 (1.1)	29 (0.6)	0.6
Symptoms of infection of the urinary tract	115 (0.9)	78 (1.1)	37 (0.7)	0.3
Diagnostics for biochemical images/examinations	108 (0.8)	77 (1.1)	31 (0.6)	0.4
Head trauma	99 (0.8)	54 (0.7)	45 (0.9)	−0.2
Other[c]	2,212 (17.2)	1,219 (15.6)	993 (19.7)	−4.2

[a]Taxi, helicopter 118, army ambulances, fire brigade, police, etc.
[b]Damages, Ent (ear, nose, and throat) problem, nephrology (enabled for transplantation), neonatology, pediatric oncology, semi-intensive therapy, maxillo facial surgery, hematology, thoracic surgery ophthalmology, heart surgery, neonatal intensive care, pediatric heart surgery, and dermatology.
[c]More than 135 cases.

The count plots for each patient for each specialty and problems are included in the Appendix, in Supplementary Figures 1, 2, respectively. The distribution of length of stay (LoS) for the patients is depicted in a histogram in Figure 4. The distribution of LoS values was found to be right-skewed, with a majority of patients having an LoS ranging from 1 to 20 days. To further explore the impact of different factors on LoS, a visualization of the dichotomous

outcome result for each factor is presented in Figure 5, while Figure 6 shows the continuous outcome for each factor. By examining these visualizations, we can gain insights into which factors may significantly impact LoS and further investigate the relationships between these factors and patient outcomes. Overall, these figures provide a clear and concise way to understand the distribution of LoS values and their relationship with different factors.

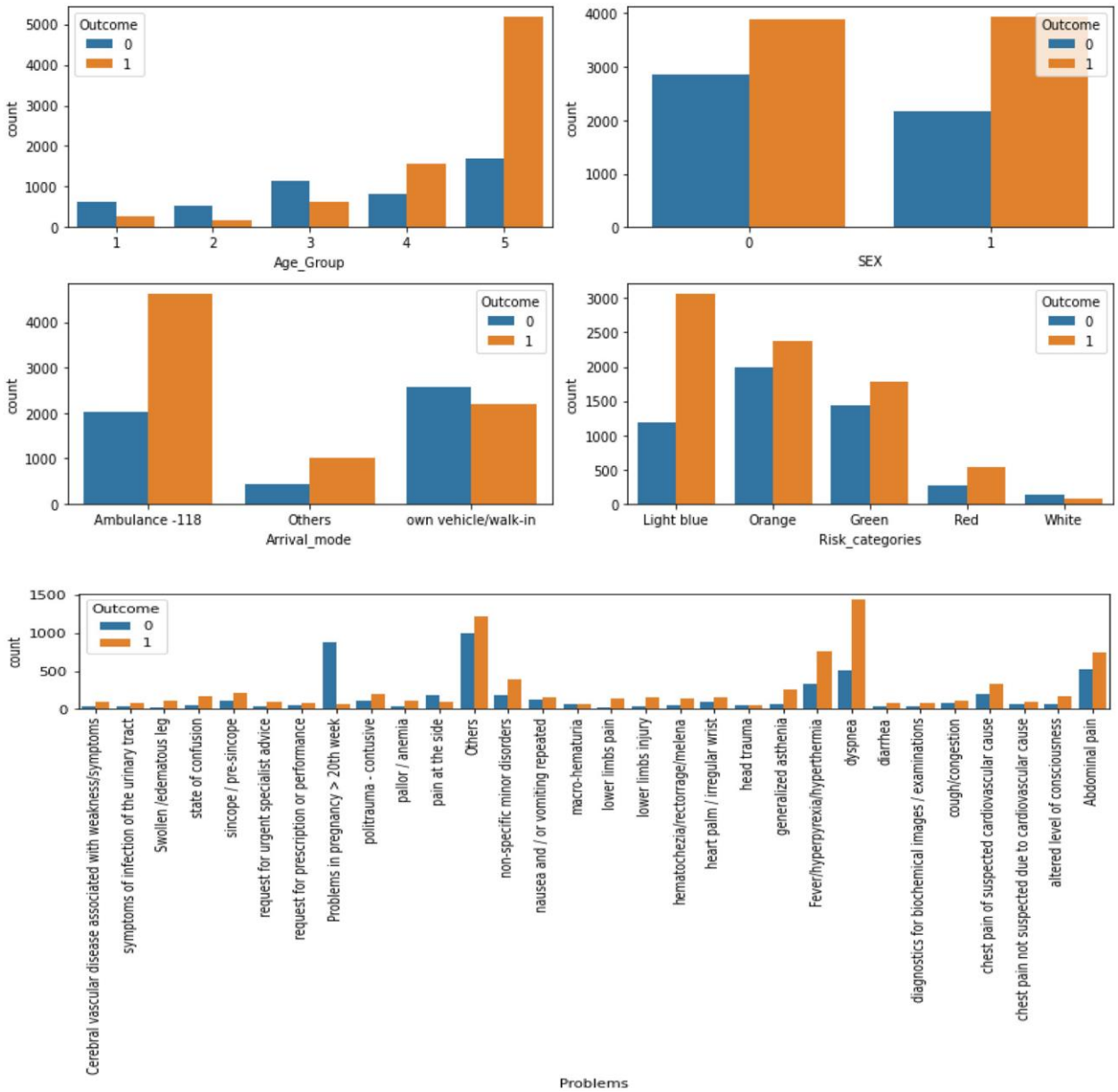


Figure 5: The results for each factor's dichotomous outcome (0, LoS < 6, without PLoS; 1, LoS ≥ 6, with PLoS).

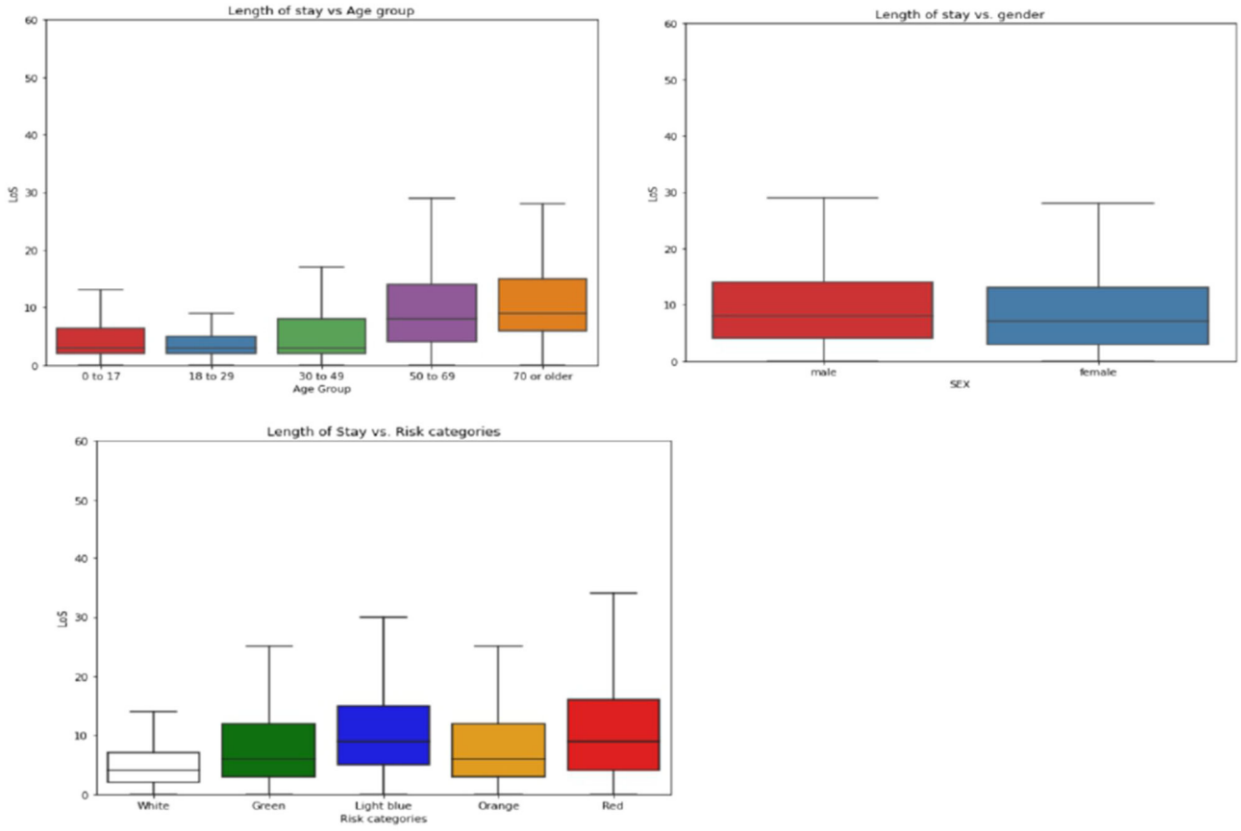


Figure 6: Boxplots of length of stay (LoS) on demographic factors, separated into two panels. (Upper panel) Shows LoS boxplots for age groups (left) and sex (right). (Lower panel) Shows the box plot for risk categories of triage evaluations.

Figure 7 displays the average LoS for each problem and specialty. The highest average LoS was observed in Intensive Care, Vascular Surgery, Nephrology, General Medicine, Gastroenterology, Infectious Diseases, Orthopedics and Traumatology, Pneumology, Geriatrics, Cardiology, Oncology, and the Coronary Unit, respectively. The average LoS was also higher for patients experiencing issues such as swollen/edematous legs, lower limb pain, generalized weakness, requests for urgent specialist advice, altered levels of consciousness, diagnostic tests for biochemical exams or images, non-specific minor disorders, dyspnea, lower limb injuries, requests for prescription refills, and pallor/anemia.

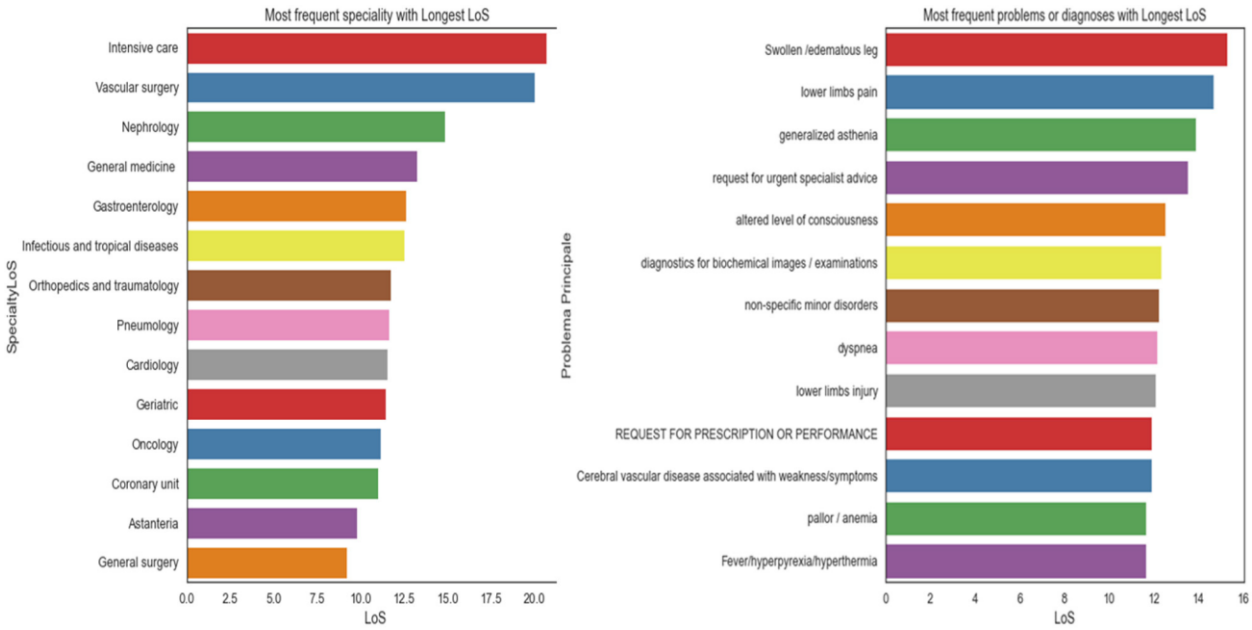


Figure 7: The average LoS for each problem (right) and for each specialty (left).

Prediction and model performance results: binary outcome;

The AUCs for all machine learning methods ranged from 0.643 for AdaBoost to 0.754 for GB (see Figure 8). GB was the best-performing classifier, followed by LoR (AUC = 0.752) and SVM (AUC = 0.726). The F1-scores ranged from 0.65 (AdaBoost) to 0.73 in GB, and 0.74 in LoR (see Table 4), indicating a high capability of these models to predict the prolonged length of stay.

TABLE 4 The prediction performance of the six classification models for PLoS prediction.

Classifier algorithms	Brier score	AUC	Precision	Recall	F1-score	Accuracy
LoR	0.182	0.752	0.75	0.75	0.74	0.75
RF	0.226	0.706	0.67	0.68	0.68	0.68
GB	0.181	0.754	0.75	0.75	0.73	0.75
SVM	0.192	0.726	0.74	0.74	0.72	0.74
AdaBoost	0.255	0.643	0.65	0.65	0.65	0.65
KNN	0.198	0.723	0.70	0.71	0.70	0.71

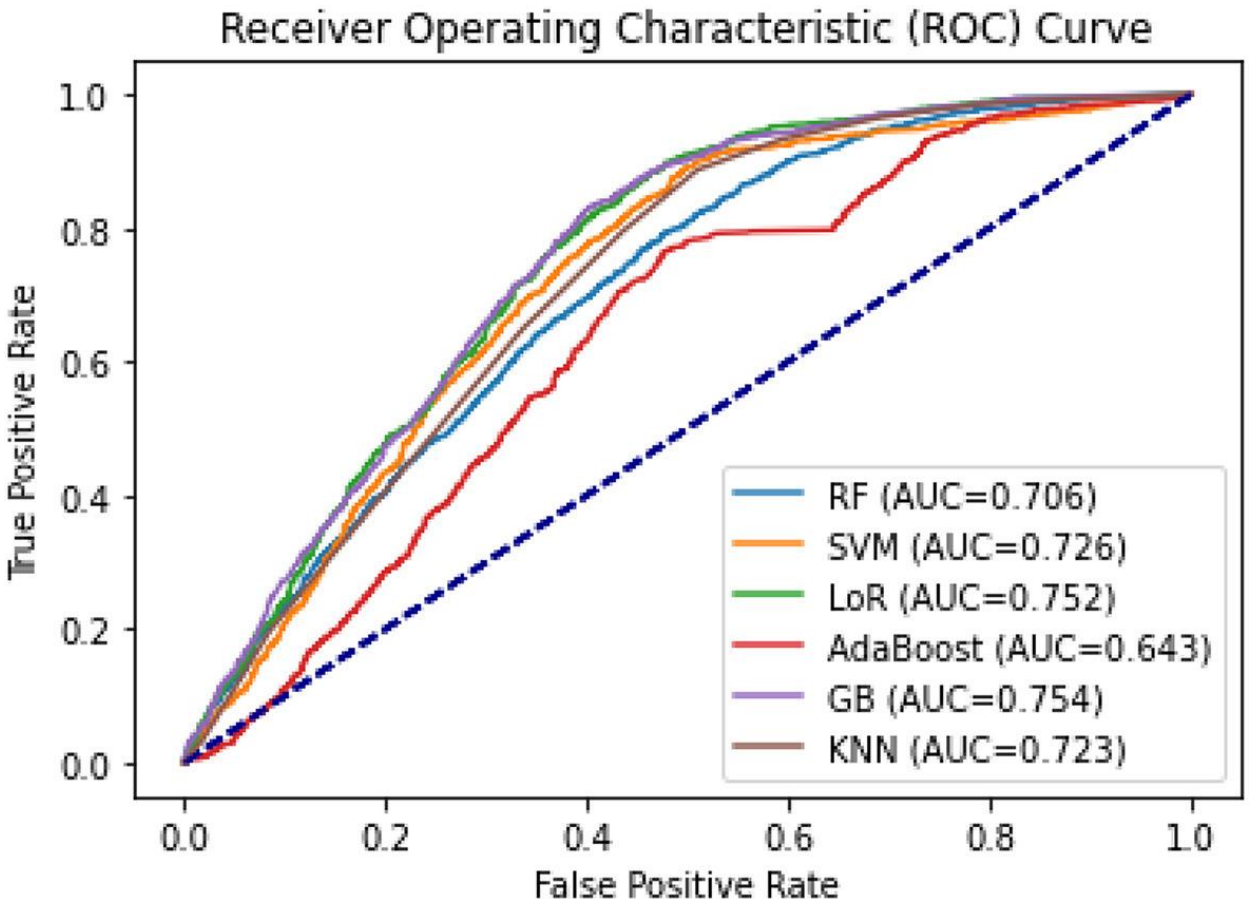


Figure: ROC curves and AUC of the six classification models for PLoS prediction.

Of the six models, the Gradient Boosting (GB) classifier demonstrated the best prediction performance in terms of accuracy (75.4%), Area Under the Curve (AUC; 0.754), and Brier score (0.181). The Logistic Regression (LoR) model had the second-best performance, with an accuracy of 75%, AUC of 0.752, and a Brier score of 0.182. Based on these results, GB and LoR were chosen as the final models due to their better performance. However, the Ada Boost model showed poor performance with the highest Brier score, lowest accuracy, and lowest AUC values. Despite attempting hyperparameter optimization, the model's accuracy did not significantly improve.

The calibration plots for each model are displayed in Figure 9. The graph shows that GB and LoR have an almost ideal calibration or optimal fit. The Random Forest (RF) and KNearest Neighbor (KNN) models are well-calibrated but tend to overestimate the probabilities of a

prolonged length of stay (PLoS) for most patients. Conversely, the Ada Boost and Support Vector Machine (SVM) models are poorly calibrated, with Ada Boost underestimating the probability of a PLoS for patients identified as low risk and overestimating it for patients in the two highest risk deciles.

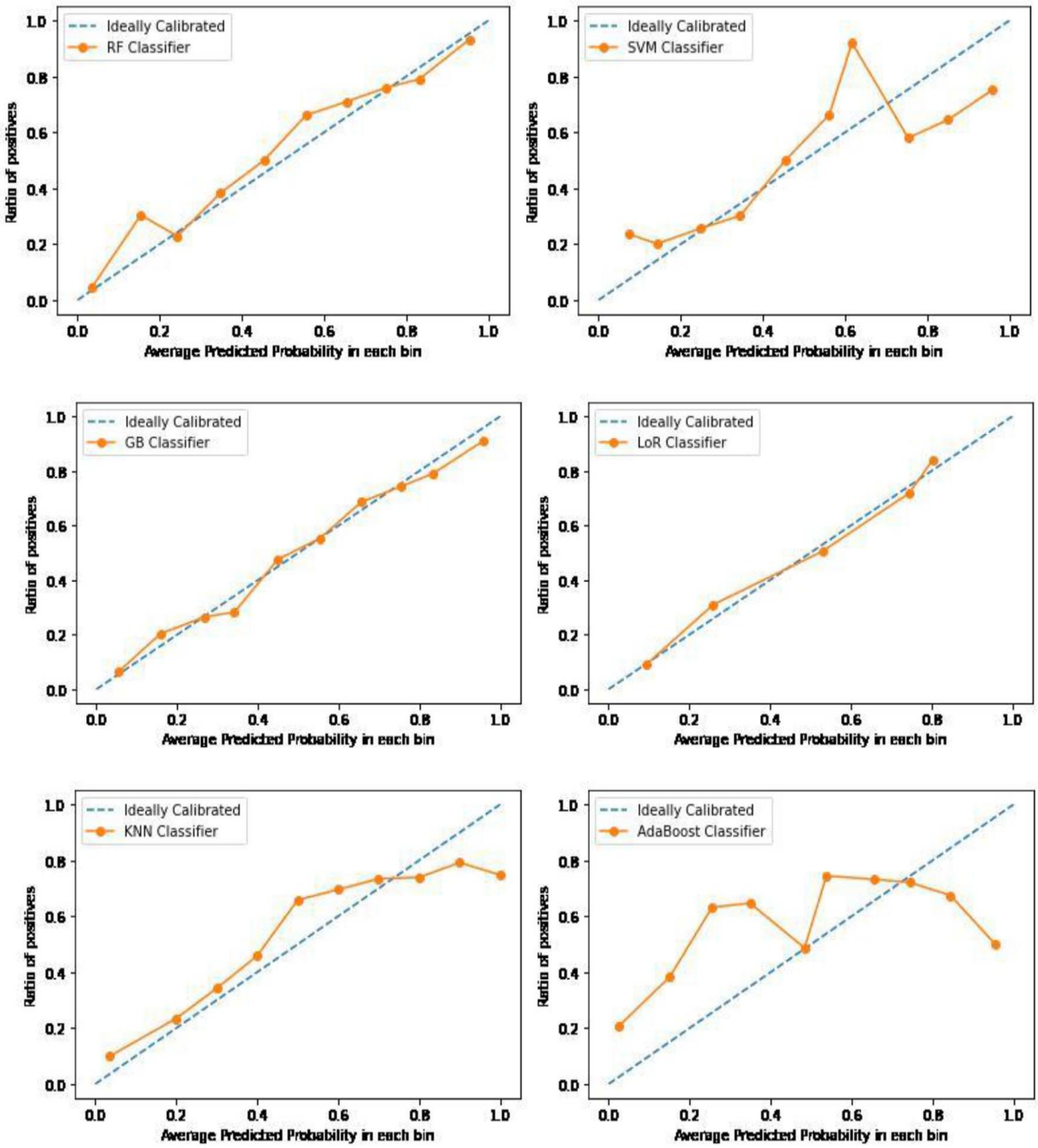


Figure: Calibration curve plots of the six classification models for PLoS prediction

The model with the highest prediction accuracy, Gradient Boosting (GB), was used to determine the relative importance of features. Figure 10 displays the results of the variable importance ranking generated by the GB model. In order of importance, the most important features were: Age Group 5 (Individuals over 70 years old), Problems in pregnancy after 20 weeks, Sex, and Age Group 4 (Individuals between 50 and 69 years old).

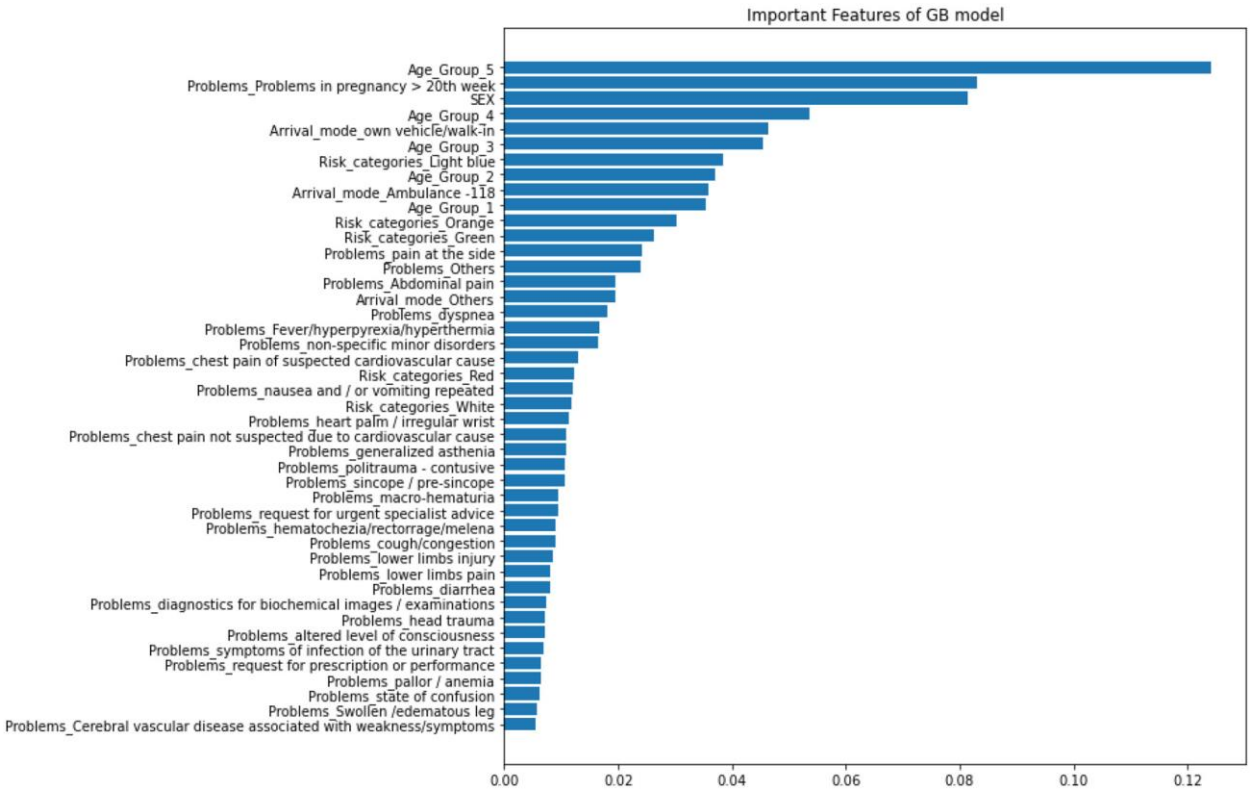


Figure 10: Gradient Boosting variable importance.

Prediction and model performance results: continuous outcome

The models used for predicting Length of Stay (LoS) were compared in Table 5, including various linear, penalized linear, and other machine learning models using different loss functions or total error measures. Ridge Regression and XGBoost Regression are identified as the best models based on their lower loss function values. The loss function or the total error performance measure is also visualized in Figure 11, where RMSE is on the left and MAE is on the right.

TABLE 5 Comparisons of classifier methods with continuous target variables for statistical and ML models applied to our datasets.

Model	Loss function			
	MSE	RMSE	MAE	MRE
Linear regression	107.045	10.346	6.671	1.283
Penalized linear models	-	-	-	-
Lasso regression	109.034	10.441	6.730	1.319
Ridge regression	**107.044**	**10.346**	**6.670**	**1.283**
Elastic net regression	109.034	10.442	6.741	1.322
Other ML learning models	-	-	-	-
Support vector regression	119.103	10.913	6.188	0.854
XGBoost regression	**107.209**	**10.354**	**6.589**	**1.213**
Random forest regressor	132.899	11.528	7.393	1.332
K-nearest neighbors regression	129.045	11.359	7.331	1.315

MSE, mean square error; RMSE, root mean square error; MAE, mean absolute error; MRE, mean relative error.
The bolded values indicate the lowest values of prediction error (e.g. Ridge and XGBoost regressions) for continuous outcomes, LoS.

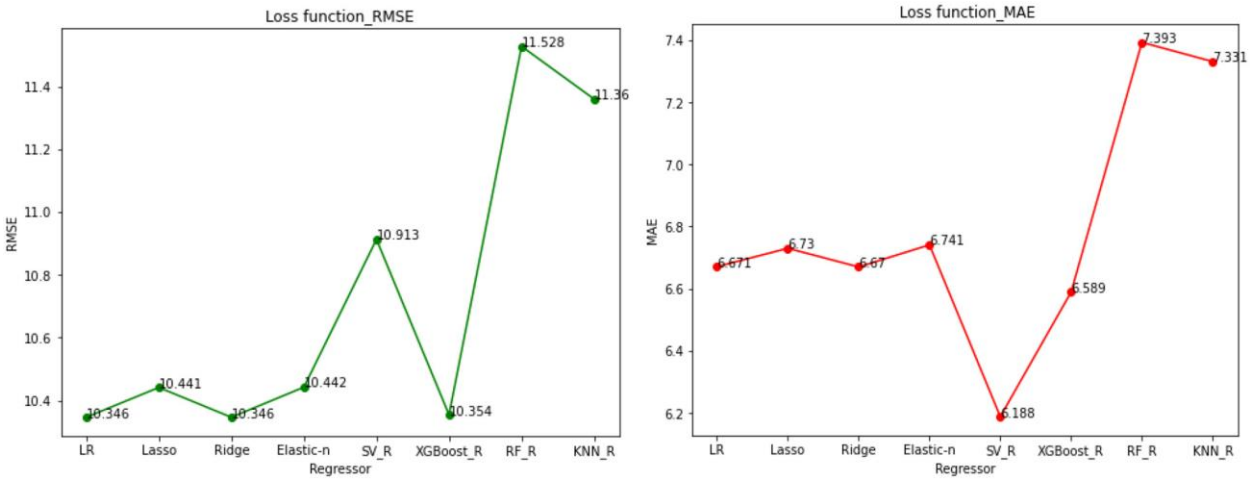


Figure 11: The loss function/total error visualization.

2.4. Discussion

In this study, we aimed to compare and evaluate predictive models using supervised machine learning algorithms for predicting prolonged length of stay in patients admitted through the emergency department (ED) in general patients settings. It is intended to promote a specific model and suggest or propose a decision-support tool as part of a predictive framework. It is well-established that reducing the length of inpatient hospital stays is one of the ways to improve the quality of life and sustainability of healthcare systems (Baek et al., 2018). Therefore, our study aims to assist physicians and doctors in making informed decisions

that enable personalized interventions and guide their decision-making process to predict hospitalizations and enhance healthcare quality.

In most PLoS prediction models, predicting the outcome relies on either classification or regression. Our study utilized two separate modeling methods to predict the outcome, employing both a dichotomous value (PLoS), and a continuous value (LoS)—that is to minimize information loss in a classification task. Adopting precise and accurate modeling techniques improves the results and interpretations. In recent years, the prediction of patient LoS for various diseases and scenarios has been extensively explored using a variety of statistical and machine learning methods such as Logistic Regression (LoR), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), decision tree-based methods, among others (Barsasella et al., 2022).

Of the six classifiers evaluated in this study (LoR, RF, SVM, GB, AdaBoost, and KNN), five of them, excluding AdaBoost, had AUCs > 0.7, suggesting them as effective tools to predict the outcome (Florkowski, 2008). The predictive performance of the classifier models was evaluated using popular statistical indicators such as accuracy, AUC, and Brier score. GB performed the best among the six classifier models, followed by LoR. AdaBoost showed poor performance as it underestimated the probability of PLoS in patients identified as low risk and overestimated it in two patient deciles classified as high risk. Similar results were observed in other studies (Alsinglawi et al., 2020b), which used ML models to predict LoS for adult ICU cardiovascular hospitalizations, with the best results obtained using the GB algorithm.

Several studies, including (Kong et al., 2020; Jo et al., 2021; Wu et al., 2021; Xiong et al., 2022), have shown that the GB classifier outperforms other algorithms in predicting PLoS, with reported accuracy, AUC, and Brier score ranging from 75.3 to 82.9%, 0.74 to 0.873, and 0.122 to 0.156, respectively. Our study's findings are consistent with these results. In contrast to some other studies, Random Forest (RF), a widely used ensemble model, has been shown to perform well in certain contexts. For instance, in Xue et al. (2022), RF achieved high accuracy, AUC, and Brier scores of 0.822, 85.8%, and 0.137, respectively, suggesting its efficacy for predicting

length of stay in hospital patients. These findings highlight the importance of carefully selecting the appropriate machine learning algorithm based on the specific data and problem being addressed. While RF may be a strong choice for certain applications, it may not necessarily be the best option in all cases. Therefore, it is crucial to systematically compare the performance of different algorithms and identify the optimal model for a given dataset. Such efforts can ultimately lead to more accurate and reliable predictions for clinical decision-making. Moreover, RF has demonstrated superior performance in predicting the outcome in various healthcare contexts. For example, RF has been shown to perform well in predicting LoS in newborns (Thompson et al., 2018), patients undergoing general surgery (Chuang et al., 2015), and individuals with COPD (chronic obstructive pulmonary disease; Luo et al.,2017). However, the results may vary depending on the specific patient population, clinical variables included in the model, and machine learning algorithm used. Moreover, we analyzed the importance of the features used in our best models, i.e., GB. In order of importance, the most important features were: Age Group 5 (Individuals over 70 years old), Problems in pregnancy after 20 weeks, Sex, and Age Group 4 (Individuals between 50 and 69 years old).

In addition, our study also aimed to predict continuous outcomes using eight ML regression models, as described in the methodology. After evaluating the models' performance, we found that Ridge and XGBoost regressions outperformed the others, resulting in lower prediction errors. Our findings align with previous studies, such as Chen and Klasky (2022), which reported similar results with lower prediction errors or loss functions. For instance, they reported the lowest mean absolute error between prediction and actual duration to be around 4 days, while our study showed a similar result of around 6 days. In addition, the XGBoost regression model also showed better results in Gabriel et al. (2023) for spine surgery LoS prediction. In another study on regression outcomes (Caetano et al., 2014), which examined the general patient population, six regression techniques were compared, including average prediction, decision trees, multiple regression, ANN ensembles, RF, and SVM. The RF regression model was found to yield the most accurate results with the lowest loss. Overall, our study adds to the existing body of literature highlighting the effectiveness of machine learning regression models in predicting

continuous outcomes in healthcare. In particular, our results demonstrate the potential of Ridge and XGBoost regressions in improving the accuracy of LoS prediction.

To summarize, selecting the most appropriate ML algorithm that matches the specific data and problem at hand and comparing the performance of different algorithms are crucial steps in identifying the optimal model for a given dataset to ensure accurate and reliable clinical decisions. The best-performing models can then be selected as the final models. As a result, GB followed by LoR is our best-performing classification model, while Ridge Regression and XGBoost Regression were the regression model choices. These final models can now be utilized to make informed decisions or derive meaningful insights for future patients. It is important to note that the choice of the optimal model may depend on various factors, such as the type of data, the problem being addressed, and the specific goals of the analysis. Therefore, it is recommended to evaluate and compare the performance of different models when developing predictive models for various clinical applications.

One of the strengths of our study was that we used all data from ED-admitted patients, so heterogeneous patients were included in the analysis. Moreover, we evaluated several ML techniques for predicting both a categorical and a continuous outcome. However, our study has some limitations that should be recognized. One limitation of the study is that vital signs for triage evaluation information and laboratory test results were not available, which is probably one of the most important indicators (Calzavacca et al., 2010); and data was only collected from one hospital so we were not able to validate the prediction model externally. Moreover, the results of this study may be biased toward other normative periods since the data were collected during the COVID-19 pandemic. Furthermore, interpreting ML results can be difficult due to the black-box nature of some models, which can make it challenging to understand the factors that contribute to the final prediction. However, linear models such as LASSO, Ridge, Elastic-Net Regression, and Logistic Regression provide regression coefficients, making them transparent and easily interpretable (Kotsiantis et al., 2006; Deo, 2015). Other techniques like feature selection and model-agnostic interpretability methods can also improve transparency.

In next Chapter four, we focused on a specific specialty or disease that is prevalent in the hospital. In addition, efforts were made to incorporate missing features such as vital signs in triage evaluation and laboratory test results. The aim is to enhance the dataset by adding more information regarding features and patients to produce better results and tackle more advanced prediction tasks such as Length of Stay (LoS) after surgeries and utilization of critical hospital resources.

2.5. Conclusions

As a result of our research, we have found that ML models are effective in predicting outcomes. Our findings showed that the GB classifier performed best, followed by LoR. These models can be utilized as a decision-support tool to inform healthcare decisions and predict new patient hospitalizations. Additionally, for continuous outcomes, Ridge regression and XGBoost regression displayed the best prediction performance with the lowest total prediction error. Healthcare providers can utilize our models to predict the hospitalization of new patients or to drive quality improvement initiatives. It is worth mentioning that this study is the first of its kind conducted in this hospital and can serve as a reference for future similar studies and provide valuable insights for informed decision-making.

References:

1. Ackroyd-Stolarz, S., Read Guernsey, J., Mackinnon, N. J., and Kovacs, G. (2011). The association between a prolonged stay in the emergency department and adverse events in older patients admitted to hospital: a retrospective cohort study. Br. Med. J. Qual. Saf. 20, 564–569. doi: 10.1136/bmjqs.2009.034926
2. Almashrafi, A., Elmontsri, M., and Aylin, P. (2016). Systematic review of factors influencing length of stay in ICU after adult cardiac surgery. BMC Health Serv. Res. 16, 318. doi: 10.1186/s12913-016-1591-3
3. Alsinglawi, B., Alnajjar, F., Mubin, O., Novoa, M., Alorjani, M., Karajeh, O., et al. (2020a). "Predicting length of stay for cardiovascular hospitalizations in the intensive care unit: machine learning approach," in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). (*IEEE*), 5442–5445.
4. Alsinglawi, B., Alnajjar, F., Mubin, O., Novoa, M., Alorjani, M., Karajeh, O., et al. (2020b). Predicting length of stay for cardiovascular hospitalizations in the intensive care unit:Machine learning approach. Annu. Int. Conf. *IEEE Eng*. Med. Biol. Soc. 2020, 5442–5445. doi: 10.1109/EMBC44109.2020.9175889
5. Asaro, P. V., Lewis, L. M., and Boxerman, S. B. (2007). The impact of input and output factors on emergency department throughput. Acad. Emerg. Med. 14, 235–242. doi: 10.1197/j.aem.2006.10.104
6. Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., Yoo, S., et al. (2018). Analysis of length of hospital stay using electronic health records: a statistical and data mining approach. PLoS ONE 13, e0195901. doi: 10.1371/journal.pone.0195901
7. Barsasella, D., Bah, K., Mishra, P., Uddin, M., Dhar, E., Suryani, D. L., et al. (2022). A machine learning model to predict length of stay and mortality among diabetes and hypertension inpatients. Medicina 58, 111568. doi: 10.3390/medicina58111568.
8. Basil, G.W., andWang,M. Y. (2019). Trends in outpatientminimally invasive spine surgery. J. Spine Surg. 5, S108–S114. doi: 10.21037/jss.2019.04.17
9. Biber, R., Bail, H. J., Sieber, C., Weis, P., Christ, M., Singler, K., et al. (2013). Correlation between age, emergency department length of stay and hospital admission rate in

emergency department patients aged 70 years. Gerontology 59, 17–22. doi: 10.1159/000342202.

10. Bo, M., Fonte, G., Pivaro, F., Bonetto, M., Comi, C., Giorgis, V., et al. (2016). Prevalence of and factors associated with prolonged length of stay in older hospitalized medical patients. Geriatr. Gerontol. Int. 16, 314–321. doi: 10.1111/ggi.12471
11. Breiman, L. (2001). Random forests. Mach. Learn. 45, 5–32. doi: 10.1023/A:1010933404324
12. Caetano N., Laureano R., and Cortez P. (2014). "A data-driven approach to predict hospital length of stay - a portuguese case study," in Proceedings of the 16th International Conference on Enterprise Information Systems - Volume 1: ICEIS (SciTePress), 407–414. doi: 10.5220/0004892204070414
13. Caetano, N., Cortez, P., and Laureano, R. M. S. (2015). Using data mining for prediction of hospital length of stay: an application of the CRISP-DM. Methodology 9, 149–166. doi: 10.1007/978-3-319-22348-3_9
14. Calzavacca, P., Licari, E., Tee, A., Egi, M., Downey, A., Quach, J., et al. (2010). The impact of rapid response system on delayed emergency team activation patient characteristics and outcomes—a follow-up study. Resuscitation 81, 31–35. doi: 10.1016/j.resuscitation.2009.09.026
15. Chen, L., and Klasky, H. (2022). Six Machine-Learning Methods for Predicting Hospital-Stay Duration for Patients With Sepsis: A Comparative Study. Mobile, AL: IEEE.
16. Chuang,M. T., Hu, Y., and Lo, C. L. (2018). Predicting the prolonged length of stay of general surgery patients: a supervised learning approach. Int. Trans. Operat. Res. 25, 75–90. doi: 10.1111/itor.12298
17. Chuang, M. T., Hu, Y. H., Tsai, C. F., Lo, C. L., and Lin, W. C. (2015). "The identification of prolonged length of stay for surgery patients," in 2015 IEEE International Conference on Systems, Man, and Cybernetics. (Hong Kong: IEEE), 3000–3003. doi: 10.1109/SMC.2015.522
18. Colella, Y., Scala, A., de Lauri, C., Bruno, F., Cesarelli, G., Ferrucci, G., et al. (2021). "Studying variables affecting the length of stay in patients with lower limb fractures by

means of Machine Learning," in 2021 5th International Conference on Medical and Health Informatics. (New York, NY: ACM), 39–43.

19. Combes, C., Kadri, F., and Chaabane, S. (2014). Predicting Hospital Length of Stay Using Regression Models: Application to Emergency Department. 10ème Conférence
20. Francophone de Modélisation, Optimisation et Simulation–MOSIM'14, Nov 2014, Nancy, France (hal-01081557) HAL (Hyper Articles en Ligne).
21. Dadeh, A., and Phunyanantakorn, P. (2020). Factors affecting length of stay in the emergency department in patients who presented with abdominal pain. Emerg. Med. Int. 2020, 1–7. doi: 10.1155/2020/5406516
22. Daghistani, T. A., Elshawi, R., Sakr, S., Ahmed, A. M., Al-Thwayee, A., Al- Mallah, M. H., et al. (2019). Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. Int. J. Cardiol. 288, 140–147. doi: 10.1016/j.ijcard.2019.01.046
23. Deo, R. C. (2015). Machine learning in medicine. Circulation 132, 1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593
24. Emori, T. G., and Gaynes, R. P. (1993). An overview of nosocomial infections, including the role of the microbiology laboratory. Clin. Microbiol. Rev. 6, 428–442. doi: 10.1128/CMR.6.4.428
25. Etu, E. E., Monplaisir, L., Arslanturk, S., Masoud, S., Aguwa, C., Markevych, I., et al. (2022). Prediction of length of stay in the emergency department for COVID-19 patients: a machine learning approach. IEEE Access 10, 42243–42251. doi: 10.1109/ACCESS.2022.3168045
26. Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin. Biochem. Rev. 29, S83–S87. Available online at: https://pubmed.ncbi.nlm.nih.gov/18852864/
27. Fridman, S. E., di Giampietro, P., Sensoli, A., Beleffi, M., Bucce, C., Salvatore, V., et al. (2022). Prediction of conventional oxygen therapy failure in COVID-19 patients with acute respiratory failure by assessing serum lactate concentration, PaO2/FiO2 ratio, and body temperature. Cureus 2022, 21987. doi: 10.7759/cureus.21987

28. Friedman, J., Latash, M. L., and Zatsiorsky, V. M. (2009). Prehension synergies: a study of digit force adjustments to the continuously varied load force exerted on a partially constrained hand-held object. Exp. Brain Res. 197, 1–13. doi: 10.1007/s00221-009-1818-1
29. Gabriel, R. A., Harjai, B., Simpson, S., Du, A. L., Tully, J. L., George, O., et al. (2023). An ensemble learning approach to improving prediction of case duration for spine surgery: algorithm development and validation. JMIR Perioper. Med. 6, e39650. doi: 10.2196/39650
30. García-González, P., Fácila Rubio, L., Montagud, V., Chacón-Hernández, N., Fabregat-Andrés, Ó., Morell, S., et al. (2014). Predictors of prolonged hospitalization in cardiology. Revista Española de Cardiología 67, 62–63. doi: 10.1016/j.recesp.2013.05.024
31. Gellman, D. D. (1974). Cost-benefit in health care: we need to know much more. Can. Med. Assoc. J. 111, 988–989.
32. Genuer, R., Poggi, J. M., and Tuleau-Malot, C. (2010). Variable selection using randomforests. Pattern Recognit. Lett. 31, 2225–2236. doi: 10.1016/j.patrec.2010.03.014
33. Gohil, R., Rishi, M., and Tan, B. H. L. (2014). Pre-operative serum albumin and neutrophil-lymphocyte ratio are associated with prolonged hospital stay following colorectal cancer surgery. Br. J. Med. Med. Res. 4, 481–487. doi: 10.9734/BJMMR/2014/5444
34. Hachesu, P. R., Ahmadi, M., Alizadeh, S., and Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthc. Inform. Res. 19, 121. doi: 10.4258/hir.2013.19. 2.121
35. Han, X., Zheng, X.,Wang, Y., Sun, X., Xiao, Y., Tang, Y., et al. (2019). Randomforest can accurately predict the development of end-stage renal disease in immunoglobulin a nephropathy patients. Ann. Transl. Med. 7, 234–234. doi: 10.21037/atm.2018.12.11
36. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. doi: 10.1007/978-0-387-84858-7

37. Jiang, X., Qu, X., and Davis, L. B. (2010). “Using data mining to analyze patient discharge data for an urban hospital,” in Proceedings of the 2010 International Conference on Data Mining (DMIN). (Las Vegas, NV), 139–144.
38. Jo, Y. Y., Han, J., Park, H. W., Jung, H., Lee, J. D., Jung, J., et al. (2021). Prediction of prolonged length of hospital stay after cancer surgery using machine learning on electronic health records: retrospective cross-sectional study. JMIR Med. Inform. 9, e23147. doi: 10.2196/23147
39. Katayama, Y., Kitamura, T., Tanaka, J., Nakao, S., Nitta, M., Fujimi, S., et al. (2021). Factors associated with prolonged hospitalization among patients transported by emergency medical services: a population-based study in Osaka, Japan. Medicine 100, e27862. doi: 10.1097/MD.0000000000027862
40. Keller, J. M., Gray, M. R., and Givens, J. A. (1985). A fuzzy Knearest neighbor algorithm. IEEE Trans. Syst. Man. Cybern. 15, 580–585. doi: 10.1109/TSMC.1985.6313426
41. Kirchebner, J., Günther, M. P., Sonnweber, M., King, A., and Lau, S. (2020). Factors and predictors of length of stay in offenders diagnosed with schizophrenia—A machine-learning-based approach. BMC Psychiatry 20, 201. doi: 10.1186/s12888-020-02612-1
42. Kong, G., Lin, K., and Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. BMC Med. Inform. Decis. Mak. 20, 251. doi: 10.1186/s12911-020-01271-2
43. Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artif. Intell. Rev. 26, 159–190. doi: 10.1007/s10462-007-9052-3
44. Krochmal, P., and Riley, T. A. (1994). Increased health care costs associated with ED overcrowding. Am. J. Emerg. Med. 12, 265–266. doi: 10.1016/0735-6757(94)90135-X
45. LaFaro, R. J., Pothula, S., Kubal, K. P., Inchiosa, M. E., Pothula, V. M., Yuan, S. C., et al. (2015). Neural network prediction of ICU length of stay following cardiac surgery based on pre-incision variables. PLoS ONE 10, e0145395. doi: 10.1371/journal.pone.0145395
46. Laky, B., Janda, M., Kondalsamy-Chennakesavan, S., Cleghorn, G., and Obermair,A. (2010). Pretreatment malnutrition and quality of life - association with prolonged length

of hospital stay among patients with gynecological cancer: a cohort study. BMC Cancer 10, 232. doi: 10.1186/1471-2407-10-232

47. Launay, C. P., Kabeshova, A., Lanoé, A., Chabot, J., Levinoff, E. J., Beauchet, O., et al. (2018). Age effect on the prediction of risk of prolonged length hospital stay in older patients visiting the emergency department: results from a large prospective geriatric cohort study. BMC Geriatr. 18, 127. doi: 10.1186/s12877-018-0820-5
48. Liew, D., Liew, D., and Kennedy, M. P. (2003). Emergency department length of stay independently predicts excess inpatient length of stay. Med. J. Aust. 179, 524–526. doi: 10.5694/j.1326-5377.2003.tb05676.x
49. Lu,M., Sajobi, T., Lucyk, K., Lorenzetti, D., and Quan, H. (2015). Systematic review of risk adjustment models of hospital length of stay (LOS). Med. Care. 53, 355–365. doi: 10.1097/MLR.0000000000000317
50. Luo, L., Lian, S., Feng, C., Huang, D., and Zhang, W. (2017). "Data mining-based detection of rapid growth in length of stay on COPD patients," in 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA) (Beijing: IEEE), 254–258. doi: 10.1109/ICBDA.2017.8078819
51. Luo, R., Tan, X., Wang, R., Qin, T., Chen, E., Liu, T. Y., et al. (2020). Accuracy prediction with non-neural model for neural architecture search. arXiv 2020. doi: 10.48550/arXiv.2007.04785
52. Marfil-Garza, B. A., Belaunzarán-Zamudio, P. F., Gulias-Herrero, A., Zuñiga, A. C., Caro-Vega, Y., Kershenobich-Stalnikowitz, D., et al. (2018). Risk factors associated with prolonged hospital length-of-stay: 18-year retrospective study of hospitalizations in a tertiary healthcare center in Mexico. PLoS ONE 13, e0207203. doi: 10.1371/journal.pone.0207203
53. Mekhaldi, R. N., Caulier, P., Chaabane, S., Chraibi, A., and Piechowiak, S. (2020). "Using machine learning models to predict the length of stay in a hospital setting," in Trends and Innovations in Information Systems and Technologies. WorldCIST 2020. Advances in Intelligent Systems and Computing, Vol. 1159, eds A. Rocha, H. Adeli, L. Reis, S.

Costanzo, I. Orovic, and F. Moreira (Cham: Springer). doi: 10.1007/978-3-030-45688-7_21

54. Morton, A., Marzban, E., Giannoulis, G., Patel, A., Aparasu, R., Kakadiaris, I. A. A., et al. (2014). "Comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients," in 2014 13th International Conference on Machine Learning and Applications (Detroit, MI: IEEE), 428–431. doi: 10.1109/ICMLA.2014.76
55. Paterson, R., MacLeod, D., Thetford, D., Beattie, A., Graham, C., Lam, S., et al. (2006). Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. Clin.Med. 6, 281–284. doi: 10.7861/clinmedicine.6-3-281
56. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,et al. (2011). Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.doi: 10.5555/1953048.2078195
57. Qi, Y. (2012). "Random forest for bioinformatics," in Ensemble Machine Learning, eds C. Zhang and Y. Ma (New York, NY: Springer), 307–323. doi: 10.1007/978-1-4419-9326-7_11
58. Rahman, M. A., Honan, B., Glanville, T., Hough, P., and Walker, K. (2020). Using data mining to predict emergency department length of stay greater than 4 hours: derivation and single-site validation of a decision tree algorithm. Emerg. Med. Australas. 32, 416–421. doi: 10.1111/1742-6723.13421
59. Sir, Ö., Hesselink, G., van den Bogaert, M., Akkermans, R. P., and Schoon, Y. (2019). Risk factors for prolonged length of stay of older patients in an academic emergency department: a retrospective cohort study. Emerg. Med. Int. 2019, 4937827. doi: 10.1155/2019/4937827
60. Song, X., Xia, C., Li, Q., Yao, C., Yao, Y., Chen, D., et al. (2020). Perioperative predictors of prolonged length of hospital stay following total knee arthroplasty: a retrospective study from a single center in China. BMC Musculoskelet. Disord. 21, 62. doi: 10.1186/s12891-020-3042-x

61. Steyerberg, E. W. (2019). Clinical Prediction Models. 2nd Edn. New York, NY: Springer-Verlag.
62. Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., et al. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21, 128–138. doi: 10.1097/EDE.0b013e3181c30fb2
63. Sud, M., Yu, B., Wijeysundera, H. C., Austin, P. C., Ko, D. T., Braga, J., et al. (2017). Associations between short or long length of stay and 30-day readmission and mortality in hospitalized patients with heart failure. JACC Heart Fail. 5, 578–588. doi: 10.1016/j.jchf.2017.03.012
64. Symum, H., and Zayas-Castro, J. L. (2020). Prediction of chronic disease-related inpatient prolonged length of stay using machine learning algorithms. Healthc. Inform. Res. 26, 20–33. doi: 10.4258/hir.2020.26.1.20
65. Tanuja, S., Acharya, D. U., and Shailesh, K. R. (2011). Comparison of different data mining techniques to predict hospital length of stay. J. Pharm. Biomed. Sci. 7.
66. Thompson, B., Elish, K. O., and Steele, R. (2018). “Machine learning-based prediction of prolonged length of stay in newborns,” in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (Orlando, FL: IEEE), 1454–1459. doi: 10.1109/ICMLA.2018.00236
67. Thongpeth, W., Lim, A., Wongpairin, A., Thongpeth, T., and Chaimontree, S. (2021). Comparison of linear, penalized linear andmachine learningmodels predicting hospital visit costs fromchronic disease in Thailand. Inform.Med.Unlocked 26, 100769. doi: 10.1016/j.imu.2021.100769
68. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. Royal Stat. Soc. 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x
69. Tsai, P. F. J., Chen, P. C., Chen, Y. Y., Song,H. Y., Lin,H.M., Lin, F.M., et al. (2016). Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. J. Healthc. Eng. 2016, 7035463. doi: 10.1155/2016/7035463
70. van der Veen, D., Remeijer, C., Fogteloo, A. J., Heringhaus, C., and de Groot, B. (2018). Independent determinants of prolonged emergency department length of stay in a

tertiary care centre: a prospective cohort study. Scand. J. Trauma Resusc. Emerg. Med. 26, 81. doi: 10.1186/s13049-018-0547-5

71. Vekaria, B., Overton, C., Wi´sniowski, A., Ahmad, S., Aparicio-Castro, A., Curran-Sebastian, J., et al. (2021). Hospital length of stay for COVID-19 patients: data-driven methods for forward planning. BMC Infect. Dis. 21, 700. doi: 10.1186/s12879-021-06371-6
72. Wolff, J., McCrone, P., Patel, A., Kaier, K., and Normann, C. (2015). Predictors of length of stay in psychiatry: analyses of electronic medical records. BMC Psychiatry 15, 238. doi: 10.1186/s12888-015-0623-6
73. Wu, J., Kong, G., Lin, Y., Chu, H., Yang, C., Shi, Y., et al. (2020). Development of a scoring tool for predicting prolonged length of hospital stay in peritoneal dialysis patients through data mining. Ann. Transl. Med. 8, 1437. doi: 10.21037/atm-20-1006
74. Wu, J., Lin, Y., Li, P., Hu, Y., Zhang, L., Kong, G., et al. (2021). Predicting prolonged length of ICU stay through machine learning. Diagnostics 11, 2242. doi: 10.3390/diagnostics11122242
75. Xiong, F., Cao, X., Shi, X., Long, Z., Liu, Y., and Lei, M. (2022). A machine learning–based model to predict early death among bone metastatic breast cancer patients: a large cohort of 16,189 patients. Front. Cell Dev. Biol. 10, 1059597. doi: 10.3389/fcell.2022.1059597
76. Xue, X., Liu, Z., Xue, T., Chen, W., and Chen, X. (2022). Machine learning for the prediction of acute kidney injury in patients after cardiac surgery. Front. Surg. 9, 946610. doi: 10.3389/fsurg.2022.946610
77. Yu, K. H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. Nat. Biomed. Eng. 2, 719–731. doi: 10.1038/s41551-018-0305-z
78. Zeleke, A. J., Moscato, S., Miglio, R., and Chiari, L. (2022). Length of stay analysis of COVID-19 hospitalizations using a count regression model and quantile regression: a study in Bologna, Italy. Int. J. Environ. Res. Public Health 19, 2224. doi: 10.3390/ijerph19042224

79. Zoller, B., Spanaus, K., Gerster, R., Fasshauer, M., Stehberger, P. A., Klinzing, S., et al. (2014). ICG-liver test versus new biomarkers as prognostic markers for prolonged length of stay in critically ill patients—a prospective study of accuracy for prediction of length of stay in the ICU. Ann. Intensive Care 4, 19. doi: 10.1186/s13613-014-0019-7

Appendix:

Table S1. Description of independent features and outcomes.

Features	Measure categories
Case ID	A unique ID for identification of inpatients
Demographic	
Age	Patient Age in classes: [0 - 17), (18 - 29), (30 - 49), (50- 69), 70+
Gender	The state of being female = 1 or male = 0
Arrival mode/admission source	Ambulance -118
	Own vehicle/walk-in
	Others
Risk categories – triaging in entrance	Red – Emergency
	Orange – indifferible urgency
	Light blue – differable urgency
	Green – minor urgency
	White – not urgency
Specialty	General medicine, geriatry , astanteria , obstetrics and gynecology , pediatrics , general surgery, infectious and tropical diseases, orthopedics and traumatology, urology, coronary unit, pediatric surgery, gastroenterology, cardiology, intensive care, pneumology, nephrology, oncology, vascular surgery
	Others (damages, details nephrology (enabled for transplantation), neonatology, pediatric oncology, semi-intensive therapy , maxillo facial surgery , hematology, thoracic surgery, ophthalmology , heart surgery , neonatal intensive care , pediatric heart surgery , and dermatology)

Most frequent problems # problems are not only diseases specific also any other problems.	Dyspnea, abdominal pain, fever/hyperpyrexia/hyperthermia, problems in pregnancy > 20th week, non-specific minor disorders ,chest pain of suspected cardiovascular cause, sincope / pre-sincope, generalized asthenia, politrauma – contusive, pain at the side, nausea and / or vomiting repeated, heart palm / irregular wrist altered level of consciousness, state of confusion, hematochezia / rectorrage / melena, lower limbs injury, cough / congestion, lower limbs pain, chest pain not suspected due to cardiovascular cause, pallor / anemia, request for urgent specialist advice, macro-hematuria, diarrhea, request for prescription or performance, swollen /edematous leg, weakness of extremities / symptoms associated with cerebrovascular disease, symptoms of infection of the urinary tract, diagnostics for biochemical images / examinations, head trauma
	Other problems (more than 135 cases, a few in numbers)
PLoS – outcome variable	LoS < 6 / LoS > = 6 days as binary outcome ; LoS [0 to 120 days) as continuous outcome

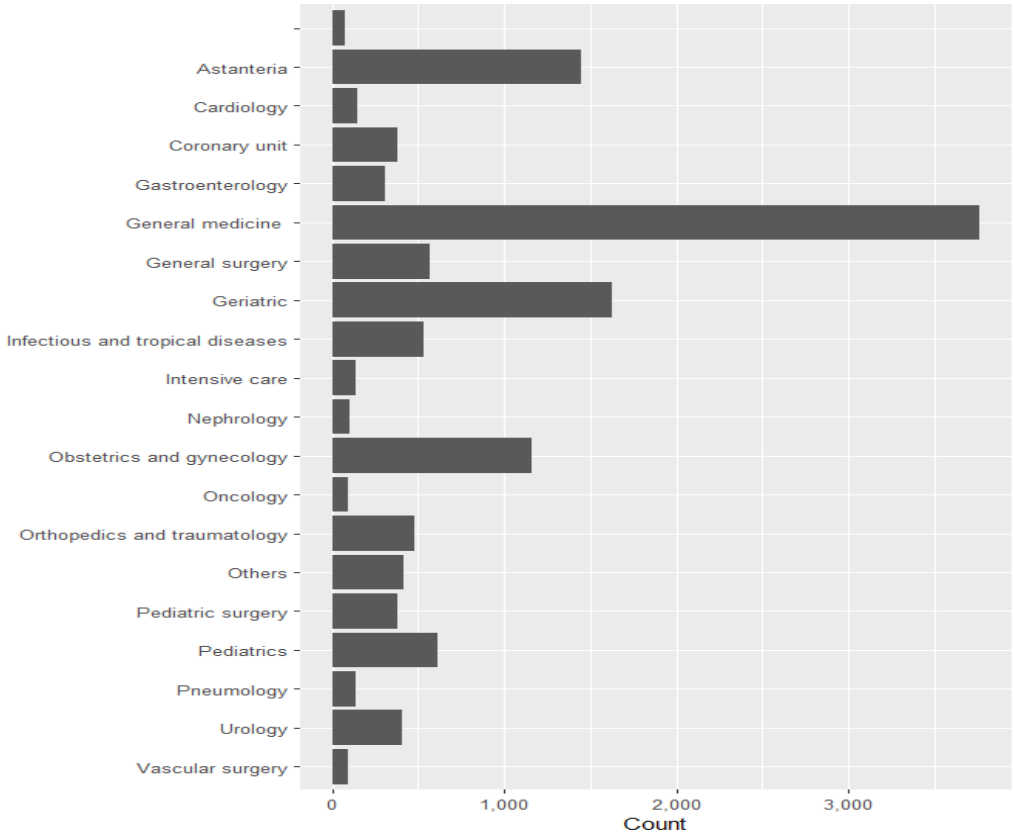


Figure S1. Frequency of type of length of stay (specialty) or Hospital wards during admission patients in 12858 patients the emergency department (ED) of the Sant'Orsola, Malpighi University Hospital of Bologna between January and October 2022.

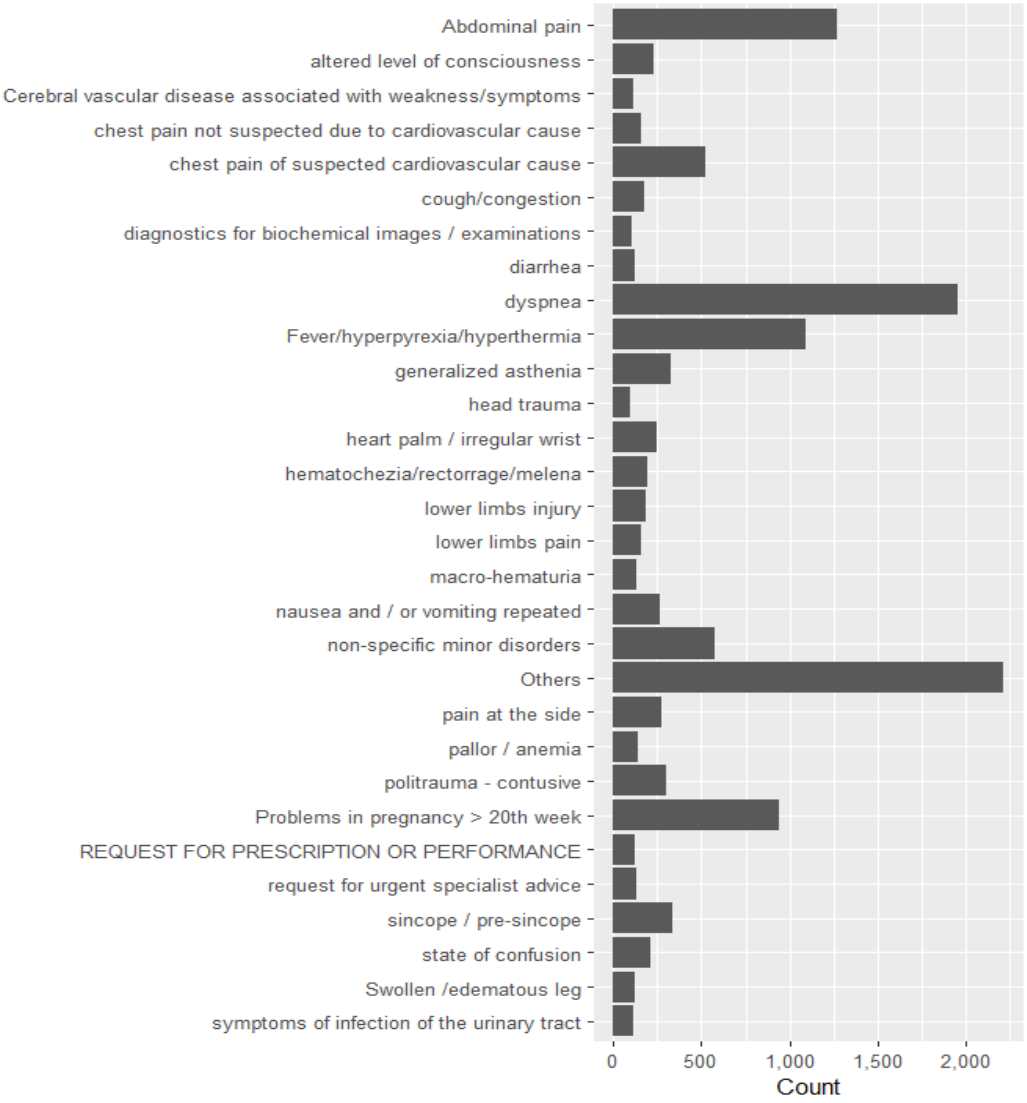


Figure S2. Frequency of most frequent problems during admission or Chief complaint for admission in 12858 patients in the emergency department (ED) of the Sant'Orsola, Malpighi University Hospital of Bologn.

CHAPTER THREE

Comparison of Nine Machine Learning Regression Models in Predicting Hospital Length of Stay for Patients Admitted to a General Medicine Department

This chapter is largely taken from our published work [∗], where we estimated and compared the predictive ability of Machine Learning (ML) regression models for the actual number of LoS days using demographic and clinical information recorded at admission at a General Medicine (GM) department.

∗ *Addisu Jember Zeleke, Pierpaolo Palumbo, Paolo Tubertini, Rossella Miglio, Lorenzo Chiari. Comparison of nine machine learning regression models in predicting hospital length of stay for patients admitted to a general medicine department. Informatics in Medicine Unlocked, Volume 47, 2024,101499*,
https://doi.org/10.1016/j.imu.2024.101499

(https://www.sciencedirect.com/science/article/pii/S2352914824000558)

Abstract

Background: The General Medicine (GM) department has the highest patient volume and heterogeneity among other hospital specialties. Examining hospitalization data closely is crucial because patients come with various conditions or traits. Length of stay (LoS) in hospitals is often used as an efficiency indicator. It is influenced by a patient's medical background, demographic factors, arrival mode at the hospital, triage evaluation, type of diseases/signs/symptoms, or administration problems. LoS is a variable that can vary widely, making it difficult to estimate it accurately, but doing so is highly beneficial. Moreover, efficiently grouping patient hospital spells or visits based on their LoS remains a significant challenge in healthcare applications.

Objectives: This study aimed to estimate and compare the predictive ability of Machine Learning (ML) regression models for the actual number of LoS days using demographic and clinical information recorded at admission.

Methods: We included patients from the GM department of the Sant'Orsola-Malpighi University Hospital in Bologna, Italy, who were admitted through the Emergency Department. The data were collected from January 1, 2022 to October 26, 2022. Nine ML regression models were used to predict LoS by analyzing historical data and patient information. The models' performance was assessed by means of root mean squared prediction error (RMSPE) and mean absolute prediction error (MAPE). Moreover, we used K-means clustering to group patients'

medical and organizational criticalities (such as diseases, signs, symptoms, and administrative issues) into four clusters. Feature Importance plots and SHAP (SHapley Additive exPlanations) were employed to identify the top important features and enhance interpretability.

Results: We analyzed 3757 eligible patients with an average LoS of 13 days, with a standard deviation of 11.8 days. We randomly divided them into a training cohort of 2630 patients (70%) and a test cohort of 1127 patients (30%). The predictive performance of the different adopted models was between 11.00 and 16.16 days for RMSPE and between 7.52 and 10.78 days for MAPE. The eXtreme Gradient Boosting Regression (XGBR) model had the lowest prediction error, both in terms of RMSPE (11.00 days) and MAE (7.52 days). Sex, arrival via own vehicle/walk-in, ambulance arrival (118), light blue risk category, age 70 or older, and orange risk category are some of the top important features. The study's limitations include the use of a single cohort for model development and testing, hindering external validation crucial for generalizability, and the omission of vital signs and laboratory results, which are key for accurately predicting hospital stay lengths.

Conclusion: The ML models evaluated in this study reported good predictive performance, with the XGBR model exhibiting the lowest prediction error. This model holds the potential to aid physicians in administering appropriate clinical interventions for patients in the GM department. This model can also help the healthcare service predict resources necessary to manage the hospitalization.

Keywords: Length of Stay; regression models; prediction; machine learning; clustering; decision-making

3.1. Introduction

3.1.1. Importance of addressing hospital Length of Stay in General Medicine patients

In General Medicine (GM), patients typically exhibit a wide range of conditions and varying characteristics, making it essential to examine hospitalization data closely. For this reason, predicting the Length of Stay (LoS) for each patient can be difficult due to the variability in their distribution. As hospitals face more pressure, due to global aging populations, with growing demand and limited resources, predicting the LoS for patients at admission has gained

popularity. A reliable prediction could aid in efficiently scheduling surgeries and providing appropriate care for patients with prolonged LoS. Ensuring appropriate hospital LoS for patients in the GM department is crucial for many reasons. Firstly, longer hospital stays are linked to a higher risk of adverse events, such as hospital-acquired infections, falls, and pressure ulcers, which can negatively impact patient outcomes and well-being [1]. Extended hospitalization can also contribute to decreased physical function and a decline in overall health status. In addition to patient benefit, monitoring LoS is important for patient satisfaction. Patients indeed prefer shorter hospital stays [2], as they are associated with faster recovery and a reduced risk of hospital-related complications. The use of resources is another critical factor, as more extended hospital stays require more nursing care, diagnostic tests, and medication administration. Extended stays can lead to increased costs for both the hospital and the patient and a higher burden on the healthcare system. However, in most cases, there is no knowledge about when GM patients will be discharged from the hospital. Accurate prediction of the LoS could hence enable hospitals to optimize healthcare resource management, resulting in better outcomes [3, 4].

In clustering, patients are partitioned into subgroups or clusters based on their similar or shared characteristics, for instance, grouping patients' problems based on how long they were hospitalized. Due to the difficulties caused by heterogeneous patient populations, it is helpful to group them into homogeneous, understandable groups. Research has shown that this approach offers several benefits in improving hospital and health facility planning and management [5]. Additionally, partitioning or grouping patients based on their LoS can provide valuable insights into patient care and outcomes [2]. Patient grouping is an advantageous method to simplify and enhance our understanding of diverse patient populations [6]. Patient diagnosis is influenced by various factors such as illness severity, medical complications, recovery rate, hospital stay, resource consumption, discharge location, and social factors. However, it can be challenging to allocate resources effectively due to the heterogeneity among patients [7]. Several patient grouping methods have been developed to address this issue. They are reported in the literature, identifying homogeneous patients groups within a given hospital population [7,8,9]. In other similar studies [11], concerning the analysis of prolonged hospitalization prediction in

patients admitted to general settings that include all departments and specialties (referred to as the all-patient model), the finding demonstrated that the XGBR model performed better. Furthermore, key features of significance or top feature importance were identified; however, in this study, we focused on a closer look at GM-specific patients, who have the highest patient volume and heterogeneities among other specialties in the hospital.

3.1.2. Related works

As part of clinical research, Machine Learning (ML) models generally aim to predict clinical outcomes based on multiple predictors. Patients can benefit from new and rapidly growing data sources by leveraging ML, Artificial Intelligence (AI), and other statistical methods. However, we had little cause to believe in advance that any particular model class would be best for this kind of problem. As such, in this paper, we systematically consider different ML regression models for predicting in-hospital LoS in GM patients.

Due to the wide range of ML regression studies conducted across various departments, specialties, diseases, and objectives, the selection of algorithms varies based on their respective performance disparities concerning data size. However, ensemble algorithms tend to outperform individual algorithms because they combine weak learning algorithms. Some similar works include the following. Siddiqa et al. [10] used several regression techniques, including multiple linear regression (MLR), decision tree regression (DTR), linear regression (LR), ridge regression (RR), eXtreme Gradient Boosting Regression (XGBR), and random forest regression (RFR) to predict LoS of inpatients. Their findings indicate that RFR was the most effective model, with an achieved mean square error (MSE) of 5 days. In another study involving 896 surgical patients, Chuang et al. [3] applied supervised ML models such as Local Gaussian Regression (LGR), support vector regression (SVR), and RFR to predict LoS. The findings indicated that the RFR model was the most accurate for predicting LoS. Kolchun et al. [12] conducted a separate study to develop a model for predicting passenger LoS following a motor vehicle accident. They evaluated several machine learning methods and found that a neural network (NN) algorithm had the lowest mean absolute error (MAE) of 2.23 days for predicting LoS. Similarly, Liu et al. [13] used data from seventeen hospitals in northern California and applied a mixture of

regression models to predict LoS in hospitals. They demonstrated that Laboratory Acute Psychological Score (LAPS) and Comorbidity Point Score (COPS) helped improve the models' efficiency.

Our paper focuses specifically on regression for predictive modeling of LoS modeled as a continuous outcome. However, the question remains as to which methods are the most effective in refining the accuracy of LoS prediction models or understanding the LoS variability among patients with different entry diagnoses.

3.1.3. Aims

This study aimed to estimate and compare the predictive ability of different machine-learning regression models for predicting the LoS of patients in the GM department. By leveraging observable characteristics of patients, the model aimed to provide accurate estimations of how long a patient is likely to stay in the hospital. Different ML approaches produce mixed results, which illustrates the need to carefully select a prediction model that effectively describes the observed data. Additionally, we had little reason to believe in advance that any particular class of model would be the most suitable for this type of study. As a result, in this article, we comprehensively consider nine ML regression models to predict the outcome of LoS. We also sought to better understand the relationship between hospital stay duration and the criticalities prior and during hospitalization. Moreover, we aimed to use unsupervised learning techniques to cluster patients into well-defined subgroups based on the reported LoS on their medical criticalities or problems. These techniques can help uncover patterns and structures in the data, leading to valuable insights for patient care and treatment.

3.2. Materials and methods

3.2.1. Study Design and Population

We screened for eligibility all the admissions to the hospital through the emergency department (ED) of the public University Hospital of Bologna Sant'Orsola-Malpighi (AOSP), Bologna, Italy between January 1, 2022, and October 26, 2022. For our analysis, we selected hospitalized GM department patients. The primary outcome of this study was hospital LoS. LoS

was calculated as the number of days between admission and discharge. Approaching the prediction of LoS as a regression task, rather than a classification task, can provide more accurate results and valuable insights for the targeted purpose. This allows for a continuous measure of LoS duration and avoids the limitations of forced categorization. This actual continuous measure of LoS can better account for the underlying complexity and individual variability in hospital stays. It will also reduce the potential bias in the modeling process. Patients whose discharge information was missing or unknown in any case were excluded from the analysis. Figure 1 illustrates the detailed ways of selecting patients and the excluded patients.

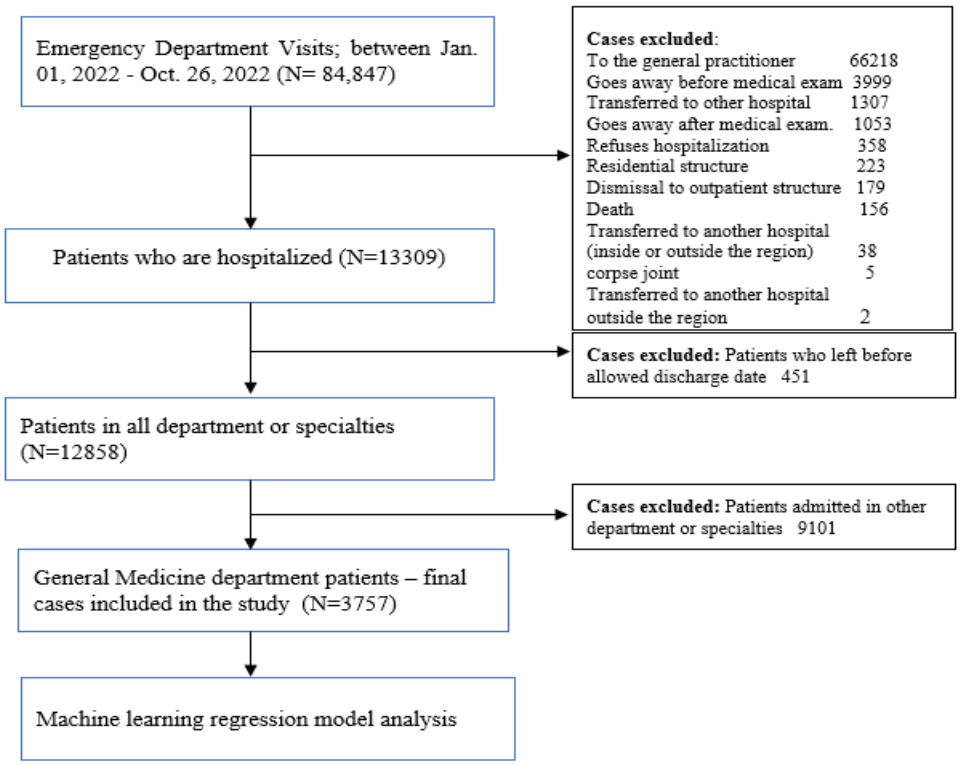


Fig 1: Flow chart of patient selection

3.2.2. Independent variables

Based on the available information reported upon admission, the independent variables that can be used to predict a patient's LoS in the GM department include:

- Demographic characteristics (gender and age): which can be used to identify patterns and trends in the data;
- Mode of arrival/source of admission (includes Ambulance -118, own vehicle/walk-in or any other means): the way a patient arrives at the hospital or the source of their admission can provide valuable insights into their condition and the potential LoS;
- Risk categories: The risk categories assigned to patients based on triage at the entrance (red, orange, light blue, green, white) can help identify patients who require more intensive care and may have a longer LoS; and
- Criticalities or problems: This includes both medical and administrative factors such as diseases, signs, symptoms, and administrative issues that can affect the patient's condition and LoS.

3.2.3. Clustering

In healthcare applications, efficiently grouping spells according to LoS remains challenging. The reason for this is primarily due to natural variability in LoS distribution. LoS-driven patient grouping could significantly improve bed allocation planning and patient admission and discharge processes. It is still a research challenge to efficiently group patients based on their LoS. Numerous clustering techniques are available for grouping patients based on similarities, including density-based clustering, k-means clustering, hierarchical clustering, and model-based clustering (such as Gaussian Mixture model).

We searched clusters on patients' criticalities, using LoS mean and interquartile range as descriptive variables for each criticality. The k-means clustering algorithm randomly selected k initial criticalities as cluster centroids. Each criticality was then assigned to the nearest centroid, which was updated based on the mean of all the criticalities assigned to it. This iterative process was repeated until either the centroids no longer moved or a predetermined number of iterations were completed. By doing so, the algorithm seeks to find the best groupings of similar criticalities in terms of their LoS mean and IQR values.

3.2.4. Regression Models

To predict the continuous outcome, we applied nine ML regression models. In-depth descriptions of these models are available in the literature [30]. Here, we provide a concise overview of each model.

Linear regression (LR) [15] predicts the outcome values using the feature values combined into a linear equation in the feature parameters. Feature parameters are chosen to maximize the likelihood of the observed outcome values or, equivalently, minimize the residuals' sum of squares.

Bayesian Ridge Regression (BRD) is a linear Bayesian regression with a Gaussian prior distribution on the feature parameters. The standard deviation λ on the prior distribution works as a regularization parameter on the Euclidean norm of the parameter vector.

Decision tree regression (DTR) [16] builds a tree-like arrangement of variables by dividing the data into smaller groups, resulting in a decision tree linked to these subsets of data.

Random forest regression (RFR) [17] is an ensemble learning technique that constructs multiple independent decision trees and aggregates their predictions. It is known to be robust to outliers and non-linear relationships in the data.

Light Gradient Boosting Regression (GBR) [18] is an ensemble learning technique that builds a model by iteratively combining weak learners to improve the accuracy of predictions. It is designed to provide fast and efficient training in decision tree models.

Linear support vector regression (SVR) [19] transforms data into a high-dimensional space using a kernel function and then finds a linear function that describes the relationship between input features and the target variable with a hyperplane.

Extreme Gradient Boosting Regression (XGBR) [20] is a DTR-based ensemble ML model (it combines the predictions from multiple machine learning algorithms to make more accurate predictions than any individual model) used to increase the speed and performance accuracy of the model.

K-Nearest Neighbors (KNN) [21] is an index-based algorithm that computes distances between instances, assigns indexes to these points, and stores the sorted distances and their indexes. It predicts the target value of a new instance as an average of the target values of k nearest training instances.

Negative binomial regression (NBR) [27] is a generalized linear model used for count outcomes, similar to Poisson regression but incorporating an extra parameter to model over-dispersion i.e. variance of the count outcomes greater that their mean.

An overview of the regression prediction model framework is shown in Figure 2.

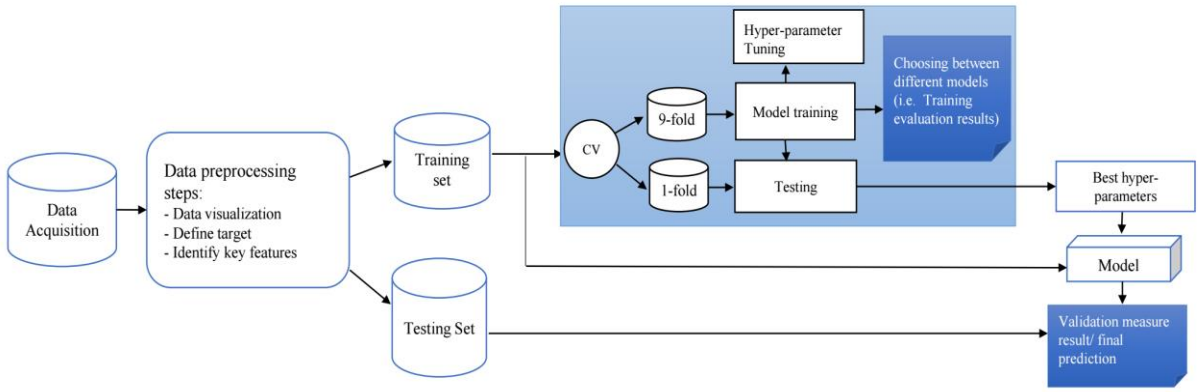


Fig 2: The predictive framework based on regression models

3.2.4.1. Parameters and hyperparameters tuning - Grid Search

In order to attain optimal performance, it was essential to determine the optimal hyper-parameters for each algorithm. A grid search was performed to identify the best set of hyper-parameters for each model. It tests all possible combinations in a parameter grid where one defines the possible values or ranges for each hyper-parameter. The grid search was performed with a 10-fold cross-validation nested within the training set. The optimal hyper-parameters settings discovered through the grid search for each algorithm are presented in Table 1.

LR	DTR	RFR	Light GBR	Linear SVR	XGBR	KNN
-	Max depth: 10 Max features: Auto Criteria: Entropy Min sample split: 4	n _estimators: 100 Max depth: 3 Max features: Auto Criteria: Entropy Min sample split: 4	Learning rate: 0.1 Max depth: 5 n_estimators: 100	Kernel: linear Gamma: 0.1 C: 1	Criterion: MSE Learning rate: 0.1 Max depth: 5 n_ estimators: 100	n_ neighbors: 10

	Min sample leaf: 2	Min sample leaf: 2				

Table 1: Optimal Hyperparameter Combinations for ML Regression Models

3.2.4.2. Model Evaluation

To assess the generalization and performance of the various models, we employed root mean squared prediction error (RMSPE), and mean absolute prediction error (MAPE), as indicated by equations (1) and (2) [26].

$$RMSPE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

$$MAPE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n} \quad (2)$$

In these equations, y and ^{y} represent the observed and predicted LoS, respectively, and n the total number of patients in the test dataset. A single performance measure is usually insufficient to generalize a model's accuracy. This is because a single measure may not fully capture the complexity and diversity of the problem at hand, so considering additional measures is necessary to gain a deeper understanding of the model's performance.

RMSPE, is the square root of the average of the squared residuals, where a residual is the difference between a predicted value and an observed value. It indicates the standard deviation of residuals. On the other hand, MAPE is computed by taking the average of all the absolute residuals. (MAPE sums up these absolute errors and divides them by the number of observations) . RMSPE and MAPE are similar, although RMSPE places a higher penalty on large prediction errors when compared to MAPE. The RMSPE metric is commonly employed in regression models as it can penalize larger errors while remaining easily interpretable.

3.2.4.3. Feature importance score and model interpretation

We calculated the importance of the features in the best-performing model using the gain score [30]. The gain score measures a feature's contribution to a model by evaluating how much it reduces the entropy or impurity of a split on that feature. This measure is commonly used for tree-based models. The higher the score, the more important the feature is for making

predictions. When a split decision is made with each feature, the gain score represents the average gain of each feature. We also used SHAP (Shapley Additive exPlanations) for model interpretation. SHAP is a Shapley-based novel approach that aids in explaining each feature's effect on model outcome [32]. In this explanation, the interplay among the local features is evaluated and subsequently aggregated to offer insights into individual predictors as well as the overarching model. Additionally, SHAP identifies predictors that degrade the model's performance and identifies high-risk and low-risk predictors. In addition, we analyzed the associations between the important features and the impact of the LoS outcome using SHAP values to improve the interpretation of the learning results of the best-achieving model.

3.2.4.4. Tools

In this study, we employed Python version 3.8 and executed it on Jupyter Notebook. We leveraged the Scikit-learn (sklearn) library to carry out the analyses.

3.3. Results

Patient selection;

Figure 1 displays a flowchart outlining the patient selection process for inclusion in the analyses according to the study eligibility criteria. From January 1 to October 26, 2022, there were 84,847 patient recorded across all departments in the ED. Upon applying the exclusion criteria, 12,858 patients were retained across all department and specialty settings. Ultimately, for the purpose of our analysis, we included 3,757 patients admitted to the GM department.

Descriptive statistics;

Out of the 3,757 patients analyzed, 50.4% were males (49.6% females) and 67.6% were aged 70 years or more. The average LoS was 13 days. The entire cohort of patients was randomly split into two groups: a training cohort of 2630 patients (70% of the total cohort) and a test cohort of 1127 patients (30% of the total cohort). Table 2 presents a comprehensive summary of the descriptive statistics.

Variables	Total (n=3757)	%	Length of Stay (in days)					
			Mean	Med	Std	Min	IQR	Max
Age, years:								
0 to 17	3	0.08	7.66	6	5.69	3	4.5 – 10	14
18 to 29	53	1.41	11.75	8	13.24	0	5 – 12	80
30 to 49	284	7.56	11.84	9	9.68	0	6 – 15	85
50 to 69	876	23.32	13.49	10	12.43	0	6 – 17	104
70 and older	2541	67.63	13.36	10	11.80	0	6 – 16	143
Gender:								
Male	1893	50.39	12.75	10	11.02	0	6 – 16	143
Female	1864	49.61	13.76	10	12.59	0	6 – 17	119
Mode of arrival:								
Ambulance	2225	59.22	12.83	10	11.32	0	6 – 16	143
Own vehicle/walk-in	1049	27.92	13.79	10	12.74	0	6 – 17	114
Others	483	12.86	14	10	12.74	0	7 – 17	119
Triage Category:								
Red	178	4.74	11.97	9	11.56	0	5 – 16	99
Orange	969	25.79	12.89	10	10.85	0	6 – 16	88
Light blue	1634	43.49	13.79	10	12.66	0	6 – 17	143
Green	936	24.91	13.01	10	11.41	0	6 – 16	114
White	40	1.06	10.80	9	9.21	0	4 - 12.3	51
Criticalities /Problems:								
Diseases/Signs/Symptoms								
Dyspnea	703	18.71	12.66	9.5	6.20	2	6 – 16	99
Fever/hyperpyrexia/hyperthermia	406	10.81	13.83	10	11.68	0	7 – 19	85
Abdominal pain	390	10.38	12.97	10	11.7	0	6 – 16	90
non-specific minor disorders	201	5.35	14.21	11	12.86	0	6 – 16	80
Generalized asthenia	155	4.13	14.90	11	16.24	0	7 – 16	143
Nausea or vomiting repeated	95	2.53	13.37	10	13.17	0	5 – 17.5	88
Syncope / presyncope	115	3.06	11.29	9	9.67	2	6.5 – 14	83
State of confusion	103	2.74	11.41	10	9.2	0	6 – 14	66
Altered level of consciousness	89	2.37	13.29	11	11.21	0	7 – 16	78
Pallor / anemia	70	1.86	10.87	9	7.1	0	6 – 13	33
Hematochezia/rectal bleeding/melena	82	2.18	11.13	8.5	9.24	2	5.3 – 13	55
CVD associated with any symptoms	79	2.10	11.34	9	7.03	1	6 – 15	30
Chest pain of suspected CV cause	77	2.05	10.44	8	7.26	2	5 – 13	43
Swollen /edematous leg	65	1.73	16.40	12	13.5	4	7 – 20	64
Diarrhea	58	1.54	11.08	9.5	6.20	2	7 – 14	29
Heart palm / irregular wrist	48	1.28	10.71	7	10.5	1	5 – 12	58
Polytraumas - contusive	48	1.28	14.25	12	10.61	1	8 – 17.3	47
Cough/congestion	47	1.25	14.31	10	11.07	0	6.5 – 21	46

Diagnostics biochemical examinations	45	1.20	12.71	11	8.92	0	6 – 17	51
Urinary tract infection symptoms	42	1.12	14.11	12.5	8.29	2	8 – 18	35
Pain at the side	39	1.04	9.35	9	5.29	1	5.5 – 12	25
Chest pain not suspected by CV cause	30	0.80	11.26	9.5	8.34	3	6.3 - 12	42
Macro-hematuria	28	0.75	14.53	10.5	12.00	2	6.8 -17	49
Lower limbs pain	24	0.64	15.87	14.5	7.50	4	10 – 23	29
Head trauma	19	0.51	11.47	8	15.04	2	6 – 10.5	71
Lower limbs injury	7	0.19	19.57	16	8.82	14	15 – 19	39
Others	594	15.81	14.43	11	13.18	0	6 – 19	119
Administration problems								
Request for urgent specialist advice	48	1.28	14.50	10.5	11.74	2	7 – 17.3	56
Request for prescription or performance	50	1.33	17.48	11	22.33	2	6.3 – 21	117

Table 2: Descriptive statistics measures of predictors and LoS.

The LoS ranged from 0 to 143 days, zero days meaning that admission and discharge occurred during the same day. The LoS values show a right-skewed distribution, with most values ranging from about 1 to 18 days (as shown in figure 3 left panel). Lower limb injury, prescription/performance request, swollen/edematous leg, lower limb pain, generalized asthenia, macro-hematuria, and request for urgent specialist advice were the top problems with the longest average LoS (see figure 3 right panel). The count plot of each predictor is shown in Appendix S1. Additionally, a granular analysis of the LoS was performed by type of disease/signs/symptoms or administration issues. Figure 4 presents the boxplot of hospital LoS for each criticality or problem. Unexpected records, i.e., outliers, were removed from the graph for effective comparative analysis.

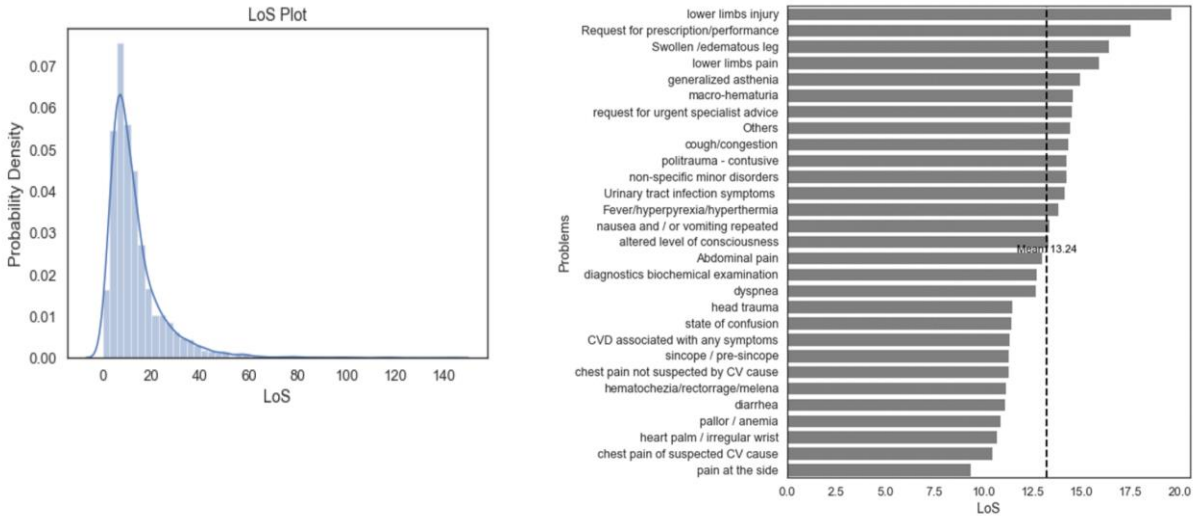


Figure 3: The LoS distribution plot (left panel) and the average LoS for each patient issue (right panel). CVD: cerebral vascular disease; CV: cardiovascular.

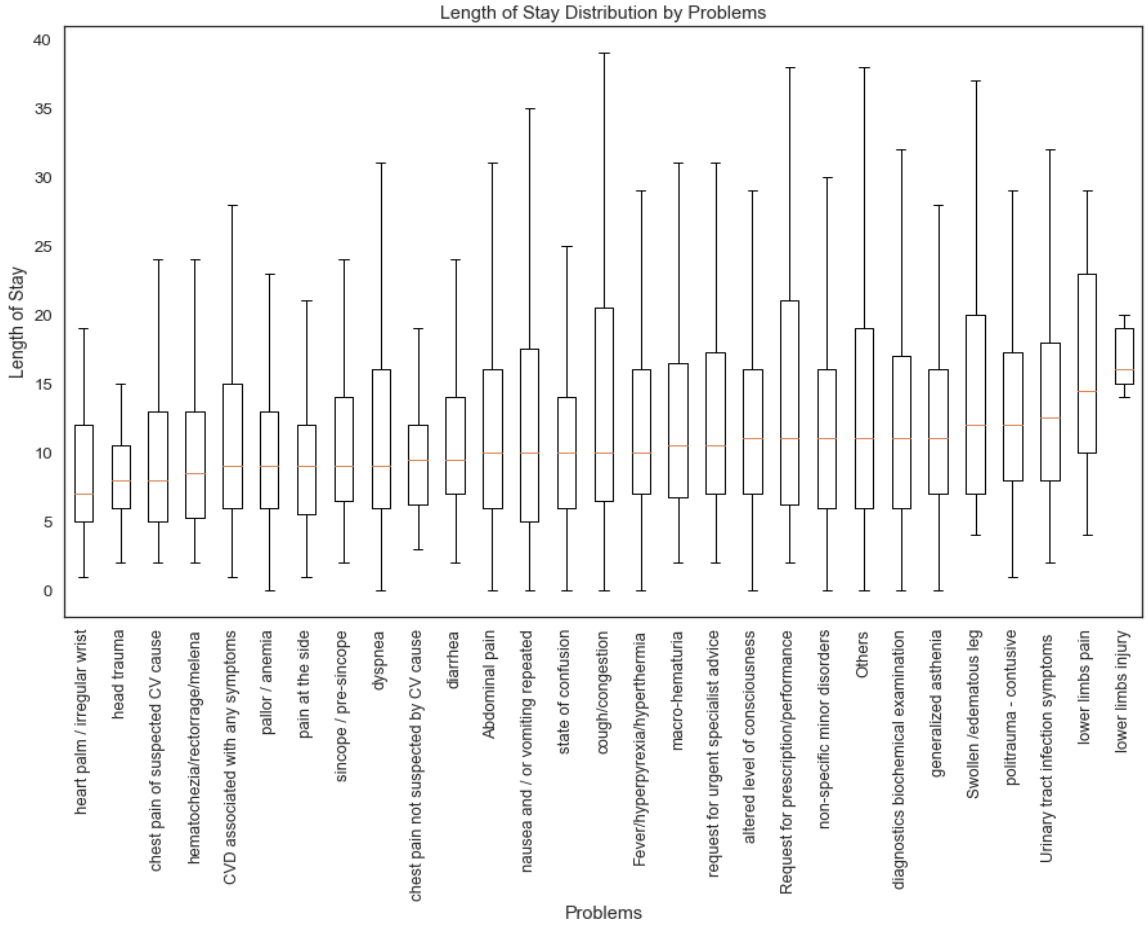


Figure 4: A granular analysis of the LoS by criticalities or problems.

Clustering;

Figure 5 depicts the result of clustering groups based on the mean and IQR of the LoS for each cluster. The plot shows the mean LoS on the horizontal axis and the IQR on the vertical axis, with k-means clustering applied to a dataset containing these values. Choosing the right number of clusters is a crucial step in k-means clustering since it impacts the quality and interpretability of the results. The Elbow Method, Silhouette Score, Silhouette Analysis, Domain Knowledge, and many other techniques can all be used to determine the number of clusters. However, we decided to select four clusters based on the data set and using domain knowledge. Patient criticalities assigned to different clusters are represented by distinct colors.

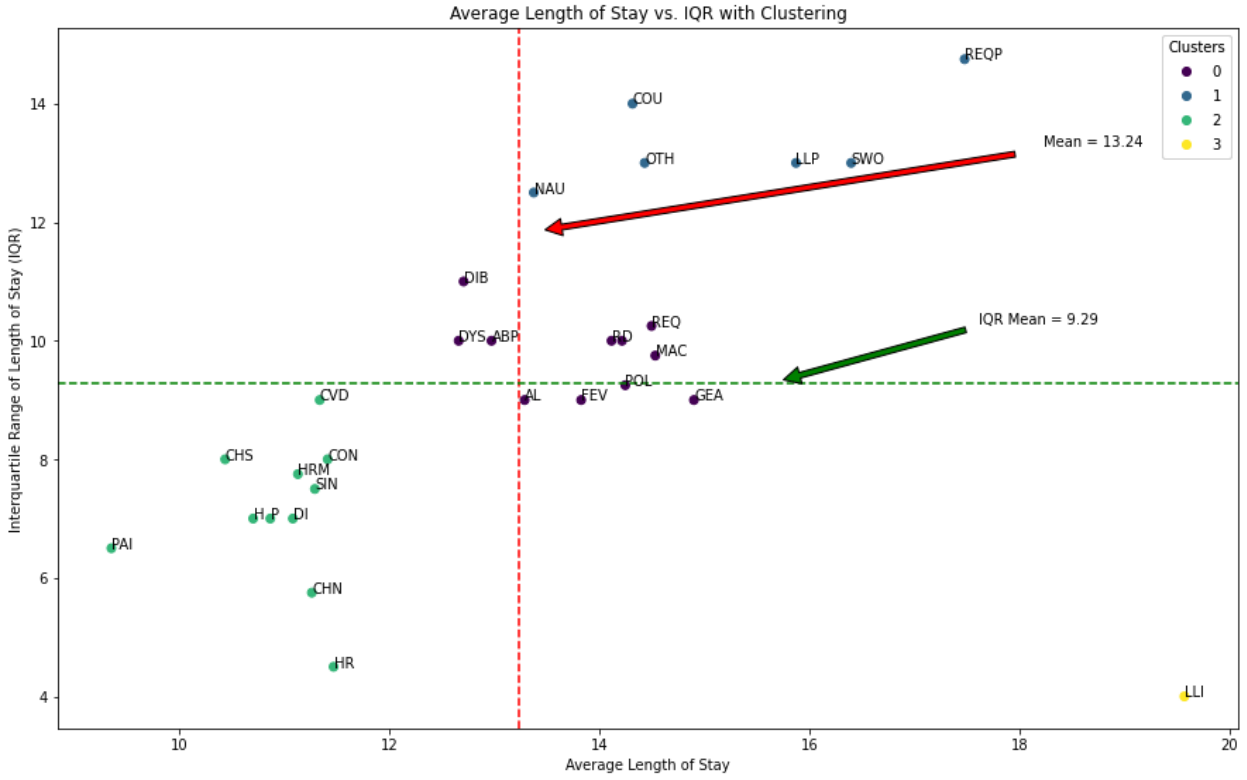


Figure 5: Clustering on patients' criticalities according to the average and IQR of LoS. CVD; CVD associated with any symptoms, R; urinary tract infection symptoms, SWO; swollen /edematous leg, CON; state of confusion, SIN; sincope / pre-sincope, REQ; request for urgent specialist advice, REQP; Request for prescription/performance, POL; politrauma – contusive, P; pallor / anemia, PAI; pain at the side, OTH; Others, D; non-specific minor disorders, NAU; nausea and / or vomiting repeated, MAC; macro-hematuria, LLP; lower limbs pain, LLI; lower limbs injury, HRM; hematochezia/rectorrage/melena, H; heart palm / irregular wrist, HR; head trauma, GEA; generalized asthenia, FEV; fever/hyperpyrexia/hyperthermia, DYS; dyspnea, DI; diarrhea, DIB;

diagnostics biochemical examination, COU; cough/congestion, CHS; chest pain of suspected CV cause, CHN; chest pain not suspected by CV cause, AL; altered level of consciousness, ABP; abdominal pain.

Regression models;

The evaluation metrics for the ML regression models are presented in Table 4. Among the various methods examined, the XGBR models exhibited superior performance on the test set, displaying the lowest prediction error with an RMSPE of 11.00 and MAPE of 7.52. These results highlight the effectiveness of XGBR in accurately predicting the LoS. While XGBR models have proven effective in solving a range of machine learning problems, it is crucial to perform careful hyperparameter tuning to attain optimal performance. On the other hand, the DTR model demonstrated the smallest prediction error in the training set but the highest RMSPE and MAPE values in the test set, suggesting poor generalization performance and overfitting to the training data, making it unsuitable for this specific problem.

Models	Training Set		Test Set	
	RMSPE	MAPE	RMSPE	MAPE
LR	11.94	7.94	11.02	7.56
Bayesian Ridge	12.03	7.98	11.04	7.53
DTR	5.74	2.06	16.16	10.78
RFR	6.94	4.23	12.38	8.48
Light GBR	10.72	7.39	11.77	7.59
Linear SVR	11.01	7.40	11.10	7.60
XGBR	11.64	7.68	**11.00**	**7.52**
KNN	10.91	7.35	12.06	8.43
NBR	11.94	7.93	11.01	7.54

Table 4: Comparisons of the ML regression models (i.e. evaluation metrics).

In figure 6, models are visualized in a plot for easy comparison. Notably, the test data show that the XGBR model performs better, suggesting it may be more reliable and generalizable to new patients or data.

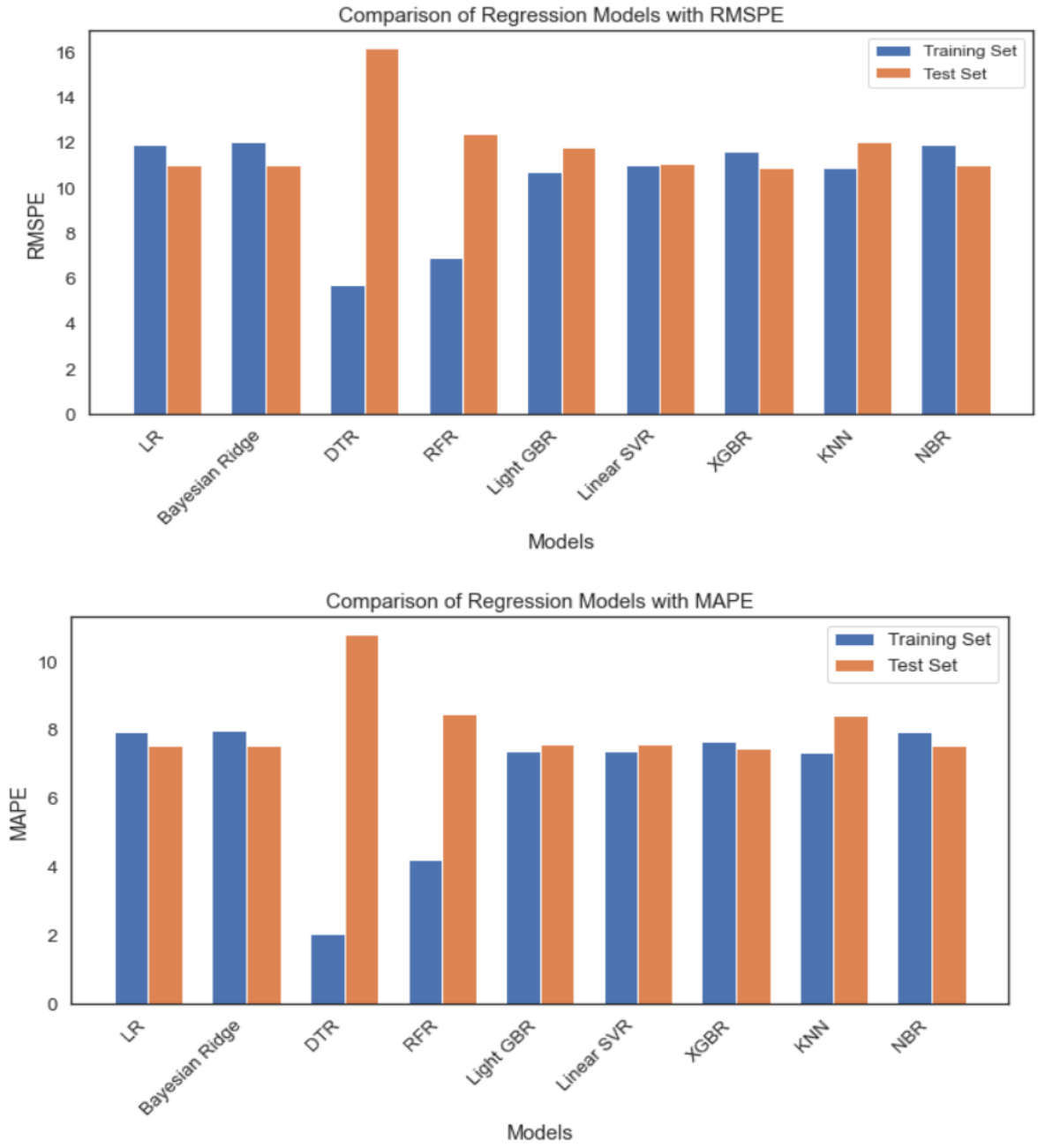


Figure 6: Comparative analysis of regression models, with RMSPE displayed on the top panel and MAPE on the bottom panel.

A calibration curve cannot be used since ML regression models do not provide direct probability estimates. However, we can still assess the calibration of predicted values in a regression model using a different approach. One common technique is to group the predicted values into bins and calculate the mean predicted value and the corresponding mean actual

value for each bin. We can then plot these values to visualize the calibration. Based on figure 7, it is evident that the optimal model (i.e., XGBR) displayed a tendency to underestimate hospital stays for cases with short LoS predictions while overestimating hospital stays for cases with long LoS predictions. For the other models, the calibration plots in the test set are presented in Appendix figure S3.

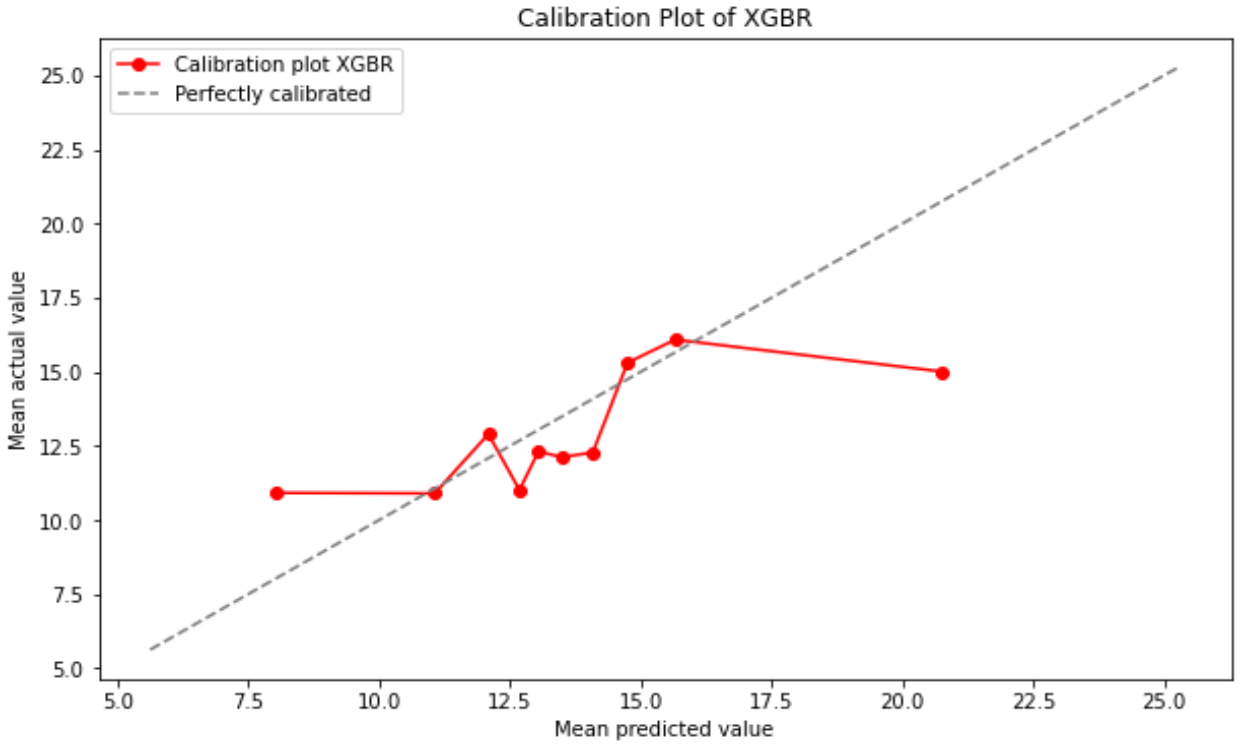


Figure 7: Calibration plots for best ML regression model

The XGBR model, selected as the best model, has feature importance scores shown in figure 8. These scores quantify the extent to which a feature contributes to the overall reduction of the objective function, such as mean squared error, in a tree-based model. The more points a feature receives, the more significant it is when making predictions. The gain score, which shows the average gain of each feature when utilized in a split decision, is one of the feature importance ratings in the model. Gain results from the feature in the tree's ability to reduce the goal function. As shown in figure S2, the heatmap plot of the similarity matrix shows the absence of collinearity between the top important features. A darker shade indicates a greater

degree of similarity between the features, whereas a lighter shade indicates a smaller degree of similarity.

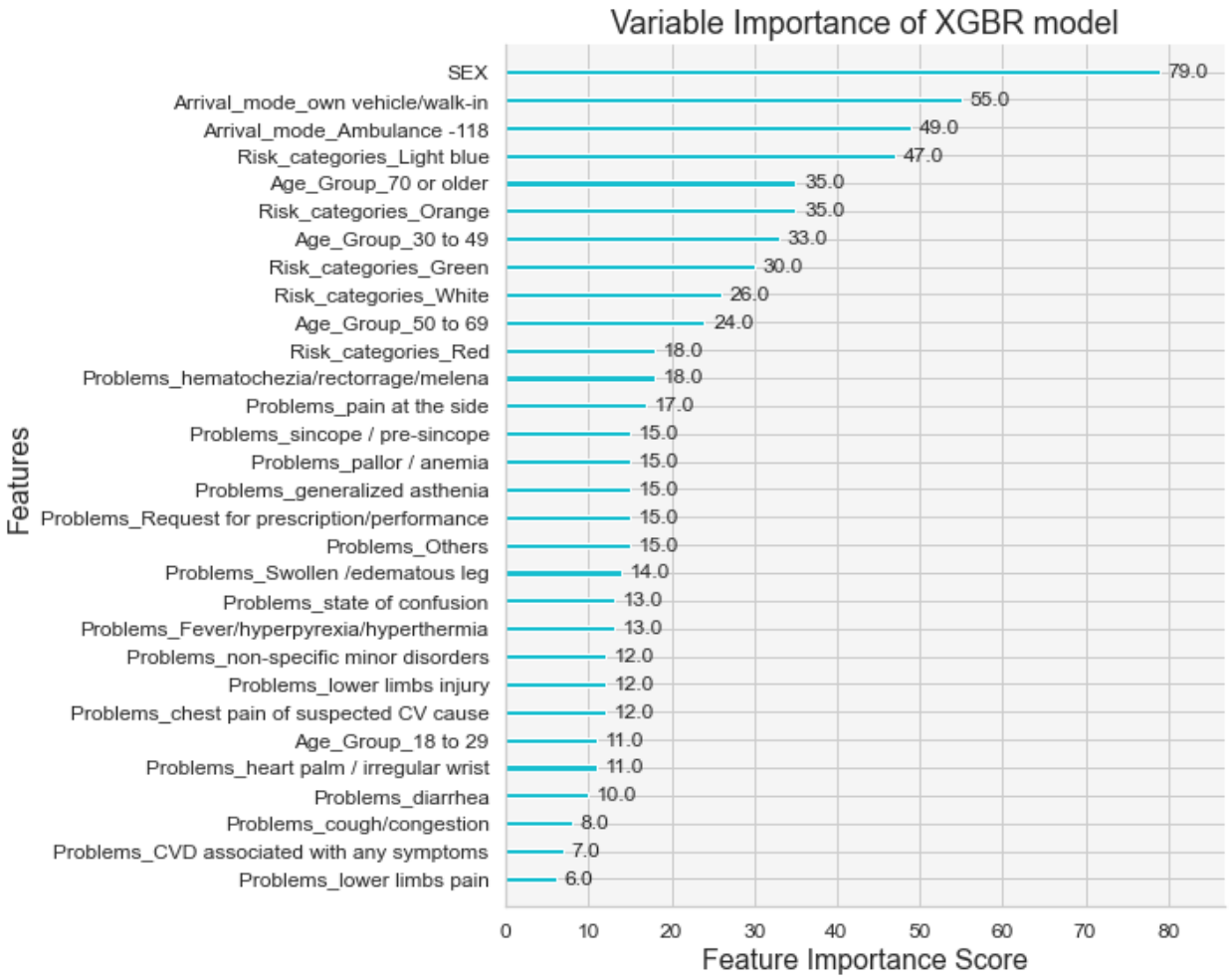


Figure 8: Feature importance score (CVD:cerebral vascular disease; CV: cardiovascular).

SHapley Additive exPlanations (SHAP) were applied to the final XGBR model to explain the effect of each feature variable on the target variable LoS. TreeExplainer was used to generate the SHAP values. A beeswarm plot was created to depict the significance of the global characteristics, as shown in figure 9. This plot shows the overall effects of each feature on the model's output. An x-axis shows the impact of a feature on prediction, while a color shows its value. A red feature has a high value and is positively correlated with the prediction, whereas a blue feature has a low value and is negatively correlated with the prediction. SHAP value plots show how each feature contributed to the model's output and can help identify the most

relevant features and their contributions to specific predictions. Both the importance score analysis and SHAP list sex, arrival to ED by ambulance, and age equal to or greater than 70 among the top five most influential variables.

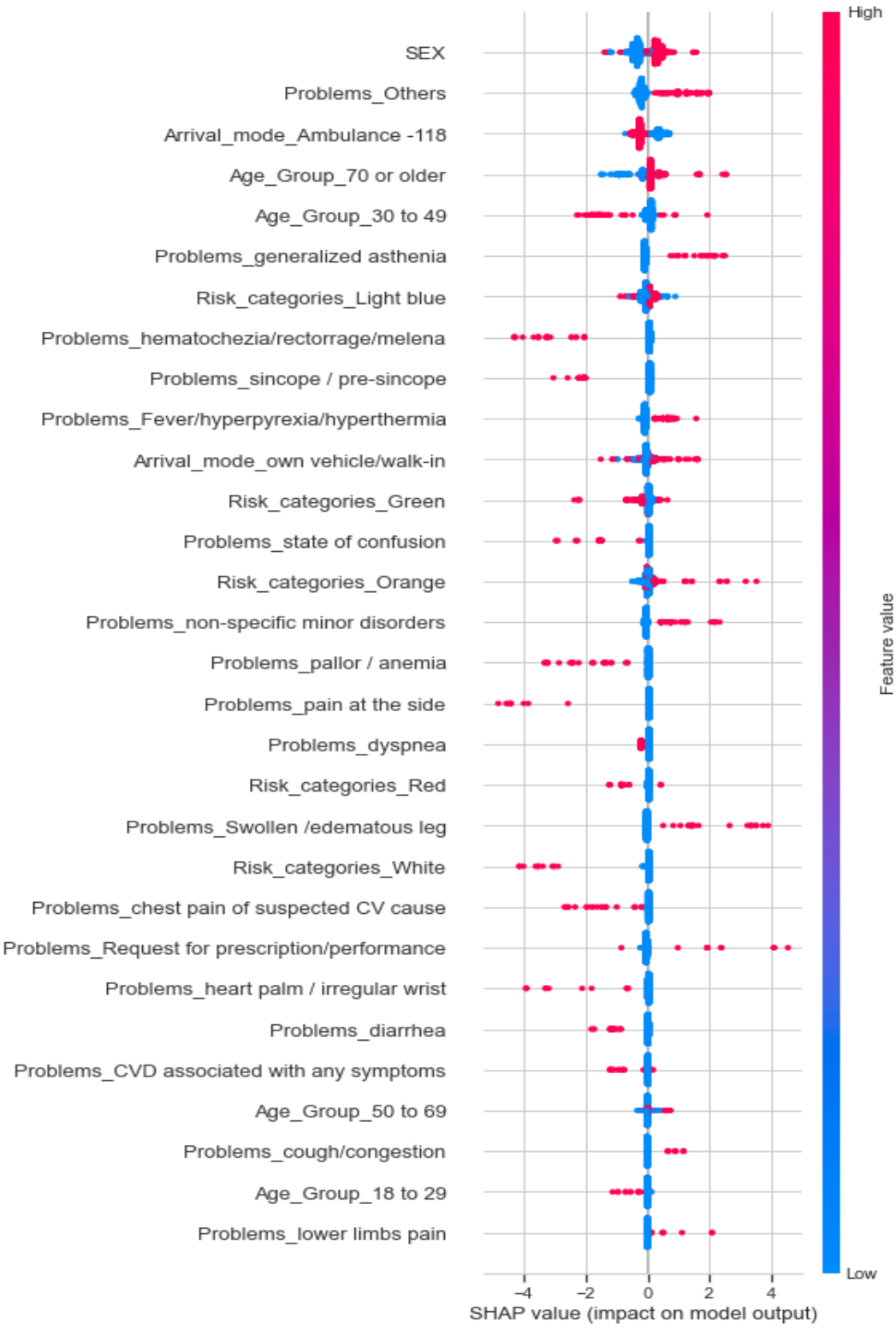


Figure 9: Beeswarm plots based on the SHAP values for model interpretations (CVD: cerebral vascular disease, CV: cardiovascular).

3.4. Discussion

This study aimed to develop a machine-learning regression model capable of predicting the length of stay (LoS) for patients in the GM department. The results of the present study showed that the XGBR model performed better than the other eight ML regression models at predicting the actual number of LoS for patients with GM-specific problems. The necessity of carefully picking a prediction model that effectively describes the observed data is demonstrated by different ML approaches giving mixed findings. Moreover, we also had little reason to believe in advance that any particular class of model would be the best choice for this type of study. As a result, in this article, we have comprehensively considered nine ML regression models to predict the LoS outcome. Furthermore, the primary objective of this study was to estimate and compare the predictive ability of different machine-learning regression models for predicting the LoS of patients in the GM department. Accurate LoS prediction models could facilitate the proactive allocation of vital healthcare resources and have the potential to reduce unwarranted hospital stays, with a particular focus on AOSP.

In a previous study [11], we analyzed the all-patient model, including patients admitted in all departments; however, in this study, we focused on GM-specific patients, who have the highest volume and heterogeneity among other specialties in the hospital. Since patients in this department have various diseases or characteristics, careful analysis of hospitalization data is essential. Although it can be challenging to estimate LoS precisely, doing so is quite advantageous. Nevertheless, healthcare applications still face a considerable barrier to properly organizing patient spells depending on their LoS. A LoS is a complex variable influenced by several clinical and social factors [29]. Modeling is one part of healthcare responses. Nowadays, healthcare systems have started focusing on efficient resource management and the prediction of feature outcomes. This is to ensure optimal levels of care, lower associated costs, and enhance patient care [22].

Before assessing the models' performance, we searched for the best hyperparameters for each model (i.e., a grid tuning hyperparameter), considering RMSPE and MAPE as loss function measures. The XGBR model outperformed the other eight models, with RMSPE of 11 days and MAPE of 7.52 days. This suggests that XGBR may be more reliable and generalizable to new data

then other regression models. As we searched for similar studies, ensemble regression models like XGBR, light GBM, and RFR usually produced superior results than single constituent models [25]. Another study [10] used six ML regression models, such as MLR, DTR, LR, RR, XGBR, and RFR, to predict LoS in a hospital. The study concluded that the RFR model outperformed the other models and achieved the lowest MSE of five days. A LoS study for patients with sepsis [23] compared six ML regression models, namely LR, RFR, KNN, NN, XGBR, and light GBR and found that light GBM showed the best model having MAE of 4 days. Similarly, in another study [24], light GBR showed a lower prediction error MAE of 3 days in inpatients.

There are several advantages of XGBR over other ML methods, such as the fact that it does not assume data distribution, uses decision trees, and can therefore be unaffected by multicollinearity. Ensemble methods, like XGBR, also automatically estimate feature importance from a trained predictive model, resulting in a boosted decision tree score for each feature. Consequently, the most important features are identified based on the chosen model, which includes sex, arrival mode (by own means/walk-in or by ambulance), triage category (light blue, orange, green, or red), age group (70 and older, 50 to 69, or 30 to 49), and specific conditions such as hematochezia/rectal bleeding/melena, pain at the side, presyncope, pallor/anemia, generalized asthenia, and prescriptions/performance requests. However, the values of performance measures can vary depending on the dataset and context. It is, therefore, important to conduct rigorous testing and validation to ensure that the chosen model is appropriate for the task at hand. Additionally, other studies may use different sets of performance measures or evaluate different subsets of regression models, which can make direct comparisons difficult.

The findings of the present study have the potential to assist in predicting the future LoS for new patients. In this way, healthcare providers or hospital administrators can estimate and manage their resources more effectively. Patients can also benefit from knowing how long they will be staying in the hospital, estimating their treatment costs, and managing their other personal affairs in a timely manner. Also, the findings can help other researchers assess the performance of different regression models. The other benefit of this study is that GM is the

most prevalent and heterogenous department, which makes it essential for an hospital to consider when deciding how long to keep new patients in the facility.

3.4.1. Limitation of the study

It is important to note that this study has several limitations. This study was limited by the fact that a single cohort was used to develop and test the model, meaning that no external validation was possible. Despite promising results, external validation is vital to ensuring the model's generalization abilities [31]. In other words, evaluating the model's performance on different datasets and populations is necessary to ensure its robustness and reliability. As another limitation, the study did not include vital signs or laboratory test results for triage assessment, which could be crucial for predicting hospital stay length.''

3.5. Conclusions

XGBR was the best-performing model out of the nine ML regression models, with the lowest RMSPE and MAPE. The proposed method performs considerably in predicting LoS, which is expected to provide valuable support for clinical decision-making. Most influencing variables for LoS prediction were sex, mode of arrival to the ED, and old age. Patients' criticalities clustered into four groups according to their LoS average and inter-quartile range. LoS prediction models could be highly advantageous to healthcare providers, as they can leverage them to accurately predict the duration of a patient's hospital stay, ensuring top-notch care and sufficient resources for all patients.

References :

1. Zolbanin, H. M., Davazdahemami, B., Delen, D., & Zadeh, A. H. (2022). Data analytics for the sustainable use of resources in hospitals: Predicting the length of stay for pa-tients with chronic diseases. Information & Management, 59(5), 103282.

2. Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S (2018) Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. PLoS ONE 13(4).

3. Chuang, M.-T., Hu, Y., & Lo, C.-L. (2018). Predicting the prolonged length of stay of general surgery patients: a supervised learning approach. International Transactions in Operational Research, 25(1), 75–90.

4. Lamere, A. T., Nguyen, S., Niu, G., Olinsky, A., & Quinn, J. (2021). Predicting the Length of Stay in Hospital Emergency Rooms in Rhode Island (pp. 35–48).

5. Sanderson H.F , Mountney L.M. (1997) The development of patient groupings for more effective management of health care The European Journal of Public Health 7, 210-214.

6. Happer P.R. (2005) A review and comparison of classification algorithms for medical decision making. Health Policy, 71, 315-331.

7. Happer P.R. (2002) A Framework for Operational Modelling of Hospital resources. Health Care Management Science, 5, 165-173.

8. Churilov L, B. A, Schwartz D, S. K , DALLY M. (2005) Data Mining with Combined Use of Optimization Techniques and Self-Organizing Maps for Improving Risk Grouping Rules: Application to Prostate Cancer Patients. Journal of Management Information Systems, 21, 85-100.

9. Costa A.X, Ridley S.A, Shahani A.K, Happer P.R & Senna V.DE. (2003) Mathematical modelling and simulation for planning critical care capacity. Anaesthesia, 58, 320-327.

10. Siddiqa, A., Abbas Zilqurnain Naqvi, S., Ahsan, M., Ditta, A., Alquhayz, H., A. Khan, M., & Adnan Khan, M. (2022). Robust Length of Stay Prediction Model for Indoor Pa-tients. Computers, Materials & Continua, 70(3), 5519–5536.

11. Zeleke AJ, Palumbo P, Tubertini P, Miglio R, Chiari (2023) L. Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis. Front Artif Intell. Jul 28;6:1179226. doi: 10.3389/frai.2023.1179226. PMID: 37588696; PMCID: PMC10426288

12. Kolcun, J. P. G., Covello, B., Gernsback, J. E., Cajigas, I., & Jagid, J. R. (2022). Ma-chine learning to predict passenger mortality and hospital length of stay following mo-tor vehicle collision. Neurosurgical Focus, 52(4), E12.

13. Liu, V., Kipnis, P., Gould, M. K., & Escobar, G. J. (2010). Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. Med-ical Care, 48(8), 739–744.

14. Esfandiari, N., Babavalian, M.R., Moghadam, A.-M.E., Tabar, V.K., 2014. Knowledge discovery in medicine: Current issue and future trend. Expert Syst. Appl. 41, 4434–4463.

15. Freedman, D. A. (2009). Statistical models: theory and practice. Cambridge University Press.

16. Loh, W. (2008). Classification and Regression Tree Methods (In Encyclopedia of Statistics in Quality and Reliability.

17. Tin Kam Ho. (n.d.). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 278–282.

18. G. Ke, M. Qi, F. Tomas et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, ACM, Long Beach, CA, USA, 2017.

19. Burges, C.J. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121–167.

20. Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

21. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.

22. Garg L., McClean S.I., Barton M., Meenan B.J., and Fullerton K. Intelligent patient management and resource planning for complex, heterogeneous, and stochastic healthcare systems. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans 42, pp. 1332–1345.

23. Chen, L., & Klasky, H. B. (2022). Six Machine-Learning Methods for Predicting Hospi-tal-Stay Duration for Patients with Sepsis: A Comparative Study. SoutheastCon 2022, 302–309.

24. Xiangrui Zeng, "Length of Stay Prediction Model of Indoor Patients Based on Light Gradient Boosting Machine", Computational Intelligence and Neuroscience, vol. 2022, Article ID 9517029, 14 pages, 2022.

25. Gabriel RA, Harjai B, Simpson S, Du AL, Tully JL, George O, Waterman R. An Ensemble Learning Approach to Improving Prediction of Case Duration for Spine Surgery: Algorithm Development and Validation. JMIR Perioper Med. 2023 Jan 26;6:e39650.

26. Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association, 78(383), 605–610.

27. Cameron, A.C. and Trivedi, P.K., Regression analysis of count data, Cambridge University Press: Cambridge, 1998.

28. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016;35(23):4124– 35.

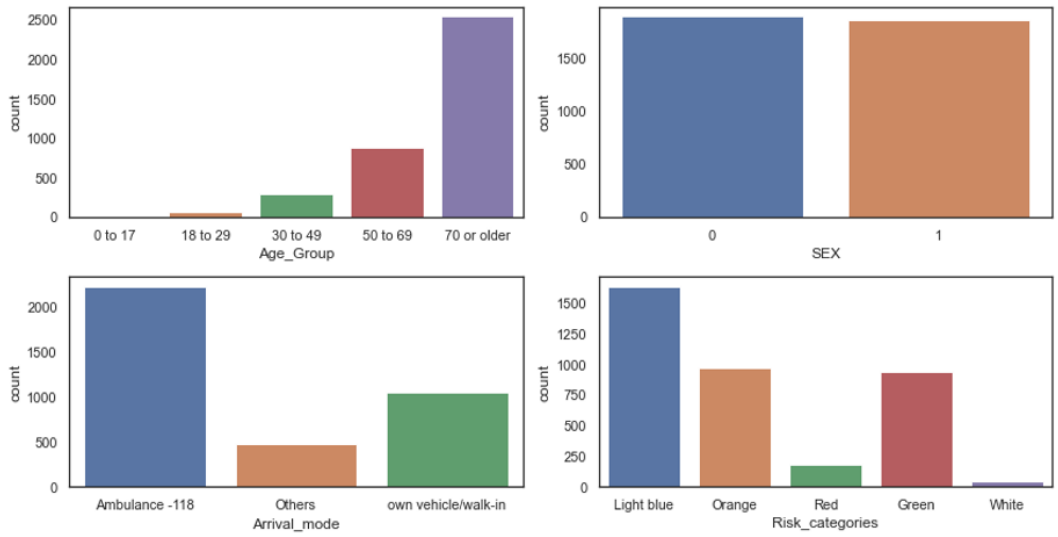
29. Shea, S., Sideli, R.V., Dumouchel, W., Pulver, G., Arons, R.R., Clayton, P.D.: Computer-generated informational messages directed to physicians: effect on length of hospital stay. J. Am. Med. Inform. Assoc. 2(1), 58–64 (1995).

30. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

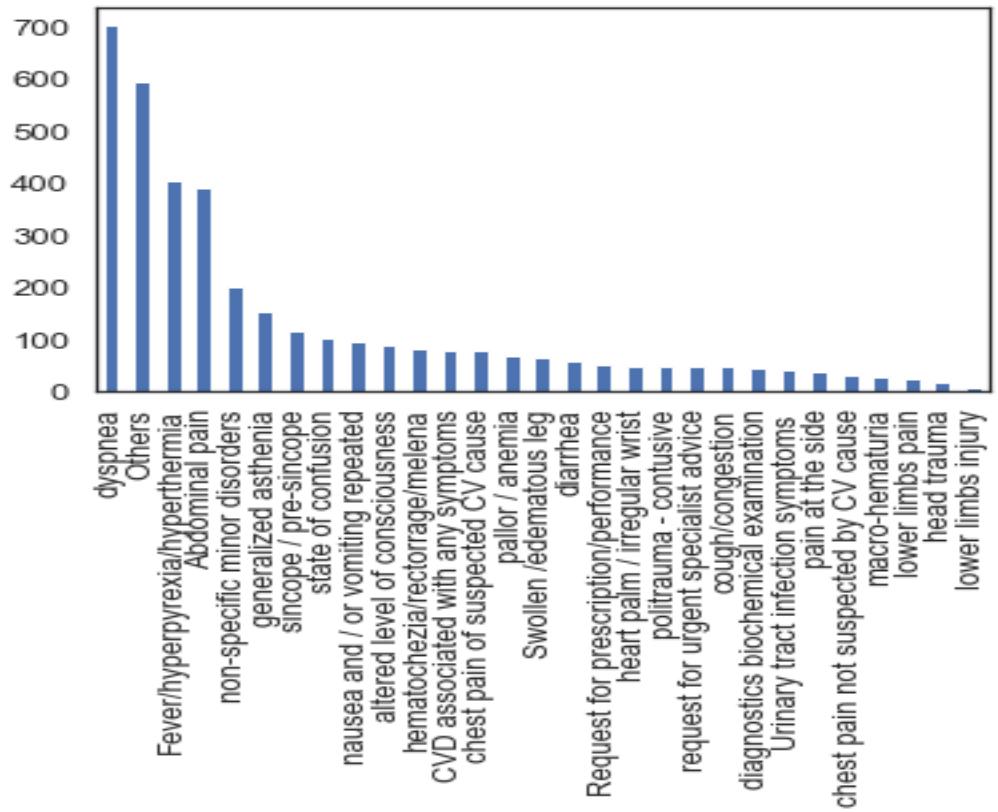
31. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016 June 22;353:i3140.

32. Lundberg, S. et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell.2, 56–67 (2020).

Appendix



A)



B)

Figure S1: Count plot of the predictors (A is for Ages and Sexes (top panel), and for Arrival Mode and Risk Categories (bottom panel), and B is for Criticalities or Problems).

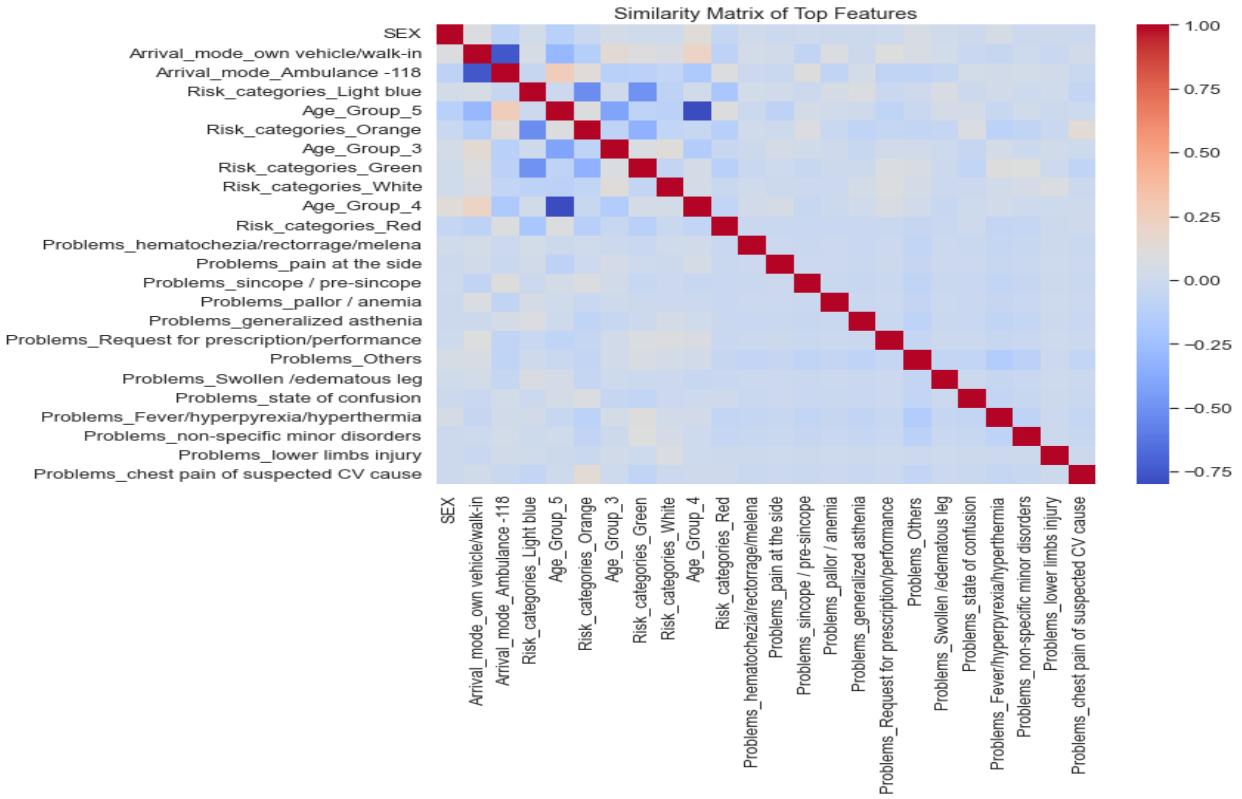


Figure S2: Correlation matrix for the most important features

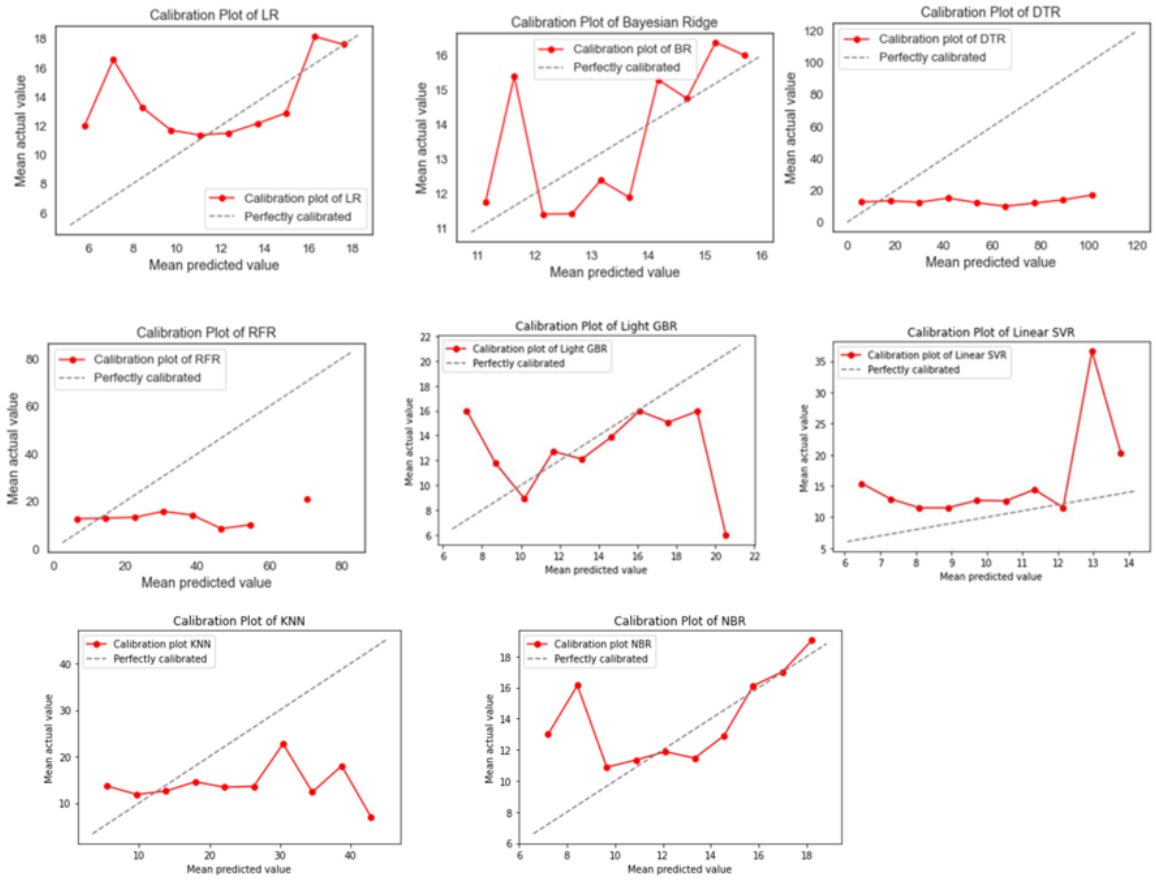


Figure S3: Calibration plots for each ML regression model

CHAPTER FOUR

Spatiotemporal heterogeneity of SARS-CoV-2 diffusion at the city level using geographically weighted Poisson regression model: The case of Bologna, Italy

---

This chapter is largely taken from our published work [∗], where we analyzed the spatiotemporal heterogeneity of SARS-CoV-2 diffusion using geographically weighted Poisson regression model.

∗Zeleke AJ, Miglio R, Palumbo P, Tubertini P, Chiari L; Bologna MODELS4COVID Study Group of the University of Bologna and the National Institute for Nuclear Physics (INFN). Spatiotemporal heterogeneity of SARS-CoV-2 diffusion at the city level using geographically weighted Poisson regression model: The case of Bologna, Italy. Geospat Health. 2022 Dec 1;17(2). doi: 10.4081/gh.2022.1145. PMID: 36468589. https://pubmed.ncbi.nlm.nih.gov/36468589/

Abstract

This work aimed to analyze the spatio-temporal patterns of the diffusion of SARS-CoV-2, the virus causing coronavirus 2019 (COVID-19), in the city of Bologna, the capital and largest city of the Emilia-Romagna Region in northern Italy. The study took place from February 1st, 2020 to November 20th, 2021 and accounted for space, sociodemographic characteristics and health conditions of the resident population. A second goal was to derive a model for the level of risk of being infected by SARS-CoV-2 and to identify and measure the place-specific factors associated with the disease and its determinants. Spatial heterogeneity was tested by comparing global Poisson regression (GPR) and local geographically weighted Poisson regression (GWPR) models. The key findings were that different city areas were impacted differently during the first three epidemic waves. The area-to-area influence was estimated to exert its effect over an area with 4.7 km radius. Spatio-temporal heterogeneity patterns were found to be independent of the sociodemographic and the clinical characteristics of the resident population. Significant single-individual risk factors for detected SARS-CoV-2 infection cases were old age, hypertension, diabetes and co-morbidities. More specifically, in the global model, the average SARS-CoV-2 infection rate decreased 0.93-fold in the 21–65 years age group compared to the >65 years age group, whereas hypertension, diabetes, and any other co-morbidities (present vs absent), increased 1.28-, 1.39- and 1.15-fold, respectively. The local GWPR model had a better

fit better than GPR. Due to the global geographical distribution of the pandemic, local estimates are essential for mitigating or strengthening security measures.

Key words: COVID-19; local regression; mapping; spatial heterogeneity; Bologna; Italy.

4.1. Introduction

4.1.1. Importance of addressing COVID-19

As noted by the World Health Organization (WHO), the rapid spread of Coronavirus disease 2019 (COVID-19), caused by the SARS-CoV-2 virus, is a global health problem (WHO, 2020a). From the first case detected in Wuhan in December 2019, COVID-19 has caused 495million incident cases and over 6.1 million deaths as of April 07th, 2022 (Worldometer, 2022). This posed significant challenges to health systems, particularly in Italy, the first European country to be infected (Specchia et al., 2021), where the overwhelming increase of patients requiring hospitalization at the start of the pandemic threatened the national health care system with collaps (Armocida et al., 2020). The COVID-19 pandemic underlines the need for a management plan to forecast hospital service demand so that healthcare facilities can efficiently make resources accessible when needed. Furthermore, older persons, including those with underlying health conditions, like cardiovascular disease, diabetes, chronic respiratory disease and cancer, are more likely to develop severe COVID-19 illness (WHO, 2020b).

4.1.2. Space-time risk models for COVID-19 are useful but rare/absent in the literature

Due to the pandemic's geographical spread, local estimates can play an essential role in mitigating or strengthening security measures. These estimates may regard the level of risk at each local area or the large-scale patterns of spreading of the contagion. However, estimates at the city level are rarely reported because official data on COVID-19 cases are only released, at least in Italy, at the provincial level (the number of people infected) or at the regional level (the number of deaths due to the virus). Furthermore, patient characteristics, such as health risk factors, pre-existing health conditions and from laboratory test results, which are helpful to estimate the level of risk in a population, are hardly ever available. Thanks to the joint effort of our Local Health Authority (AUSL Bologna), an Italian research hospital (IRCCS Policlinico

Sant'Orsola-Malpighi) and the Municipality of Bologna, we could access these data in a common framework, after identification, linkage, and anonymization of the relevant data sources. This allowed us to estimate the level of risk in each city area and the space-and-time patterns of SARS-CoV h-2 diffusion across areas. It also permitted investigation how demographic inequalities, health risk factors and cases at different periods are associated with the infection dynamics. Spatial models are important tools for studying the geographic relationship between many explanatory variables and disease outbreaks (Mollalo et al., 2015; Mollalo et al., 2016). Spatial analysis of small geographical areas can provide useful information about at-risk areas, but it is not yet possible to investigate geographical variation or spatial analysis at the level of a single city.

4.1.3. Statistical methods to derive space-time models

Most studies in medical research are based on classical regression models like ordinary least square (OLS) regression and generalized linear models (GLM) (Choi et al., 2017; Takele et al., 2019; Zheng et al., 2017) but these classical models produce bias by the resulting, average parameters over the whole study area without considering geographical variation. This kind of global regression models cannot detect non-stationary phenomena and may thus obscure differences in relationships between predictors and the outcome variable. Based on results from global models in which non-stationary is present but not detected, public policy inferences will be variable and may even be relatively poor in specific local/regional cases (Ali et al., 2007). Local geographically weighted Poisson regression (GWPR) modelling, on the other hand, calculates local regression coefficients to help healthcare professionals better understand how the effects of independent factors vary depending on where they are located (Haque et al., 2012; Matthews et al., 2012). This allowed us to investigate possible geographical variations in disease infection rates and other health problems accounting for the level of risk at the local, resident population level (Nakaya et al., 2005; Yang et al., 2012; Zhou et al., 2015). The unobserved spatial heterogeneity cannot be directly accounted for in standard models or global regression models since the estimated parameters are fixed and represent a mean relationship between dependent and explanatory variables. However, a GWPR model is best for relaxing the fixed association assumption between the response and the explanatory variables thereby

capturing spatial heterogeneity. Recent studies have shown that the applications of GWPR in a wide variety of fields, including but not limited to, analysis of the effects of demographic and socio-economic factors on the spatial variation of COVID-19 (Chen et al., 2022; Shawky et al., 2021; Zhang et al., 2021), health and other disease analysis (Bui et al., 2018 ; Chen et al., 2010; Goovaerts, 2005; Nakaya et al., 2005; Poliart et al., 2021; Yang et al., 2009), traffic crash modelling (Li et al., 2013; Zhao et al., 2004), population density and housing (Mennis and Jordan, 2005; Chen et al., 2017), poverty mapping (Benson et al., 2005; Loubert et al., 2018) and diseases mapping (Nakaya et al., 2005). In most of these studies, however, one of the challenges is the presentation and synthesis of the large number of "mappable" results generated by local GWR models.

4.1.4. Risk factors for COVID-19 infection

Recent studies referred to by the U.S. Centers for Disease Control and Prevention (CDC) have shown that the elderly and individuals with underlying co-morbidities, such as cardiovascular disease, diabetes and hypertension, are more susceptible to complications and possibly death due to COVID-19 (CDC COVID-19 Response Team, 2020). These groups also stand an increased risk of becoming infected (CDC, 2020).

Infections are more likely to occur in those with diabetes who are older and have poor glycemic control (Klekotka et al., 2015; Umpierrez et al., 2002). In the absence of publicly available findings, it is clear that diabetes patients are more likely to contract COVID-19 infection (Choi et al., 2016). Hypertensive patients had a higher rate of COVID-19 infection in a trial of 41 patients conducted in a Wuhan hospital (Huang et al., 2020); a larger analysis of 138 hospitalized patients confirmed these results (Wang et al., 2020). There is a heterogeneous nature to COVID-19 transmission processes over space (Wang et al., 2020; Chen et al., 2021). Variables such as local demographics, risk factors, epidemic waves and socio-economic characteristics, play a role in COVID-19 transmission (Dowd et al., 2020; Gatto et al., 2020; Pedro et al., 2020; Qiu et al., 2020). With the help of local modelling, researchers have been able to estimate the geographic variation in the relation between the outcome variable and explanatory factors more accurately. However, in some circumstances, predictors in the GWPR

model might not all have spatial effects, so there is a need to check the spatial variability of the predictors in advance.

4.1.5. Aims

In this study, we aimed to analyze space and time patterns of SARS-CoV-2 diffusion across the 90 statistical area of Bologna, Italy, from February 1st, 2020 to November 20th, 2021, accounting for the space-referenced sociodemographic characteristics and health conditions of the resident population. Notably, we take an interest in how well global and local GPWR models explain the variations in disease infection rates, which means that using the global model in the absence of variation and the local one if variations were found, comparing the results for optimal results. The secondary goal of this study was to derive a model for estimating the level of risk of infection with SARS-CoV-2, based on sociodemographic factors and health conditions and to identify and measure the place-specific factors associated with diseases and their determinants. That that end, COVID-19 infection levels were compared with respect to sociodemographic and clinical features at different time points over the entire study period.

4.2. Materials and Methods

4.2.1. Study area and population

Our study area refers to the city of Bologna, the capital and largest city of the Emilia-Romagna Region in northern Italy. It is the seventh most populous city in Italy, with about 400,000 inhabitants and 150 different nationalities, covering an area of 140.9 km2 (Ufficio Statistica Regionale-Regione Emilia Romagna, 2019). Figure 1 shows the subdivision of the municipal territory into 90 statistical areas, pre-defined by the Statistical Bureau of the Bologna Municipality. This study obtained data with permission from the Local Health Unit (AUSL) of the Municipality of Bologna. The reference population consisted of all residents in the city as of February 1st, 2020 (Figure 1).

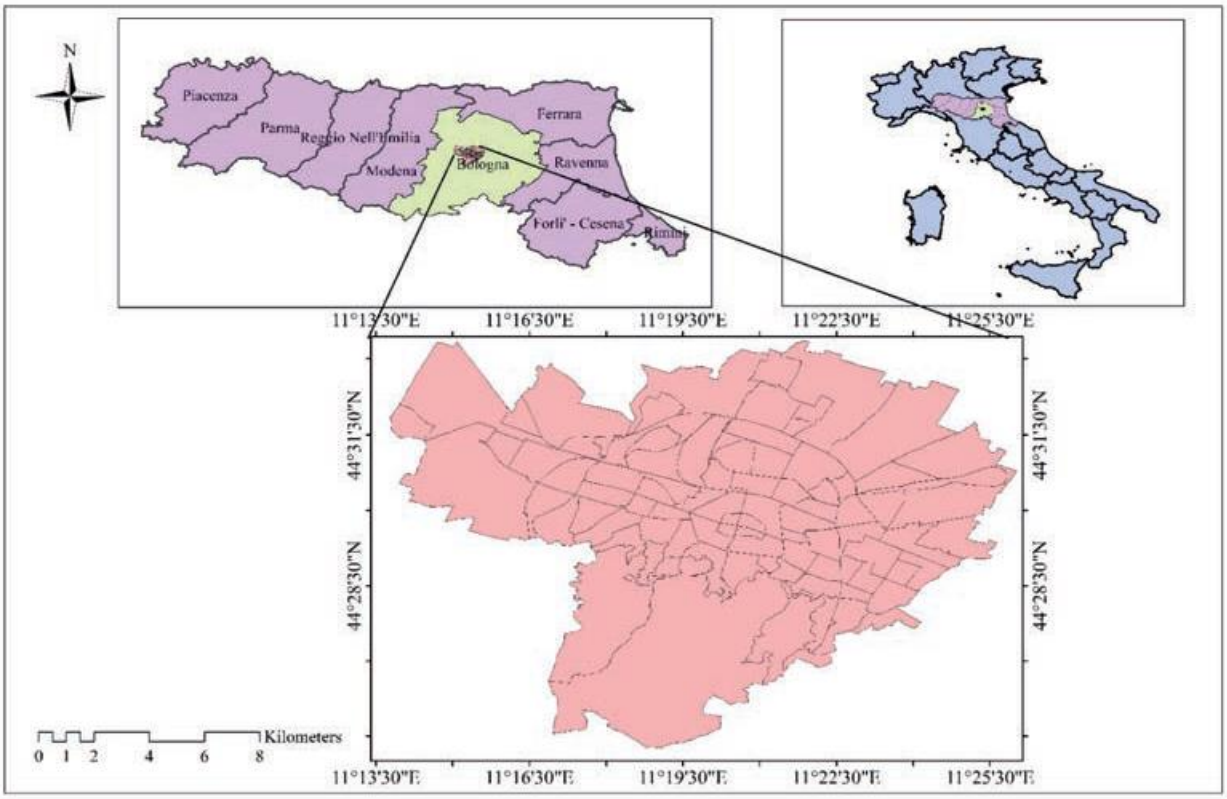


Figure 1. Location of the study area. The city of Bologna, here divided into 90 statistical areas, is the capital of the Emilia-Romagna Region (top-left panel).

4.2.2. Study variables

The infection rate of COVID-19 in each of the 90 statistical areas was taken as the response (outcome) variable); as usual, the population size was used as offset in the following models. The individual COVID-19 case status was classed as 1 = Yes and 0 = No. Some explanatory variables were referred to as population characteristics and others as area characteristics. In this study, the former comprised demographics, health-related characteristics, age (divided into <20, 20-65 and >65 years), gender, family (divided according to size (n=1), (n=2), (n=3) and (n=≥4), presence of co-morbidities (hypertension, diabetes or other). When tested for COVID-19 infection, people had documented their coexistent or preconditioned medical history according to the Index of Coexistent Disease (ICED).

The area variables included data about the 90 statistical areas of the city of Bologna. In particular, we used the total number of residents as well as the position (latitude and longitude) of the centroid of each statistical area. The minimum geo-point Y-coordinate (latitude) was 44.4592, with the maximum = 44.5463, while the minimum X-coordinate (longitude) was 11.2359, with the maximum = 11.3638. The COVID-19 case counts (all individual information) were aggregated by statistical area (into the 90 areas).

4.2.3. Study periods

The observation period extended from February 1st, 2020 to November 20th, 2021.

As in our previous study (Zeleke et al., 2022), we partitioned our observation periods according to epidemic waves assuming that a wave starts with more than 500 total hospitalizations and ends when the hospitalization rate drops below 500. Using this cutoff, we chose a day in the stationary phase between the second and the third wave located in the middle of this period. We chose a middle point to divide the second from the third wave since hospitalization in that period never dropped to less than 500. Using this criterion, four wave periods were identified (Figure 2).

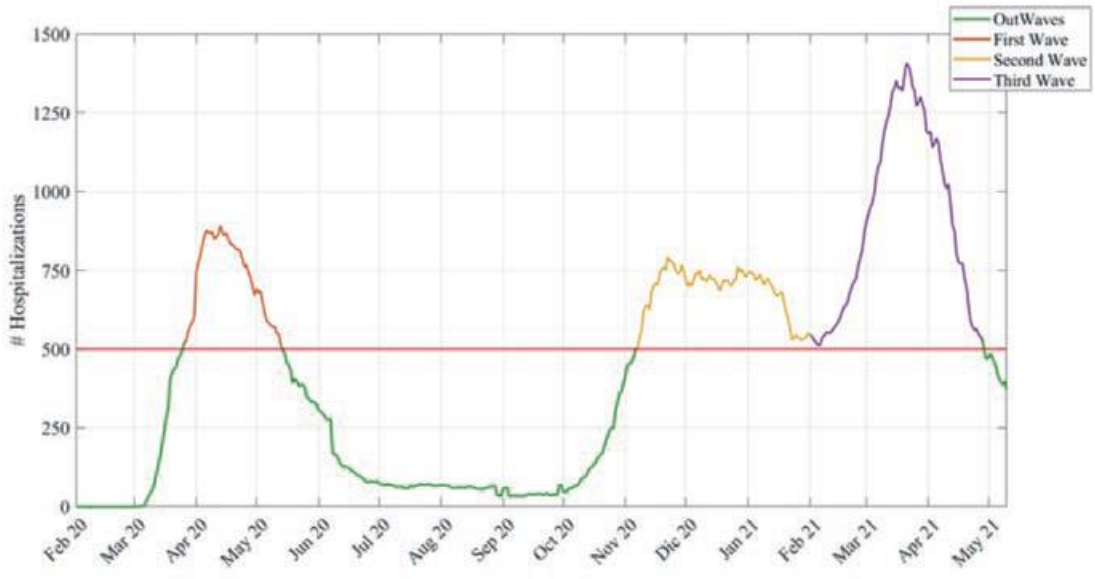


Figure 2. Number of hospitalizations over the study period and the empirical definition of infection waves. First Wave → 26 March 2020 - 13 May 2020 (48 days); Second Wave →07

November 2020- 01February 2021 (86 days); Third Wave →02February 2021 – 28 April 2021 (85 days); Out Waves → other periods.

4.2.4. Statistical modelling

Estimation of infection rates;

COVID -19 infection rate each local statistical area was calculated using the following formula for:

$$IR_i = \frac{C_i}{P_i}$$

where IRj is the estimated infection rate for statistical area unit i ; Ci the number of observed cases in the statistical area unit; and Pi is the total number of people living in the ith statistical area.

Comparison of COVID-19 infection rate predictors during the different waves;

COVID-19 infection rates were compared during different time points over the entire study period based on sociodemographic features, clinical features and case data. The characteristics and dates for each wave are listed in Figure 2.

Spatial distribution of predictors and the vaccination status by area;

We looked at how each sociodemographic variable was distributed and classified in each statistical area in the city. Additionally, we tried to determine the vaccination status of the population in the areas. However, because individual vaccination varies over time and can be partial or complete, it would not be appropriate to use the COVID-19 infection rate as a measure of vaccination status. Instead, the time-to-event model, i.e. the survival analysis according to the Cox proportional-hazards (COXPH) model (Cox, 1972), was preferable since it also helped control the time-varying confounding factors and examine the relationship between a predictor variable and the survival time.

4.2.4.1. Testing for spatial variability or spatial non-stationarity

The variability of the local coefficient over space can be used to examine the plausibility of the stationarity assumption held in global regression. As suggested by Nakaya (2014) for the GWR4 Windows application, we performed a model comparison for each variable coefficient to determine its geographic variability based on the difference in model comparison indicator between the GWR model and the switched GWPR model (Diff of Criterion), which indicates whether the variable is a local variable (locally fixed) or a global variable (locally variable). For example, to compare the variability of the kth varying coefficient, we compared the fitted GWPR - the original model, with the switched model, which keeps the remaining coefficients unchanged from the fitted GWPR. When the corrected Akaike Information Criterion (AICc) comparison suggests that the original GWPR model outperforms a switched GWPR model, it could be inferred that the kth coefficient changes significantly with space (the DIFF of Criterion becomes negative). In a switched GWR model that is well fitted, there is no spatial variability in the local coefficients (the DIFF of Criterion becomes positive). Thus, when this indicator gives positive numbers, there is no spatial variability, which means that the global estimate can be considered; when negative, however, the DIFF of Criterion is a sign of spatial variability or non-stationarity.

Similarly, many other authors (Brunsdon et al., 1996; Fotheringham et al., 2003; Nakaya et al., 2005) used the Monte Carlo test. Multiscale GWR (MGWR) 2.2 (Oshan et al., 2019)allows running a Monte Carlo test to determine whether the spatial variability of the local estimates is due to sampling variation or other intrinsic processes. The Monte Carlo test calculates local parameter estimates once and then calculates new local parameter estimates after randomly rearranging the data points to see if the variability of each parameter surface is due to chance (i.e. using the 0.05 significance level, non-significant values ($p > 0.05$) are considered as absence of spatial heterogeneity, while other significant values indicate spatial heterogeneity. Therefore local coefficient estimation for a variable is the next step. In our analysis, we considered both criteria to test for spatial non-stationarity.

4.2.4.2. Global Poisson Regression model (GPR)

A local regression model can evaluate the presence of changes in the importance of different variables over space, while global regression models express the relationship between the dependent and independent variables by a constant mean value across the study area. The relationship between COVID-19 infection rate and available covariates were considered using global and local regression models. The classical GLM Poisson regression model (Fotheringham et al., 2013) is formulated as follows:

$$ln(E(IR_i|X)) = \beta_0 + \sum_k^p \beta_k x_{k,i}$$

where IRi is the value of the outcome variable (here the infection rate of COVID-19) in each statistical area of i; β0 the global Intercept; xk is the kth explanatory variable and βk are the parameters corresponding to the explanatory variables.

4.2.4.3. Local GWPR

Fotheringham et al. (2003) introduced GWPR to describe a family of regression models in which the coefficients are allowed to vary spatially and focus attention on local variations. Unlike the global model, the linear GWR is a non-stationary regression model with an additional function of spatial location, where the estimated coefficient parameters vary over space. The equation of this fitted model is as follows:

$$ln(E(IR_i|X) = \beta_0(u_i, v_i) + \sum_k^p \beta_k(u_i, v_i) x_{ik}$$

where IRi is the value of the outcome variable at the coordinate location i; (ui, vi); the two-dimensional geographical coordinates of the centroid of for each statistical area; β0 and βk represent the local estimated intercept and effect of variable j for location i, respectively. The iterative reweighted OLS method was used to calibrate the GWPR model using the GWR4.0 software (Nakaya et al., 2014). The parameter estimation value of sample i is given by (Fotheringham et al., 2003):

$$\hat{\beta}(u_i, v_i) = (X^T w(u_i, v_i) X)^{-1} X^T w(u_i, v_i) IR$$

Where $\hat{\beta}$ (ui, vi) is the vector of the local parameters in each statistical area of i; X the matrix of the independent variables with a column of 1s for the intercept; and w (ui, vi) is the spatial weight matrix that can be presented as:

$$w(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & w_{in} \end{bmatrix}$$

where wij is the weight given to each statistical area of j during the calibrating procedure for the statistical area i.

The calculation of GWPR coefficients consists of two major steps, the first of which is choosing a proper kernel function to express the spatial relationship between the observed units. and the second the selection of the optimal bandwidth for which the spatial weight matrix calculation could contribute to a better fit.

4.2.4.4. Kernel function and bandwidth selection

Using a kernel regression method to calibrate the model to predict smoothed geographical parameter variations with a distance based weighting scheme (Nakaya et al., 2005) is a key step in the development of the GWPR model. There are five kernel functions used for GWR or GWPR modelling in different research areas: Box-car, bi-square, tri-cube, exponential and Gaussian weighted function (Li et al., 2013). To obtain the most accurate possible estimation of spatial heterogeneity, bi-square was chosen after careful analysis and comparison of our dataset. It was expressed as follows:

$$w_{ij} = \begin{cases} [1 - (||u_i - v_j||/b)^2]^2 if\ ||u_i - v_j|| < b_i \\ 0 \qquad\qquad otherwise \end{cases}$$

where wij is the spatial/geographical weight; j a specific point in space where data are observe; i any point in space for which parameters are estimated; and ||ui – vj|| is the Euclidean distance between observation i and regression location j. The kernel size is controlled by the parameter bi (also known as bandwidth).

According to Fotheringham et al. (2003), bandwidth is defined as the optimum distance at which a location can still influence the observed location. Its size is important during model calibration since it determines the weights for the nearest spatial units. The optimal bandwidth was chosen using a bi-square adaptive kernel method. The GWR 4.0 software (Nakaya et al., 2014) provides a setting for optimum bandwidth selection in the golden section search algorithm. The bandwidth of the kernel might be fixed and the distance determines the size of the kernel to the point of interest. The kernel is the same at any point in space, or adaptive, where the size of the kernel is determined by the number of neighbours at the point of interest. The adaptive kernel (bi-square function) was employed in this paper as it allows the weighting method to vary spatially according to the density of data. Several studies found that adaptive (optimal) bandwidth performs better than fixed bandwidth (Braun and Rousson, 2000; Davies and Lawson, 2019; Wang and Li, 2017).

4.2.4.5. Spatial heterogeneity maps

After implementing the GWPR model in MGWR2.2 and GWR4 software, the estimated local coefficients were calculated along with significance values for each one (local t-values) in each statistical area of Bologna. Finally, the spatial heterogeneity map of local coefficients and local t-values were plotted to show the non-stationarity and spatial variability of the parameters. The t-statistic used to examine the significance of the local β coefficients for the jth parameter of the model was the following:

$$t_j(u_i) = \frac{\beta_j(u_i, v_i)}{se(\beta_j(u_i, v_i))}$$

where βj is the model's jth estimated parameter; the estimated parameter standard error; and (ui, vi)the geographical coordinates of reference point i.

4.2.4.6. Comparison of GWPR and GPR models

According to Nakaya et al. (2005, 2014), we used the following statistical approaches or bandwidth selection parameters to compare the performance and goodness of fit of the GWPR and GPR models: AICc, Bayesian information criterion (BIC), deviance and percent deviance

explained. The model indicates the best model performance with the lowest AICc (Hadayeghi et al., 2010; Fotheringham et al., 2003; Nakaya et al., 2005).

4.3. Results

Descriptive statistics

Figure 3 shows the aggregated COVID-19 infection rate for each of the 90 statistical areas from February 1st, 2020, to November 20th, 2021. The aggregation is based on the number of persons who lived in each of the 90 statistical areas at the time. Increasing infection is illustrated in the figure by the shift of shading from yellow to dark brown. The higher values were in the south-eastern districts and the lower in the north-western districts (Figure 3).

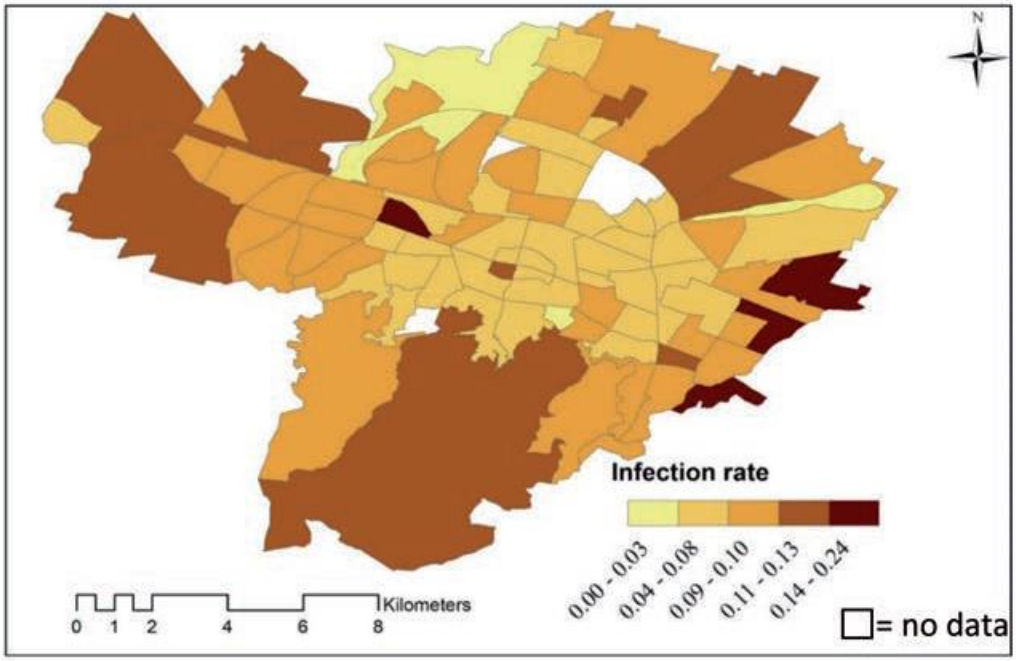


Figure 3. Distribution of the COVID-19 infection rate across Bologna' 90 statistical areas as of November 20th, 2021.

The number of people subjected to tests for COVID-19 in Bologna's 90 statistical areas between February 1st, 2020, and November 20th, 2021 were 154,544 and the cumulative positive rate amounted to 38,579 persons on the latter date, which represent 9.89 % of Bologna's population. The number of COVID-19 cases ranged from 1 to 55 in the different statistical areas (Table 1).

Table 1. Variables used in the analysis.

Variable	COVID infection (no.)	Percentage (%)	Infection rate (per 1,000 residents)
Age			
0-21	7,396	19.2	48
21-65	24,565	63.7	159
>65*	6,618	17.1	43
Gender			
Male	18,638	48.3	121
Female*	19,941	51.7	129
Family size			
1	8,462	21.9	55
2	8,364	21.7	54
3	8,661	22.5	56
≥4	13,092	33.9	85
Co-morbidity			
Any co-morbidity (yes/no*)	(7,388/31,191)	(19.2/80.8)	(48/202)
Hypertension (yes/no*)	(3,681/34,898)	(9.5/90.5)	(24/226)
Diabetes (yes/no)	(1,540/37,039)	(4.0/96.0)	(10/240)
Period			
First wave*	1,509	3.9	10
Second wave	15,696	40.7	102
Third wave	14,238	36.9	92
Out wave	7,136	18.5	46

*Indicates the reference categories of the covariates.

Comparison of COVID-19 infection predictors during the different waves

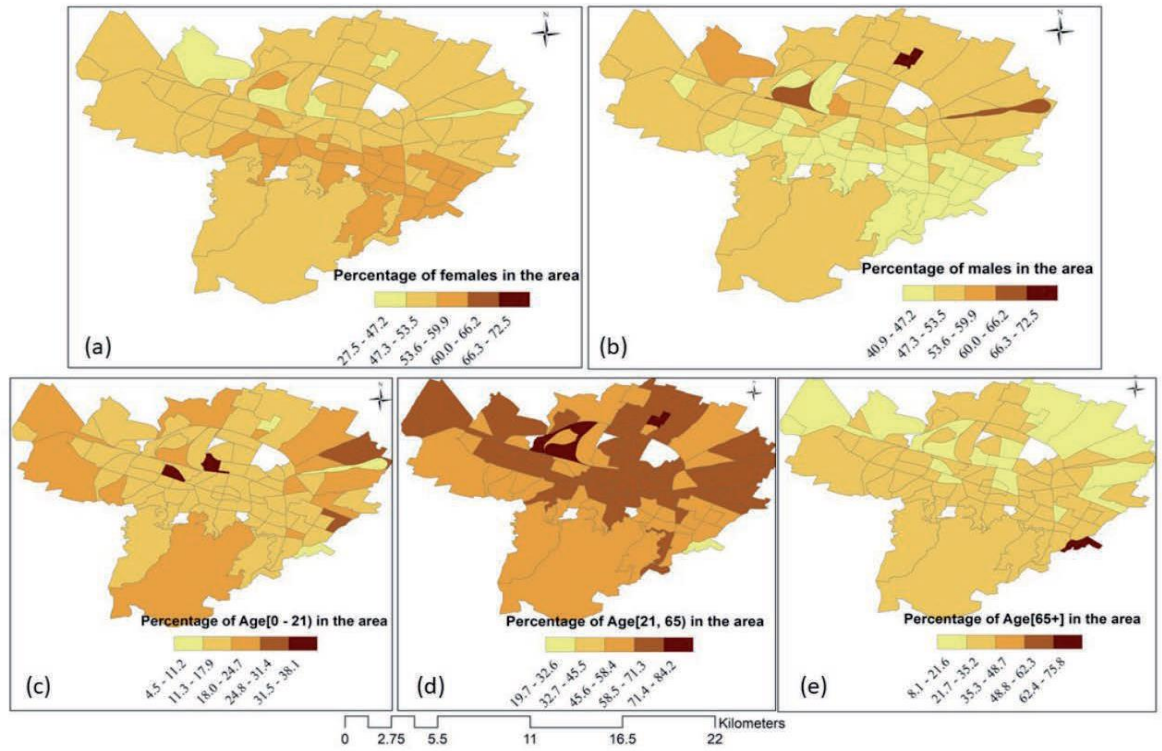
The descriptive comparison of COVID-19 cases during the different time points or epidemic waves over the entire study period, based on sociodemographic features, clinical features and case data, is shown in Table 2.

Table 2. Comparison of variables during the different COVID-19 waves in Bologna.

Variable	Total cases no. (%)	First wave no. (%)	Second wave no. (%)	Third wave no. (%)	Out waves no. (%)
Age					
0-21	7,396(19.2)	56(3.7)	2,524(16.1)	3,033(21.3)	1,783(25.0)
21-65	24,565(63.7)	1,053(69.8)	10,326(65.8)	8,856(62.2)	4,330(60.7)
≥65	656,618(17.1)	400(26.5)	2,846(18.1)	2,349(16.5)	1,023(14.3)
Gender					
Male	18,638(48.3)	594(39.4)	7,542(48.1)	6,944(48.8)	3,558(49.9)
Female	19,941(51.7)	915(60.6)	8,154(51.9)	7,294(51.2)	3,578(50.1)
Family size					
1	8,462(21.9)	513(34.0)	3,583(22.8)	2,887(20.3)	1,479(20.7)
2	8,364(21.7)	413(27.4)	3,538(22.5)	2,986(21.0)	1,427(20.0)
3	8,661(22.5)	273(18.1)	3,466(22.1)	3,341(23.5)	1,581(22.2)
≥4	1,3092(33.9)	310(20.5)	5,109(32.6)	5,024(35.2)	2,649(37.1)
Co-morbidity					
Any co-morbidity	7,388(19.2)	385(25.5)	3152(20.1)	2,678(18.8)	1,173(16.4)
None	31,191(80.8)	1,124(74.5)	12,544(79.9)	1,1560(81.2)	5,963(83.6)
Hypertension	3681(9.5)	219(14.5)	1,658(10.6)	1,297(9.1)	507(7.1)
Normal	34,898(90.5)	1,290(85.5)	14,038(89.4)	12,941(90.9)	6,629(92.9)
Diabetes	1,540(4.0)	86(5.7)	662(4.2)	582(4.1)	210(2.9)
Normal	37,039(96.0)	14,23(94.3)	15,034(95.8)	13,656(95.9)	6,926(97.1)
Total	38579(100)	1509(100)	15696(100)	14238(100)	7136(100)

Spatial distribution of predictors in the area

To begin, we examined how the population was distributed or varied within each demographic covariate category for each statistical area. As shown in Figure 4, there is a variation of people living in the area for each sociodemographic category.



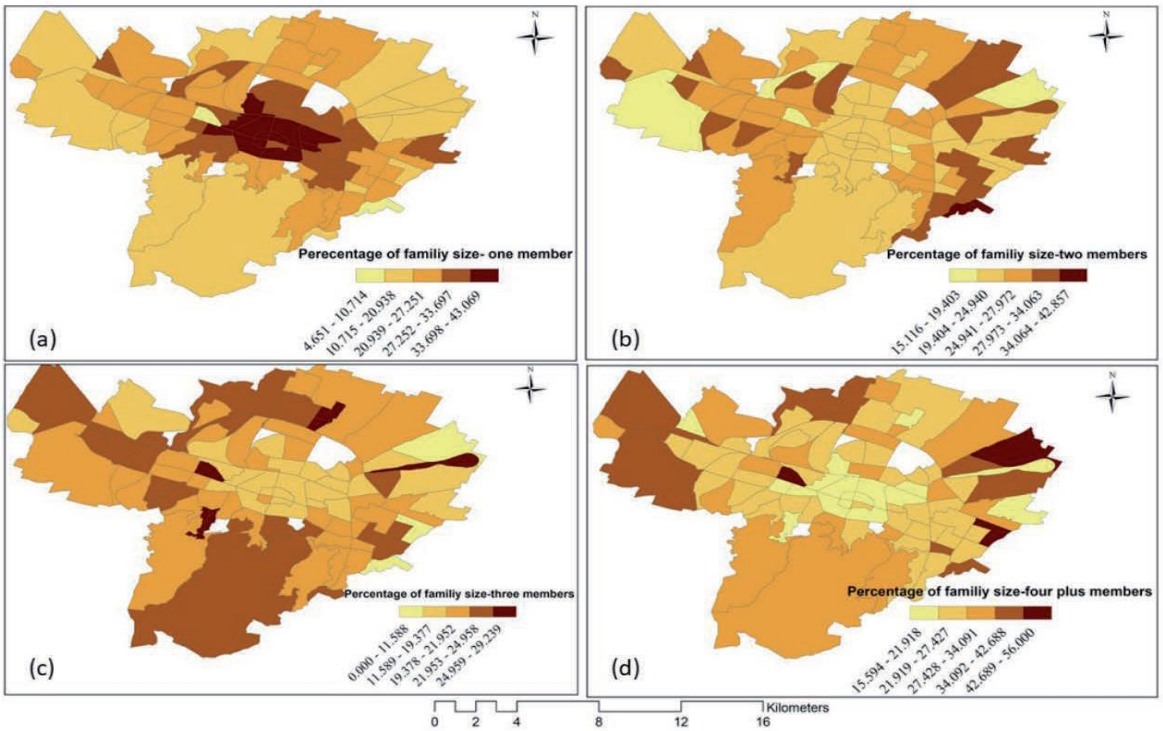


Figure 4. The spatial distribution of sociodemographic characteristics--percentage of residents living in the study area. White area = no data.

Vaccination status

As summed up in Table 3, people were said to be fully vaccinated when they had received their first and second dose of any vaccine by Pfizer, Moderna Biotech Spain, AstraZeneca S.P.A or one dose of Johnson and Johnson. Those who had received only one dose of Pfizer, Moderna Biotech Spain, or AstraZeneca S.P.A. were termed partially vaccinated individuals. Spatial distribution of Vaccination status in the 90 statistical areas of the population of Bologna are shown in Figure 5.

Table 3. Vaccination status of the Bologna population as of November 20th, 2021.

Status	People (no.)	Percentage (%)
Unvaccinated	161,561	41.22
Fully vaccinated (with or without booster)	210,180	53.63
Partially vaccinated	20,176	5.15

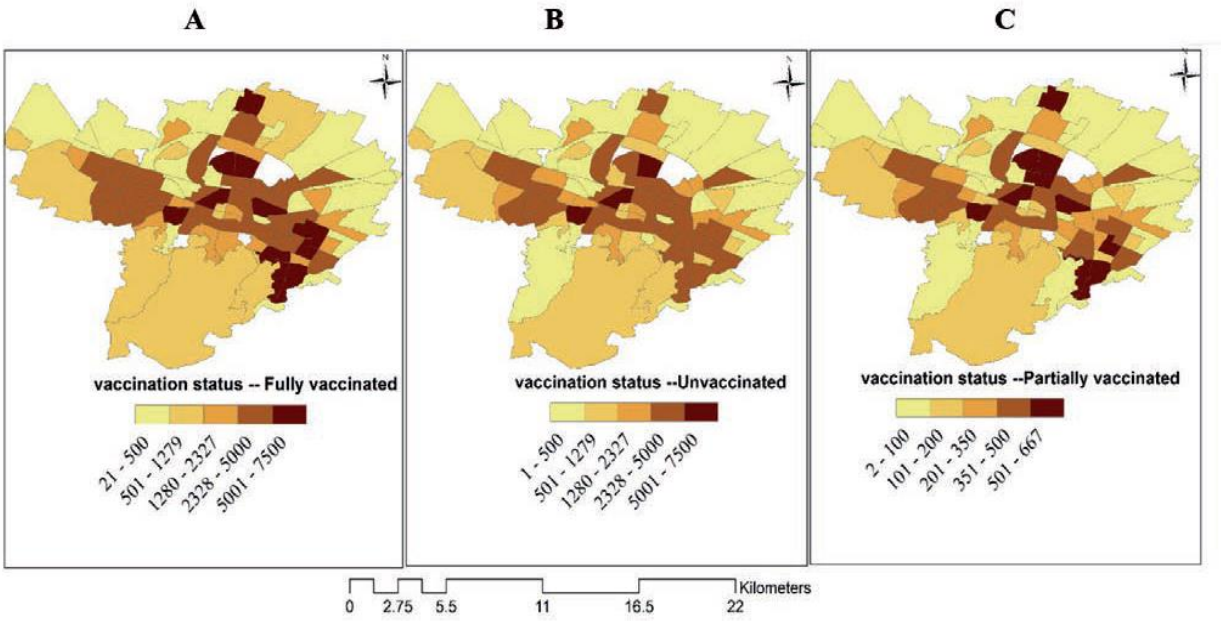


Figure 5. The spatial distribution of vaccinated people in the 90 statistical areas of Bologna. A) fully vaccinated; B) unvaccinated; C) partially vaccinated. White area = no data.

Testing for spatial variability

The results of the test for spatial variability using DIFF of Criterion and Monte Carlo Test are shown in Table S1 in Supplementary Materials. The Intercept, Second Wave, Third Wave,

and Out Waves variables had negative values (based on DIFF of criterion result), indicating that those variables had significant local variation in the area. Similarly, the Intercept, Second Wave, Third Wave and Out Waves variables had a smaller p-value ($p<0.05$) in the Monte Carlo Test, indicating that the parameter estimates had significant local variation. As a result, a spatial variability local (varying) model was used for those parameters. In contrast, gender, age group, all family size categories, hypertension, diabetes, and co-morbidity variables showed positive values (see Table 4). Similarly, the p-values for variables >0.05, indicating that the parameter estimates do not exhibit significant local variation so that we used the global (fixed) model for the interpretation of the model result.

Table 4. Summary statistics of the regression models.

Global model-GPR				Local model-GWPR				
Variable	Estimate	Rate ratio	P-value	Mean	STD	Min	Median	Max
Intercept	-2.601	0.074	<0.0001*	-2.653	0.050	-2.750	-2.650	-2.539
Male	0.014	1.014	0.166	-	-	-	-	-
Age [0–21)	-0.008	0.992	0.705	-	-	-	-	-
Age [21-65)	-0.068	0.934	<0.0001*	-	-	-	-	-
Family size-1	-0.007	0.993	0.655	-	-	-	-	-
Family size-2	-0.015	0.985	0.323	-	-	-	-	-
Family size-3	0.024	1.024	0.083	-	-	-	-	-
Hypertension	0.243	1.275	<0.0001*	-	-	-	-	-
Diabetes	0.329	1.389	<0.0001*	-	-	-	-	-
Any co-morbidity	0.139	1.149	<0.0001*	-	-	-	-	-
Second Wave	1.734	5.663	<0.0001*	1.786	0.064	1.656	1.778	1.815
Third Wave	1.656	5.238	<0.0001*	1.707	0.049	1.610	1.693	1.815
Out Waves	1.079	2.941	<0.0001*	1.138	0.087	0.894	1.153	1.260

GPR=global Poisson regression; GWPR=geographically weighted Poisson regression; STD=standard deviation; *Results statistically significant. Bandwidth size: 4689. The estimates are calculated for the reference categories (age 65+, family size - 4+, hypertension-no, diabetes-no, without any co-morbidities). As defined in the methodology, bandwidth is the optimum distance between a location and an observed location at which it is still possible to influence the location. The area-to-area influence was estimated to exert its effect over an area with 4.7 km radius.

GPR estimates

Table 4 shows, based on the global GPR model result, that the intercept and age group (21–65 years) are statistically significant variables (at p=0.05) that were negatively related to COVID-19 infection rates. In contrast, hypertension, diabetes, and any comorbidity, second wave, third wave, and out waves were statistically significant variables (at p=0.05) that positively affected the COVID-19 infection rate, while sex, the 0-20 years age group and family size did not significantly affect the COVID-19 infection rate. Comparisons were always made with reference categories. Table 4 further shows that, globally, the estimated rate ratios for the intercept, the 21-65 age group, hypertension (present vs absent), diabetes (present vs absent),

any comorbidity (present vs absent), the second wave, third wave and out waves were 0.074, 0.934, 1.275, 1.389, 1.149, 5.663 ,5.238 and 2.941, respectively, indicating that the average infection rates of COVID-19 decreased by 0.934-fold in the 21 – 65 years age group compared to the ≥65 years age group, while hypertension increased by 1.275-fold when compared to absence of hypertension and diabetes increased by 1.389-fold when compared to absence of diabetes, and the presence of any comorbidity increased by 1.149-fold when compared to a general absence of comorbidities. The infection rate of COVID-19 spiked (increased 5.663-fold) during the second wave compared to the first wave, 5.238-fold during third wave compared to the first wave and 2.941-fold during out waves compared to the first wave.

GWPR estimates

Based on the minimal AICc value, 4689 was calculated as the optimal bandwidth for the GWPR model. Table 4 shows the summary of parameter estimates in the local GWPR models. The effect of some covariates varies across the study area. As a result, it is important to map the local parameter estimated to see where there is significant spatial heterogeneity between the independent and the dependent variables. The estimated coefficient of the local parameters are described by the summaries of central tendency and dispersion indices expressed by the mean, the standard the deviation, the minimum, the median and the maximum. Only four variables, including the intercept were finally accepted for GWPR calibration. The results of this model indicated a goodness-of-fit (R2) of 62.1% with an AICc = 6076.

Spatial heterogeneity maps

After estimating the GWPR parameters, the map of spatial variability for each GWPR parameter and a significant explanatory variables map was created by ArcMap10.8. Jenks' Natural Breaks algorithm (Slocum,1999) were used to classify each category. Class breaks are created in a manner that maximizes the differences between classes while grouping similar values. The features are classified into classes whose boundaries are determined by the differences in the data values. Local parameter estimate (beta-local coefficients) map plots for each variable that showed a significant variability are shown in Figure 6, and it is identified that the parameters have an obvious pattern of spatial non-stationarity. The values of local

coefficients for the relationship between predictors and COVID-19 infection rate were not equal in all locations, with the direction and intensity of relationships changing depending on the geographical coordinates. The parameter of waves, the second wave ranges from 1.66 to 1.92, the third wave ranges from 1.61 to 1.93, and out waves ranges from 0.93 to 1.30. All the parameter signs are positive, indicating that all positively impacted the COVID-19 infection rates in the area. Whereas the baseline or intercept parameter value ranges from -2.74 to -2.54, the distribution implied that besides these three predictors, other factors were bound up with the COVID-19 infection rate. The colour code for the local parameter (the waves parameter) map plot indicates the estimated risk level for COVID-19 infection, where the high risk is represented by brown and decreases towards yellow (Figure 7).

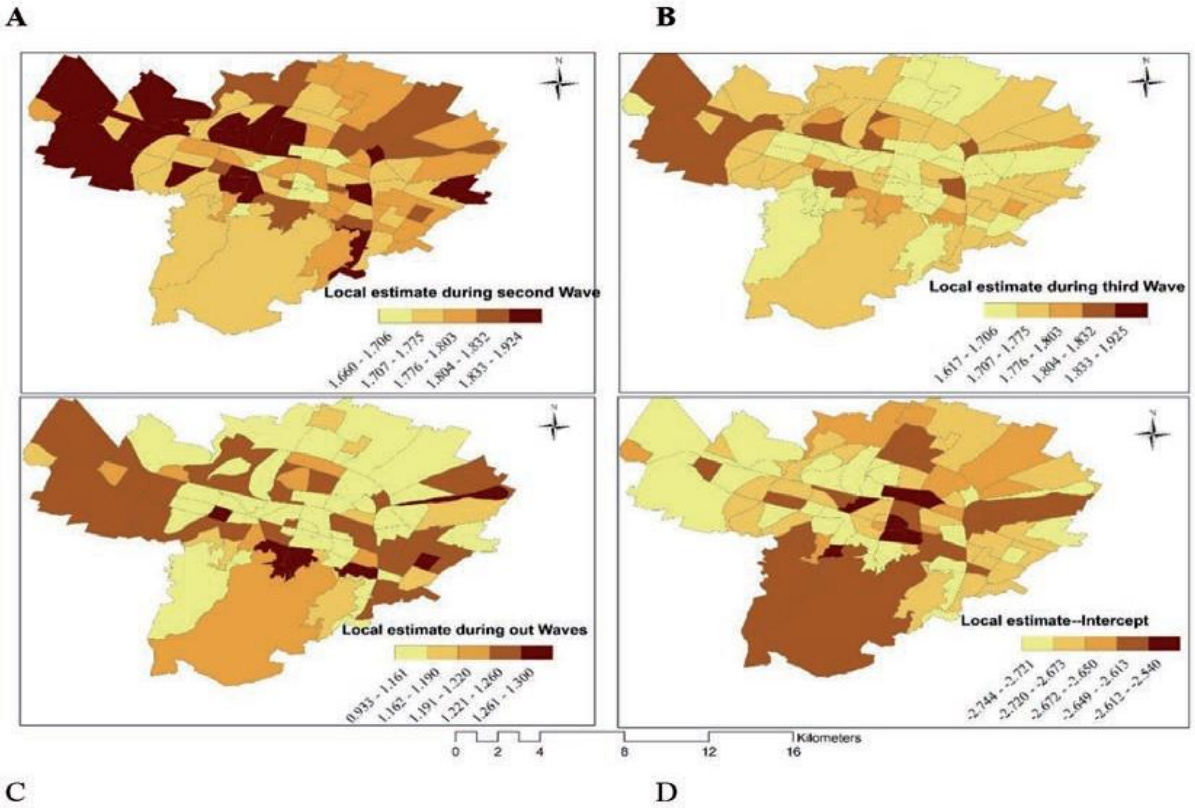


Figure 6. Spatial heterogeneity maps -the effects of predictors in describing COVID-19 infection rates at local level. Local parameter estimate /beta-local coefficients map plots; A) Second Wave; B) Third Wave; C) Out Waves; D) Intercept that showed aspatial variability.

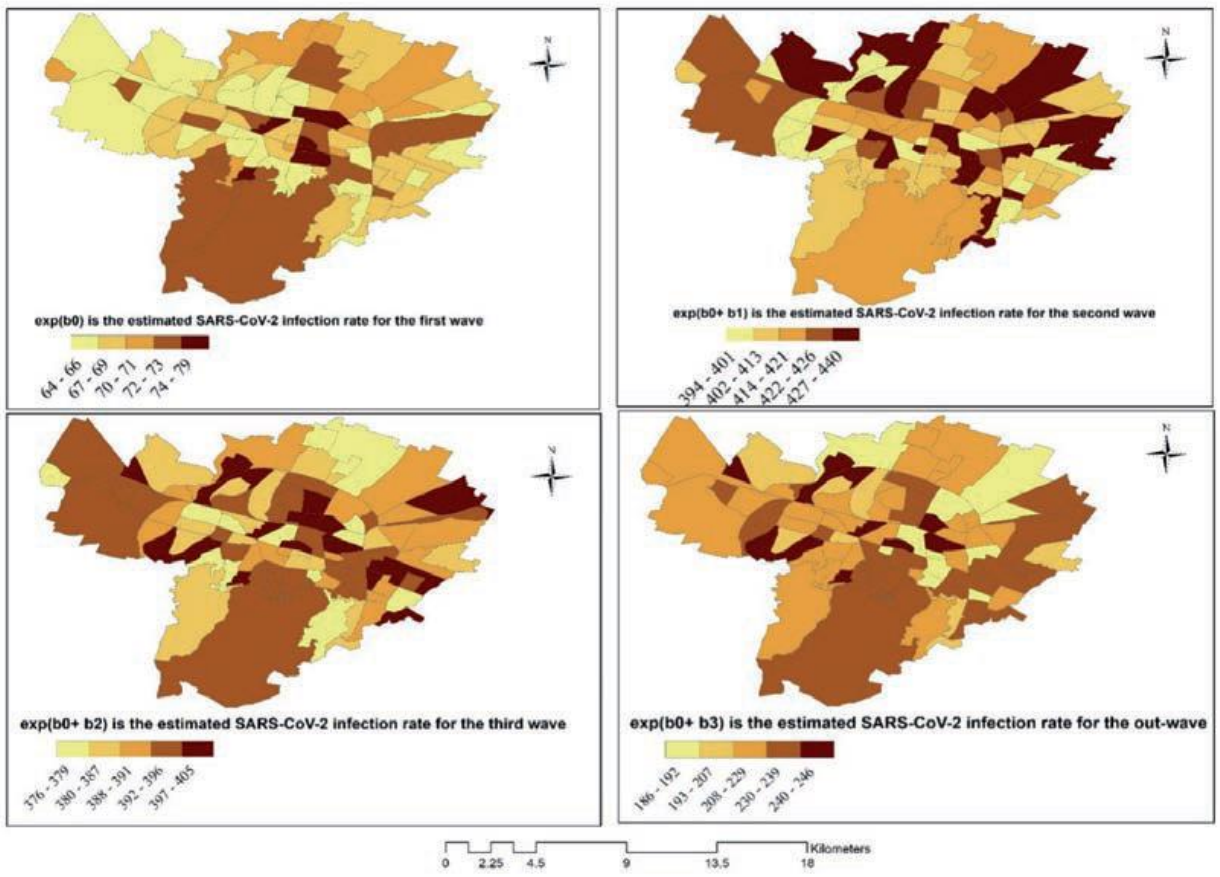


Figure 7. Estimated SARS-CoV-2 infection rates (per 1,000 inhabitants) for each statistical area and Wave. Estimated infection rates were calculated as exp(beta_i,0+beta_i,j), for j=2ndWave, 3rdWave, and Out Wave.

Local estimate significance measure - the local t-value map plots are shown in Fig S1 in Supplementary Materials. A map showing a local parameter estimate significance measure - the spatial heterogeneity of local t-values estimated using the GWPR at 95% and 99% significance levels. For each of the 90 statistical area, local t-values greater than +1.96 and +2.58 indicate a positive and significant relationship between the predictors and COVID-19 infection rate; local t-values estimated between -1.96 and +1.96 indicate there is no significant relationship between the predictors and COVID-19 infection rate; and local t-values less than − 1.96 and − 2.58 indicate that there is a negative and statistically significant relationship between the predictors and COVID-19 infection rate in each area. As shown in fig S1. in Supplementary Materials, all the Wave parameters are greater than +1.96 and +2.58, indicating that the t-value for this relationship is positive and statistically significant. However, we did not find a local t-value

between -1.96 and +1.96, and the intercept parameter estimated value was lower than -1.96 and -2.58. This indicates that the predictors and COVID-19 infection rate have a negative and statistically significant relationship. According to the local R2 value in Fig S1 in Supplementary Materials top (right), the GWPR model fits better in the darker area (R2>0.68).

Comparison of GWPR and GPR models

The performance of both GPR and local GWPR models is shown in Table 5. The GWPR model showed a smaller AICc, and larger percent deviance (it is equivalent to R2, which explains how predictors affect outcome variables), so this was chosen as the best model. In our result, the best bandwidth size was 4689 with a minimum AICc of 6076. The results demonstrate that GWPR predicts COVID-19 infection rates in particular areas more accurately than GPR. In GWPR, the variability of COVID-19 infection rates across the statistical areas was better captured by the spatial heterogeneity in the data (Table 5).

Table 5. Comparison of model performance in the estimation of the effects of the explanatory variables with respect to the COVID-19 infection rate.

Model	AICc	Percent deviance explained
GPR model	6144	0.611
GWPR model	6076	0.621
Difference	68	0.010

Best bandwidth size: 4689 with minimum AICc : 6076.

4.4. Discussion

This is the first population-based study to look at how well GPR and local GWPR models can explain the variations of COVID-19 infection rates in a large city based on sociodemographic factors, health risk and different time periods-waves as explanatory variables. Different city areas were impacted differently during different epidemic waves. The area-to-area influence was estimated to be confined within a 4.7 km radius and these spatiotemporal heterogeneity patterns were independent of the sociodemographic and clinical characteristics of the resident population. Presence of hypertension, diabetes and any co-morbidity were found to be significant, individual risk factors.

We used GIS and geographical data to explore the potential risk factors of COVID-19 infection rates in relation to sociodemographic factors, health risk factors, and epidemic waves. We compared the local GWPR model with a traditional GLM Poisson regression model to find

the best fit for exploring the association between the available predictors and COVID-19 infection rates. The results showed that calibration of the GWPR model led to an improved fit compared to GLM Poisson regression.

The heterogeneous nature of COVID-19 transmission processes over space has been observed previously (Chen et al., 2021; Wang et al., 2020). Several factors contribute to COVID-19 transmission, including local demographic inequalities, health risk factors, different time periods or epidemic waves, and socioeconomic characteristics (Gatto et al., 2020; Qiu et al., 2019; Pedro et al., 2020). Recent studies have found that the geo-epidemiological distribution of epidemic waves varies between countries, even between areas of limited size (Gaudart et al., 2021). Geo-epidemiological analyses based on available data can therefore contribute greatly to the development of effective health policies (Hou et al., 2021; Srivastava et al., 2020) by focusing on the most impacted areas and developing tailored interventions.

Individuals, particularly the elderly, and if they have underlying co-morbidities, such as cardiovascular disease, diabetes and hypertension, are at higher risk for complications from COVID-19 (Klekotka et al., 2015). In this research, key findings, based on GPR, were that sociodemographic, health risk factors, and epidemic waves were found to impact on COVID-19 infection rate, and based on the local model, COVID-19 infection rate varied geographically within the study area during epidemic Waves. More specifically, in the global model, the average COVID-19 infection rate decreased in age group (21 - 65) compared to age group (65+), which is consistent with most previous findings (Dowd et al., 2020; Shawky et al., 2021; Sun et al., 2020); whereas the average COVID-19 infection rate increased in hypertension – yes compared to hypertension – no, this is in line with other study (Landstra et al., 2021), however, the opposite is observed in other studies (Fresán et al., 2021; Jayaswal et al., 2021); the average COVID-19 infection rate increased in diabetes – yes compared to diabetes - no, this is in line with most previous studies (Landstra et al., 2021; Jayaswal et al., 2021), also one review article (Nassar et al., 2021) confirmed that diabetes mellitus is associated with a significant risk of complications, extended hospital stays, and mortality in COVID-19 infected patients; the average COVID-19 infection rate increased in any comorbidities - yes compared to any comorbidities - no, this is in line with most previous studies (Landstra et al., 2021); and the average COVID-19

infection rate increased significantly during the Second and Third Wave compared to the First Wave, also moderately increased during Out Waves compared to the First Wave, this is in line with other study (Gaudart et al., 2020), in the Second and Third waves, it spiked significantly and showed spatial variability.

The period with the highest SARS-CoV-2 infection rate was the Second Wave (07 November 2020- 01 February 2021, 15696 cases). The rate was similar but lower during the 3rdWave (02 February 2021 – 28 April 2021, 14238 cases) and lowest during the First Wave 26 March 2020 - 13 May 2020 ,1509 cases). The low rate during the First Wave is probably due to insufficient diagnostic capabilities at the beginning of the pandemic. Mostly, during the Second Wave, was the hardest hit and spiked significantly. This was due in part to the relaxation of the severe lockdown measures (Magnavita et al., 2020). During these two waves (the second and the third), Italian cases reached their highest peak. However, after the Omicron variant, daily cases also spiked, even if the death rates were very low. The number of patients having health risk factors or proportion of the health risk factors (hypertension - yes, diabetes - yes, any comorbidities - yes) was very small compared to no categories like 9.5 %, 4 %, and 19.2 % respectively. As a result, we did not observe spatial variabilities in the area, whereas sex-male, age group [0, 20], and all three family size categories did not significantly affect the COVID-19 infection rate.

In contrast to global modelling, where the relationship between variables is constant across the study area, the results from GWPR overcame the disadvantages and drawbacks of global modelling. From the GWPR result, we only observed spatial variability in four parameters (the intercepts and the three Waves). For those parameters from spatial heterogeneity maps, COVID-19 infection rates were not equal in all locations, and it is identified that the parameters have obvious patterns of geographic variation. All the parameter signs of Waves were positive in all statistical areas indicating that all have positive impacts on the COVID-19 infection rates in the area, whereas the baseline or intercept signs were negative in all areas, implying that besides these three predictors, there were still other factors bound up with the COVID-19 infection rate. Showing epidemic trends in advance may provide information about geographic

risks, social determinants, and health factors influencing COVID-19 transmission and how to respond to it.

Methodologically (Brunsdon et al., 1996; Fotheringham et al., 2003; Nakaya et al., 2005), developed GWPR into a convenient yet powerful technique for exploring spatial non-stationarity and providing mappable statistics to visualize the spatial patterns of relationships between dependent and independent variables. Recent studies have suggested that GWPR is one of the geostatistical methods that should be promoted in health studies, given the locality of health outcomes (Young et al., 2010). Furthermore, GWPR can potentially contribute significantly to health research, allowing researchers to understand etiology and spatial processes better, providing specific results beyond global models to facilitate place-specific health policy formation, and allowing scholars to investigate questions that cannot be answered with traditional (global) analytical models. Finally, as we mentioned above, GWPR is useful in a wide range of disciplines where spatial data are used and in applications where spatial non-stationarity is suspected (and should be confirmed).

Spatial modelling of disease epidemics can be useful for assessing where and why disease hotspots and clusters exist and providing explicit information about the spatial variation in disease incidence and transmission. However we cannot clearly identify outbreaks and clusters. Possible reasons: i) by visual inspection outbreaks/clusters are not evident because of the coarse spatiotemporal resolution (90 areas x 4 waves); ii) the range of human mobility (and, in turn, virus diffusion) is comparable to the size of the city of Bologna; iii) other analytical methods are needed for cluster identification. Similarly, little is known about how sociodemographic, epidemic, and risk factors affected COVID-19 infection rates within the Bologna area. As such, this study aimed to provide insights into the relationship between predictors and disease infection rates at the local level. Before our study , more detailed information about individual subjects(including pre-existing medical conditions, and detailed demographic information) with COVID-19 were not accessible in Bologna, thus no spatial or geographical analysis was undertaken so far. As a result, this research could be the first and most valuable.

As a research paper, we need to be aware of several limitations - the study involved a single center at a city level, so the results cannot be generalized across all areas in Italy due to heterogeneity in practice, building infrastructure, population, and many other factors. Furthermore, obtaining all covariates other than COVID-19 case number was impossible. Furthermore, we could not obtain socio-economic information, detailed clinical characteristics, or detailed information about medical conditions and laboratory tests for the patients in our study area in Bologna, which is likely to be one of the most influential factors affecting the rate of COVID-19 infection. Despite these limitations, our study can help in two ways: i) the results provide a clear understanding of the relationships between predictor characteristics and disease infection rates, both globally and at the local level in Bologna; ii) studies at the local level can also provide insight into national policies aimed at improving the health and welfare of citizens. Also the ideas presented here may provide a basis for evaluating public health policies using spatial model results.

4.5. Conclusions

Based on our results, the local GWPR model fits better than the traditional GLM – global. Different city areas were hit differently by SARS-CoV-2 infection during different epidemic waves. The area-to-area influence was estimated to be confined within a 4.7 km radius. These spatiotemporal heterogeneity patterns were independent of the sociodemographic and clinical characteristics of the resident population. Significant single-individual risk factors for detected SARS-CoV-2 infection were hypertension, diabetes, and comorbidities. On the basis of the global model results, sociodemographic factors, health risk factors, or pre-existing conditions, epidemic waves played a role in the COVID-19 infection rate variation, and the local model results confirmed that the COVID-19 infection rate did vary geographically during epidemic waves. When analyzing the effectiveness of preventative programs and other health services, it is important to keep this spatial variability in mind. Moreover, we believe that this study can be used as a guide for other countries to understand the local spread of COVID-19, and the information we have obtained and the method we used could serve as a tool or as a reference in the event of similar epidemics in the future.

References:

1. Ali K, Partridge MD, Olfert MR, 2007. Can geographically weighted regressions improve regional analysis and policy making? Int Reg Sci Rev 30:300-31.
2. Armocida B, Formenti B, Ussai S, Palestra F, Missoni E, 2020. The Italian health system and the COVID-19 challenge. Lancet Public Health 5:e253.
3. Benson T, Chamberlin J, Rhinehart I, 2005. An investigation of the spatial determinants of the local prevalence of poverty in rural Malawi. International Food Policy Research Institute. Washington DC, USA. Available from: http://gisweb.ciat. cgiar.org/povertymapping/download/Malawi.pdf
4. Braun JW, Rousson V, 2000. An autocorrelation criterion for bandwidth selection in nonparametric regression. J Statist Computat Simulat 68:89-101.
5. Brunsdon C, Fotheringham AS, Charlton ME 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. Geogr Anal 28:281–98.
6. Bui LV, Mor Z, Chemtob D, Ha ST, Levine H, 2018. Use of Geographically Weighted Poisson Regression to examine the effect of distance on Tuberculosis incidence: A case study in Nam Dinh, Vietnam. PLoS One 13:e0207068.
7. CDC COVID-19 Response Team, 2020. Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) – United States, February 12-March 16, 2020. MMWR Morb Mortal Wkly Rep 69:343–6.
8. CDC, 2020. Coronavirus COVID-19: symptoms of coronavirus. Centers for Disease Control and Prevention. Accessed April 18, 2020. Available from: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html
9. Chen J, Liu L, Zhou S, Xiao L, Jiang C, 2017. Spatial variation relationship between floating population and residential burglary: a case study from ZG, China. ISPRS Int J Geo-Inf 6:246.
10. Chen M, Chen Y, Wilson JP, Tan H, Chu T, 2022. Using an eigenvector spatial filtering-based spatially varying coefficient model to analyze the spatial heterogeneity of COVID-19 and its influencing factors in mainland China. ISPRS Int J Geo-Inf 11:67.

11. Chen VY-J, Wu P-C, Yang T-C, Su H-J, 2010. Examining non-stationary effects of social determinants on cardiovascular mortality after cold surges in Taiwan. Sci Total Environ 408:2042– 9.
12. Chen YY, Assefa Y, 2021. The heterogeneity of the COVID-19 pandemic and national responses: an explanatory mixed-methods study. BMC Public Health 21:835.
13. Choi M, Lee M, Lee MJ, Jung D, 2017. Physical activity, quality of life and successful aging among community-dwelling older adults. Int Nurs Rev 64:396–404.
14. Choi WS, Kang CI, Kim Y, Choi JP, Joh JS, Shin HS, Kim G, Peck KR, Chung DR, Kim HO, Song SH, Kim YR, Sohn KM, Jung Y, Bang JH, Kim NJ, Lee KS, Jeong HW, Rhee JY, Kim ES, Woo H, Oh WS, Huh K, Lee YH, Song JY, Lee J, Lee CS, Kim BN, Choi YH, Jeong SJ, Lee JS, Yoon JH, Wi YM, Joung MK, Park SY, Lee SH, Jung SI, Kim SW, Lee JH, Lee H, Ki HK, Kim YS; Korean Society of Infectious Diseases, 2016. Clinical presentation and outcomes of Middle East respiratory syndrome in the Republic of Korea. Infect Chemother 48:118–26.
15. Cox D, 1972. Regression models and life tables (with discussion). J Roy Stat Soc 34:187–202.
16. Davies TM, Lawson AB, 2019. An evaluation of likelihood-based bandwidth selectors for spatial and spatiotemporal kernel estimates. J Statist Computat Simulat 89:1131-52.
17. Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, Liu Y, Mills MC, 2020. Demographic science aids in understanding the spread and fatality rates of COVID-19. Proc Natl Acad Sci U S A 117:9696-9698.
18. Fotheringham AS, Brunsdon C, Charlton M. 2003. Geographically weighted regression: The analysis of spatially varying relationships John Wiley & Sons.
19. Fotheringham AS, Kelly MH, Charlton M, 2013. The demographic impacts of the Irish famine: Towards a greater geographical understanding. Transact Inst Bri Geograph 38:221-237.
20. Fresán U, Guevara M, Trobajo-Sanmartín C, Burgui C, Ezpeleta C, Castilla J, 2021. Hypertension and related comorbidities as potential risk factors for COVID-19 hospitalization and severity: a prospective population-based cohort study. J Clin Med 10:1194.

21. Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, Rinaldo A, 2020. Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures. Proc Natl Acad Sci U S A 117:10484-10491.
22. Gaudart J, Landier J, Huiart L, Legendre E, Lehot L, Bendiane MK, Chiche L, Petitjean A, Mosnier E, Kirakoya- Samadoulougou F, Demongeot J, Piarroux R, Rebaudet S, 2021. Factors associated with the spatial heterogeneity of the first wave of COVID19 in France: a nationwide geo-epidemiological study. Lancet Public Health 6:e222-e231.
23. Goovaerts, P. 2005. Analysis and detection of health disparities using geostatistics and a space-time information system: The case of prostate cancer mortality in the United States, 1970- 1994. Proceedings of GIS Planet 2005. (May 30-June 2, 2005, Estoril, Portugal).
24. Hadayeghi A, Amer S. Shalaby, Bhagwant N. Persaud. 2009. Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. Accid Anal Prev 42:676-88.
25. Haque U, Scott LM, Hashizume M, Fisher E, Haque R, Yamamoto T, Glass GE, 2012. Modelling malaria treatment practices in Bangladesh using spatial statistics. Malar J 201211:63.
26. Hou X, Gao S, Li Q, Kang Y, Chen N, Chen K, Rao J, Ellenberg JS, Patz JA, 2021. Intra county modelling of COVID-19 infection with human mobility: assessing spatial heterogeneity with business traffic, age, and race. Proc Natl Acad Sci U S A 118:e2020524118.
27. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B, 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395:497-506.
28. Jayaswal SK, Singh S, Malik PS, Venigalla SK, Gupta P, Samaga SN, Hota RN, Bhatia SS, Gupta I, 2021. Detrimental effect of diabetes and hypertension on the severity and mortality of COVID-19 infection: A multi-center case-control study from India. Diabetes Metab Syndr 15:102248.

29. Klekotka RB, Mizgała E, Król W. 2015. The ethology of lower respiratory tract infections in people with diabetes. Pneumonol Alergol Pol 2015;83:401-8.
30. Landstra CP, de Koning EJP, 2021. COVID-19 and Diabetes: Understanding the Interrelationship and Risks for a Severe Course. Front Endocrinol (Lausanne) 12:649525.
31. Li Z, Wang W, Liu P, Bigham JM, Ragland DR, 2013. Using Geographically Weighted Poisson Regression for county-level crash modelling in California, Safety Sci 58:89-97.
32. Loubert L, 3.27 - Using GIS to Understand Schools and Neighborhoods. Comprehensive Geographic Information Systems. Elsevier, 2018, pp. 422-439.
33. Magnavita N, Sacco A, Chirico F, 2020. Covid-19 pandemic in Italy: pros and cons. Zdrowie Publicznei Zarządzanie 18:32-35.
34. Matthews SA, Yang TC. Mapping the results of local statistics: Using geographically weighted regression. Demogr Res 2012 26:151-166.
35. Mennis JL, Jordan LM, 2005. The distribution of environmental equity: exploring spatial non-stationarity in multivariate models of air toxic releases. Ann Ass Am Geographers doi:10.1111/j.1467-8306.2005.00459.x.
36. Mollalo A, Alimohammadi A, Shirzadi MR, Malek MR, 2015. Geographic information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, north-east of Iran. Zoonoses Public Health 62:18-28.
37. Mollalo A, Khodabandehloo E 2016. Zoonotic cutaneous leishmaniasis in northeastern Iran: a GIS-based spatio-temporal multicriteria decision-making approach. Epidemiol Infect 44:2217-29.
38. Nakaya T, Fotheringham AS, Brunsdon C, Charlton M, 2005. Geographically weighted Poisson regression for disease association mapping. Stat Med 24:2695-717.
39. Nakaya T. 2014. GWR4 User Manual: Windows application for geographically weighted regression modelling [Internet]. Available from: https://gwr.maynoothuniversity.ie/wp-content/uploads/2013/04/GWR4_Manual.pdf.

40. Nassar M, Daoud A, Nso N, Medina L, Ghernautan V, Bhangoo H, Nyein A, Mohamed M, Alqassieh A, Soliman K, Alfishawy M, Sachmechi I, Misra A, 2021. Diabetes Mellitus and COVID- 19: Review Article. Diabetes Metab Syndr 15:102268.
41. Oshan TM, Li Z, Kang W, Wolf LJ, Fotheringham AS, 2019. mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale. ISPRS Int J Geo-Inf 8:269.
42. Pedro SA, Ndjomatchoua FT, Jentsch P, Tchuenche JM, Anand M and Bauch CT. 2020 Conditions for a Second Wave of COVID-19 Due to Interactions Between Disease Dynamics and Social Processes. Front Phys doi: 10.3389/fphy.2020. 574514.
43. Poliart A, Kirakoya-Samadoulougou F, Ouédraogo M, Collart P, Dubourg D, Samadoulougou S, 2021. Using geographically weighted Poisson regression to examine the association between socio-economic factors and hysterectomy incidence in Wallonia, Belgium. BMC Womens Health 21:373.
44. Qiu Y, Chen X, Shi W, 2020. Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID- 19) in China. J Popul Econ 33:1127-1172.
45. Shawky M., Abdullah K, Alkhattab S, Adham S, Peter A. 2021. Sociodemographic determinants of COVID-19 incidence rates in Oman: Geospatial modelling using multiscale geographically weighted regression (MGWR). Sustain Cities Soc 65:102627.
46. Slocum, TA .1999. Thematic Cartography and Visualization. Prentice-Hall.
47. Specchia ML, Di Pilla A, Sapienza M, Riccardi MT, Cicchetti A, Damiani G, Group IR, 2021. Dealing with COVID-19 Epidemic in Italy: Responses from Regional Organizational Models during the First Phase of the Epidemic. Int J Environ Res Public Health 18:5008.
48. Srivastava A, Chowell G. 2020. Understanding Spatial Heterogeneity of COVID-19 Pandemic Using Shape Analysis of Growth Rate Curves. medRxiv [Preprint]. doi: 10.1101/2020.05.25.20112433.
49. Sun P, Lu X, Xu C, Sun W, Pan B. 2020. Understanding of COVID-19 based on current evidence. J Med Virol 2020;92:548-551.
50. Takele K, Zewotir T, Ndanguza D, 2019. Understanding correlates of child stunting in Ethiopia using generalized linear mixed models. BMC Public Health 19:626.

51. Ufficio Statistica Regionale Regione Emilia Romagna. 10 April 2019. Available from: https://sociale.regione.emiliaromagna.it/notizie/2019/aprile/la-fotografia deicittadinistranieri-a-bologna-150-le-nazionalita (in Italian).
52. Umpierrez GE, Isaacs SD, Bazargan N, You X, Thaler LM, Kitabchi AE.2002. Hyperglycemia: an independent marker of in-hospital mortality in patients with undiagnosed diabetes. J Clin Endocrinol Metab 87:978-82.
53. Wang Y and Teunis P. 2020. Strongly heterogeneous transmission of COVID-19 in mainland China: Local and regional variation. Front Med 7:329.
54. Wang W, Li D, 2017. Structure identification and variable selection in geographically weighted regression models. J Statist Computat Simulat 87:2050-2068.
55. WHO, 2020a. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). Available from: https://www.who.int/docs/default-source/coronaviruse/whochina- joint-mission-on-covid-19-final-report.pdf.
56. WHO, 2020b. Coronavirus disease (COVID-19) outbreak. Available from: https://www.who.int/westernpacific/healthtopics/ coronavirus Worldometers, "COVID-19 CORONAVIRUS PANDEMIC," 2021. Accessed: 07-April-2022. Available: https://www.worldometers.info/coronavirus/
57. Yang T-C, Teng HW, Haran M. 2009. The impacts of social capital on infant mortality in the U.S.: A spatial investigation. Applied Spatial Analysis and Policy 2:211-227.
58. Yang TC, Matthews SA, 2012. Understanding the non-stationary associations between distrust of the health care system, health conditions, and self-rated health in the elderly: a geographically weighted regression approach. Health Place 18:576-85.
59. Young, L.J., Gotway, C.A., 2010. Using geostatistical methods in the analysis of public health data: the final frontier. geoENV VII-Geostatistics for Environmental Applications 16, 89–98.
60. Zeleke AJ, Moscato S, Miglio R, Chiari L, 2022. Length of stay analysis of COVID-19 hospitalizations using a count regression model and quantile regression: a study in Bologna, Italy. Int J Environ Res Public Health 19:2224.

61. Zhang H, Liu Y, Chen F, Mi B, Zeng L, Pei L, 2021.The effect of sociodemographic factors on COVID-19 incidence of 342 cities in China: a geographically weighted regression model analysis. BMC Infect Dis 21:428.
62. Zhao F, Park N, 2004. Using geographically weighted regression models to estimate annual average daily traffic. In: Transportation Research Record 1879. Washington, DC: Transportation Research Board, National Research Council: 99-107.
63. Zheng X, Qin G, Tu D, 2017. A generalized partially linear mean-covariance regression model for longitudinal proportional data, with applications to the analysis of quality of life data from cancer clinical trials. Stat Med 36:1884-1894.
64. Zhou YB, Wang QX, Liang S, Gong YH, Yang MX, Chen Y, Nie SJ, Nan L, Yang AH, Liao Q, Yang Y, Song XX, Jiang QW, 2015. Geographical variations in risk factors associated with HIV infection among drug users in a prefecture in Southwest China. Infect Dis Poverty 4:38.

Appendix - Supplementary materials

Table S1. Test of spatial stationarity - Monto Carlo Test for Spatial Variability (p-value), and DiFF of Criterion., Figure S1: Local estimate significance measure - Local t-value map plots.

Table S1. Test spatial stationarity - Monto Carlo Test for Spatial Variability(p-valve), and DIFF of Criterion.

Variable	p -valve	DIFF of Criterion
Intercept	0.038	-26.595
Sex_ m	0.740	5.621
Age_ group [0 – 21)	0.063	4.667
Age_group2 [21 - 65)	0.059	0.413
FS_1	0.643	5.720
FS_2	0.551	5.213
FS_3	0.076	5.961
HYP_ yes	0.110	5.371
DIAB_ yes	0.439	2.493
AC_ yes	0.741	3.612
Second Wave	0.001	-24.472
Third Wave	0.001	-9.963
Out Waves	0.001	-13.829

The italicized findings showed that the predictors vary spatially. FS, family size ; HYP, hypertension ; DIAB, diabetics; AC, Any comorbidities.

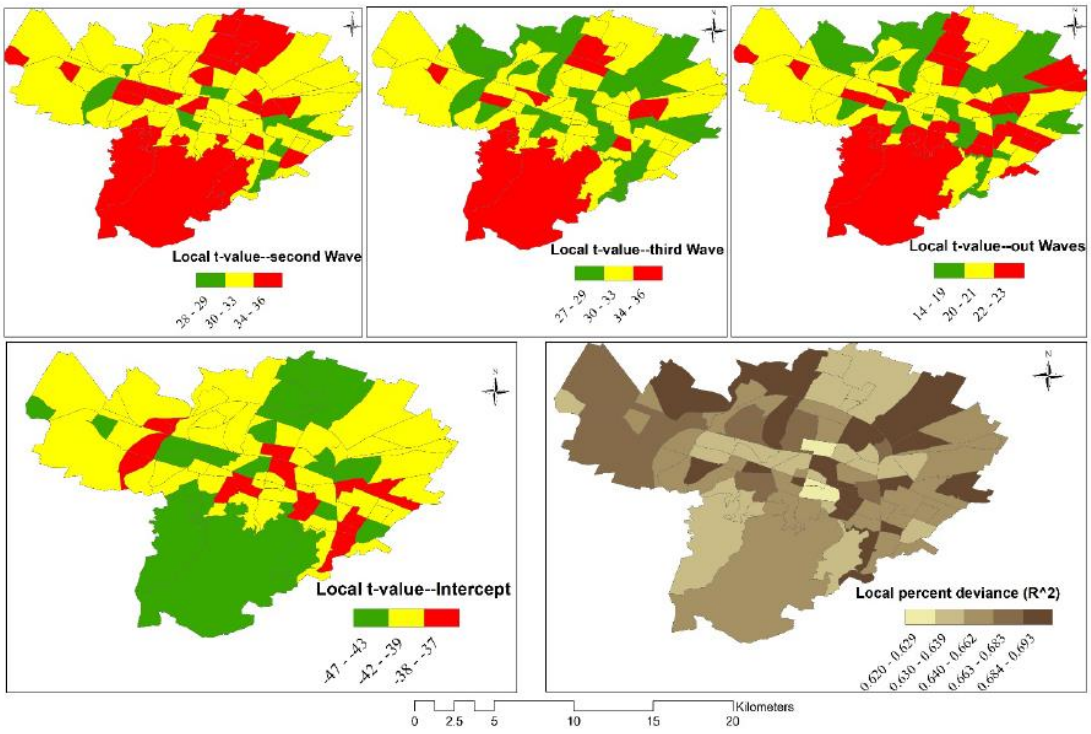


Figure S1. Local estimate significancy measure - Local t-value (Secondwave (top left) , Thirdwave (top middle Out Waves(top right), Intercept (bottom left)), and the spatial distribution of local percent deviance(R^2) of GWPR (bottom right).

CHAPTER FIVE

Length of Stay Analysis of COVID-19 Hospitalizations Using a Count Regression Model and Quantile Regression: A Study in Bologna, Italy

---

This chapter is largely taken from our published work [∗], where we analyzed the LoS using a count regression model and quantile regression when COVID-19 levels were high.

∗*Zeleke AJ, Moscato S, Miglio R, Chiari L. Length of Stay Analysis of COVID-19 Hospitalizations Using a Count Regression Model and Quantile Regression: A Study in Bologna, Italy. Int J Environ Res Public Health. 2022 Feb 16;19(4):2224. doi: 10.3390/ijerph19042224. PMID: 35206411; PMCID: PMC8871974. https://pubmed.ncbi.nlm.nih.gov/35206411/*

Abstract

This work aimed to identify and explore the hospital admission risk factors associated with the length of stay (LoS) by applying a relatively novel statistical method for count data using predictors among COVID-19 patients in Bologna, Italy. The second goal of this study was to model the LoS of COVID patients to understand which covariates significantly influenced it and identify the potential risk factors associated with LoS in Bolognese hospitals from 1 February 2020 to 10 May 2021. The clinical settings we focused on were the Intensive Care Unit (ICU) and ordinary hospitalization, including low-intensity stays. We used Poisson, negative binomial (NB), Hurdle–Poisson, and Hurdle–NB regression models to model the LoS. The fitted models were compared using the Akaike information criterion (AIC), Vuong's test criteria, and Rootograms. We also used quantile regression to model the effects of covariates on the quantile values of the response variable (LoS) using a Poisson distribution, and to explore a range of conditional quantile functions, thereby exposing various forms of conditional heterogeneity and controlling for unobserved individual characteristics. Based on the chosen performance criteria, Hurdle–NB provided the best fit. As an output from the model, we found significant changes in average LoS for each predictor. Compared with ordinary hospitalization and low-intensity stays, the ICU setting increased the average LoS by 1.84-fold. Being hospitalized in long-term hospitals was another contributing factor for LoS, increasing the average LoS by 1.58 compared with regular hospitals. When compared with the age group [50, 60) chosen as the reference, the average LoS

decreased in the age groups [0, 10), [30, 40), and [40, 50), and increased in the oldest age group [80, 102). Compared with the second wave, which was chosen as the reference, the third wave did not significantly affect the average LoS, whereas it increased by 1.11-fold during the first wave and decreased by 0.77-fold during out-wave periods. The results of the quantile regression showed that covariates related to the ICU setting, hospitals with longer hospitalization, the first wave, and the out-waves were statistically significant for all the modeled quantiles. The results obtained from our study can help us to focus on the risk factors that lead to an increased LoS among COVID-19 patients and benchmark different models that can be adopted for these analyses.

Keywords: count data model; length of stay; COVID-19; generalized linear model; Hurdle model; Vuong test; AIC; Rootograms; quantile regression

5.1. Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by the novel coronavirus SARS-CoV-2 virus [1], characterized as a pandemic in March 2020 by the World Health Organization [2]. Since the first case detected in Wuhan, dated December 2019, COVID-19 has caused 275 million cases and over 5 million deaths up to December 2021 [3]. Health systems have been significantly challenged by the spread of COVID-19, especially Italy, which was the first European country to be affected by COVID-19 [4] and which faced an overwhelming increase in COVID-19 patients who needed hospitalization at the beginning of the pandemic, risking the collapse of the system [5]. The COVID-19 pandemic highlighted the need for a management strategy to predict the demand for hospital services so that healthcare facilities can efficiently make resources available when needed.

The length of stay (LoS) is a key indicator of how efficiently hospitals are being managed and is used to assess the efficiency of hospital management and patient quality of care, and for functional evaluations. A shorter stay means that more beds are available for more patients and reduces hospital resource consumption; thus, it corresponds to a decrease in health-related

expenditure [6]. Reducing LoS has been linked to lower risks of opportunistic infections and medication side effects [7], and improved treatment outcomes and lower mortality rates. In addition, a shorter hospital stay reduces the burden of medical fees and increases bed turnover, and thus increases hospitals' profit margins [7] while lowering the overall social costs [8–10]. Researchers must determine which characteristics are associated with longer or shorter hospital stays in patients [11].

A recent systematic review and meta-analysis [12] research compared the LoS of COVID-19 patients in China versus the rest of the world (Europe, the U.S., and the UK) using 52 studies. By combining information from different studies, a substantial difference in LoS between China (longer LoS) and other locations (shorter LoS) was found. There is also little evidence concerning the impact of the study period, age, and disease severity on LoS. Another retrospective study in Vietnam found that age was significantly associated with a long hospital stay in COVID-19 patients [13]; other studies in China found that the average age of patients with a prolonged LoS was higher than that of patients with an average LoS [14]. A better understanding of the COVID-19 pandemic's spillover effects between waves might help healthcare professionals better support at-risk persons during the pandemic. Study results have indicated that there is a minor difference in overall LoS before and after the years of the COVID-19 pandemic according to a territory-wide retrospective cohort study conducted in Hong Kong on 28-day in-hospital mortality and LoS. In 2020, patients triaged as critical in the emergency department (ED) had a 2.71-day shorter LoS than those in 2019. LoS declined significantly for patients who died in 2020, when their hospital stay was 2.31 days shorter on average than in 2019 [15].

Through univariate and multivariate logistic regressions models, a retrospective analysis of patients hospitalized with COVID-19 from a single hospital in Hefei, China, was carried out to explore the risk variables associated with a prolonged hospital LoS. The findings revealed that patients with a prolonged LoS had a higher average age than those with a median LoS [13]. A time-to-event analysis implementing a Cox proportional hazard model was applied to identify the factors associated with the LoS. Factors linked to a longer hospital stay should be considered when arranging bed strength on a contingency basis, and the estimated LoS of patients was required in order to model bed occupancy and make contingency plans. Artificial Intelligence

(AI) researchers developed algorithms in order to model bed occupancy and make contingency plans [16]. Another study in South Korea used multivariate negative binomial and gamma regression models to determine the influencing factors of LoS [17]. The results showed that the average LoS in hospitals for COVID-19 patients was 5.5 days, and the average LoS also tended to be higher for older patients, except for the group aged 65 years or older, who had a shorter hospital stay than the others. A study in the UK used an Accelerated Failure Time (AFT) survival model, a truncation corrected (TC) method, both with underlying Weibull distributions, and a multi-state (MS) survival model to analyze LoS data [18]. The results showed that all methods produced similar estimates of LoS for the overall hospital stay, considering ordinary hospitalization only: 8.4, 9.1, and 8.0 days for AFT, TC, and MS, respectively. This study also tried to mention the complexities and partiality of many data sources, as well as the continuously developing nature of the COVID-19 pandemic that suggests approaches should be based on a variety of analytical methodologies across multiple datasets. From a management point of view, research in Dubai used AI-based modeling by using Decision Tree prediction models to forecast the LoS and risk of mortality accurately. In principle, these smart models might equip front-line clinicians with the tools they need to improve management techniques and save lives [19].

Researchers have been focusing on modeling count data due to observed problems with the commonly used analytic methods. For example, ordinary least squares (OLS) regression is often inappropriate because the count data violate the underlying assumptions of OLS regression (i.e., the normality of the residuals), resulting inaccurate coefficient estimates and standard errors, and misleading p-values and consequent confidence intervals [20,21].

Only recently, several studies have examined the effective management of LoS, and several statistical models have been proposed to analyze the data with count-type outcomes (such as LoS in the hospital, number of antenatal care visits, number of purchases made, and many more) with different settings and predictors [22,23]. There are, however, different characteristics in different count data; therefore, certain count data models cannot be used with them. Often, the first model used to analyze count outcomes is the Poisson model [10,20] but, in practice, these two assumptions (i.e., the variance being equal to mean, and independent events) [10,20] are typically violated. Overdispersion is generally always the result of any

assumptions being violated [20,24,25], and alternate strategies for modeling count data, such as negative binomial regression [20,24,25], should be considered. Aside from overdispersion, many empirical counts contain a large number of zero observations, which can be modeled using Hurdle models [26]. A Hurdle model combines a binary count outcome (LoS = 0 vs. LoS > 0) with a truncated model for results above the hurdle (LoS > 0). Excess zeros are modeled using the Hurdle model. In some cases, these zeros cannot be ignored, since they are critical and meaningful. To assess the fitted count regression model, researchers have used a graphical approach called Rootograms [27,28].

Another statistical method used to characterize the effects of covariates on a given phenomenon is to study the quantiles' distribution instead of modeling the mean. Quantile regression models estimate the relationship between the qth quantile of the response y and the covariate x [29]. Quantile regression models at multiple percentiles can help compare the changes associated with covariates across the distribution of the phenomenon under study [30]. Machado et al. [31] extended quantile regression models for count data. This approach has already been applied in a study [32] to evaluate the LoS of diabetes patients.

This study aimed to identify and explore the hospital admission risk factors associated with LoS in COVID-19 patients in Bologna, Italy, by applying a relatively novel statistical method (i.e., count regression and quantile regression models) using the currently available predictors. The second goal of this study was to model the LoS of COVID patients in Bologna hospitals from 1 February to 10 May 2021 in order to determine which covariates had a significant impact on it and to discover the potential risk factors linked with LoS.

5.2. Materials and Methods

5.2.1. Population

This study obtained data with permission from the local health authority (AUSL) of Bologna. Data were retrieved from 1 February 2020 to 10 May 2021. Our dataset referred to COVID-19 patients both in the ICU (intensive plus sub-intensive) and ordinary settings in 7 Bolognese hospitals. We also categorized the hospital stays as "regular" for 5 hospitals, and "long-term"

for 2 hospitals. This distinction is important because hospitals providing extended hospitalization have mainly rehabilitation purposes.

5.2.2. Database Preprocessing

Individual patients were considered on the basis of their unique ID, and we merged the repeated IDs (patients who went to the hospital more than one time) via the following procedure:

- Detect repeated ID;
- Order rows referring to the same ID chronologically;
- Check if two consecutive rows differ for >3 days
- If they differed for > 3 days, the two rows were treated as two different stays. The second stay was removed from further analysis;
- If they did not differ for >3 days, merge multiple rows by using:
- Hospital: prevalent hospital
- Setting: add a variable with two levels: “low-intensity” if the patient was only hospitalized in the low-intensity setting and “ICU” if the patient was in the intensive care unit at least one time.

We categorized the hospital stay information as regular hospitalization and long-term

hospitalization based on the prevalent hospital.

5.2.3. Outcome and Covariates

The study’s main outcome was the LoS (a non-negative integer), defined as the number of days between inpatient admission and hospital exit or discharge, and it was the target outcome variable for which this study aimed to identify a proper count regression model. The explanatory variables were:

Clinical setting: ICU setting (intensive care + sub-intensive) of COVID-19 patients, and ordinary and low-intensity COVID-19 stays.

Age: An integer representing patient age in years, grouped in 10-year categories: [0, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80), and [80, 102).

Waves: We defined waves based on an empirical approach.

We assumed that a wave started with more than 500 total hospitalizations and ended when the hospitalizations dropped below 500 (see Figure 1). Using this cut-off, we chose a day in the stationary phase between the second and the third wave located in the middle of this period. We chose a middle point to divide the second from the third wave, since hospitalization in that period never dropped to less than 500. Using this criterion, these were the dates and durations of the waves:

First wave →26 March 2020–13 May 2020 (48 days)

Second wave →7 November 2020–1 February 2021 (86 days)

Third wave → 2 February 2021–28 April 2021 (85 days)

Out-waves → other periods.

Figure 1: Number of hospitalizations over the study period and the empirical definitions of the waves.

Hospital stay: A factor containing hospital name and categorized as long-term or regular hospitalization.

5.2.4. Ethical Considerations

Ethical approval for the study was obtained from the University of Bologna's Ethical Committee (approval number 283066, 5 October 2021).

5.3. Statistical Analysis

5.3.1. Poisson and Hurdle Models

The generalized linear model (GLM) is a framework for regression models with a wide range of outcome variable types [33]. A p-value of ≤ 0.05 was deemed to be significant. The outcome variable, LoS, is a count type, and Poisson regression is the most common type used to analyze this kind of data. However, in practice, the Poisson assumptions are usually violated (there is overdispersion), and it was revealed that the Poisson model did not fit the data, so alternative models such as the negative binomial model were considered [20,24,25]. Both the Poisson and NB models were implemented in R Studio with R version 4.0.4. [34] by the *glm* function in the stats package and the *glm.nb* function in the *MASS* package.

In addition to overdispersion, many count data have more zero observations. As a result, if the proportion of zeros in the response variable is not high, the Hurdle model would be a preferable option [26].

This is a two-stage model:

1. Zero vs. non-zero (logistic regression);
2. Regression of counts > 0 (zero-truncated Poisson or NB).

It combines a count data model $f_{count}(y; x, \beta)$ (left-truncated at $y = 1$ and a zero Hurdle model $f_{zero}(y; z, \gamma)$ (right-censored at y = 1):

$$f_{hurdle}(y; z, \beta, \gamma) = \begin{cases} f_{zero}(0; z, \gamma) & if\ y = 0 \\ (1 - f_{zero}(0; z, \gamma)) f_{count}(y; x, \beta) / (1 - f_{count}(0; x, \beta)) & if\ y > 0 \end{cases}$$

where y is the value of the dependent variable, z is a vector denoting the predictor variable in the zero Hurdle model, x represents a vector denoting the predictor variable in the count data model, g is a vector of the coefficients related to z, and b denotes a vector of

coefficients related to x. fzero is a probability density function of y = 0, typically modeled with binary logistic regression, where all counts greater than 0 are given a value of 1 and 0 otherwise [26]. In R Studio (R version 4.0.4), Hurdle count data models are fitted with the hurdle function of the *pscl* package [35].

5.3.2. Model Selection and Comparison

We compared the performance of various count models to determine which one was the best based on the following criteria:

1. Akaike Information Criterion (AIC):

For non-nested models, the standard approach is to use information criteria such as the AIC [36,37] and the Bayesian Information Criteria (BIC) [37]. The AIC estimates the quality of each model based on the formula: $\mathrm{AIC} = 2\mathrm{k} - 2\log\mathrm{L}(\theta)$, where L($\theta$) is the maximized likelihood function for the estimated model and k is the number of parameters in the model. Given a set of candidate models for the data, the preferred model has the minimum AIC value [36].

2. Vuong test:

The purpose of the Vuong test [38] is to compare two models (that are not nested) fitted to the same data by maximum likelihood, and it is based on a comparison of the predicted probabilities of two models that are not nested. Specifically, it tests the null hypothesis that the two models fit the data equally well. A large positive test statistic provides evidence of the superiority of Model 1 over Model 2, while a large negative test statistic is evidence of the superiority of Model 2 over Model 1, under the null hypothesis that the models are indistinguishable.

Let p_1 be the predicted probabilities from Model 1, evaluated conditionally on the estimated MLEs. Let p_2 be the corresponding probabilities from Model 2. The Vuong statistic is:

$$Vuong = (s_m)^{-1}\sqrt{n}\,m$$

Where $m = log(p_1) - log(p_2)$, s_m is the sample standard deviation of m, and N is the sample size [9]. We compared all the count models in R Studio with R version 4.0.4. using the *pscl* package [35].

3. Rootograms:

Rootograms is a graphical approach used to assess the fit of a count regression model. It is useful for diagnosing and treating issues such as overdispersion and/or excess zeros in count data models [27,28]. We visualized the plot in R Studio with R version 4.0.4. using the *countreg* package [27,28].

5.3.3. Quantile Regression Model

A quantile regression model can be used to explore the relationships between the quantiles of the response distribution and the available covariates. By comparing such quantiles, we can obtain a more complete picture of the conditional distribution than we can with regression models that consider the mean. In addition, quantile regression allows researchers to explore a range of conditional quantile functions, thereby exposing various forms of conditional heterogeneity and controlling for unobserved individual characteristics.

With a Poisson distribution, quantile regression models were used to model the effects of covariates on the quantile values of the response variable, LoS. A linear quantile regression model was fitted to the log-transformed response. The modeled quantiles were 0.25, 0.5, 0.75, 0.8, 0.9, and 0.95. We used the *quantreg* package in R [39].

5.4. Results

Descriptive statistics;

Descriptive statistics were used to summarize the baseline characteristics of the study population. In our setting , a total of 13,200 COVID-19 patients were included in the study. The median length of stay during the study period was 6 days. Table 1 summarizes the descriptive statistics and the characteristics of the variables considered in the study. Figure 2 shows the frequency distribution of LoS, which was right-skewed.

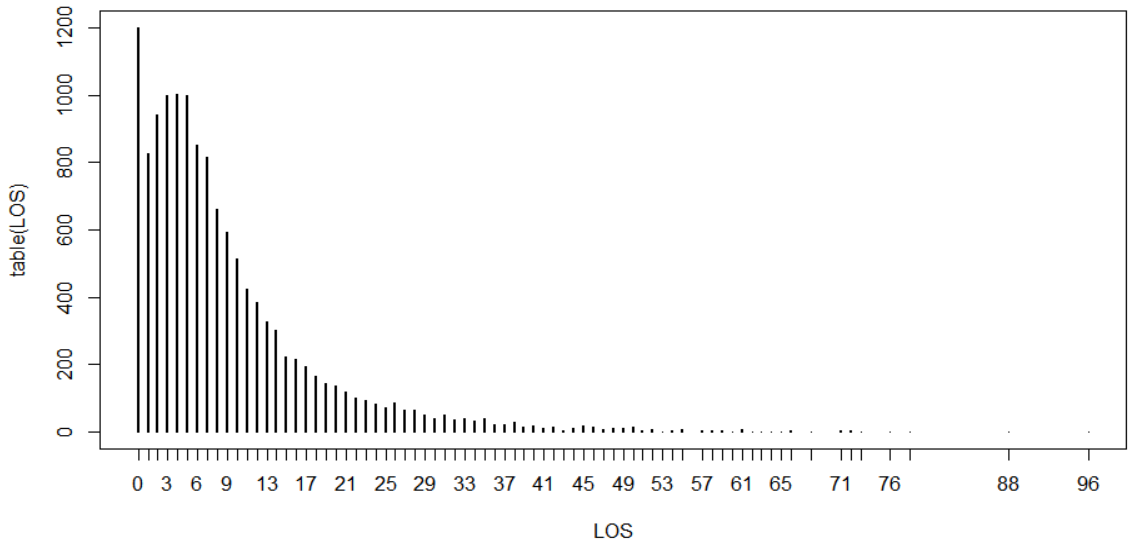


Figure 2. Frequency distribution of LoS.

Table 1. Descriptive statistics, frequency, and characteristics of the study sample (N = 13,203).

Predictors	N	%
Clinical setting	-	-
(Ordinary hospital) *	9818	74.36
(ICU, intensive + sub-intensive)	3385	25.64
Hospital-stay	-	-
Regular *	12,064	91.37
Long-term	1139	8.63
Age	-	-
[0–10)	412	3.12
[10–20)	181	1.37
[20–30)	167	1.26
[30–40)	401	3.04
[40–50)	1082	8.20
[50–60) *	1881	14.25
[60–70)	2237	16.94
[70–80)	2847	21.56
[80–102)	3991	30.23
Wave	-	-
First wave	2391	18.11
Second wave *	3369	25.52
Third wave	4173	31.61
Out-waves	3270	24.77

* Reference categories.

Figure 3 shows the variation in LoS for each predictor. In Figure 4, the plots of LoS = = 0 (true) vs. LoS > 0 (false) are displayed for each predictor. True values (in black) indicate that zeros are infrequent except for the age groups [0, 10) and [10, 20). False values (in white)

indicate that LoS values greater than zero are more frequent, and these are shown in the count boxplot for each predictor

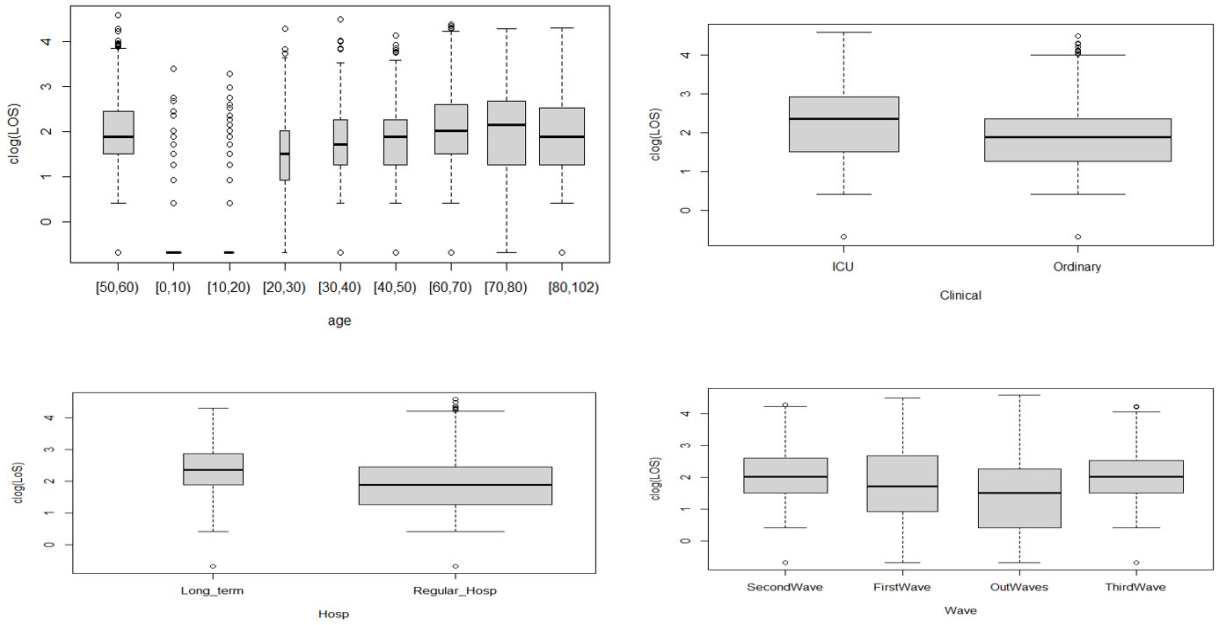
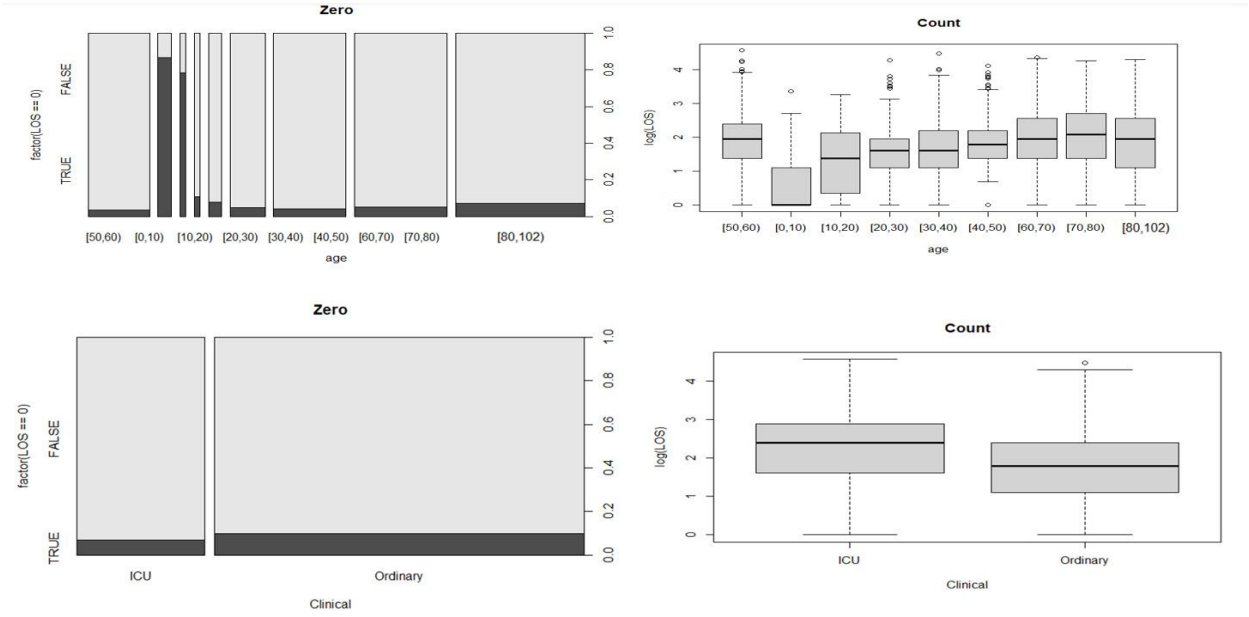


Figure 3. LoS plotted against the predictors used. All plots show the variation in LoS when predictors were changed.



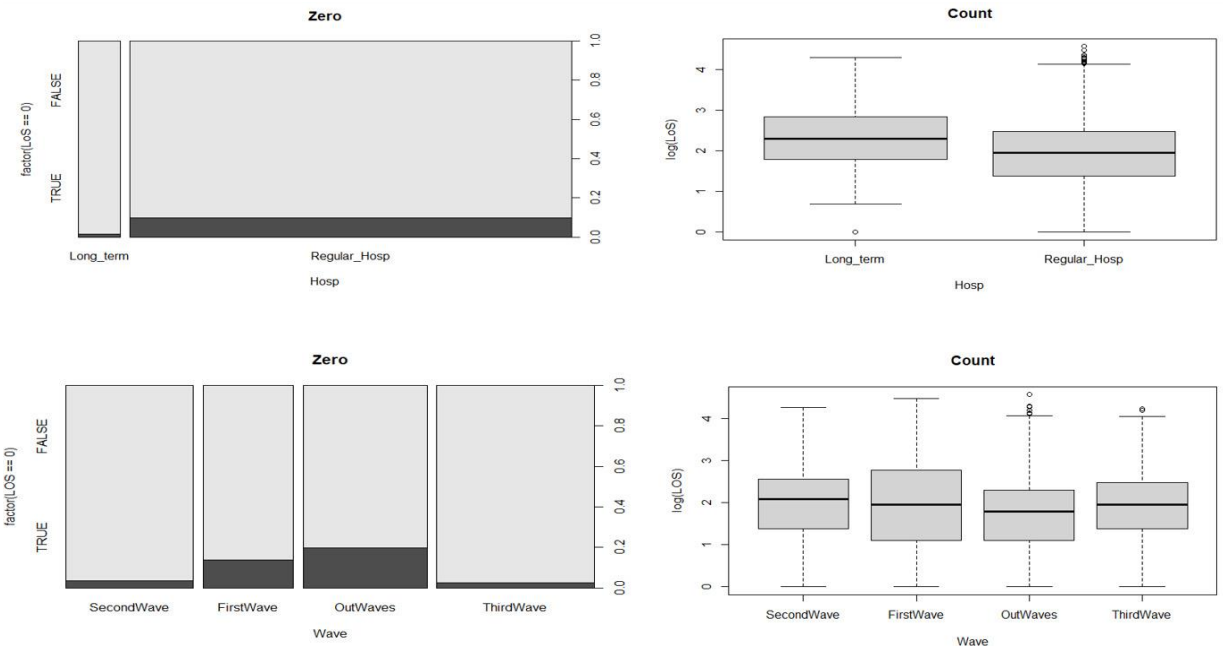


Figure 4: Plot of (LoS = 0 vs LoS > 0) for each predictor

Count Regression Models:

Results of the Poisson and Hurdle Models;

Table 2 shows the values of the parameter estimates (and the standard errors in the brackets) of the Poisson and Negative binomial (NB) models. All predictors were significant for the Poisson models, but there was overdispersion and poor data fit. In contrast, the NB model fitted data well compared with the Poisson model, and all predictors were significant, except for the third wave category among the wave predictors (see the Table 2).

Table 2. Summary of estimates and standard errors (in brackets) of the Poisson and NB model regression coefficients.

	Poisson Model	NB Model
(Intercept)	2.00 *** (0.01)	1.97 *** (0.03)
Clinical–ordinary	-	-
Clinical–ICU	0.55 *** (0.01)	0.57 *** (0.02)
Regular	-	-
Long-term	0.46 *** (0.01)	0.48 *** (0.03)
Age [0, 10)	−2.73 *** (0.08)	−2.79 *** (0.09)
Age [10, 20)	−1.70 *** (0.07)	−1.78 *** (0.10)
Age [20, 30)	−0.21 *** (0.03)	−0.18 * (0.07)
Age [30, 40)	−0.17 *** (0.02)	−0.15 ** (0.05)
Age [40, 50)	−0.15*** (0.01)	−0.14 *** (0.03)
Age [50, 60)	-	-
Age [60, 70)	0.10 *** (0.01)	0.09 *** (0.03)
Age [70, 80)	0.13 *** (0.01)	0.14 *** (0.03)
Age [80, 102)	0.09 *** (0.01)	0.11 *** (0.03)
First wave	0.11 *** (0.01)	0.11 *** (0.02)
Second wave	-	-
Third wave	−0.05 *** (0.01)	−0.04 (0.02)
Out-wave	−0.27 *** (0.01)	−0.26 *** (0.02)
AIC	131,254.90	82,120.2
BIC	131,359.73	82,232.52
Pseudo R^2	0.77	0.23

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 3 reports the coefficients of the fitted models. The second column of the table reports the count process of the Hurdle–Poisson model; the third column shows the count process of the Hurdle–NB model. In Section A, only the truncated process, i.e., positive counts, has been fitted. In Section B, zero counts were fitted through a binary logit model.

Table 3. Coefficients of the Hurdle–Poisson and Hurdle–NB models for the count process.

	Hurdle–Poisson Model	Hurdle–NB Model
	(A) Truncated count process	
	Estimate (Std. Error)	Estimate (Std. Error)
(Intercept)	2.01 *** (0.01)	1.91 *** (0.03)
Clinical–ordinary	-	-
Clinical–ICU	0.56 *** (0.01)	0.61 *** (0.02)
Regular	-	-
Long-term	0.42 *** (0.01)	0.46 *** (0.03)
Age [0, 10)	−0.98 *** (0.01)	−1.23 *** (0.15)
Age [10, 20)	−0.42 *** (0.07)	−0.45 ** (0.15)
Age [20, 30)	−0.16 *** (0.03)	−0.14 (0.08)
Age [30, 40)	−0.13 *** (0.02)	−0.13 * (0.05)
Age [40, 50)	−0.14 *** (0.01)	−0.14 *** (0.04)
Age [50, 60)	-	-
Age [60, 70)	0.11 *** (0.01)	0.11 *** (0.03)
Age [70, 80)	0.15 *** (0.01)	0.17 *** (0.03)
Age [80, 102)	0.13 *** (0.01)	0.17 *** (0.03)
First wave	0.17 *** (0.01)	0.19 *** (0.02)
Second wave	-	-
Third wave	−0.06 *** (0.01)	−0.05 * (0.02)
Out-wave	−0.19 *** (0.01)	−0.18 *** (0.02)
Log(theta)	-	0.436 *** (0.019)
	(B) Zero count process	
	Estimate (Std.Error)	Estimate (Std. Error)
(Intercept)	3.86 *** (0.16)	3.86 *** (0.16)
Clinical–ordinary	-	-
Clinical–ICU	−0.11 (0.09)	−0.11 (0.09)
Regular	-	-
Long-term	1.29 *** (0.25)	1.29 *** (0.25)
Age [0, 10)	−4.62 *** (0.19)	−4.62 *** (0.19)
Age [10, 20)	−4.10 *** (0.22)	−4.10 *** (0.22)
Age [20, 30)	−0.93 ** (0.28)	−0.93 ** (0.28)
Age [30, 40)	−0.71 ** (0.23)	−0.71 ** (0.23)
Age [40, 50)	−0.35 (0.19)	−0.35 (0.19)
Age [50, 60)	-	-
Age [60, 70)	−0.16 (0.17)	−0.16 (0.17)
Age [70, 80)	−0.45 ** (0.15)	−0.45 ** (0.15)
Age [80, 102)	−0.83 *** (0.14)	−0.83 *** (0.14)
First wave	−1.08 *** (0.12)	−1.08 *** (0.12)
Second wave	-	-
Third wave	0.26 (0.14)	0.26 (0.14)
Out-wave	−1.20 *** (0.11)	−1.20 *** (0.11)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Model Selection and Comparison;

A structured model selection and comparison process is crucial to avoid misleading results and interpretations. The final interpretations were based on the selected model.

Vuong Test

The comparison between the NB and Poisson models had a Vuong test statistic of 39.50 with a p-value < 0.0001, indicating that the NB model provided a better fit. Similar results were obtained for the other models. From Table 4, we can conclude that the preferable model is the

Hurdle–NB model because it has the highest value of the Vuong test statistics and a significant *p-value* (see Table 4).

Table 4. Model comparison using the Vuong test.

Model Comparison	Vuong Test Statistic	*p*-Value	Preferred Model
NB vs. Poisson	39.50	<0.0001	NB
Hurdle–Poisson vs. Poisson	21.20	<0.0001	Hurdle–Poisson
Hurdle–NB vs. Poisson	**39.81**	<0.0001	**Hurdle–NB**
Hurdle–Poisson vs. NB	−35.66	<0.0001	NB
Hurdle–NB vs. NB	10.17	<0.0001	Hurdle–NB
Hurdle–NB vs. Hurdle–Poisson	36.48	<0.0001	Hurdle–NB

Bold indicates the preferred model.

AIC/BIC

The Poisson regression model had the largest AIC value but the worst fit to the data. Hurdle–Poisson had the second largest AIC, whereas Hurdle–NB had the smallest AIC value and hence was the best choice for this dataset (see Table 5).

Table 5. Model comparison using AIC/BIC.

Model	AIC	BIC
Negative Binomial (NB)	82,120.2	82,232.5
Hurdle–NB	**81,353.9**	**81,571.1**
Poisson	131,254.9	131,359.7
Hurdle–Poisson	122,505.8	122,715.4

Bold indicates the preferred model.

Rootograms

The hanging Rootograms for the Hurdle–NB showed a much better fit with the data than other models. The second best was NB, but some of the zero counts were underfitted (under the line) and the LoS counts were also overfitted (over the line), as in [25], the line at 0 allows us to easily visualize where the model is over- or underfitted. At 0, it fits perfectly by design. The line above indicates that the variability of our observed data is much greater than what a Poisson model would predict. The red line is the theoretical Poisson fit. “Hanging” from each point on the red line is a bar, the height of which represents the difference between the expected and observed counts (see Figure 5).

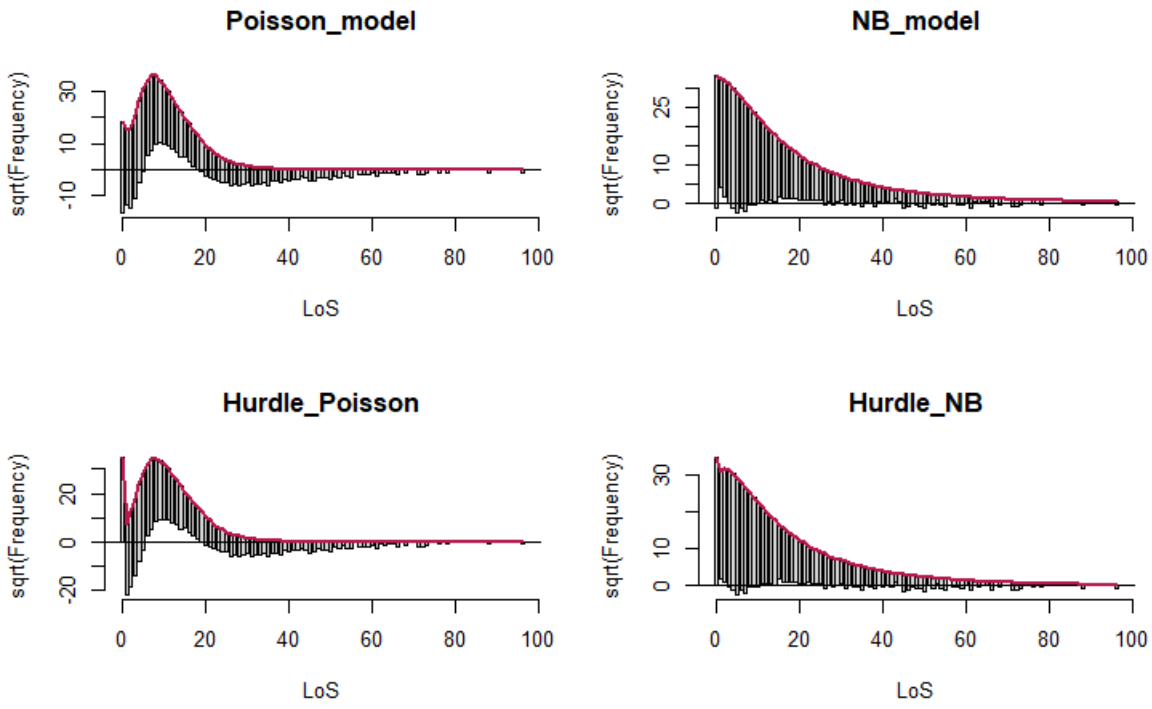


Figure 5. Rootogram plots

The regression exponential coefficients (exp(coef)), i.e., the incidence rate ratios (IRR), give information about the increase in the expected LoS for a unit of increase in a given predictor. Hence, an IRR greater (smaller) than the one obtained for the predictors suggests longer (shorter) expected LoSs for the selected models.

Interpretation of NB_IRR (see Table 6): The baseline or the average intercept LoS count is 7 days. The other exponentiated coefficients are interpreted multiplicatively. On average, the ICU setting increased LoS by 1.77-fold compared with ordinary hospitalization and low-intensity stays. Long-term hospitals increased the average LoS by 1.62 compared with regular-stay hospitals. Average LoS decreased by 0.06-, 0.17-, 0.84-, 0.86-, and 0.87-fold in the age groups of [0, 10), [10, 20), [20, 30), [30, 40), and [40, 50), respectively, compared with the intermediate age group of [50, 60), and increased by 1.15-, 1.12-, and 1.11-fold in the age groups of [60, 70), [70, 80), and [80, 102), respectively. The third wave did not have a significant effect on the average LoS, which increased by 1.11-fold during the first wave and decreased by 0.77-fold during the out-wave compared with the second wave (chosen as the reference level).

Table 6. Estimates of coefficients: exponential (coef).

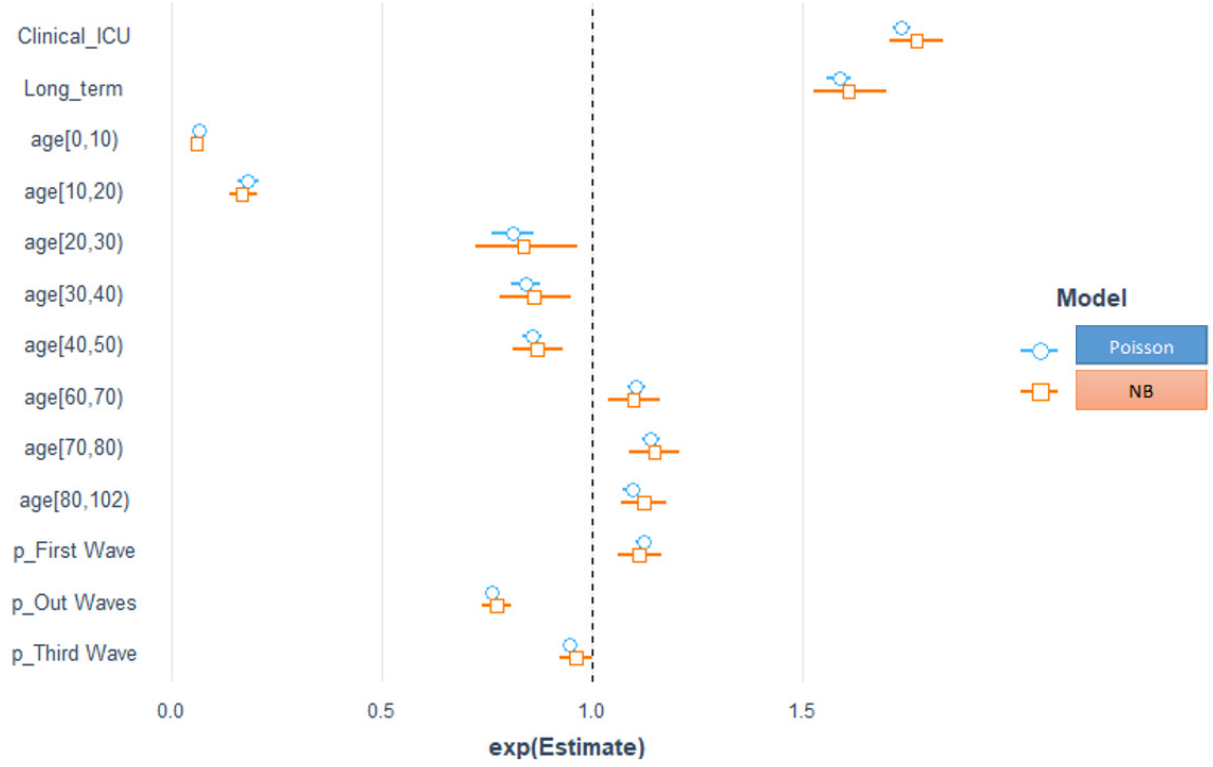
	Hurdle-NB		NB
	Count_Model	Zero_Hurdle_Model	-
Intercept	6.72	47.57	7.20
Clinical–ICU	1.84	***0.90***	1.77
Long-term	1.58	3.61	1.61
Age [0, 10)	0.29	0.01	0.06
Age [10, 20)	0.64	0.02	0.17
Age [20, 30)	***0.87***	0.40	0.84
Age [30, 40)	0.88	0.49	0.86
Age [40, 50)	0.87	***0.71***	0.87
Age [60, 70)	1.12	***0.85***	1.10
Age [70, 80)	1.19	0.64	1.15
Age [80, 102)	1.19	0.44	1.12
First wave	1.21	0.34	1.11
Third wave	0.95	***1.30***	***0.96***
Out-wave	0.84	0.30	0.77

Bold italic font indicates the non-significant variables.

Interpretation (Hurdle–NB, zero-Hurdle model) (see Table 6): The baseline or the intercept odds having a positive count vs. zero were 47.57. The odds of having a positive count vs. zero in the ICU setting did not significantly affect ordinary hospitalization and low-intensity stays. The odds of having a positive count vs. zero increased by 3.61-fold for every unit of increase in long-term stays in hospital compared with regular stays in hospital. The odds of having a positive count vs. zero decreased by 0.34- and 0.39-fold during the first wave and out-wave compared with the second wave, whereas the third wave did not have a significant effect. The odds of having a positive count vs. zero decreased by 0.01-, 0.02-, 0.40-, 0.49-, 0.64-, and 0.44-fold in the age groups of [0, 10), [10, 20), [20, 30), [30, 40), [70, 80), and [80, 102), respectively, whereas the age groups of [40, 50), [60, 70) did not have a significant effect.

Interpretation (Hurdle–NB, positive count model or truncated Poisson) (see Table 6): Given the positive response (among those with positive counts), the average count was 6.72. The average LoS increased by 1.84-fold in the ICU setting compared with ordinary hospitalization and low-intensity stays among those with positive counts. The average LoS increased in the age groups of [60, 70), [70, 80), and 80+ by 1.12-, 1.19-, and 1.19-fold, respectively, and decreased by 0.29-, 0.64-, 0.88-, and 0.87-fold in the age groups of [0, 10), [10, 20), [30, 40), and [40, 50), respectively, whereas the age group of [20, 30) did not have a

significant effect on the average LoS of those who had positive counts or a LoS greater than 0, all compared with the [50, 60) age group. The average LoS increased by 1.58-fold in long-term stays compared with regular hospitalization among those with positive counts. The average LoS increased by 1.21-fold during the first wave and decreased during the third wave and out waves by 0.95 and 0.84 times, respectively, for those with positive counts, compared with the second wave (chosen as the reference level). The exponentials of the coefficients or estimates for the Poisson model and NB model, and the Hurdle–NB model for count and zero Hurdle effect plots are shown in Figure 6.



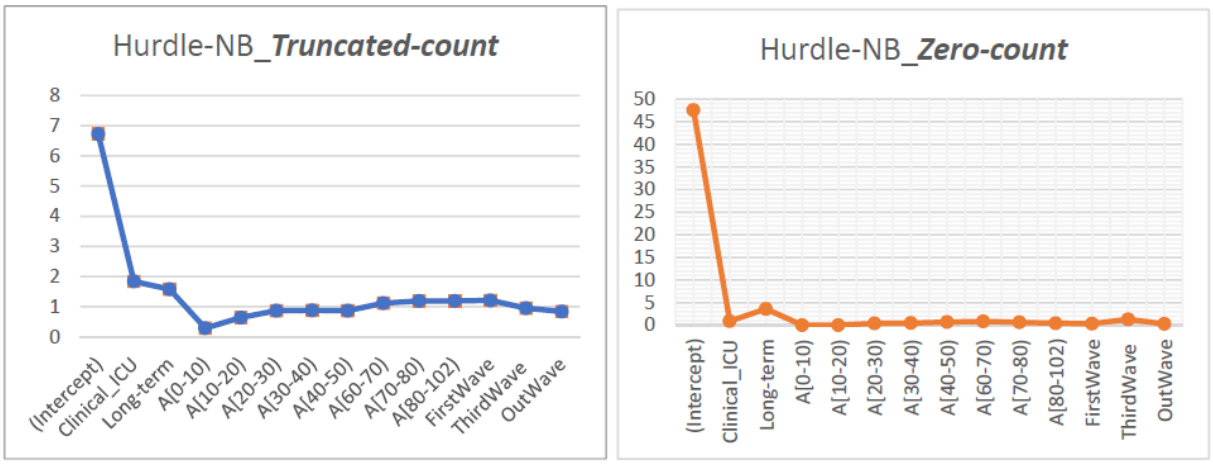


Figure 6. Visualization of IRR for the Poisson and NB models (top panel), and the Hurdle NB truncated (bottom left) and zero-count effect plots (bottom right).

Quantile Regression Models;

The results of the quantile regression models are reported in Table 7. The intercept coefficients increased with the progression of the quantiles, ranging from 1.42 to 2.82. Covariates related to the ICU setting, hospitals with longer hospitalization, the first wave, and the out-waves were found to be statistically significant for all the modeled quantiles. Specifically, the coefficients for ICU settings, the first wave, and the out-wave were higher for higher quantile values, while coefficients for longer hospitalization and the third wave became lower as the quantile values increased. Differences in the third wave's level coefficients were not statistically significant from the second wave's values, except for the 0.8, 0.9, and 0.95 quantiles. Moreover, the coefficients were positive for ICU and longer hospitalization, while those for the out-wave were negative. The first wave's coefficients were negative for the 0.25 and 0.5 quantiles, and they became positive for the higher quantile values. The age group covariate had a significant impact on the LoS. Estimated values for younger age groups ([0, 10) to [40, 50)) were found to be negative for all the quantiles, while most older groups presented positive estimates. The trend of coefficients for all the levels related to the age groups was to increase as the quantile values increased. A graphical depiction is presented in Figure 7.

Table 7. Quantile regression models for count data.

Quantile Regression Models						
	Estimate (Std. Error)					
	0.25	0.5	0.75	0.8	0.9	0.95
(Intercept)	1.42 (0.03) ***	1.80 (0.02) ***	2–16 (0.03) ***	2.25 (0.02) ***	2.55 (0.04) ***	2.82 (0.03) ***
Clinical–Ordinary						
Clinical–ICU	0.36 (0.03) ***	0.55 (0.02) ***	0.64 (0.02) ***	0.64 (0.02) ***	0.65 (0.03) ***	0.66 (0.03) ***
Regular						
Long-term	0.46 (0.03) ***	0.46 (0.02) ***	0.46 (0.04) ***	0.46 (0.03) ***	0.43 (0.04) ***	0.32 (0.05) **
Age [0, 10)	−4.84 (147.26)	−4.27 (142.09)	−4.26 (0.016) ***	−4.32 (0.15) ***	−3.27 (0.34) ***	−2.38 (0.57) **
Age [10, 20)	−4.08 (239.49)	−3.60 (0.21) ***	−3.64 (0.22) ***	−3.19 (1.12) **	−1.52 (0.16) ***	−0.98 (0.30) *
Age [20, 30)	−0.40 (0.16)	−0.33 (0.08) **	−0.31 (0.11) **	−0.31 (0.09) **	−0.19 (0.10)	−0.02 (0.10)
Age [30, 40)	−0.27 (0.07) *	−0.18 (0.05) **	−0.19 (0.06) **	−0.18 (0.05) **	−0.16 (0.10)	−0.14 (0.06)
Age [40, 50)	−0.08 (0.04)	−0.13 (0.03) **	−0.14 (0.03) **	−0.16 (0.03) **	−0.16 (0.05) *	−0.18 (0.05) **
Age [50, 60)						
Age [60, 70)	0.009 (0.03)	0.05 (0.03)	0.11 (0.03) **	0.11 (0.02)	0.14 (0.03) **	0.17 (0.05) *
Age [70, 80)	−0.07 (0.04)	0.10 (0.03) **	0.21 (0.03) **	0.22 (0.03) **	0.24 (0.03) **	0.21 (0.04) **
Age [80, 102)	−0.26 (0.03) **	0.02 (0.03)	0.21 (0.03) **	0.24 (0.03) ***	0.27 (0.04) **	0.30 (0.03) ***
First wave	−0.48 (0.04) ***	−0.13 (0.03) **	0.14 (0.03) **	0.17 (0.03) **	0.29 (0.04) **	0.34 (0.05) **
Second wave						
Third wave	0.04 (0.03)	−0.02 (0.02)	−0.03 (0.02)	−0.05 (0.02)	−0.07 (0.03)	−0.12 (0.04) *
Out-wave	−0.57 (0.04) ***	−0.36 (0.03) ***	−0.21 (0.03) **	−0.19 (0.03) **	−0.14 (0.04) **	−0.13 (0.04) *

$p < 0.05$, * $p < 0.01$, ** $p < 0.001$, *** $p = 0.0001$.

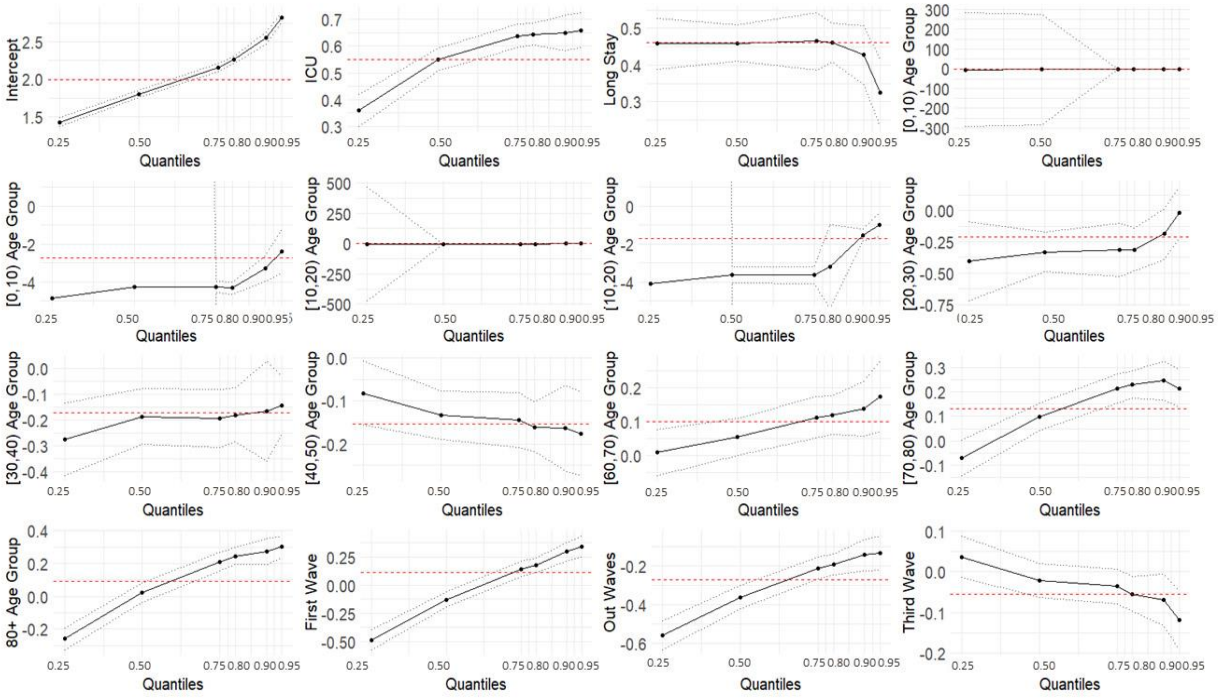


Figure 7. Quantile regression models.

5.5. Discussion

Researchers in the medical field are currently working to improve the quality and efficiency of health care systems and services in various ways, with LoS [26] being one of the

efficiency and quality indicators. To the best of our knowledge, no previous study has examined the LoS in Bologna, Italy, among COVID-19 patients, and this is the first study to consider analyzing LoS using the best count fit model and comparing several models. This study also aimed to explore the hospital admission risk factors associated with LoS and presents a relatively novel method for modeling LoS using predictors.

The necessity of carefully picking a model that effectively describes the observed count data is demonstrated by the fact that the different statistical techniques produce mixed findings. First, we used the Poisson, negative binomial, Hurdle–Poisson, and Hurdle–NB regression models to model the effects of covariates on LoS. Second, we used quantile regression to model the impact of the variables on the quantile values of the response variable LoS. Compared with Poisson regression, the fit of the NB regression better tolerated the overdispersion in the data [20,24,25]. In addition to the overdispersion, the Hurdle models accounted for zero LoS more thoroughly [26]. The Hurdle–NB model was finally chosen on the basis of three criteria, including the AIC, the Vuong test, and Rootograms to understand the impacts of the predictors on the average LoS.

In our analysis, the median LoS was 6 days, which is comparable with the value reported in a similar study in Europe, USA, and the UK, but shorter than one in a study on China [12]. In this systematic review, the median hospital LoS ranged from 4 to 53 days within China (45 studies surveyed), and from 4 to 21 days outside China (eight studies surveyed). Similarly, a shorter LoS was also documented by the ISARIC (International Severe Acute Respiratory and Emerging Infections Consortium) report [40]. This report (which included data from 25 countries) described a median and IQR LoS of 4 and 1–9 days, respectively, which are substantially lower than in our study. Differences in LoS duration between nations can be explained by the different policies or strategies applied to control COVID-19 infection, or by the different population samples involved in the studies. Knowing the LoS or other adverse events in advance can help to health care systems organize the allocation of limited resources more efficiently.

Our results in Bologna are entirely consistent with those from studies which observed that older age ($\geq$ 60 years) was associated with prolonged LoS [41–43]. Patients with ICU

admissions had a longer LoS than those with only ordinary hospitalization, as these patients might need additional treatment or time when their disease reaches severe stages and they need more complex treatments. Staying in a long-term hospital was another contributing factor for a LoS longer than that in regular hospitals. This might be explained by the higher percentage of surgical operations and transfer rates, as well as restricted antibiotic use compared with other patients.

Usually, public hospitals manage acute patients' LoS efficiently [44] in Italy. However, due to the new variants of the virus that causes COVID-19, the ability to assess or predict LoS will be more and more crucial in the future as a higher LoS is also associated with higher costs [45] and reduced capacity for other sanitary needs [46]. From this perspective, it is worth noting that we found that the LoS in Bologna was higher during the first wave—when the government proclaimed a state of emergency in response to an increase in the number of infection cases—and peaked around April 2021. However, during the third wave and out-wave, there was a drop, indicating that the chosen policies and strategies of the government and the health department, together with better clinical knowledge of the disease, have had a positive impact on the management of the patients.

We also used quantile regression to model the effects of the covariates on the quantile values of the response variable LoS, using a Poisson distribution, and to explore a range of conditional quantile functions, thereby exposing various forms of conditional heterogeneity and controlling for unobserved individual characteristics. The results from the quantile regression showed that the covariates related to the ICU setting, long-term hospitals, age groups, the first wave, and the out-waves were statistically significant for all the modeled quantiles. Specifically, for the ICU setting, the coefficients were higher for higher quantile values than ordinary hospitalization, meaning a longer LoS, and the effect was more pronounced for the higher quantiles. All the age groups, except for the [40, 50) group, showed an increasing trend with the quantiles. The wave also played a role: during the first wave, short LoSs were shorter (negative coefficients) and long LoSs were longer (positive coefficients) compared with those of the second wave. LoSs recorded in the out-wave period were shorter for all the quantiles (negative coefficients), and for the third wave, longer LoSs shortened (negative coefficients). The

coefficients for long-term hospitalization became closer to zero as the quantile values increased, representing a less marked difference with respect to regular hospitalization.

The quantile regression models provided a more comprehensive overview of the effect of the covariates on a given phenomenon. As appears evident from this work, the same covariate can have a non-constant effect on different quantiles. This is valuable information for better modeling the LoS, which would not be controlled by simply evaluating the effect of the covariates on the mean values.

One limitation of our study is the absence of information on additional demographic and socio-economic variables, clinical characteristics, medical conditions, and laboratory tests for the patients, which might be among the most important factors that tend to increase the LoS. However, we have now started to analyze an extended, non-aggregated version of this dataset including all these variables.

COVID-19 cases are complex and still increasing worldwide, and it is difficult to predict when it will stop completely. Countries should plan and prepare for the worst-case scenarios, such as when different variants come into play (as in the case of Omicron at this time). In Italy, we have now entered into a fourth wave, so the wise use of limited health care resources is one of the most important priorities. Studies such as the present one might help health policymakers and managers to better plan the logistics of hospital settings, define priorities, and carry out a more accurate cost analysis. In addition, the information we have obtained and the method we used might serve as a tool or as a reference in the case of similar epidemics in the future.

5.6. Conclusions

Hurdle–NB had an excellent fit to the Bolognese COVID-19 data based on the AIC, the Vuong test, and Rootograms. On the basis of this model, we found a significant change in the average LoS with clinical setting, age group, type of hospital, and time in the pandemic (i.e., different waves). Moreover, these covariates, assessed via quantile regression models, showed a different impact on different quantiles and not only on the mean LoS values. The results obtained from our study can help to shed light on the risk factors that led to increased LoS

among COVID-19 patients in Bologna, Italy. Hence, they could contribute to reducing unnecessary spending or arrangements, and help in identifying and managing prioritized settings for health professionals and policymakers, given the impact and the economic burden of the pandemic.

References:

1. World Health Organization. Coronavirus Disease (COVID-19). 2020. Available online: https://www.who.int/health-topics/coronavirus#tab=tab_1 (accessed on 25 October 2021).
2. World Health Organization. WHO Director-General's Opening Remarks at the Media Briefing on COVID-19. 2020. Available online: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-mediabriefing-on-covid-19---11-march-2020 (accessed on 25 October 2021).
3. Worldometers. COVID-19 Coronavirus Pandemic. 2021. Available online: https://www.worldometers.info/coronavirus/ (accessed on 21 December 2021).
4. Specchia, M.; Di Pilla, A.; Sapienza, M.; Riccardi, M.; Cicchetti, A.; Damiani, G. Dealing with COVID-19 Epidemic in Italy: Responses from Regional Organizational Models during the First Phase of the Epidemic. Int. J. Environ. Res. Public Health 2021, 18, 5008.
5. Armocida, B.; Formenti, B.; Ussai, S.; Palestra, F.; Missoni, E. The Italian health system and the COVID-19 challenge. Lancet Public Health 2020, 5, e253.
6. Feng, C.X.; Li, L. Modeling Zero Inflation and Overdispersion in the Length of Hospital Stay for Patients with Ischaemic Heart Disease. In Advanced Statistical Methods in Data Science; Chen, D.G., Chen, J., Lu, X., Yi, G., Yu, H., Eds.; ICSA Book Series in Statistics; Springer: Singapore, 2016.
7. Baek, H.; Cho, M.; Kim, S.; Hwang, H.; Song, M.; Yoo, S. Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. PLoS ONE 2018, 13, e0195901.
8. Bueno, H.; Ross, J.S.; Wang, Y.; Chen, J.; Vidan, M.T.; Normand, S.L.; Curtis, J.P.; Drye, E.E.; Lichtman, J.H.; Keenan, P.S.; et al. Trends in length of stay and short-term outcomes among Medicare patients hospitalized for heart failure, 1993–2006. JAMA 2010, 303, 2141–2147.
9. Rotter, T.; Kinsman, L.; James, E.; Machotta, A.; Gothe, H.; Willis, J.; Snow, P.; Kugler, J. Clinical pathways: Effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane Database Syst. Rev. 2010, 3, CD006632.
10. Williams, R. Models for Count Outcomes; University of Notre Dame: Notre Dame, IN, USA, 2016.
11. Dodd, S.; Bassi, A.; Bodger, K.; Williamson, P. A comparison of multivariable regression models to analyse cost data. J. Eval. Clin. Pract. 2006, 12, 76–86.
12. Rees, E.M.; Nightingale, E.S.; Jafari, Y.;Waterlow, N.R.; Clifford, S.; Pearson, C.A.; CMMIDWorking Group; Jombart, T.; Procter, S.R.; Knight, G.M. COVID-19 Length of Hospital Stay: A Systematic Review and Data Synthesis. BMC Med. 2020, 18, 270.
13. Tai, P.Q.; Toan, D.T.T.; Son, D.T.; Van Hoang, T.H.; Minh, L.N.; Hung, L.X.; Van Toan, N.; Hoat, L.N.; Luong, N.K.; Khue, L.N.; et al. Factors associated with the duration of hospitalization among COVID-19 patients in Vietnam: A survival analysis. Epidemiol. Infect. 2020, 148, e114.

14. Guo, A.; Lu, J.; Tan, H.; Kuang, Z.; Luo, Y.; Yang, T.; Xu, J.; Yu, J.; Wen, C.; Shen, A. Risk factors on admission associated with hospital length of stay in patients with COVID-19: A retrospective cohort study. Sci. Rep. 2021, 11, 7310.
15. Xiong, X.;Wai, A.;Wong, J.; Tang, E.; Chu, O.;Wong, C.; Rainer, T.H. Impact of VaryingWave Periods of COVID-19 on in-Hospital Mortality and Length of Stay for Admission through Emergency Department: A Territory-Wide Observational Cohort Study. Influenza Other Respir. Viruses 2020, 16, 193–203.
16. Thiruvengadam, G.; Lakshmi, M.; Ramanujam, R. A Study of Factors Affecting the Length of Hospital Stay of COVID-19 Patients by Cox-Proportional Hazard Model in a South Indian Tertiary Care Hospital. J. Prim. Care Community Health 2021, 12, 21501327211000231.
17. Jang, S.Y.; Seon, J.Y.; Yoon, S.J.; Park, S.Y.; Lee, S.H.; Oh, I.H. Comorbidities and Factors Determining Medical Expenses and Length of Stay for Admitted COVID-19 Patients in Korea. Risk Manag. Healthc. Policy 2021, 14, 2021–2033.
18. Vekaria, B.; Overton, C.; Wi´sniowski, A.; Ahmad, S.; Aparicio-Castro, A.; Curran-Sebastian, J.; Eddleston, J.; Hanley, N.A.; House, T.; Kim, J.; et al. Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning. BMC Infect. Dis. 2021, 21, 700.
19. Mahboub, B.; Bataineh, M.T.A.; Alshraideh, H.; Hamoudi, R.; Salameh, L.; Shamayleh, A. Prediction of COVID-19 Hospital Length of Stay and Risk of Death Using Artificial Intelligence-Based Modeling. Front. Med. 2021, 8, 592336.
20. Karazsia, B.T.; Van Dulmen, M.H. Regression models for count data: Illustrations using longitudinal predictors of childhood injury. J. Pediatr. Psychol. 2008, 33, 1076–1084.
21. Hammer, T.H.; Landau, J.C. Methodological issues in the use of absence data. J. Appl. Psychol. 1981, 66, 574–581.
22. Bhowmik, K.R.; Das, S.; Islam, M.A. Modelling the number of antenatal care visits in Bangladesh to determine the risk factors for reduced antenatal care attendance. PLoS ONE 2020, 15, e0228215.
23. Shaaban, A.N.; Peleteiro, B.; Martins, M.R.O. Statistical models for analyzing count data: Predictors of length of stay among HIV patients in Portugal using a multilevel model. BMC Health Serv. Res. 2021, 21, 372.
24. Cameron, A.C.; Trivedi, P.K. Econometric models based on count data. Comparisons and applications of some estimators and tests. J. Appl. Econ. 1986, 1, 29–53.
25. Cameron, A.C.; Trivedi, P.K. Regression Analysis of Count Data; Cambridge University Press: Cambridge, UK, 1998.
26. Mullahy, J. Specification and Testing of Some Modified Count Data Models. J. Econom. 1986, 33, 341–365.
27. Christian, K.; Achim, Z. Visualizing Count Data Regressions Using Rootograms. Am. Stat. 2016, 70, 296–303.
28. Zeileis, A.; Kleiber, C.; Jackman, S. Regression Models for Count Data in R. J. Stat. Softw. 2008, 27, 1–25.

29. Koenker, R.; Hallock, K.F. Quantile Regression. J. Econ. Perspect. 2001, 15, 143–156.
30. Kaufman, B.G.; Klemish, D.; Kassner, C.T.; Reiter, J.P.; Li, F.; Harker, M.; O'Brien, E.C.; Taylor, D.H.; Bhavsar, N.A. Predicting Length of Hospice Stay: An Application of Quantile Regression. J. Palliat. Med. 2018, 21, 1131–1136.
31. Machado, J.A.F.; Silva, J.M.C.S. Quantiles for Counts. J. Am. Stat. Assoc. 2005, 100, 1226–1237.
32. Sudarat, S.; Taweesak, C. Quantile Regression: An Alternative Statistical Model for Predicting Length of Stay of Diabetes Patients. In Proceedings of the International Conference on Natural Science, Engineering, and Technology, Sapporo, Japan, 19–21 January 2019.
33. Nelder, J.A.;Wedderburn, R.W.M. Generalized Linear Models. J. R. Stat. Soc. 1972, 135, 370–384.
34. R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2018. Available online: https://www.R-project.org (accessed on 22 December 2021)
35. Jackman, S. Pscl: Classes and Methods for R Developed in the Political Science Com-Putational Laboratory, Stanford University; R Package Version 0.95; Department of Political Science, Stanford University: Stanford, CA, USA, 2008.
36. Bozdogan, H. Akaike's information criterion and recent developments in information complexity. J. Math. Psychol. 2000, 44, 62–91.
37. Vrieze, S.I. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol. Methods 2012, 17, 228–243.
38. Vuong, Q.H. Likelihood ratio tests for model selection and non-nested hypotheses. Econom. J. Econom. Soc. 1989, 57, 307–333.
39. Geraci, M. rq.counts: Quantile Regression for Counts. 2019. Available online: https://rdrr.io/rforge/Qtools/man/rq.counts.html (accessed on 20 October 2021).
40. International Severe Acute Respiratory and Emerging Infections Consortium (ISARIC). COVID-19 Report. 2020. Available online: https://media.tghn.org/medialibrary/2020/05/ISARIC_Data_Platform_COVID19_Report_27APR20.pdf (accessed on 22 December 2021).
41. Shanley, L.A.; Lin, H.; Flores, G. Factors associated with length of stay for pediatric asthma hospitalizations. J. Asthma 2015, 52, 471–477.
42. Gardner, R.; Smith, G.A.; Chany, A.M.; Fernandez, S.A.; McKenzie, L.B. Factors associated with hospital length of stay and hospital charges of motor vehicle crash related hospitalizations among children in the United States. Arch. Pediatr. Adolesc. Med. 2007, 161, 889–895.
43. Matin, B.K.; Rezaei, S.; Karyani, A.K. Factors associated with length of stay and hospital charges among pediatric burn injury in Kermanshah, West of Iran. Int. J. Pediatr. 2015, 3, 403–409.

44. Hospital Performance: Length of Stay. Available online: https://www.audit.vic.gov.au/report/hospital-performance-length-stay (accessed on 22 December 2021).
45. Fine, M.J.; Pratt, H.M.; Obrosky, D.S.; Lave, J.R.; McIntosh, L.J.; Singer, D.E.; Coley, C.M.; Kapoor, W.N. Relation between Length of Hospital Stay and Costs of Care for Patients with Community-acquired Pneumonia. Am. J. Med. 2000, 109, 378–385.
46. White, B.A.; Biddinger, P.D.; Chang, Y.; Grabowski, B.; Carignan, S.; Brown, D.F. Boarding Inpatients in the Emergency Department Increases Discharged Patient Length of Stay. J. Emerg. Med. 2013, 44, 230–235.

CHAPTER SIX

Predicting the risk of nonsurgical premature menopause (NSPM) in survivors of childhood cancer in the presence of competing risks

This chapter is primarily rooted in my visiting PhD student project at the School of Public Health, University of Alberta, Canada. As of now, our work has not been finalized, we are still working on it, we applied and independently validated risk prediction models for nonsurgical premature menopause (NSPM) with competing risks among survivors of childhood cancer over time.

6.1. Introduction

6.1.1. Importance of addressing the risk of nonsurgical premature menopause (NSPM) in survivors of childhood cancer

The landscape of childhood cancer survivorship has witnessed a substantial improvement in recent decades, leading to a surge in the prevalence of chronic conditions among survivors. Notably, female childhood cancer survivors face an elevated risk of experiencing nonsurgical premature menopause (NSPM) attributable to treatment-related toxicities. NSPM occurs when ovarian function persists for a minimum of 5 years post-cancer diagnosis, yet menopause occurs naturally before reaching the age of 40. This condition can significantly impede the quality of life and curtail the potential reproductive years for affected individuals. During treatment, patients are exposed to chemotherapy and radiation designed to kill cancer cells; however, these harsh exposures can also damage the surrounding healthy cells. The main task is to develop a landmark predictive model that will predict the probability of cancer patients developing NSPM based on the type and magnitude of exposure. This probability is a crucial component in determining if and when fertility procedures are needed. Medical progress in treating childhood cancer has notably elevated survival rates, yet the life-saving interventions involved frequently entail enduring side effects. Among these, a significant impact is observed on reproductive health, specifically in the form of NSPM. Recognizing the risk factors associated with NSPM becomes pivotal for implementing preventive measures or timely interventions. Our objective was to facilitate informed discussions among physicians, patients, and their families regarding the imperative nature of fertility preservation measures.

We aimed to achieve this by developing predictive algorithms capable of assessing the likelihood of patients developing NSPM.

6.1.2. Methodological Review/Predictive modeling of the risk of NSPM

Extensive research has been undertaken on the various risk factors for the development of NSPM in CCSs following cancer treatment completion (Green,2009). Currently, researchers have been focusing on improving cancer survivors' health and well-being by generating new knowledge, developing and testing supportive care interventions, and evaluating novel models of care. Identifying treatment risk factors associated with compromised reproductive function following cancer treatment or therapy has received particular attention from researchers (Oktem,2018; Thomas,2015). The landmark model by (Van Houwelingen, 2007); this model does not use all available information but rather estimates the survival probability from a specified point in time, known as the "landmark time." For each landmark time, new survival predictions are made based on data up to that point, ignoring the data collected after the landmark time. This model is useful for dynamic prediction, where predictions are updated at several time points as new information becomes available. The regression calibration model by Ye (Ye, 2008);This model does not account for dropout information, which means it does not adjust for data that might be missing due to subjects leaving a study prematurely. When some variables are measured with error, the regression coefficients are adjusted to better reflect the true relationships between variables. The Joint model (Rizopoulos, 2011; Proust-Lima, 2009);This model enables the simultaneous estimation and modelling of both the change over time in a repeated measures outcome (longitudinal process) and the time until an event of interest occurs (survival process), allowing for the assessment of the relationship between the two processes within a single unified framework. A notable study (Cindy, 2023) developed and validated risk prediction models for age-specific primary ovarian insufficiency (POI) in long-term survivors of childhood cancer. In this study, POI is defined as the cessation of gonadal function before the age of 40 years and includes both AOF and NSPM. SPM was treated as a competing risk because once a female childhood cancer survivor experiences SPM, she is no longer at risk of developing other conditions. While a study might be designed to specifically observe a particular event within a population, the attainment of this objective can be hindered by the

occurrence of other events that impede the individual from witnessing the intended event of interest. These intervening events are referred to as competing risk events (Austin,2017).

Survival analysis is used for the examination of time-to-event data, where subjects are expected to undergo only one type of event during follow-up, such as death from any form of cancer. In contrast, real-life scenarios may involve subjects potentially experiencing more than one type of a particular event. Events of this nature arise in clinical research when there are multiple possible outcomes during the follow-up period for survival data, and the occurrence of one outcome of interest may be prevented by another; this is referred to as a competing event . The probabilities associated with these events are termed competing risks, and models for competing risks (Satagopan, 2004; Haller, 2013) are the optimal choice for analyzing such data. The primary concern revolves around how to address individuals with competing risks. This necessitates the use of a critical or appropriate statistical methodology to obtain an accurate message or interpretation for the problem or research question at hand. Furthermore, in cases where data typically exhibit a high rate of censoring, the reliability of traditional models, such as the Cox proportional hazards (PH) regression model, may be compromised (Hsich, 2011). In certain studies, all covariates are measured at baseline, and none of them are considered time-varying covariates, even though their effects may change over time. Consequently, more flexible models are required to accommodate such situations. The regression models for competing risks data have been extensively reviewed elsewhere (Kim, 2007; Tai, 2011). Similarly in a specific way; Random survival forest (Ishwaran, 2008), Cause-specific hazard model (Prentice,1978), Fine-Gray regression model(Fine and Gray ,1999) and time-specific Logistic regression with competing risks (Vock,2016; Robins,1995) were the most frequently used models for this kind of setting.

6.1.3. Aims

Predict the risk factors associated with NSPM before prespecified ages among female childhood cancer survivors, enabling early intervention and improved quality of life. As a late effect of cancer treatments, NSPM can manifest at any point during a survivor's young adulthood, especially after reaching the 5-year survival milestone. The main objective is to

facilitate deeper conversations among physicians, patients, and their families regarding the significance of fertility preservation. To achieve this, we will design landmark predictive models that assess the likelihood of patients developing NSPM.

6.2. Materials and methods

6.2.1. Study design and population

Data was acquired for 7,891 female participants of the Childhood Cancer Survivor Study (CCSS), an ongoing multi-institutional North American cohort study of 5-year survivors of childhood cancer treated between 1970-1999. Female survivors aged ≥18 years at their latest follow-up with self-reported menstrual history(MH) information and free of subsequent malignant neoplasms (SMNs) within 5 years of diagnosis were included. In CCSS, follow-up concluded on 11/25/2016.

6.2.2. Outcome variables

The Outcome variables are the ovarian status & age at event. By assessing the risk of developing menopause at different pre-specified ages, clinicians and female childhood cancer survivors can take action to reduce the potential impact of the late effects of cancer treatment based on their age rather than by time since diagnosis. Individuals are at risk for NSPM until age 40 (or until experiencing a competing event) according to the age at which they completed the most recent survey with MH information.

Data from responses to the survey questions assisted in classifying patients as having acute ovarian failure (AOF), NSPM, surgical premature menopause (SPM) or neither. It is possible to determine a positive diagnosis of AOF by comparing dates and answers indicating the loss of menstruation within 5 years of diagnosis. Individuals with NSPM are those who retained normal ovarian function for at least 5 years after treatment, but did not menstruate for at least 6 months before age 40. As an outcome event, NSPM diagnoses patients with normal ovarian function after 5 years of treatment but no menstruation for 6 months before turning 40, and is censored if the patient exceeded 40 years old without experiencing menopause, or was no longer under follow-up, and had a competing risk for SPM or premature menopause.

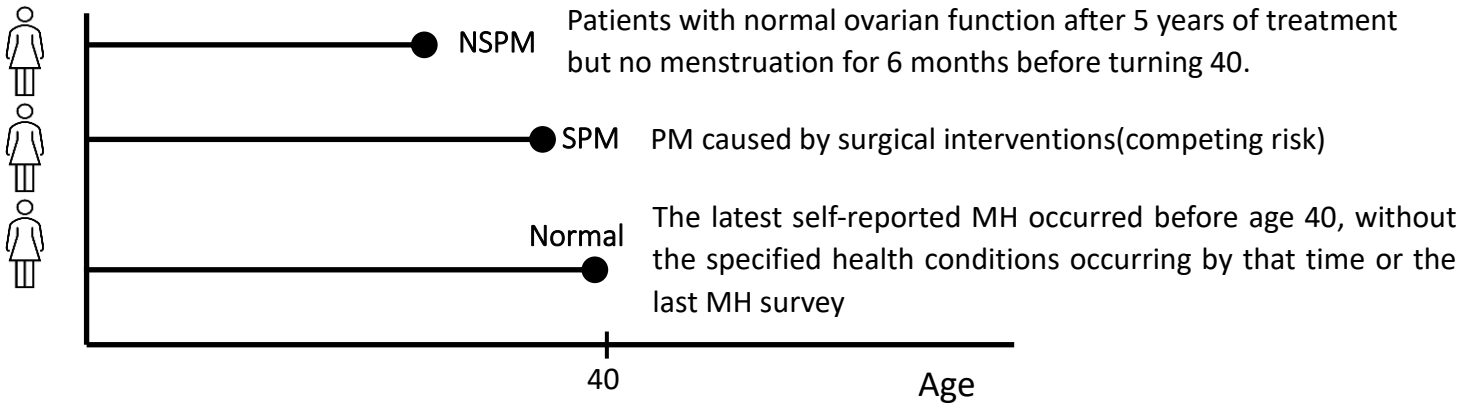


Figure 1: Ovarian status: Normal or other than SPM and NSPM, NSPM as event of interest , and SPM as a competing risk. The focus here lies in exploring effective strategies for managing individuals who are susceptible to competing risks.

Participants were considered at risk for NSPM from their entry into the study (5 years post-diagnosis) up to the point they either developed NSPM, SPM, reached the age of 40, or were no longer followed up. Individuals are at risk for NSPM until age 40 (or until experiencing a competing event) according to the age at which they completed the most recent survey with Menstrual History (MH) information.

6.2.3. Independent or risk variables

Candidate predictors for considered model are shown in Table 1. For the ten alkylating agents and their derived cumulative dose Cyclophosphamide equivalent dose (CED), the information would be redundant if included all of them in one model, therefore we should apply two separate model, one with CED and another ten individual alkylating agents.

The risk factors (treatment) include:

Demographic;

- Age at the event or attained age,
- Age at diagnosis,
- Race/ethnicity (White, Black, Other)

Primary cancer diagnosis or type (8 levels);

- leukemia, central nervous system cancers, neuroblastoma, non-Hodgkin lymphoma, Hodgkin lymphoma, kidney cancer, bone tumors, soft-tissue sarcoma

Bone marrow transplant (BMT)(Yes/No);

Chemotherapy;

- Alkylating agents,
- Heavy metals,
- Non-classical alkylators,

Radiation;

- Ovarian radiation,
- Doses of abdominal/pelvic,
- Total Body Irradiation (TBI),

Categorical variables	
Diagnosis (8 levels):	
	Leukemia(reference),
	Central nervous system cancers,
	Neuroblastoma,
	Non-Hodgkin lymphoma,
	Hodgkin lymphoma,
	Kidney cancer,
	Bone tumors,
	Soft-tissue sarcoma
Race (3 levels) :	
	White(reference),
	Black,
	Other
BMT	
	Yes
	No
Continuous variables:	
Age at diagnosis	
Irradiation dose (Gy)	
• total body irradiation dose (tbidose)	
• minimum ovary radiation dose (minovary)	
• radiation dose to pituitary(pitdose)	
Alkylating agents' doses (g/m2)	
Cyclophosphamide equivalent dose (CED) vs	BCNU(Carmustine), Busulfan, CCNU(Lomustine), Chlorambucil, Cyclophosphamide, Ifosfamide, Melphalan , Nitrogen Mustard, Procarbazine, Thiotepa
Other chemotherapy agents' doses (g/m2)	
• Carboplatin	
• Cis_Platinum	
• Bleomycin	
• Daunorubicin	
• Doxorubicin	
• Epirubicin	
• Idarubicin	
• Methotrexate	
• Mitoxantrone	
• VM 26(Teniposide)	
• VP 16(Etoposide)	
Interaction (for Case-specific hazard and Fine-Gray regression models):	
• Age at Diagnosis and BMT	
• Age at Diagnosis and Minimum Ovarian RT Dose	

Table1 : Candidate predictors for model development

6.2.4. Model development

Late-onset post-cancer treatment complications, known as PM, can occur at any point in the young adulthood of survivors after reaching the 5-year survival mark. Our focus lies in

predicting the occurrence of NSPM over time. During model estimate, we evaluated potential treatment exposure predictors and other types and doses for both chemotherapy and radiation, we had also other menstrual history Information. Recognizing the limitations of relying on a singular model, we utilized a comprehensive approach to evaluate the risk of NSPM having the competing risk. The regression models for competing risks data have been extensively reviewed elsewhere (Kim, 2007; Ishwaran, 2008; Tai, 2011). We provided a brief overview as follows:

1. Cause-specific hazard (CSH) model:
2. Fine-Gray regression (FGR) model,
3. Random Survival Forest with Competing Risks (RSF-CR),

6.2.5. Model Performance and Evaluation

Model performance and accuracy were measured using the time-specific area under the ROC curve (AUCt), the time-specific average positive predictive value (APt), and calibration curves on both the training and validation sets. We also applied a scaled Brier score for each model to compare their performance. The evaluation of corresponding risk prediction models developed using time-to-event data typically requires weighting observations with inverse probability-of-censoring weights to account for censoring, e.g., the time-specific area under the ROC curve (Heagerty,2000), the time-specific average positive predictive value (Yuan,2018) and the Brier score(Steyerberg, 2010).

6.3. Preliminary results

The distribution of ovarian status focuses on individuals within the most fertile age range, specifically between the ages of 21 and 40. However, we have also incorporated data from individuals in other age groups for comprehensive model estimate. This inclusive approach allows for a more robust and representative analysis, ensuring that the model accounts for a broader spectrum of ages and ovarian statuses, thus enhancing its applicability and reliability across diverse age cohorts.

Cut-off age years	Normal	NSPM	SPM	Total
21	153	15	3	171
22	168	10	2	180
23	191	12	1	204
24	185	21	4	210
25	236	18	5	259
26	233	8	12	253
27	219	9	7	235
28	261	12	13	286
29	291	10	12	313
30	280	23	19	322
31	293	13	20	326
32	298	20	24	342
33	244	17	14	275
34	310	12	25	347
35	266	22	24	312
36	279	13	23	315
37	257	12	14	283
38	218	16	24	258
39	205	14	15	234
Total	4587	277	261	5125

Table 2: Ovarian status distribution at different age thresholds(i.e. fertile age groups)

Characteristic	Overall count
Age at Cancer Diagnosis	
< 5	2317
5 – 9	1789
10 – 14	1506
≥ 15	1723
Primary cancer diagnosis/cancer type	
Leukemia	2416
HD	1075
Kidney (Wilms)	879
Bone cancer	746
CNS	720
Neuroblastoma	604
HNL	497
Soft tissue sarcoma	398
Cyclophosphamide Equivalent Dose, mg/m2	
Median(IQR)	5430(6361)
None(dose=0)	3959
< 4000	999
4000 - 7999	778
≥ 8000	847
Missing	752
Ovarian Radiation Dose, Gy	
None(Gy = 0)	3626
< 5	972
6 - 20	559
≥ 20	1550
Missing	650
Bone Marrow Transplant	
No	7154
Yes	176

Table 3: Characteristics of the Survivor Cohort

Figure 2 displays non-parametric estimates of cumulative incidence functions (CIF) for ovarian status, y-axis the proportion of event and In the x-axis, time = age (years). The CIF probability of PM is higher than that of SPM. There is a consistent pattern or change after the age of 40 and up to around 10 years thereafter.

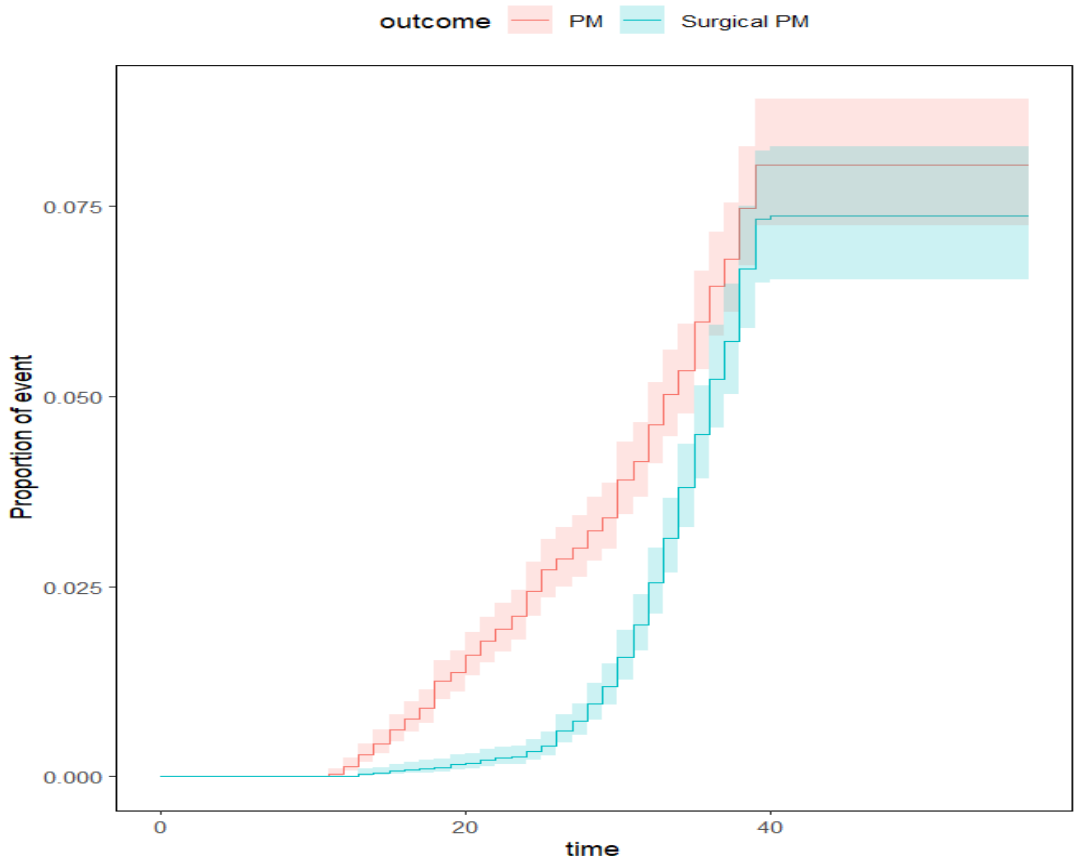


Figure 2: Cumulative incidence function curves for events, y-axis the proportion of event and In the x-axis, time = age (years).

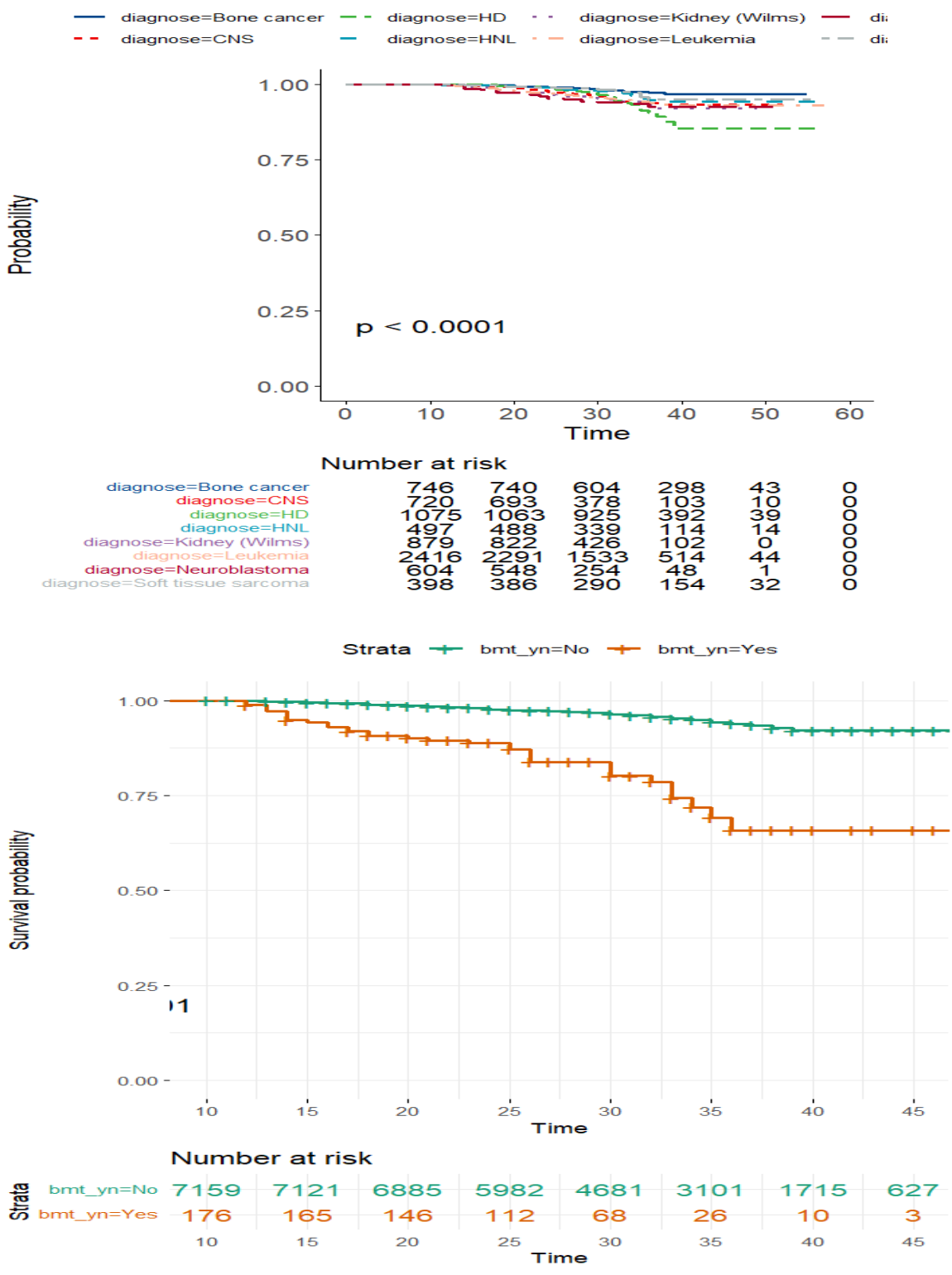


Figure 3: Survival plot with a table showing the number at risk over time: The y-axis It represents the proportion of patients surviving at any given time point after diagnosis. The number at risk shows the number of patients ('at risk') who have not yet had the event. We

utilized cumulative incidence function (CIF) instead of Kaplan-Meier (KM) curve in the presence of competing risks. Figure 3 displayed the count of individuals at risk for each feature category.

Table 3 presents Hazard Ratios (HR) along with their corresponding confidence intervals and p-values derived from the CSH model for HR (multivariate) and FG for HR (competing risks multivariate) outcomes. The univariate HR examines each variable independently to assess its association with the risk of the event without considering potential confounding effects from other variables. It addresses the question: "How does this individual factor relate to the risk of the event?" On the other hand, the multivariate HR is computed by incorporating all variables into a single model. This analytical approach helps to account for confounding variables, offering estimates of each variable's impact while holding all other factors constant. The inquiry shifts to: "How does this factor influence the risk of the event when accounting for all other factors under consideration?" The term "all" in this context encompasses the entire study population or sample analyzed in this statistical output. This variable indicates whether an individual has experienced the event of interest, irrespective of menopause type.

Dependent: Surv(age, event == 1)		all	HR (univariable)	HR (multivariable)	HR (competing risks multivariable)
age_dx	Mean ± SD	8.5 ± 5.9	0.98 (0.96-0.99, p=.010)	0.98 (0.96-1.00, p=.117)	0.98 (0.97-1.00, p=.110)
minovary	Mean ± SD	75.1 ± 308.2	1.00 (1.00-1.00, p<.001)	1.00 (1.00-1.00, p<.001)	1.00 (1.00-1.00, p<.001)
ced5	Mean ± SD	3592.9 ± 6458.2	1.00 (1.00-1.00, p=.007)	1.00 (1.00-1.00, p=.017)	1.00 (1.00-1.00, p=.012)
carboplatin_yn	No	7217 (98.4%)			
	Yes	118 (1.6%)	1.65 (0.73-3.69, p=.227)	0.97 (0.41-2.28, p=.938)	0.97 (0.40-2.36, p=.940)
cis_platinum_yn	No	6887 (93.9%)			
	Yes	448 (6.1%)	0.63 (0.37-1.07, p=.085)	0.55 (0.31-0.96, p=.035)	0.55 (0.32-0.95, p=.032)
bleomycin_yn	No	6767 (92.3%)			
	Yes	568 (7.7%)	1.02 (0.72-1.43, p=.919)	1.34 (0.90-1.98, p=.150)	1.33 (0.91-1.94, p=.140)
daunorubicin_yn	No	6251 (85.2%)			
	Yes	1084 (14.8%)	1.22 (0.93-1.61, p=.147)	0.89 (0.65-1.22, p=.475)	0.89 (0.66-1.20, p=.460)
doxorubicin_yn	No	4683 (63.8%)			
	Yes	2652 (36.2%)	0.92 (0.75-1.14, p=.439)	0.88 (0.69-1.12, p=.285)	0.88 (0.68-1.13, p=.300)
idarubicin_yn	No	7288 (99.4%)			
	Yes	47 (0.6%)	2.96 (1.11-7.95, p=.031)	1.39 (0.45-4.24, p=.567)	1.38 (0.39-4.94, p=.620)
mitoxantrone_yn	No	7308 (99.6%)			
	Yes	27 (0.4%)	3.91 (1.62-9.44, p=.002)	2.49 (0.93-6.67, p=.070)	2.48 (0.86-7.16, p=.094)
vm_26_yn	No	7035 (95.9%)			
	Yes	300 (4.1%)	1.51 (0.97-2.35, p=.066)	1.41 (0.88-2.25, p=.151)	1.41 (0.88-2.26, p=.160)
vp_16_yn	No	6431 (87.7%)			
	Yes	904 (12.3%)	1.72 (1.30-2.28, p<.001)	1.19 (0.83-1.69, p=.339)	1.19 (0.84-1.68, p=.330)
bmt_yn	No	7159 (97.6%)			
	Yes	176 (2.4%)	5.94 (4.17-8.47, p<.001)	5.59 (2.87-10.89, p<.001)	5.57 (2.73-11.34, p<.001)
race_3	Black	443 (6.0%)			
	Other	474 (6.5%)	1.05 (0.60-1.84, p=.854)	1.01 (0.58-1.77, p=.963)	1.02 (0.58-1.77, p=.960)
	White	6418 (87.5%)	0.88 (0.57-1.34, p=.539)	0.88 (0.57-1.34, p=.551)	0.88 (0.59-1.32, p=.540)
smoke_100	No	5481 (74.7%)			
	Yes	1854 (25.3%)	0.96 (0.77-1.21, p=.754)	0.99 (0.79-1.25, p=.933)	0.99 (0.78-1.25, p=.930)
age_dx:bmt_yn	No		0.97 (0.95-0.99, p<.001)		
age_dx:bmt_yn	Yes		1.08 (1.05-1.12, p<.001)	0.96 (0.90-1.02, p=.217)	0.96 (0.90-1.02, p=.210)
age_dx:minovary			1.00 (1.00-1.00, p<.001)	1.00 (1.00-1.00, p=.006)	1.00 (1.00-1.00, p=.011)

n=7335, events=391, Likelihood ratio test=143.96 on 18 df(p<.001)

Table 3: Hazard ratios (HR) with their confidence intervals, and p-values in the CS and FG results

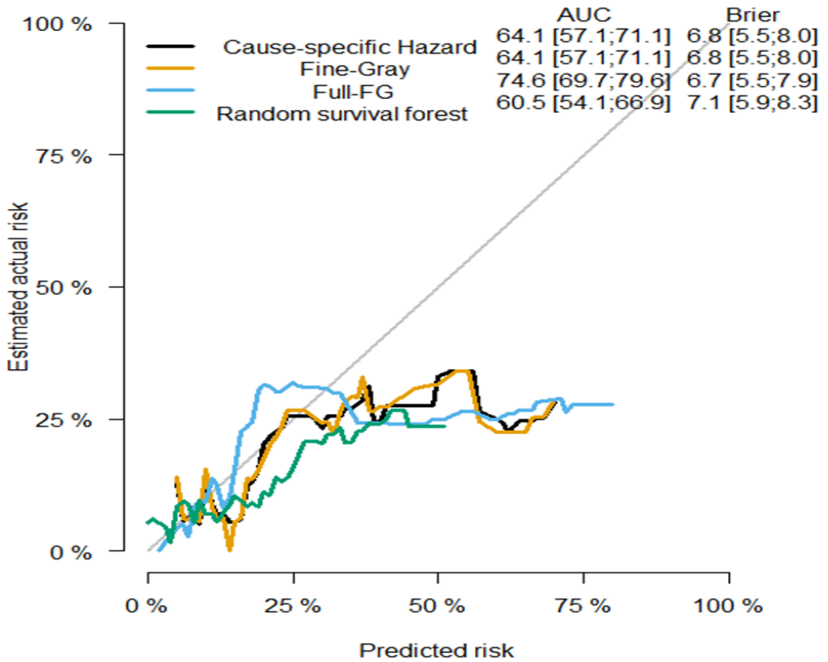


Figure 4: Calibration curves for the cause-specific hazards model, Fine-Gray model , Fine-Gray with full model, and the Random survival forest model. AUC and Brier score are expressed as the point estimates and 95% confidence intervals. Cancer treatment and cancer type are highly correlated, so we excluded the cancer type from the analysis of CSH and FGR, while including the full FGR and RSF models.

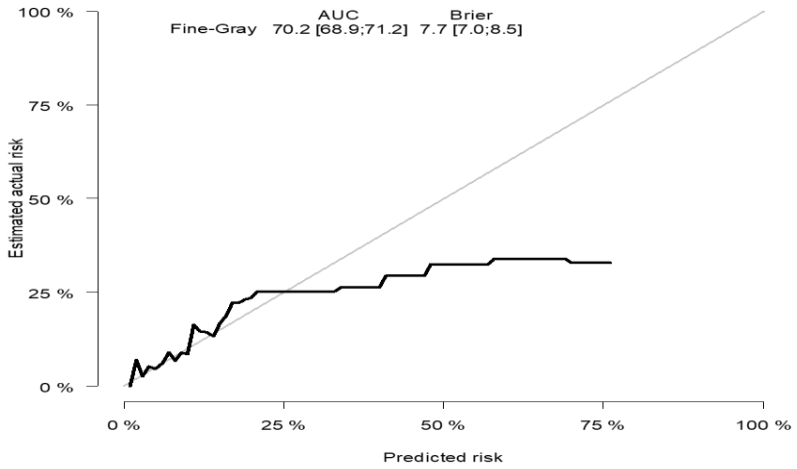


Figure 5 : Calibration curves obtained by using cross-validation method, which shows the calibration plots for the full Fine-Gray model at the age of 40.

Figure 6 showed the Brier scores for the models. The RSF model is represented by a dashed line, culminating in the highest Brier score at the final time point, potentially suggesting reduced accuracy over time compared to the others. The Cause-specific Hazard and Fine-Gray models exhibit similar performance, with their lines almost overlapping. As time advances, the Brier scores for all models rise, signifying an increased difficulty in accurately predicting the event as age goes on.

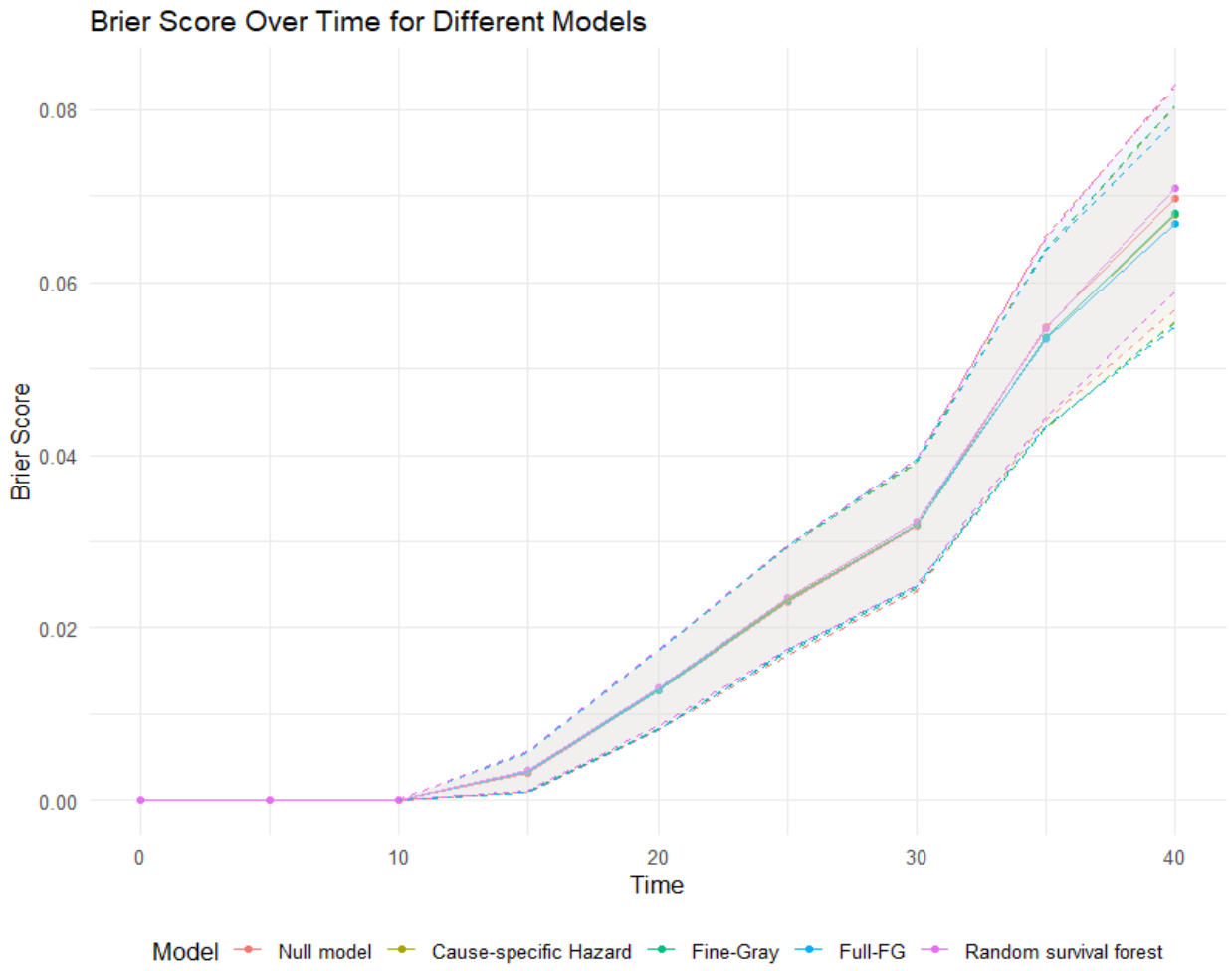


Figure 6: Brier score for the models

We will conduct a more in-depth analysis, compare models, and engage in further discussion.

References:

1. Oktem, O., Kim, S. S., Selek, U., Schatmann, G. & Urman, B. Ovarian and Uterine Functions in Female Survivors of Childhood Cancers. The Oncologist 23, 214–224; 10.1634/theoncologist.2017-0201 (2018).
2. Thomas-Teinturier, C. et al. Ovarian reserve after treatment with alkylating agents during childhood. Human reproduction (Oxford, England) 30, 1437–1446; 10.1093/humrep/dev060 (2015).
3. Van H., H.C., Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics*, vol. 34, no. 1, 2007, pp. 70–85. *JSTOR*,
4. Proust-Lima C, Taylor JM, Williams SG, Ankerst DP, Liu N, Kestin LL, Bae K, Sandler HM. Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. International Journal of Radiation Oncology* Biology* Physics. 2008; 72(3):782–791.
5. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978;34:541–554.
6. Cindy Im, Zhe Lu, Sogol Mostoufi-Moab, Angela Delaney, Lin Yu, Jessica L Baedke, Yutong Han, Yadav Sapkota, Yutaka Yasui, Eric J Chow, Rebecca M Howell, Smita Bhatia, Melissa M Hudson, Kirsten K Ness, Gregory T Armstrong, Paul C Nathan, Yan Yuan, Development and validation of age-specific risk prediction models for primary ovarian insufficiency in long-term survivors of childhood cancer: a report from the Childhood Cancer Survivor Study and St Jude Lifetime Cohort, The Lancet Oncology, Volume 24, Issue 12,2023, Pages 1434-1442, ISSN 1470-2045, https://doi.org/10.1016/S1470-2045(23)00510-7
7. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology. 2010;21(1):128-138.
8. Fine, Jason P., and Robert J. Gray. “A Proportional Hazards Model for the Subdistribution of a Competing Risk.” Journal of the American Statistical Association, vol. 94, no. 446, 1999, pp. 496–509. JSTOR, https://doi.org/10.2307/2670170
9. Green DM, Sklar CA, Boice JD, et al. Ovarian Failure and Reproductive Outcomes After Childhood Cancer Treatment: Results from the Childhood Cancer Survivor Study. Journal of Clinical Oncology. 2009;27(14):2374-2381.
10. Yuan Y, Zhou QM, Li B, Cai H, Chow EJ, Armstrong GT. A threshold-free summary index of prediction accuracy for censored time to event data. Statistics in Medicine. 2018;37(10):1671-1681
11. Rizopoulos et al. (2011). Biometrics, 67(3), 819–829.
12. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risks. Statistics in Medicine. 2017;36:4391-4400.
13. Ishwaran et al. (2008). The Annals of Applied Statistics, 2(3), 841–860.

14. Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. Biometrics. 2000;56:337-344.
15. Ye W, Lin X, and Taylor JM (2008). Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach. Biometrics, 64(4), 1238–1246.
16. Tai BC, Grundy R, Machin D. On the importance of accounting for competing risks in pediatric brain cancer: II. Regression modeling and sample size. Int J Radiat Oncol Biol Phys. 2011;79:1139–1146.
17. Robins JM, Rotnitzky A. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. Journal of the American Statistical Association. 1995;90(429):122-129. doi:10.1080/01621459.1995.10476494
18. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circ Cardiovasc Qual Outcomes 2011;4:39-45.
19. Satagopan JM, Ben-Porat L, Berwick M, et al. A note on competing risks in survival data analysis. Br J Cancer 2004;91:1229-35.
20. Haller B, Schmidt G, Ulm K. Applying competing risks regression models: an overview. Lifetime Data Anal 2013;19:33-58.
21. Vock DM, Wolfson J, Bandyopadhyay S, et al. Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. Journal of Biomedical Informatics. 2016;61:119-131. doi:10.1016/j.jbi.2016.03.009
22. Kim HT. Cumulative incidence in competing risks data and competing risks regression analysis. Clin Cancer Res. 2007;13:559–565.

CHAPTER SEVEN

Overall Conclusion and Recommendation

Essentially, these studies support the idea that data is a powerful ally in improving patient outcomes and more effectively providing healthcare services in healthcare delivery. The healthcare industry is poised to usher in a new era of precision, efficiency, and resilience by embracing and refining these data-driven approaches. The comprehensive nature of these studies demonstrates the versatility and effectiveness of statistical and machine learning models in addressing diverse healthcare challenges. The findings not only contribute to the academic understanding of these issues but also have practical implications for improving patient outcomes, resource allocation, and decision-making in healthcare settings. The integration of data-driven approaches into healthcare research and practice is crucial for advancing the field and enhancing the overall quality of healthcare services. The implementation of these models into healthcare decision-making processes represents a crucial step toward realizing the full benefits of the big data era in healthcare.

The following summarizes the usefulness of these studies:

1. Unlocking Insights for Patient Wellbeing:

The application of predictive models in nonsurgical premature menopause, emergency department admissions, and specialized departments exemplifies how data-driven insights can unlock a deeper understanding of patient trajectories. By combining patient-generated information with official records, these models provide a comprehensive view of health journeys, aiding in the identification of risks and improving overall patient wellbeing.

2. Optimizing Healthcare Operations:

The studies on Length of Stay (LoS) prediction showcase the potential of machine learning to optimize healthcare operations. By accurately forecasting LoS and PLoS, these models act as valuable decision-support tools, allowing healthcare providers to allocate resources efficiently, streamline processes, and enhance the overall performance of healthcare systems.

3. Navigating Public Health Challenges:

The exploration of spatio-temporal patterns in the diffusion of SARS-CoV-2 demonstrates the broader societal impact of data-driven approaches. By accounting for space, sociodemographic characteristics, and health conditions, the study not only provides valuable insights into the dynamics of a pandemic but also informs targeted interventions, contributing to public health resilience.

4. Tailoring Care in the Face of Global Health Crises:

The investigation into COVID-19-related hospital admissions exemplifies the adaptability of statistical methods to address evolving health crises. By identifying risk factors and modeling the length of stay, these studies contribute essential knowledge for tailoring care strategies, optimizing hospital resources, and managing patient flow during a global health emergency.

5. The Future Landscape of Healthcare:

As we stand at the intersection of healthcare and the big data era, the demonstrated potential of advanced statistical and machine learning models offers a glimpse into the future landscape of healthcare. The integration of these models into routine decision-making processes has the potential to revolutionize patient care, improve operational efficiency, and foster a healthcare ecosystem that is both responsive and patient-centric.