Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics and Management Science

Predicting the length of stay of nonsurgical patients

Jaco de Hoog (572567jh)

Supervisor:	dr. Paul Bouman
Second assessor:	dr. Olga Kuryatnikova
Date final version:	22nd July 2024

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Abstract

A length of stay (LoS) prediction for a patient provides the capacity managers in a hospital an indication how long a bed could be occupied. If it is possible to construct accurate predictions for the LoS, hospitals can better plan and manage their bed occupancy, which can lead to more efficient use of available resources. Eventually, capacity managers are able to anticipate when a bed becomes available again, which could smoothen the patient flow in the hospital. This research focuses on deriving the most determining factors for the LoS and evaluating whether it is possible to predict the LoS of nonsurgical patients. The data used in this research contains the patient admissions of a large hospital organization in the Netherlands. In order to evaluate the predictability of the LoS, several statistical- and machine learning models are applied. The results of this empirical research indicate that the admission type, the assigned specialism, and the number of mutations are important factors for determining the LoS of nonsurgical patients. It is concluded that for a large part of the nonsurgical admissions it is possible to predict the LoS. The machine learning models Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM) are substantially better able to predict the LoS than the considered statistical models. Despite the fact that SVM is the most accurate model in regression, XGBoost is considered to be the most suitable model for a practical application, because the XGBoost model performs similarly to SVM, requires less computation time, and offers more flexibility.

Contents

1 Introduction **5**
- 1.1 Aim of the research . . . 6
- 1.2 Relevance and structure of the research . . . 7

2 Literature Review **7**
- 2.1 Determining factors length of stay . . . 8
- 2.2 Modelling length of stay . . . 9
- 2.3 Machine learning in healthcare . . . 10

3 Data **12**
- 3.1 Data source . . . 12
- 3.2 Descriptive statistics . . . 13

4 Methodology **16**
- 4.1 Effects on length of stay . . . 16
  - 4.1.1 Length of stay as a duration . . . 17
- 4.2 Predicting length of stay . . . 20
  - 4.2.1 Statistical models . . . 20
  - 4.2.2 Machine learning models . . . 21
  - 4.2.3 Training-test procedure . . . 23
  - 4.2.4 Performance evaluation . . . 25
  - 4.2.5 Model interpretation . . . 25

5 Results **26**
5.1 Modelling results 26
5.1.1 Linear regression models 26
5.1.2 Duration models 29
5.2 Prediction results 33
5.2.1 Classification 33
5.2.2 Regression 35
5.3 Prediction diagnostics 38
5.4 Possible explanations prediction deviations 42

6 Conclusion **43**

References **45**

A Estimated coefficients Cox's proportional hazards model **50**

B Hyperparameter tuning **51**

C Prediction diagnostics Random Forest and Support Vector Machine **54**

D Switched hyperparameters **55**

List of abbreviations

AFT accelerated failure time. 17–19

AI artificial intelligence. 10

ANN Artificial Neural Network. 11, 21

BGA brief geriatric assessment. 11

BN Bayesian Network. 11

BPTT backpropagation through time. 22

CABG coronary artery bypass graft. 9, 10

CDF cumulative distribution function. 17, 19

COPD chronic obstructive pulmonary disease. 8, 14, 15, 28, 32, 51

ED emergency department. 5, 8, 9

EHR electronic health record. 7, 11

GDP Gross Domestic Product. 5

GLM generalized linear model. 10, 20, 21, 33–39

HiX Healthcare information eXchange. 7, 17, 37, 39, 42–45, 54

ICU intensive care unit. 8, 10, 14, 15, 28, 32, 51

IQR interquartile range. 13

LIME Local Interpretable Model-agnostic Explanations. 11, 12, 25, 26

LoS length of stay. 1, 6–21, 23–45, 50, 51, 54–56

LSTM Long Short-Term Memory. 22

MAE mean absolute error. 22, 25, 35, 36, 38, 39, 56

MLE maximum likelihood estimation. 19

MLP multilayer perceptron. 11

MSE mean squared error. 22

OLS ordinary least squares. 9, 16, 28, 29

PDF probability density function. 19

PH proportional hazards. 10, 16–19, 29, 30, 33, 50

QP quadratic programming. 22, 54

RF Random Forest. 1, 11, 21, 34–40, 44, 52, 54, 55

RNN Recurrent Neural Network. 21, 22, 24, 34–36, 38, 39, 52, 53, 55, 56

SHAP SHapley Additive exPlanations. 11, 12, 25, 26, 41, 42, 44

SVM Support Vector Machine. 1, 11, 21–23, 34–40, 44, 53–55

WAPE weighted absolute percentage error. 25, 35, 36, 38, 39, 56

XGBoost eXtreme Gradient Boosting. 1, 21, 33–42, 44, 52, 54–56

Nonsurgical specialisms			
CAR	Cardiology	**MDL**	Gastrointestinal diseases
DER	Dermatology	**NEU**	Neurology
GER	Geriatrics	**PSY**	Psychiatry
INT	Internal medicine	**REU**	Rheumatology
KIN	Pediatrics	**REV**	Rehabilitation
LON	Pulmonary medicine		

1 Introduction

The surge of demand for healthcare services together with the increase in expenditures results in added pressure on the hospitals in the Netherlands (Hulshof et al., 2012). In total, there were 2.8 million patient admissions in the Netherlands in 2021 (CBS, 2023a). Due to the aging of the population, the demand for healthcare is increasing strongly (Rijksoverheid, 2023). In 2022, the total costs of health and welfare in the Netherlands were equal to 126.2 billion euros (CBS, 2023b). This came down to 13.2% of the Dutch Gross Domestic Product (GDP). It is expected that the costs of health and welfare keep increasing and that the total costs will cover more than 18% of the Dutch GDP in 2060 (Aalbers & Roos, 2022).

To improve patient care and limit operational costs, an efficient occupancy of the admission ward is crucial for hospitals. For patient admissions, a distinction can be made between elective and emergency admissions. These patient admissions are assigned to a surgical or a nonsurgical ward, which also makes it possible to differentiate between surgical and nonsurgical admissions, respectively.

The elective patients have an appointment with a doctor at the outpatient clinic. During this appointment, the doctor can decide that the patient needs to be admitted to the hospital or requires a surgery. If it is decided that a surgery is required, the patient is put on a waiting list to receive a surgery. Eventually, surgery is scheduled for this patient and the patient will be admitted to a surgical ward on the scheduled date. This admission is considered a surgical admission. For patients belonging to the nonsurgical patient flow, the doctor decides that the patient needs to be admitted to the hospital. Usually, a nonsurgical patient is admitted within a few days or weeks and gets assigned to a nonsurgical ward.

An emergency patient comes in through the emergency department (ED), where a doctor can decide that the patient needs to be admitted or requires a surgery. In case of an emergency patient, the admission process is similar for both the surgical and nonsurgical patient flow. In large hospitals, the admitted patient goes to the acute admission ward, where the patient usually stays a maximum of 48 hours. If it turns out that the patient needs to stay longer than these 48 hours, the patient is admitted to another ward. This can be either a surgical or a nonsurgical ward depending on the fact whether a surgery is required. The admission process described above does not apply to smaller hospitals, because these hospitals generally do not have an acute admission ward. In these smaller hospitals, the patient is assigned directly to the appropriate ward. In addition, for nonsurgical emergency admissions in larger hospitals, it is not strictly necessary that the patient is first admitted to the acute admission ward and the patient could thus be assigned directly to a nonsurgical ward. It should be noted that unplanned admissions that unexpectedly require surgery after being assigned to a nonsurgical ward still belong to the nonsurgical patient flow.

It is clear that nonsurgical admissions consist of admissions that initially do not require surgery. The nonsurgical patient flow is an important component in healthcare, and it includes the following specialisms: cardiology (CAR), dermatology (DER), geriatrics (GER), internal medicine (INT), pediatrics (KIN), pulmonary medicine (LON), gastrointestinal diseases (MDL), neurology (NEU), psychiatry (PSY), rheumatology (REU), and rehabilitation (REV). The nonsurgical admissions are an unpredictable element in capacity management, because these

admissions often arise from emergency situations. Scheduled nonsurgical admissions are usually planned at short notice, which makes it important to have an idea when a bed could become available. In addition, it is difficult to estimate how long a nonsurgical patient needs to stay in the hospital, because the patient's condition is not always clear at time of admission. Hence, gaining insight into the length of stay (LoS) of nonsurgical patients is interesting, as it could eventually help optimize bed occupancy and streamline patient flow.

1.1 Aim of the research

This research focuses on analyzing and predicting the LoS of nonsurgical patients. In general, the LoS refers to the amount of time the patient stays in a hospital during a single admission (Huntley et al., 1998). This research considers the LoS as the number of hours that a patient who is allocated to a bed stays in a hospital during an admission. The aim of this research is to answer the following central research question:

"Is it possible to predict a nonsurgical patient's length of stay, and which of the considered predictive models is the most accurate?"

In order to determine whether it is possible to predict the LoS, several statistical models and machine learning techniques are considered. In addition to determining the most accurate predictive model, the suitability of the model for an application in healthcare is taken into account. Although a machine learning model may be able to achieve higher accuracy, a regression model is often preferred in healthcare. This is due to the straightforward interpretation of the coefficients, making it understandable how the prediction is made. This research evaluates whether the possible increase in accuracy due to the use of a machine learning model is substantial, such that a machine learning model could be preferred over a statistical model despite the fact that machine learning models generally lack interpretability.

During this research it is evaluated how the predictable and unpredictable components of the LoS can be distinguished. Certain factors that can cause a longer LoS only become known during the admission, e.g. a complication that occurs or a change in diagnosis. However, these factors cannot be included in a prediction framework for determining the expected LoS in advance. Hence, this research determines which possible factors are known at time of admission that can possibly explain the LoS and which of these variables are the most important for constructing a prediction. By considering various possible explanatory variables, insight is offered into how these variables influence the LoS, making it clear which variables should be included in a predictive model.

In order to assess the reliability of the LoS predictions, the predictive power of the models is evaluated based on several performance measures. In this research, predicting the LoS is both considered as a classification problem and a regression problem. It is evaluated how accurately the different models can classify the LoS in short, medium, and long LoS. This way, it is examined whether it is possible to provide a 'rough' estimate for the LoS with high accuracy. When considering LoS as a regression problem, the deviation from the actual LoS is determined. If hospitals choose to use exact predictions, an insight is provided in what the possible deviation from the actual LoS could be.

1.2 Relevance and structure of the research

This research is commissioned by ChipSoft, which is a company that develops innovative healthcare IT. The software provided by ChipSoft, called Healthcare information eXchange (HiX), is suitable for different types of healthcare institutions and manages patient-oriented treatment- and registration procedures. Since 2018 ChipSoft is the Dutch market leader in providing electronic health records (EHRs) (Medisch Contact, 2018), and in 2021 already 70% of the hospitals in the Netherlands was using HiX (M&I/Partners, 2021). The insights of this research are relevant for ChipSoft, because if it turns out that it is possible to predict the LoS, it gives ChipSoft the possibility to incorporate a prediction framework into their software. This allows ChipSoft to advise on the expected LoS, which could make it easier for hospitals to manage capacity.

The findings of this research are also of great importance for hospitals. By gaining insight into the expected LoS, hospitals can better plan and manage their bed occupancy. This enables them to make more efficient use of available resources, such as nurses and medical equipment, resulting in better utilization of available capacity. If capacity managers of the hospital are provided with an indication of the possible LoS, they are able to anticipate when a bed becomes approximately available again. This could smoothen the patient flow in the hospital, as patients are eventually admitted to the appropriate department more quickly. This could reduce waiting times and improve the quality of care.

Besides, predicting the LoS can also contribute to improved patient care and a better hospital experience of the patient. By setting realistic expectations about the length of hospital stay, patients and their families are able to prepare better for the recovery process and become more involved in treatment planning.

By evaluating whether it is possible to predict the LoS of the nonsurgical patient flow, this research contributes to existing literature. Until now, the LoS was often predicted for one specialism only (see e.g. Alsinglawi et al. (2022), Daghistani et al. (2019), and Launay et al. (2015)). This research differentiates from earlier research by considering a whole patient flow. Especially, the LoS of nonsurgical patients is predicted, which to our knowledge has not been done before.

The remainder of this research is structured as follows. First, in Section 2 relevant findings of earlier research are discussed. Subsequently, Section 3 describes the data and provides an initial analysis of the LoS. An extensive description of the applied methods is provided in Section 4. Then, Section 5 discusses the results. Section 6 concludes with the most important findings and some suggestions for future research.

2 Literature Review

In this section relevant findings of earlier research are discussed, which could corroborate the choices and considerations of this research. First, in Section 2.1 determining factors of the LoS are mentioned. After that, Section 2.2 considers statistical models that could explain and/or predict the LoS. Lastly, in Section 2.3 the application of machine learning in healthcare is discussed.

2.1 Determining factors length of stay

The LoS of a patient can be influenced by various factors. The **diagnosis** can certainly explain a large part of the LoS, because it reflects why the patient has been admitted and it represents the patient's condition. However, at time of admission, the diagnosis has not necessarily been determined yet. In addition, it is possible that no diagnosis is established at all during the admission, because the exact condition of the patient is unclear.

Besides the diagnosis, several other factors that are not yet known at time of admission can describe a large part of the LoS. A **change in diagnosis** could represent a diagnostic error. Hautz et al. (2019) evaluated the effect of such diagnostic discrepancies and concluded that these errors are associated with a longer LoS. In addition, it is possible that there is a **change in specialism**. Often this means that the patient has been transferred to a higher level of care, e.g. a intra-hospital transfer to the intensive care unit (ICU). Escobar et al. (2011) found that patients who have been transferred to the ICU have a longer LoS than patients who have not been transferred. Furthermore, Sykora et al. (2020) concluded that an early transfer to a higher level of care leads to a significant increase in LoS.

As mentioned earlier in Section 1, unplanned admissions that unexpectedly necessitate surgery during the admission process are also classified as nonsurgical admissions. The surgery itself and the possible recovery period take time and therefore affect the LoS. It is also stated by Aghajani and Kargari (2016) that the **number of surgeries** is an important factor in determining the LoS. During a surgery, but also at other moments during a hospital admission, a **complication**, i.e. an unfavorable result of a disease, health condition, or treatment, can occur. It is found by McAleese and Odling-Smee (1994) that complications double the average LoS. The authors calculated a numerical ratio for surgical complications that reflects the severity of the complication and increases the LoS.

It is interesting for hospitals to be able to predict the LoS at time of admission. Consequently, it is not possible to include the variables described above in a prediction framework, because these variables are not yet known at time of admission. However, there are plenty of variables directly available at time of admission and these variables can be of interest in explaining and/or predicting the LoS. For example, at time of admission, the patient gets assigned to a **specialism** and it is immediately determined whether it is an **emergency**. Caetano et al. (2014) found that the specialism is one of the most influential factors in predicting the LoS of inpatient admissions. It is concluded by Nippak et al. (2014) that the time spent in the ED is correlated with the total LoS of a patient. Liew et al. (2003) stated that the time spent in the ED can accurately predict the LoS. Hence, the fact that an admission is considered emergency could be an important factor in explaining the LoS.

In addition, a patient's medical history could have an impact on the LoS. Turgeman et al. (2017) found that the LoS depends on the **number of previous admissions**, the **time that has elapsed since the last discharge**, and the **total LoS** of the previous admissions added together. It is studied by Wang et al. (2014) which factors cause a prolonged LoS after acute exacerbation of chronic obstructive pulmonary disease (COPD). They found that having a **previous diagnosis** related to COPD in the past 12 months was significantly associated with a LoS of more than 11 days. Moreover, having one or more **comorbidities**, e.g. heart failure,

cardiac arrhythmias, or diabetes, also had a significant effect on having a LoS of more than 11 days (Wang et al., 2014).

The **date and time of admission** are not directly related to the patient, but nevertheless have a major impact on the LoS. Earnest et al. (2006) found that patients who are admitted during weekends, during holidays, and after office hours have a longer LoS in general. In addition, Ryan et al. (2010) observed that patients who are admitted during the weekend faced longer waiting times until they received their medical procedures. This directly leads to a longer LoS, because it takes more time before the treatment actually starts.

2.2 Modelling length of stay

By modelling the relationship between LoS and other variables the determining factors of the LoS can be derived. The most widely studied technique to model the relationship between two or more variables is linear regression. To estimate the linear relationship between the variables, the method of ordinary least squares (OLS) can be used. If the normality assumptions hold, the OLS estimator is the best linear unbiased estimator (Heij, 2004). However, the estimated coefficients are biased and inconsistent if these assumptions are violated (Heij, 2004).

It is probable that the LoS data does not satisfy the normality assumptions. LoS data in hospitals often exhibits a right-skewed distribution, because a small number of patients require a significantly longer stay compared to the majority of admitted patients (Harerimana et al., 2021; Ma et al., 2020). Factors such as severe illnesses, complications, or the need for extensive treatment can contribute to a longer LoS, which creates a skewed distribution where the mean is inflated due to the presence of these outliers. As a result, the OLS method becomes inconsistent and provides biased coefficient estimates. This encourages to apply a different model and use a different estimation method.

Despite the fact that the normality assumptions probably do not hold, Yoon et al. (2003) applied a linear regression model to derive the determining factors for the LoS of patients admitted to the ED. Although the model identified determining factors of the LoS and there was no significant evidence of multicollinearity, the R^2 value of 0.384 suggested that other unconsidered factors may explain a large part of the variability in LoS. Moreover, Austin et al. (2002) used a linear regression model to determine the relationship between patient characteristics and LoS. In their research, the performance of the linear regression model is compared to several other statistical strategies for analyzing the LoS of patients undergoing coronary artery bypass graft (CABG) surgeries. The main goal of Austin et al. (2002) was to illustrate that the significance of the association between LoS and patient characteristics depends on the chosen statistical model. However, the presence of heteroscedasticity and the non-normality of the error term made it difficult to draw conclusions about the coefficient estimates in the linear model. In addition, this also raised the question whether linear regression is suitable for explaining the relationship between LoS and patient characteristics.

LoS refers to the *duration* of a patient hospitalization, i.e. time between consecutive admission and discharge times over a given time period. A duration model, which is part of survival analysis, can be used to analyze the expected duration of time until an event occurs (Heij, 2004). In case of modelling the LoS, the event represents the moment of discharge from the hospital.

In a duration model, the distribution of the baseline hazard rate can be specified so that the model is able to control skewed data. Moreover, duration models are capable of incorporating censored data. These censored observations are valuable, because the fact that they passed a specific duration without encountering an event holds meaningful information (Heij, 2004).

Ravangard et al. (2011) applied duration models to determine the factors influencing the LoS of patients in a women hospital in Iran. In their research, several duration models assuming different distributions for the hazard rate and Cox's proportional hazards (PH) model (Cox, 1972) were compared. They found that the parametric model assuming the gamma distribution provided the best fit for modelling the LoS. Additionally, the proportional hazards assumption of Cox's PH model was violated, as a result of which the results of Cox's PH model were unreliable and parametric models should have been preferred. Cox's PH model is also used by Austin et al. (2002). Despite the fact that the coefficients estimated by Cox's PH model corresponded to the coefficients of the generalized linear models (GLMs), Cox's PH model poorly predicted the LoS compared to the other models.

GLMs expand upon the traditional linear model framework (McCullagh, 2019). These GLMs provide flexibility in modeling various types of dependent variables by allowing different specifications for the distribution of the dependent variable. As mentioned earlier, the LoS data is probably right-skewed and therefore almost certainly does not follow a normal distribution. As a result, it is highly probable that a GLM provides a better fit than the standard linear model.

Besides linear models and Cox's PH model, Austin et al. (2002) applied GLMs with a logarithmic link function and Poisson, negative binomial, and gamma distributions. They stated that an analyst should be encouraged to examine different modelling frameworks to determine which model provides the best fit and has its assumptions satisfied. In their research, it is concluded that GLMs with a logarithmic link function should be seriously considered for LoS predictions of CABG surgery patients. Verburg et al. (2014) compared the predictive performance of different statistical models, which included GLMs with a logarithmic link function and Gaussian, Poisson, negative binomial, and gamma distributions. Despite the fact that GLMs with a logarithmic link function were slightly better able to predict untransformed ICU LoS than the other models, all models considered by Verburg et al. (2014) performed poorly. As a result, it is concluded by Verburg et al. (2014) that their considered models are unsuitable to predict the LoS of ICUs and that the models should not be utilized for designing policies regarding unplanned ICU admissions.

2.3 Machine learning in healthcare

Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and models allowing computers to learn from data and make predictions or decisions without being explicitly programmed for each task (Samuel, 1959). Despite the skepticism about the practical application of machine learning in healthcare and the fact that the results are difficult to interpret, the use of machine learning approaches in healthcare increases substantially (Habehh & Gohel, 2021). Machine learning can be applied in various ways in healthcare. For example, Escobar et al. (2016) used machine learning to identify patients with a higher risk of being transferred to the ICU. Furthermore, Ardila et al. (2019) applied machine learning

techniques in order to detect potential early signs of lung cancer.

Besides these applications of machine learning in healthcare, a lot of research has been conducted on predicting the LoS among patients. A distinction can be made between studies examining patient populations belonging to a specific specialism, e.g. cardiology, geriatrics, or oncology, and studies including all patients admissions at a hospital.

Daghistani et al. (2019) evaluated several machine learning models for predicting the LoS among cardiac patients. In their research, Random Forest (RF) achieved the highest accuracy and also outperformed all models based on the other performance measures. However, Daghistani et al. (2019) emphasize that the data came originated from a single health organization and hence applying the RF model in other health organizations could lead to different results. Launay et al. (2015) used Artificial Neural Networks (ANNs) to predict the LoS of geriatric patients who have been admitted to the emergency department. They used the 10-item brief geriatric assessment (BGA), which is a comprehensive evaluation designed to assess various aspects of an older person's well-being, in order to predict the LoS by using ANNs. Launay et al. (2015) concluded that an ANN with a modified multilayer perceptron (MLP) is able to accurately predict the LoS, and that the chronic conditions of the patient are the main contributors to the predictive accuracy. In the research of Alsinglawi et al. (2022), the RF model achieved the highest accuracy for predicting the LoS of lung cancer patients. Alsinglawi et al. (2022) used the explainable technique SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) to identify the variables that contributed the most to predicting the LoS.

Contrary to the research described above, Rajkomar et al. (2018) considered all patient admissions of two American academic medical centers and constructed a deep learning model to predict the LoS. Their deep learning approach integrates EHR data, and they concluded that their approach creates accurate and scalable predictions. Steele and Thompson (2019) considered elective admissions, i.e. an admission where the day and the time is scheduled, and did not focus on a specific specialism. Caetano et al. (2014) used all patient admissions of a Portuguese hospital. In their research, several machine learning models were compared based on their predictive performance. The RF model achieved the best values for all considered performance measures. A whole different approach has been implemented by Azari et al. (2012). They applied k-means clustering during the training phase to construct training sets from dissimilar clusters. Their results showed that this approach led to a more distinctive set of training rules being selected, which resulted in better predictive performance. In the research of Azari et al. (2012), a Bayesian Network (BN) and a Support Vector Machine (SVM) had the best performance overall.

Hospitals remain reluctant to apply machine learning techniques, because the *black box* of the models causes uncertainty about the realization of the prediction. Especially in cases regarding clinical decision making, the application of machine learning models is lagging behind, because the consequences of possible misclassifications could be disastrous (Mozaffari-Kermani et al., 2014). Statistical models, which are discussed in Section 2.2, are often preferred over machine learning models for their interpretation, even though their predictive performance is commonly disappointing. Post-hoc explanation methods, e.g. Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), make it possible to

extract information about the realization of a prediction by a machine learning model. This way, it becomes more clear how a prediction is constructed by a machine learning model and which variables have had the biggest influence on the prediction. Hence, these explanation methods could be valuable for a hospital wanting to implement machine learning models for prediction, because they open the so-called black box and provide more clarity about the predictions. LIME generates explanations for individual predictions of machine learning models by approximating their behavior locally with interpretable surrogate models (Ribeiro et al., 2016). This makes it possible for users to gain understanding about the reasoning behind individual predictions without needing to comprehend the complex processes ongoing within the model. SHAP is a game theoretic approach to determine the relative contribution of each variable to the prediction. It considers all possible combinations of variables and determines each variable its contribution by evaluating how the addition or removal of a particular variable affects the model its prediction when combined with other variables (Lundberg & Lee, 2017).

3 Data

This section mentions the data source in Section 3.1 and discusses the descriptive statistics in Section 3.2.

3.1 Data source

The data that is used in this research is obtained from a large hospital organization in the Netherlands. The data contains all patient admissions over the period from 01-01-2016 until 27-02-2020. Each observation in the data corresponds to a mutation, which is a change in the admission. Such changes occur when e.g. a patient is transferred to another bed. In total, these mutations correspond to 433069 patient admissions. These admissions do not represent unique patients. The data covers a period of more than four years, hence it occurs that certain patients are admitted multiple times during this time period.

This research focuses on the admissions of nonsurgical patients, which are admitted patients without a planned operation. This means that it is not necessarily the case that a nonsurgical admission is not planned, because a patient can have a planned admission for treatment that does not require a surgery. In total, the data contains 279616 admissions belonging to the nonsurgical patient flow, which is approximately 65% of all patient admissions. This research only takes into account the nonsurgical admissions that are allocated to a bed, and thus not nonsurgical admissions that are allocated to e.g. a chair. For hospitals, it is especially interesting to gain insight into the expected time a patient takes up a bed. If hospitals are able to predict how long a bed is occupied, they can take this into account when scheduling and/or admitting a new patient. First, the admissions for which it is known that the patient is allocated to a chair are removed from the data. For the remaining admissions, it is assumed that the patient needs to have a LoS of at least an hour to have been actually allocated to a bed. This assumption is necessary, because the bed type was not clear for each patient. This assumption also applies to patients for whom it is indicated that they are assigned to a bed, because it is unlikely that a patient occupying a bed will have a LoS shorter than one hour. In total, presumably 211591 of

the 279616 patients belonging to the nonsurgical patient flow have been allocated to a bed, and are therefore considered in this research.

3.2 Descriptive statistics

Table 1 presents statistics regarding the LoS per nonsurgical specialism. It can be seen that there are substantially less patient admissions belonging to the specialisms dermatology (DER), psychiatry (PSY), and rehabilitation (REV) than admissions belonging to the other specialisms. Most of the nonsurgical patient admissions belong to the internal medicine (INT) specialism. Psychiatry has the highest average LoS among all specialisms. However, the mean LoS of the psychiatry specialism is heavily influenced by outliers as the median of this specialism is much lower. For almost all specialisms the mean LoS strongly deviates from the median LoS, which indicates that the LoS data of these specialisms contains outliers. Only the specialisms dermatology and rehabilitation have a mean LoS similar to the median LoS.

A large standard deviation indicates the degree of dispersion in the data by indicating how much the observed values deviate from the mean. Table 1 shows that almost all specialisms have large standard deviations. This indicates again that the LoS data contains outliers. The psychiatry specialism has the largest standard deviation. The interquartile range (IQR) is a measure of spread for the middle half of the data as it is the difference between the first quartile (25th percentile) and third quartile (75th percentile). It is striking that the psychiatry specialism has a relatively small IQR. This shows that the middle 50% of the LoS data of patient admissions belonging to the psychiatry specialism is relatively densely spread around the median.

It stands out that the minimum LoS of almost all specialisms is equal to one hour. This is not surprising, because it is stated in Section 3.1 that it is assumed that a patient needs to have a LoS of at least an hour to have been allocated to a bed. The high maximum values also suggest that outliers are present in the data. The highest maximum is attained by the psychiatry specialism with a maximum equal to 5063.65 hours, which is approximately 211 days.

Table 1: Length of stay (LoS) statistics per nonsurgical specialism.

		LoS (hours)					
Specialism	# Observations	Mean	Median	St. dev.	IQR	Min.	Max.
CAR	47761	37.89	7.15	80.63	25.50	1.00	1870.53
DER	7	52.71	46.45	58.61	81.79	1.40	141.90
GER	3311	137.47	45.80	194.65	202.25	1.00	2148.87
INT	58700	38.35	4.00	99.19	16.87	1.00	2662.87
KIN	16073	60.14	22.57	127.21	55.00	1.00	2619.20
LON	22830	62.40	8.53	111.46	86.39	1.00	2876.93
MDL	43736	19.97	2.48	64.09	1.20	1.00	3057.17
NEU	14760	76.62	27.03	125.18	68.71	1.00	3940.32
PSY	672	189.00	3.74	627.50	1.60	1.75	5063.65
REU	3687	4.89	2.50	17.66	1.20	1.00	365.42
REV	54	4.80	4.83	1.22	1.30	2.02	8.28

Note. IQR refers to the interquartile range.

Table 2 shows the descriptive statistics of the explanatory variables considered in this research. This table contains both variables known at time of admission and variables that become

known during the admission. The variables that become known during the admission, cannot be included in a prediction framework, but could explain part of the LoS. All variables are included in the descriptive models that are used to derive the determining factors of the LoS. In the predictive models, however, only the variables known at time of admission and the assigned specialism are included. Besides presenting the descriptive statistics for each variable over all nonsurgical patients, Table 2 also shows the descriptive statistics for the different LoS classes. By also showing the descriptive statistics per class, it is possible to compare the values for the explanatory variables across the LoS classifications and evaluate whether these values differ. In this research the admissions are divided into the following three classes: short LoS ($<$ 12 hours), medium LoS (12 hours $\leq$ LoS $\leq$ 96 hours), and long LoS ($>$ 96 hours). In the remainder of this research, the same partition of LoS classes is used.

It directly stands out that the LoS of nonsurgical patients is most often shorter than 12 hours, as 62.08% of the observations belongs to the short LoS class. The medium and long LoS classes contain 25.12% and 12.80% of all nonsurgical patient admissions, respectively. The average age of patients having a LoS longer than 4 days is substantially higher than the age of the patients belonging to the other classes. Nonsurgical patients with a LoS longer than 12 hours almost always arise from a clinical admission, i.e. an admission where the patient stays a certain period in the hospital for observation, treatment, or recovery. Nonsurgical admissions with a short LoS are most often of the day admission type, which is understandable as day admissions generally take less than a day. Almost all observational and polyclinical admissions, i.e. admissions to observe the patient's condition and admissions for a medical consult or short treatment by a doctor, respectively, have a short LoS.

A longer LoS is often the result of an emergency admission, because the majority of patient admissions belonging to the medium and long LoS classes arise from an emergency situation. In addition, nonsurgical patients with a medium or long LoS are relatively more often admitted during the weekend and/or outside office hours than patients with a short LoS.

It is not immediately possible to deduce whether having a condition considered as a risk factor leads to a longer LoS. Only patients having a long LoS seem to have COPD relatively more often. It is surprising that patients with a short LoS have been admitted more often to the hospital in the past 3, 6, and 12 months than the patients with a medium or long LoS. On the other hand, the average total LoS over the past 3, 6, and 12 months of patients with a medium or long LoS is much higher than for patients with a short LoS. However, it must be noted that it is difficult to draw conclusions about the average total LoS over the past 3, 6, and 12 months, because the standard deviation is very large, which indicates that the total LoS varies greatly among the nonsurgical patient admissions.

During a nonsurgical admission, it is not often the case that surgery turns out to be necessary, as also becomes clear from the average number of surgeries presented in Table 2. It does not necessarily seem to be the case that the emergence of a complication leads to a longer LoS, because for all LoS classes the rate for the occurrence of a complication is equal to around 22%. For patients with a long LoS, the number of mutations is higher than for the other classes. It becomes clear that a transfer to an ICU almost never happens during a nonsurgical admission. Besides, the average number of diagnosis changes is also low.

Table 2: Descriptive statistics explanatory variables.

Variable	Total n = 211591	LoS class: Short (< 12 hours) n = 131355	LoS class: Medium (12-96 hours) n = 53160	LoS class: Long (> 96 hours) n = 27076
Known at time of admission				
Age	59.71 ± 22.06	60.45 ± 18.41	54.89 ± 27.82	65.58 ± 23.55
Is man	103569 (48.95%)	60890 (46.36%)	28653 (53.9%)	14026 (51.8%)
Admission type				
Day admission	87552 (41.38%)	87277 (66.44%)	265 (0.50%)	10 (0.04%)
Clinical admission	92436 (43.69%)	13460 (10.25%)	51913 (97.65%)	27063 (99.95%)
Observational admission	12115 (5.73%)	11135 (8.48%)	980 (1.84%)	0 (0.00%)
Polyclinical admission	19488 (9.21%)	19483 (14.83%)	2 (0.00%)	3 (0.01%)
Admission information				
Is emergency	85222 (40.28%)	22974 (17.49%)	38957 (73.28%)	23291 (86.02%)
Is during weekend	57432 (27.14%)	30063 (22.89%)	17546 (33.01%)	9823 (36.28%)
Is outside office hours	45501 (21.5%)	12397 (9.44%)	22051 (41.48%)	11053 (40.82%)
Risk factors				
Cardiac arrhythmias	25679 (12.14%)	15947 (12.14%)	6053 (11.39%)	3679 (13.59%)
COPD	28218 (13.34%)	16025 (12.2%)	6555 (12.33%)	5638 (20.82%)
Obese	1823 (0.86%)	1054 (0.8%)	448 (0.84%)	321 (1.19%)
Diabetes	13857 (6.55%)	7722 (5.88%)	3587 (6.75%)	2548 (9.41%)
Previous admissions				
# Admissions past 3 months	1.29 ± 2.42	1.73 ± 2.82	0.56 ± 1.27	0.63 ± 1.31
# Admissions past 6 months	1.96 ± 3.75	2.63 ± 4.4	0.81 ± 1.83	0.94 ± 1.96
# Admissions past 12 months	2.75 ± 5.51	3.69 ± 6.5	1.12 ± 2.58	1.34 ± 2.84
# Past complications	0.83 ± 3.80	1.12 ± 4.61	0.34 ± 1.53	0.41 ± 1.90
Total LoS past 3 months	25.57 ± 82.92	20.66 ± 66.78	26.49 ± 87.68	47.63 ± 128.05
Total LoS past 6 months	39.63 ± 118.04	33.59 ± 99.89	39.56 ± 123.25	69.09 ± 172.08
Total LoS past 12 months	56.27 ± 153.77	48.94 ± 131.26	54.74 ± 160.96	94.83 ± 220.40
Becomes known during admission				
# Surgeries	0.10 ± 0.33	0.07 ± 0.26	0.15 ± 0.37	0.17 ± 0.47
Complication occurred	47467 (22.43%)	30089 (22.91%)	11478 (21.59%)	5900 (21.79%)
# Mutations	0.55 ± 1.18	0.06 ± 0.28	0.85 ± 0.97	2.38 ± 1.98
Mutation to ICU	407 (0.19%)	40 (0.03%)	156 (0.29%)	211 (0.78%)
# Diagnosis changes	0.03 ± 0.18	0.00 ± 0.05	0.03 ± 0.18	0.16 ± 0.42

Note. Values are mean ± standard deviation, n (%).

4 Methodology

This section discusses all applied methods. Section 4.1 starts by describing the models that are used to derive the determining factors of the LoS. After that, Section 4.2 states the statistical- and machine learning models for which their predictive performance is evaluated.

4.1 Effects on length of stay

To find out the variables influencing the LoS, both linear regression models and duration models are considered. The advantage of linear regression models is the fact that the estimated coefficients are rather straightforward to interpret. In this research, both a standard linear regression model using the LoS as dependent variable, and a linear regression model using the log LoS are applied. The data regarding the LoS is right-skewed, as a result of which the distribution of the residuals does not follow a normal distribution. A log-transformation of the LoS could bring the distribution of the dependent variable closer to a normal distribution. In the standard linear regression model, each coefficient can be interpreted as the change in LoS for a unit increase in the corresponding variable, holding all other explanatory variables constant. If the log LoS is used as dependent variable, each coefficient represents a percentage change in LoS for a unit increase in the corresponding variable, holding all other explanatory variables constant.

In a duration model, the dependent variable is a *duration*, which is a measure for the amount of time until a certain event happens (Heij, 2004). The LoS of a patient is a duration, because it measures the time until a certain event happens, namely, discharge from the hospital. Using a duration model could be a more appropriate approach than using a linear model and estimating the effects by the method of OLS. As mentioned earlier, the LoS data is right-skewed. As a result, the assumption of normally distributed error terms is probably violated and the OLS method becomes inconsistent (Heij, 2004). Duration models do not assume normality of the error terms and are better suited for the typically skewed distribution of the durations. Especially, Cox's PH model (Cox, 1972) allows for more flexibility, as the underlying distribution for the durations can remain unspecified. LoS data is often right censored, because certain patients have been admitted to the hospital but have not yet been discharged. OLS does not account for censored data, which leads to biased estimates and incorrect standard errors (Heij, 2004). Duration models, on the other hand, can appropriately handle censored data and eventually provide unbiased estimates.

However, the interpretation of the estimated coefficients by a duration model is a bit cumbersome. In case of Cox's PH model, the exponent of the estimated coefficients represent the effect of an increase in an explanatory variable on the conditional probability of discharge from the hospital. For this model, a positive coefficient means that as the value of an explanatory variable increases, the hazard rate of a patient's LoS increases, which corresponds to an increased probability of discharge from the hospital. Likewise, a negative coefficient indicates that an increase in the explanatory variable has an increasing effect on the LoS. If an underlying distribution is assumed for the durations, the interpretation of the estimated coefficients depends on the functional form and corresponding parameters of the baseline hazard rate function.

Despite the fact that a duration model might be more suitable for modeling the LoS, it is

decided to also present the results of the linear regression models, because the estimated coefficients in these models directly show how the LoS is influenced by a change in the explanatory variables. This is mainly interesting for hospitals, because it provides insight into how a possible change during the admission could affect the LoS. By considering the results of both linear regression models and duration models, an informed decision can be made about which variables are the most determining factors of the LoS. This is mainly important for ChipSoft, because it provides clarity which variables should always be included in a predictive model, if they would decide to implement a prediction framework for the nonsurgical patient flow in HiX.

4.1.1 Length of stay as a duration

In a duration model, the *hazard rate* represents the probability that the duration ends now, conditional on the fact that is has not ended before (Kiefer, 1988). Hence in this research, the hazard rate is the chance that a patient gets discharged from the hospital, conditional on the fact that the patient is still staying in the hospital. A high hazard rate means that the patient has a high probability to be discharged from the hospital at that moment in time. The duration model estimates the hazard rate based on the durations $y_1, ..., y_n$, where n is the number of durations in the data. These durations are assumed to follow a distribution with density $f(t)$ and cumulative distribution function (CDF) F(t). Assuming that the hazard rate is the same for each individual i, it holds that the hazard rate $\lambda(t)$ and *survival function* $S(t)$ are given by

$$\lambda(t) = \lim_{\delta \to 0} \frac{P[t < y_i \leq t + \delta | y_i > t]}{\delta}, \text{ and} \tag{1}$$

$$S(t) = P[y_i > t] = 1 - F(t), \tag{2}$$

respectively. Usually, only the hazard rate is estimated, because it is possible to derive the density and survival function from the hazard rate. It holds that

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d \log(S(t))}{dt}, \tag{3}$$

and therefore it follows that the density and survival function can be obtained by

$$f(t) = \lambda(t) S(t), \text{ and} \tag{4}$$

$$S(t) = e^{-\int_0^t \lambda(s) d(s)}. \tag{5}$$

Duration model types. It is likely that the hazard rate varies per individual. Namely, the LoS of a patient can be heavily dependent on the patient's individual history. A patient who has spent more time in the hospital in the past and has possibly underlying ailments, is more likely to have a longer LoS and therefore a different hazard rate. The type of the duration model depends on the effect of the explanatory variables on the hazard rate and survival function. A distinction can be made between PH models and accelerated failure time (AFT) models.

In a PH model the hazard rate is directly modified as a product of the baseline hazard and the individual specific effect (Breslow, 1975). A unit increase in one of the explanatory variables is multiplicative with respect to the hazard rate. In a PH model, the hazard rate corresponding

to individual i is expressed as $\lambda_i(t) = g_i\lambda(t)$, where $g_i > 0$ is the individual-specific effect. The survival function corresponding to this hazard rate is defined by $S_i(t) = [S(t)]^{g_i}$. Often $g_i = e^{x_i'\beta}$ is used, where x_i represents the explanatory variables corresponding to observation i. This functional form is also used in this research. It follows that the hazard rate $\lambda_i(t)$ of a PH model increases for large values of $x_i'\beta$. Besides, the survival function $S_i(t)$ decreases for large values of $x_i'\beta$ if it holds that $0 < S(t) < 1$. This means that larger values of $x_i'\beta$ correspond to shorter durations.

In an AFT model, the explanatory variables accelerate or decelerate the time until discharge (Kalbfleisch & Prentice, 2011). This is due to the fact that the time in an AFT model is scaled by the factor $g_i = e^{x_i'\beta}$. As a result, the survival function becomes $S_i(t) = S(g_it) = S(e^{x_i'\beta}t)$. The hazard rate is therefore given by $\lambda_i(t) = g_i\lambda(g_it) = e^{x_i'\beta}\lambda(e^{x_i'\beta}t)$. It follows that if $x_i'\beta$ is positive, the time until discharge for individual i is effectively reduced (accelerated), leading to a higher hazard rate. Conversely, if $x_i'\beta$ is negative, the time until discharge is prolonged (decelerated), leading to a lower hazard rate.

Baseline hazard rate functions. For the durations several distributions can be assumed, leading to different specifications of the baseline hazard rate $\lambda(t)$. It is possible to leave the specification of the distribution unspecified, which leads to Cox's PH model (Cox, 1972). Cox's PH model assumes proportional hazards, i.e. a unit increase in one of the variables has a multiplicative effect with respect to the hazard rate. The Schoenfeld residuals (Schoenfeld, 1982) are used to test whether the assumption of Cox's PH model holds. If this assumption does not hold, the estimates by Cox's PH model are unreliable, which should make parametric models preferable.

Besides Cox's PH model, the following parametric distributions for the baseline hazard rate are considered: log-normal, log-logistic, exponential, Weibull, and gamma distribution. Table 3 summarizes the functional forms for the baseline hazard rate when these different distributions are assumed. In addition, the model type corresponding to the parametric distribution is shown.

For the log-normal distribution it holds that the hazard rate increases until some point in time, after which the hazard rate starts to decrease (Kurniasari et al., 2019). The log-logistic distribution has a shape similar to the log-normal distribution. For the log-logistic distribution it can be explicitly evaluated that the contribution of a right-censored observation to the likelihood is equal to the value of the survivor function at time of censoring (Bennett, 1983). This is not possible for the log-normal distribution and hence the log-logistic is likely to be more suitable for analyzing duration data. The baseline hazard is constant in case the exponential form is used (Hosmer & Lemeshow, 1999). This means that the initial probability of discharge is the same for each patient, regardless of the current LoS. Hence, applying the exponential form for the hazard rate may be inappropriate, because the exponential distribution is unable to control data containing both short and long durations (Kiefer, 1988). Using the exponential distribution leads to a PH model, because the baseline hazard rate is not affected by a change of explanatory variables. The Weibull distribution is a generalization of the exponential distribution. Namely, the exponential distribution is equivalent to the Weibull distribution if $\alpha = 1$. Hence if it is the case that $\alpha = 1$, using the Weibull distribution leads to a PH model. If $\alpha \neq 1$, the model aligns with the AFT model its flexibility in modelling different times of discharge. The

Weibull distribution is particularly effective modelling monotonic hazard rates, i.e. hazard rates that exponentially increase or decrease over time (Jiang & Murthy, 2011). The shape of the gamma baseline hazard depends on the shape parameter γ. If $\gamma = 1$, the gamma hazard function corresponds to the constant hazard function of the exponential distribution. If $\gamma > 1$, the hazard function is concave and increasing. For $\gamma < 1$, the hazard function is convex and decreasing. In case of the gamma hazard function, $\Gamma_t(\gamma) = \int_0^t x^{\gamma-1}e^{-x}dx$ represents the incomplete gamma function.

Table 3: Baseline hazard rate functions when different distributions are assumed.

Distribution	Baseline hazard rate	Parameter restrictions	Model type
Log-normal	$\lambda(t) = \frac{\phi(\log t;\mu,\sigma)}{1-\Phi(\log t;\mu,\sigma)}$	$\sigma > 0$	AFT
Log-logistic	$\lambda(t) = \frac{\beta/\alpha(t/\alpha)^{\beta-1}}{1+(t/\alpha)^{\beta}}$	$\alpha > 0, \beta > 0$	AFT
Exponential	$\lambda(t) = \gamma$		PH
Weibull	$\lambda(t) = \gamma\alpha t^{\alpha-1}$	$\gamma > 0$, $\alpha > 0$	AFT/PH
Gamma	$\lambda(t) = \frac{t^{\gamma-1}e^{-t}}{\Gamma(\gamma)-\Gamma_t(\gamma)}$	$\gamma > 0$	AFT/PH

Note. The expressions ϕ and Φ represent the probability density function (PDF) and cumulative distribution function (CDF), respectively, of a normal distribution with mean μ and standard deviation σ. AFT refers to accelerated failure time, while PH refers to proportional hazards.

The duration models using various baseline hazard rate functions are compared based on their log-likelihood value. The model achieving the highest log-likelihood value provides the best fit for explaining the LoS. Hence, only the estimates of the best fitting duration model are presented.

Parameter estimation. To estimate the parameters of a duration model, maximum likelihood estimation (MLE) is performed. A distinction is made between observed durations that are finished and durations that are censored. The finished durations represent patients who have been discharged from the hospital, while the censored durations refer to patients who have been admitted but not yet discharged. In the log-likelihood expression, the finished durations are indicated by $z_i = 1$ and the unfinished durations by $z_i = 0$, respectively. It holds that the probability that duration i is not yet finished is equal to $P(z_i = 0) = S_i(y_i)$. The expression $p_i(y_i) = \lambda_i(y_i)S_i(y_i)$ represents the density of the finished durations. It follows that the log-likelihood function is defined by

$$\begin{aligned}
\log(L) &= \sum_{\{i;z_i=1\}} \log(p_i(y_i)) + \sum_{\{i;z_i=0\}} \log(S_i(y_i)) \\
&= \sum_{\{i;z_i=1\}} \log(\lambda_i(y_i)) + \sum_{i=1}^{n} \log(S_i(y_i)) \\
&= \sum_{\{i;z_i=1\}} \big(x_i'\beta + \log(\lambda(y_i))\big) - \sum_{i=1}^{n} \Big(e^{x_i'\beta} \int_0^{y_i} \lambda(t)dt\Big).
\end{aligned} \tag{6}$$

It is used that $S_i(t) = [S(t)]^{e^{x_i'\beta}}$, such that $\log S_i(y_i) = e^{x_i'\beta} \log S(y_i) = e^{x_i'\beta} \int_0^{y_i} \lambda(t)dt$. Note that it is assumed that the n durations are mutually independent.

4.2 Predicting length of stay

In order to evaluate to what extent it is possible to predict a patient's LoS, several regression and machine learning models are considered. This way, it is possible to compare the predictive performance of the different models and conclude which model is the most suitable for an application in healthcare. As it might be difficult to predict the exact LoS of a patient, it is first evaluated whether the different models can accurately classify the admissions in the three LoS classes as introduced in Section 3.2. Besides classifying the LoS in the three different classes, the LoS is exactly predicted and the deviation from the actual LoS is examined.

4.2.1 Statistical models

As discussed earlier in Section 4.1, this research uses a duration model to derive the determining factors of the LoS. The best fitted duration model is also considered in the prediction framework. In case of the duration model, the LoS is directly predicted. To assess the accuracy of this model for classifying the LoS, it is determined for the predicted LoS which class it would fall into.

In case of classifying the LoS, multinomial logistic regression is considered, while linear regression is applied to directly predict the LoS. Besides using the original LoS as dependent variable, a linear regression on the log-scaled LoS is considered. This ensures that the predicted LoS is non-negative, such that a more realistic prediction can be formed. In addition, this could show how a log-transformation possibly improves the predictive performance. In Section 2.2 it is already stated that linear models possibly do not provide a good fit for modelling the LoS. However, it is interesting to include these models in the prediction framework as baseline models, so that it is possible to evaluate how the accuracy can increase when different models are applied. The advantage of a linear regression model is that the estimated coefficients are interpretable as the increase in LoS when a given variable increases by one unit. However, as also mentioned in Section 4.1, the conclusions drawn from the linear regression model are probably questionable due to the LoS data being skewed and heavily influenced by outliers.

A GLM is a flexible generalization of the linear regression model (McCullagh, 2019). In the standard linear regression model, the dependent variable is assumed to be normally distributed. GLMs generalize the linear model by allowing different probability distributions, belonging to the exponential family, for the dependent variable. A so-called *link function* allows the linear model to be connected to the dependent variable by relating the mean of the dependent variable to the linear combination of explanatory variables. In this research, GLMs with a logarithmic link function and assuming the Poisson and negative binomial distribution are considered. These distributions probably fit the LoS data better. The predictions by a GLM assuming the Poisson or negative binomial distribution are always positive, which probably leads to more realistic predictions than the predictions of the standard linear regression model, because the standard linear regression model can provide negative predicted values for the LoS. In case of the Poisson distribution, the variance of the conditional distribution increases when the mean of the conditional distribution increases, because the variance is equal to its mean. In reality this assumption is often violated, and overdispersion is observed, i.e. the presence of greater variability than would be expected based on the assumed distribution. The negative binomial distribution generalizes the Poisson distribution by incorporating overdispersion. The coefficients estimated

by GLMs with a logarithmic link function can be interpreted as the logarithm of the relative change in average LoS when a given predictor variable increases by one unit.

4.2.2 Machine learning models

Besides the statistical models, several machine learning models are considered, so that it can be evaluated to what extent it is possible to predict the LoS. In addition, by evaluating various predictive models it is possible to mutually compare the performance and conclude which model is the most accurate. In this research, RF, eXtreme Gradient Boosting (XGBoost), a Recurrent Neural Network (RNN), and SVM are considered.

RF (Breiman, 2001) is a supervised machine learning method that generates numerous decision trees during the training phase to form an ensemble. During the training phase the general technique of bootstrap aggregating is applied by RF. In this bagging approach, a random sample with replacement of the training set is repeatedly selected and fitted to the decision trees. Using a single training set could result in strongly correlated trees, which leads to a high sensitivity to noise. Bootstrapping de-correlates the trees by showing them different training sets, which leads to better performance, because the variance decreases without an increase in bias. After training, predictions for unseen data can be constructed by taking the average of the predictions from the individual trees. In case of classification, the majority vote of the trees is used. Besides this original bagging procedure for trees, RF uses another bagging scheme that selects a random subset of the variables for each bootstrap sample. Randomly selecting a subset of variables ensures that the trees are uncorrelated in an ordinary bootstrap sample. Namely, if random variable selection is not applied, the variables that are strong predictors of the dependent variable are selected in many trees, which results in correlated trees.

Boosting algorithms try to construct an accurate classifier from multiple weak classifiers. It works by making a series of weak models, each one trying to fix the mistakes of the previous model. This process continues until either all the training data is predicted correctly or the maximum number of models allowed is reached. A widely applied boosting algorithm is the gradient boosting framework by Friedman (2001), where each classifier corrects the error of the predecessor. The XGBoost algorithm developed by Chen and Guestrin (2016) is an efficient and scalable implementation of the gradient boosting algorithm. In general, XGBoost has great predictive performance and is much faster than standard gradient boosting. In the XGBoost algorithm, decision trees are sequentially created. The variables allocated to a decision tree are assigned weights. If the predicted result is wrong, the corresponding variables get assigned a larger weight in the succeeding decision tree. As a result, each successive decision tree is able to learn from the mistakes of its predecessor, such that the combined result of the individual classifiers is a strong and accurate prediction.

A RNN is a bi-directional ANN, which means that some nodes in the RNN have connections that loop back on themselves allowing them to maintain a memory of previous inputs. RNNs introduce the concept of memory by including the dependency between data points. This means that RNNs have the ability to learn recurring patterns, because the network can be trained to retain concepts based on context. The hidden state of a RNN serves as memory and contains information about the network's previous inputs. Each time new information is assigned, the

hidden state is updated based on the current input and the previous hidden state. The back-propagation through time (BPTT) algorithm (Werbos, 1988) is used to train RNNs. During the so-called forward pass the RNN processes the input along with the hidden state of the previous input to produce an output and update the hidden state. Subsequently, during the so-called backward pass the predicted output is compared to the actual target output using a loss function and the calculated error is then backpropagated through time. When BPTT is applied, gradients are computed to update the model parameters. However, during BPTT the vanishing gradient problem can be encountered, which is a major disadvantage of RNNs. The problem is that the gradients that are used to derive the weights for updating the model parameters become very small, as a result of which the network is unable to learn long-term dependencies.

The RNN considered in this research consists of an embedding layer, a Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) layer, a dropout layer, and an output layer. The embedding layer converts input sequences of integers into dense vectors of fixed size. The input dimension for this layer is equal to the number of observations to consider. The second layer is an LSTM layer, which is a type of RNN that can capture long-term dependencies and mitigate the vanishing gradient problem often encountered in traditional RNNs. The LSTM layer processes the embedded sequences and captures temporal dependencies, where the output represents the learned patterns in the input data. The input in the LSTM layer flows in only one direction, which can be either forwards or backwards. The dropout layer prevents overfitting and enhances the generalization ability of the model. This layer sets a fraction of the input units equal to zero during training, which reduces the model its reliance on specific neurons. The output layer constructs an output based on the learned patterns. In case of classification, the output layer uses a sigmoid activation function, while the linear activation function is used for regression. The Adam optimizer (Kingma & Ba, 2014) is used to adjust the learning rate during training. For classification, the categorical crossentropy loss function is used and the accuracy is the evaluation metric. In case of regression, the mean squared error (MSE) and mean absolute error (MAE) are used as loss function and evaluation metric, respectively.

SVM (Vapnik, 1995) is a supervised machine learning algorithm which can be used for both classification and regression. In case of a classification problem, SVM tries to find the optimal hyperplane that is able to separate the different classes in the feature space. A kernel function is used to map the original feature space into a higher-dimensional space, which allows SVM to find a hyperplane in this higher-dimensional space, even if the data is not linearly separable in the original space. In this research, the Gaussian radial basis function kernel is used. For each class, the data point that is closest to the hyperplane is called a Support Vector. The hyperplane tries to maximize the margin between these Support Vectors. Hence, the margin represents the confidence with which SVM is able to classify new data points. For classification, SVM involves solving a quadratic programming (QP) problem with a number of constraint proportional to the number of observations in the training set. In case of a regression problem, SVM aims to find a find a function that does not deviate more than a specified margin, *epsilon*, from the actual target values. This leads to an optimization problem that includes both a penalty for errors, i.e. deviations outside the epsilon margin, and a cost regularization term. The resulting QP problem can be more complex, as it involves additional variables and constraints compared to

the classification problem.

4.2.3 Training-test procedure

The LoS data being right-skewed could cause the models to overfit the large outliers. Due to overfitting these outliers, the admissions with a short LoS will be greatly overestimated. Most of the admissions have a short LoS, as a result of which most of the observations will be inaccurately predicted. In addition, it is not in the interest of hospitals to predict these large outliers, because these observations are the exception rather than the rule. Hospitals are mainly interested in the LoS of regular patients, because being able to accurately predict the LoS of regular patient admissions can lead to a more streamlined patient flow. Hospitals can account for the few outliers by keeping a fixed percentage of beds available for any long-term patients. However, overestimating the LoS for regular patients can result in planned admissions being postponed while in reality there are beds available. This obstructs patient flow and could result in an inefficient occupancy of the admission ward.

Hence, before the training-test procedure is started, the observations with outlying LoS values are removed. The observations for which the z-score is larger than three, i.e. $z = \frac{LoS_i - \mu}{\sigma} > 3$, where LoS_i represents the LoS of admission i, μ is the mean LoS, and σ is the standard deviation of the LoS data, are removed. Note, that it is not necessary to take the absolute value of the z-score to detect the outlying observations, because the LoS data is right-skewed and there are no observations that have an unusually small LoS due to the fact that admissions with a LoS of less than an hour are already removed (see Section 3.1).

The data is randomly split in a training and test set. 80% of the data is used for training and the remaining 20% is used as test data. When splitting the data, it is ensured that both the training and the test set contain equal proportions of short, medium, and long LoS observations. In case of the statistical models, the relationship between the dependent variable and explanatory variables is estimated based on the complete training data. Predictions for the dependent variable in the test data are then constructed by substituting the values of the explanatory variables of the test data in the estimated regression equation. The training-test procedure is different for the machine learning models and is described in the next two paragraphs.

Training phase. First, the continuous variables are scaled. Scaling ensures that all variables contribute equally to the models, preventing variables with larger ranges from dominating the learning process. It could improve the performance and convergence speed of optimization algorithms, and it could enhance the effectiveness of distance-based algorithms. Especially the performance of SVM depends heavily on scaling. SVM is a distance-based machine learning approach, where large values can dominate and/or skew the algorithm if scaling is not applied. The continuous explanatory variables are scaled by min-max normalization:

$$X_j^{\text{scaled}} = \frac{X_j - \min(X_j)}{\max(X_j) - \min(X_j)} \qquad j = 1, ..., m, \tag{7}$$

where m is the number of continuous explanatory variables. As all continuous dependent variables in the data are non-negative, the scaled variables are defined over the range [0, 1]. Hence,

the continuous explanatory variables are then defined over the same range as the binary explanatory variables.

The dependent variable, i.e. the LoS, is scaled by a log-transformation to ensure that the predicted value is always strictly positive when transformed back. This makes sure that the LoS prediction is somewhat realistic, because a negative LoS is practically impossible. The log-transformation of the LoS is given by

$$LoS^{\text{scaled}} = \frac{\log_{10}(LoS)}{\log_{10}(\max(LoS))}. \tag{8}$$

This log-transformation also ensures that the LoS is defined over the same [0, 1]-range as the explanatory variables.

During the training phase, a grid search using stratified k-fold cross-validation is performed in order to tune the hyperparameters for each machine learning algorithm. In this research, k is set equal to five. Using only five folds could cause an increase in bias and variance. However, applying 5-fold cross-validation should lead to relatively stable results and ensures a decrease in computational costs (Raschka, 2018). In this application of stratified 5-fold cross-validation the training data is split in five folds, where admissions with a short, medium, and long LoS are equally distributed over the five folds. In cross-validation, one of the folds is used as validation set, while the other folds are used to train the model. Hence, in total five iterations are performed for each combination of hyperparameters. The hyperparameters of the most accurate model over the five validation sets are selected as the best. In Appendix B, the tuned hyperparameters are discussed and the used grids are presented. After having found the best hyperparameters for each algorithm, the models are trained using the complete training dataset. In case of the RNN, the batch size is equal to 2500 and the number of epochs is equal to 10 when tuning the hyperparameters. When training the RNN on the full training dataset, the batch size is set to 1250 and the number of epochs is equal to 20.

In this research, predicting the LoS is considered as both a classification problem and a regression problem. It could be possible that the best hyperparameters for the models when considering it as a classification problem deviate from the best hyperparameters for regression. Hence, the hyperparameters for the models are determined separately for the classification and regression problems. In Appendix D, it is examined whether using the best hyperparameters for the classification problem affects the performance of the machine learning models in regression, and it is evaluated whether the best hyperparameters in regression have an impact on the accuracy when they are used as hyperparameters for the models in classification.

Test phase. After determining the best hyperparameters and having trained the models on the complete training data, the predictive performance of the different models is evaluated on the test data. The model that achieves the best values for the different performance measures can be considered as the most accurate predictive model for the patient admissions data considered in this research. However, this model may not be the most suitable for an application in healthcare due to e.g. the lack of interpretability. In addition, it could be that the most accurate model in this research is not able to accurately predict another dataset.

4.2.4 Performance evaluation

For all considered models, it is evaluated how accurately they can classify the patient admissions in short, medium, and long LoS. The overall accuracy of each model is simply calculated by dividing the number of correctly classified instances by the total number of classifications made. The precision and recall are determined for each class separately. Considering the precision and recall per class provides detailed and balanced understanding of model performance, especially in the presence of class imbalance. The precision per class measures the accuracy of the positive predictions for that class, while the recall per class measures the ability of the model to identify all true instances of that class.

Besides classifying the patient admissions in short, medium, and long LoS, predicting the LoS is also considered as a regression problem. This way it is possible to derive the deviation of the predicted LoS from the true LoS for both the complete sample and the different classes individually. In order to evaluate the predictive performance of the different models, the MAE and weighted absolute percentage error (WAPE) are considered. The MAE measures the average absolute deviation between the predicted value $\hat{y}_i$ and the actual value y_i. For n predicted data points, the MAE is given by

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}. \tag{9}$$

The WAPE is a performance measure that determines the average percentage difference between the predicted value and the true value, while also weighing the error over the total LoS of all admissions together. Incorporating these weights facilitates a more refined assessment of the precision in predicting. The WAPE is defined by

$$\text{WAPE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} y_i} \times 100\%. \tag{10}$$

Note that here the denominator does not represent the absolute value of the true value, because taking the absolute value is redundant due to the LoS always being strictly positive.

4.2.5 Model interpretation

As mentioned earlier in Section 4.2.1 the interpretation of the estimated coefficients for a linear regression model is straightforward, as a result of which a linear model is often preferred over machine learning models. This is due to the fact that it is clear for linear models how the prediction is constructed. For machine learning models it is not immediately clear how the prediction is made. As a result, hospitals are reluctant with implementing machine learning models for predicting the LoS, even though it is possible that these machine learning models are better able to predict the LoS. As discussed earlier in Section 2.3, explanation methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), make it possible to find out which variables have had the biggest impact on the realization of the prediction. Despite the fact that this does not reflect exactly how the prediction is made, it is interesting for hospitals to have insight into the factors influencing the LoS prediction the most. As a result, the output of the machine learning models is somewhat interpretable.

In this research the post-hoc explanation method SHAP is used to derive which variables

have had the biggest influence on the predictions. As introduced earlier in Section 2.3, SHAP determines the relative contribution of each variable to the prediction by evaluating how the addition or removal of a particular variable affects the prediction. SHAP is preferred over LIME, because SHAP can provide a globally consistent explanation, while LIME cannot (Rathi, 2019). This allows users to understand the overall behavior of the model across multiple predictions, in addition to the local explanation why a specific prediction is made. If the input data is slightly changed, the explanations change in a consistent and predictable manner and hence the method of SHAP is considered consistent.

5 Results

In this section the results are discussed. First, Section 5.1 presents the estimation results of the linear regression models and a duration model. In Section 5.2 the performance of the different predictive models is evaluated in both classification and regression. Subsequently, Section 5.3 provides insight into the reliability of the predictions. Section 5.4 ends by providing some possible explanations for the prediction deviations.

5.1 Modelling results

5.1.1 Linear regression models

In Table 4 the estimated coefficients by the standard linear regression model and the linear regression model with the log LoS as dependent variable are presented. The R^2 is equal to 0.467 for the standard linear regression model, and equal to 0.791 for the linear regression model using the log LoS as dependent variable. This indicates that the linear regression model using the log LoS is more suitable.

It directly stands out that in both models many variables are statistically significant. The admission type and specialism seem to explain a large part of the LoS. The coefficients of the admission type are estimated with respect to the day admission type. Both the standard linear regression model and the linear regression model using the log LoS as dependent variable estimated that the clinical admissions and observational admissions have a longer LoS in general. For both models, the estimated effect of a polyclinical admission is negative with respect to the day admissions. The coefficients of the specialism are estimated with respect to the cardiology (CAR) specialism. Noticeable is the fact that, in case of the standard linear regression model, all specialisms, except the dermatology (DER) specialism, have a significantly positive effect on the LoS with respect to the cardiology specialism. This is surprising, because in Table 1 of Section 3.2 it can be seen that some of these specialisms have a lower mean LoS than the cardiology specialism. It is also striking that, in case of the linear regression model using the log LoS as dependent variable, the psychiatry (PSY) specialism has a significantly negative effect with respect to the cardiology specialism. In the standard linear regression model, this variable has the largest positive coefficient.

It is remarkable that the coefficients for the admission information have opposing signs in the two models. In the standard linear regression model a weekend admission does not have a statistically significant effect. This coefficient is statistically significant in the linear regression

model using the log LoS as dependent variable. However, it can be argued that this variable has little effect on the LoS, because a weekend admission in general only leads to 1.8% longer LoS compared to an admission on a working day according to the linear regression model using the log LoS as dependent variable.

Some risk factors and certain information of previous admissions have statistically significant effects according to the two models. However, it is difficult to conclude which of these variables is the most determining for the LoS, because the significance of these variables differs per model.

Certain variables that only become known during the admission also seem to be able to explain part of the LoS. In particular, the number of mutations has a large significant effect in both models. The standard linear regression model shows that an additional mutation leads to a 48.828 hours longer LoS. The linear regression model with the log LoS as dependent variable shows that the LoS increases with nearly 50% if an additional mutation occurs. It is remarkable that the standard linear regression model estimates that an additional diagnosis change leads to a 33.581 hours longer LoS, while the linear regression model with the log LoS as dependent variable estimates a relatively small effect for this variable.

Based on the results of the linear regression models it is not necessarily directly possible to deduce which variables are the most determining for the LoS. The **admission type** and the **specialism** appear to be able to explain a large part of the LoS. In addition, certain **risk factors** and **information of previous admissions** could determine the LoS. There are also certain factors that only become known during the admission that influence the LoS of a patient. In particular, the **number of mutations** seems to have a major impact on the LoS.

Table 4: Estimated coefficients by the standard linear regression model and the linear regression model with the log LoS as dependent variable.

	Model	
Variable	Linear regression	Linear regression (log LoS)
Intercept	-39.015 (0.969)***	0.302 (0.009)***
Known at time of admission		
Age	0.187 (0.012)***	0.006 (0.000)***
Is man	0.289 (0.338)	0.011 (0.003)***
Admission type		
Day admission		
Clinical admission	28.958 (0.659)***	2.006 (0.006)***
Observational admission	13.326 (1.016) ***	1.016 (0.010)***
Polyclinical admission	-3.167 (0.632)***	-0.843 (0.006)***
Specialism		
Cardiology (CAR)		
Dermatology (DER)	23.152 (29.024)	0.405 (0.282)
Geriatrics (GER)	106.434 (1.424)***	1.260 (0.014)***
Internal medicine (INT)	39.241 (0.604)***	0.698 (0.006)***
Pediatrics (KIN)	46.386 (1.051)***	1.053 (0.010)***
Pulmonary medicine (LON)	23.255 (0.679)***	0.321 (0.007)***

Gastrointestinal diseases (MDL)	28.904 (0.620)***	0.335 (0.006)***
Neurology (NEU)	11.033 (0.784)***	0.477 (0.008)***
Psychiatry (PSY)	193.968 (3.100)***	-0.839 (0.03)***
Rheumatology (REU)	30.635 (1.383)***	0.425 (0.013)***
Rehabilitation (REV)	38.574 (10.484)***	0.869 (0.102)***
Admission information		
Is emergency	9.239 (0.649)***	-0.076 (0.006)***
Is during weekend	-0.088 (0.384)	0.018 (0.004)***
Is outside office hours	-2.530 (0.463)***	0.176 (0.005)***
Risk factors		
Cardiac arrhythmias	-1.975 (0.519)***	-0.042 (0.005)***
COPD	2.235 (0.506)***	0.080 (0.005)***
Obese	3.065 (1.815)	0.014 (0.018)
Diabetes	0.814 (0.685)	0.033 (0.007)***
Previous admissions		
Number of previous complications	0.022 (0.052)	-0.002 (0.001)***
# Admissions past 3 months	-1.891 (0.211)***	-0.013 (0.002)***
# Admissions past 6 months	0.286 (0.208)	-0.003 (0.002)
# Admissions past 12 months	-0.276 (0.091)**	-0.003 (0.001)***
Total LoS past 3 months	0.053 (0.003)***	0.001 (0.000)***
Total LoS past 6 months	0.001 (0.003)	0.000 (0.000)
Total LoS past 12 months	0.011 (0.002)***	0.000 (0.000)***
Becomes known during admission		
# Surgeries	-0.095 (0.627)	0.229 (0.006)***
Complication occurred	0.473 (0.447)	0.006 (0.004)
# Mutations	48.828 (0.184)***	0.479 (0.002)***
Mutation to ICU	0.749 (3.910)	-0.092 (0.038)*
# Diagnosis changes	33.581 (1.027)***	0.024 (0.010)*
R^2	0.467	0.791

Note. Standard errors are in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In order to assess the suitability of the linear regression models, the OLS assumptions are verified. The Breusch-Pagan test (Breusch & Pagan, 1979) is used to test the null hypothesis of homoskedasticity against the alternative of heteroskedasticity. To test for the presence of serial correlation, the Breusch-Godfrey test (Breusch, 1978; Godfrey, 1978) is applied. To assess the assumption of normally distributed error terms, it is tested on the basis of the Jarque-Bera test (Jarque & Bera, 1980) whether the residuals approximately follow a normal distribution.

Table 5 shows the test results of the Breasch-Pagan, Breusch-Godfrey, and Jarque-Bera tests for the different linear regression models. For the Breusch-Pagan test the number of degrees of freedom is equal to 34. In case of the Breusch-Godfrey test, the number of lags to be included for the residuals in the auxiliary regression is set equal to 3. This choice is based on a plot of the autocorrelations, where only the autocorrelations up to the third order were found to be different from zero.

The results of the Breusch-Pagan test show that the null hypothesis of homoskedasticity is rejected for both the standard linear regression model and the linear regression model using the log LoS as dependent variable. Heteroskedasticity causes inconsistency of the covariance matrix, as a result of which the significance of the coefficients is not exact. The Breusch-Godfrey test results show the presence of serial correlation in both models. This causes underestimation of the standard errors, which leads to exaggeration of the coefficients their significance. Despite the presence of heteroskedasticity and serial correlation in both linear regression models, the method of OLS remains unbiased and consistent. However, OLS is not efficient anymore. The Jarque-Bera test presents that the residuals in both models are not normally distributed. This does not have impact on the estimation of the coefficients, but it leads to inexact p-values.

Table 5: Test-statistic and p-value of the Breusch-Pagan, Breusch-Godfrey, and Jarque-Bera test for the different linear regression models.

	Model	
Test	Standard linear regression	Linear regression log LoS
Breusch-Pagan	9684.30 (< 0.001)	35441.00 (< 0.001)
Breusch-Godfrey	180.83 (< 0.001)	763.34 (< 0.001)
Jarque-Bera	1368290675.00 (< 0.001)	86085.00 (< 0.001)

Note. P-values of the tests are presented in parentheses.

These test results indicate that a linear regression model may not be particularly suitable for modeling the LoS. The method of OLS remains unbiased and consistent, as a result of which the estimated coefficients are still accurate. However, due to the fact that not all assumptions hold, the standard errors are incorrect and the significance of the coefficients is exaggerated. For this reason, there is a small caveat to be made, because it could be that certain coefficients are not supposed to be significant. Nevertheless, many coefficients are highly significant, so that they would probably also have been significant if the assumptions of OLS did hold.

5.1.2 Duration models

Table 6 presents the log-likelihood value for the duration models assuming different distributions. The model that achieves the highest log-likelihood value provides the best fit for explaining the LoS. Table 6 shows that the duration model using the log-logistic distribution for the hazard rate achieves the highest log-likelihood value. It is notable that the log-likelihood value of Cox's PH model is much lower than the log-likelihood of the parametric models. This could indicate that Cox's PH does not provide a good fit and that its estimates are probably unreliable.

Table 6: Log-likelihood value for the duration models assuming different distributions.

Model	Log-likelihood
Log-normal	-718845.50
Log-logistic	-707220.60
Exponential	-748006.60
Weibull	-732129.40
Gamma	-722442.30
Cox's PH	-2225360.00

As mentioned earlier in Section 4.1.1, Cox's PH model assumes proportional hazards, and a violation of this assumption leads to unreliable coefficient estimates by Cox's PH model. The test results of Cox's PH model indicate that the assumption of proportional hazards has been rejected (degrees of freedom = 34; $\chi^2 = 53480.18$; $p < 0.001$). Hence, Cox's PH model provides wrong coefficients and wrong standard errors. The parametric model using the log-logistic distribution for the hazard rate should be preferred. As a result, only the coefficients estimated by the duration model using the log-logistic distribution for the hazard are discussed.

In Table 7 the estimation results of the parametric model using the log-logistic distribution are shown. For this duration model, the estimated scale parameter is equal to $\hat{\alpha} = 1.65$, and the estimated shape parameter is given by $\hat{\beta} = 2.60$. The coefficients are estimated by only considering finished durations, because the data did not contain any censored observations. As mentioned earlier in Section 4.1, the interpretation of the estimated coefficients depends on the functional form of the baseline hazard rate. In case of the log-logistic distribution, the estimated coefficients represent the effect of the explanatory variables on the log of the scale parameter. The scale parameter determines the shape and the behavior of the hazard rate function. For a lower scale parameter value, the event, i.e. discharge from the hospital, happens sooner on average. Therefore, a negative coefficient represents an increase in the hazard rate, which means that the conditional probability of discharge from the hospital increases and that the respective patient has a shorter LoS. Likewise, a positive coefficient corresponds to an increase in LoS, because the hazard rate decreases. This means that the signs of the estimated coefficients by the duration model using the log-logistic distribution have the same interpretation as in the linear regression models. In this duration model, however, it is difficult to deduce the actual effect of the explanatory variables on the LoS, because the magnitude of the coefficients cannot be directly interpreted.

It is immediately noticeable that almost all coefficients estimated by the duration model assuming the log-logistic distribution are statistically significant. In addition, it can be deduced that the signs of the estimated coefficients by the duration model using the log-logistic distribution in general correspond to the signs of the coefficients estimated by the linear regression models. The intercept is significant and corresponds to the log of the estimated scale parameter $\hat{\alpha}$. The admission type is important for explaining the LoS. The coefficients for the clinical and observational admission types are significantly positive, which means that these admission types have a higher survival probability to stay in the hospital than the day admission type. This is not surprising, because clinical admissions in general have a longer LoS due to the fact that they stay for a longer period in the hospital for observation, treatment, or recovery. The assigned specialism seems to explain a large part of the LoS. With respect to the cardiology (CAR) specialism, almost all specialisms have a significantly positive coefficient, which means that these specialisms have a lower probability of discharge than patients belonging to the cardiology specialism. Only the coefficient of the psychiatry (PSY) specialism has a negative sign, indicating that patients belonging to this specialism are more likely to have a shorter LoS.

It is not surprising that weekend admissions and admissions outside office hours have significantly positive coefficients, because it was also expected that these admissions would have a longer LoS in general. It is remarkable that the coefficient for emergency admissions is insig-

nificant. In Section 5.1.1 it is shown that in both linear regression models this coefficient was significant. However, the coefficients for emergency admissions in the linear regression models had opposing signs, as a result of which the actual effect of an emergency admission was not necessarily clear.

Certain risk factors and some information regarding previous admissions can explain part of a patient's LoS. However, it should be noted that the coefficients for these variables are really small, which makes it difficult to draw conclusions about the actual effect on the hazard rate. The variables that become known during the admission are also significant, and therefore can explain part of a patient's LoS. In particular, similar to the result of the linear regression models, the number of mutations seems to be important for determining the LoS.

Due to the fact that almost all coefficients are significant, it is not possible to directly deduce what the most determining factors of the LoS are. It is also not possible to determine based on the magnitude of the coefficients which variables are the most important for explaining the LoS, because the variables do not have the same range. The **admission type** and assigned **specialism** seem to be able to explain a large part of the LoS. It is not necessarily unexpected that possible **changes during the admission** explain a large part of the LoS. However, this information is of course not known at time of admission, as a result of which the corresponding variables cannot be included in a prediction framework.

In conclusion, both the linear regression models and the duration model indicate that the **admission type** and the assigned **specialism** are import factors for explaining the LoS. The admission information, risk factors and information about previous admissions also seem to explain part of the LoS. However, it is not necessarily clear which respective variables have the biggest influence on the LoS, because the significance and direction of the sign of the coefficients sometimes differ in the considered models. Certain variables that become known during admission can also have an impact on the LoS. Especially, the **number of mutations** is considered by all models as a determining factor of the LoS.

Table 7: Estimation results of the duration model using the log-logistic distribution for the hazard rate.

Variable	Estimated coefficient	Standard error	P-value
Intercept	0.502	0.009	< 0.001
Known at time of admission			
Age	0.005	0.000	< 0.001
Is man	0.006	0.003	0.039
Admission type			
Day admission			
Clinical admission	1.996	0.006	< 0.001
Observational admission	0.851	0.009	< 0.001
Polyclinical admission	-0.835	0.005	< 0.001
Specialism			
Cardiology (CAR)			

Dermatology (DER)	0.497	0.328	0.130
Geriatrics (GER)	1.128	0.014	< 0.001
Internal medicine (INT)	0.549	0.006	< 0.001
Pediatrics (KIN)	0.831	0.009	< 0.001
Pulmonary medicine (LON)	0.281	0.006	< 0.001
Gastrointestinal diseases (MDL)	0.190	0.006	< 0.001
Neurology (NEU)	0.297	0.007	< 0.001
Psychiatry (PSY)	-1.559	0.026	< 0.001
Rheumatology (REU)	0.293	0.011	< 0.001
Rehabilitation (REV)	0.719	0.078	< 0.001
Admission information			
Is emergency	0.007	0.006	0.232
Is during weekend	0.026	0.003	< 0.001
Is outside office hours	0.173	0.004	< 0.001
Risk factors			
Cardiac arrhythmias	-0.024	0.004	< 0.001
COPD	0.069	0.004	< 0.001
Obese	0.011	0.016	0.497
Diabetes	0.027	0.006	< 0.001
Previous admissions			
Number of previous complications	-0.002	0.000	< 0.001
# Admissions past 3 months	-0.010	0.002	< 0.001
# Admissions past 6 months	-0.003	0.002	0.046
# Admissions past 12 months	-0.002	0.001	0.038
Total LoS past 3 months	0.001	0.000	< 0.001
Total LoS past 6 months	0.000	0.000	0.364
Total LoS past 12 months	0.000	0.000	< 0.001
Becomes known during admission			
# Surgeries	0.203	0.006	< 0.001
Complication occurred	0.009	0.004	0.013
# Mutations	0.474	0.002	< 0.001
Mutation to ICU	-0.188	0.041	< 0.001
# Diagnosis changes	0.053	0.010	< 0.001

In order to assess the goodness-of-fit of the various duration models, the Cox-Snell residuals (Cox & Snell, 1968) are calculated and the survival functions are estimated by the Kaplan-Meier method (Kaplan & Meier, 1958). The cumulative hazard functions are derived from the estimated survival functions. Figure 1 shows the estimated cumulative hazard plotted against the Cox-Snell residuals for the duration models assuming different distributions. The plot of the cumulative hazard against the residual values should be a straight line through the origin with a slope of one (Klein et al., 2003). Hence, models for which the Cox-Snell residuals cause a deviation from this linear line do not provide a good fit for modelling the LoS data. From Figure 1 it becomes clear that the duration model using the log-logistic distribution for the hazard rate

provides the best fit for explaining the LoS, because the Cox-Snell residuals of this model cause the smallest deviation from the bisector. In addition, it is evident that Cox's PH model provides the worst fit of all models.

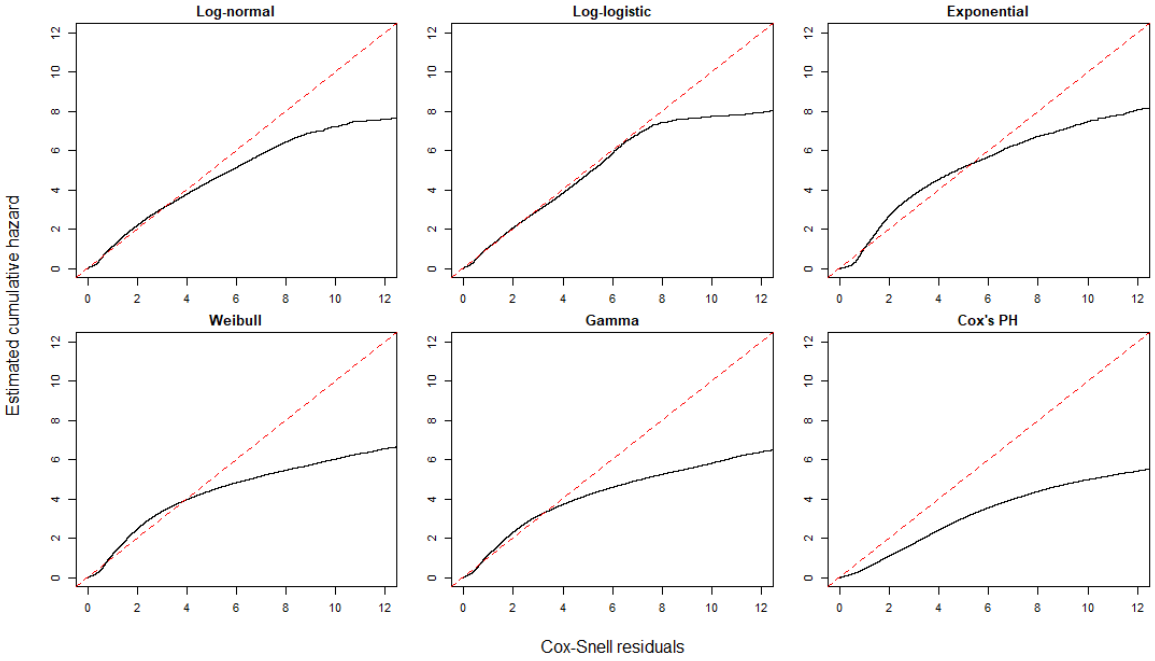


Figure 1: Estimated cumulative hazard against Cox-Snell residuals for the duration models assuming different distributions.

5.2 Prediction results

This section presents the prediction results, where first the classification performance is discussed in Section 5.2.1, and after that the regression results are evaluated in Section 5.2.2. As mentioned earlier in Section 4.2.3, the admissions with an extremely large LoS are removed from the data. In total, 3809 observations are classified as outlying observations, which is approximately 1.8% of the data considered in this research. After removing these observations, 207782 admissions remain, of which 131355 (63.22%), 53160 (25.58%), and 23267 (11.20%) admissions belong to the short, medium, and long LoS class, respectively.

5.2.1 Classification

Table 8 shows the performance results of the different models for classifying the LoS in the short, medium, and long LoS classes. Besides the overall accuracy, this table presents the precision and recall for the different classes. The highest overall accuracy is obtained by XGBoost.

In case of the statistical models, the highest accuracy is obtained by the multinomial logit model. This model achieves an accuracy of 83.77%, which is substantially higher than always predicting the majority class, namely, the short LoS class that contains 63.22% of the observations. The obtained accuracy by the GLMs and the duration model using the log-logistic distribution is also considerably higher than always predicting the majority class, but a little bit lower than the accuracy achieved by the multinomial logit model. The GLM using the negative

binomial distribution and the duration model using the log-logistic distribution for the hazard rate detect substantially less observations belonging to the long LoS class than the multinomial logit model and the GLM using the Poisson distribution, as becomes clear from the recall scores for the long LoS class. The precision of the GLMs and the duration model is higher than the precision of the multinomial logit model for the short LoS class, which means that the percentage of observations classified as having a short LoS that actually have a short LoS is larger for these models. However, this is accompanied by a lower recall, which indicates that these models detect less observations belonging to the short LoS class. The GLM with the negative binomial distribution and the duration model achieve a really high recall for the medium LoS class. However, as the precision of these models is around 60%, approximately 40% of the observations classified as having a medium LoS actually has a short or long LoS.

XGBoost achieves an overall accuracy of 84.05%, and is therefore the most accurate model for classification in this research. This model is best able to detect the admissions with a long LoS, as XGBoost achieves the highest recall for the long LoS class. However, it should be noted that the recall of 33.42% is still relatively low. The XGBoost model is precise in classifying the short and medium LoS classes, because this model achieves a relatively high precision for these classes.

Table 8: Classification performance of the different predictive models.

Model	Performance measure	Total	LoS class		
			Short	Medium	Long
Statistical models					
	Accuracy	83.77%			
Multinomial logit	Precision		97.00%	64.93%	56.10%
	Recall		92.54%	86.31%	28.48%
	Accuracy	82.87%			
GLM Poisson	Precision		98.96%	62.28%	54.95%
	Recall		89.99%	86.97%	33.29%
	Accuracy	82.79%			
GLM negative binomial	Precision		98.91%	61.14%	58.24%
	Recall		90.19%	91.40%	21.34%
	Accuracy	82.39%			
Duration log-logistic	Precision		99.18%	59.77%	64.36%
	Recall		89.71%	96.07%	9.78%
Machine learning models					
	Accuracy	83.93%			
RF	Precision		97.31%	64.99%	56.75%
	Recall		92.50%	86.63%	29.38%
	Accuracy	84.05%			
XGBoost	Precision		96.46%	66.47%	55.12%
	Recall		93.37%	83.19%	33.42%
	Accuracy	82.61%			
RNN	Precision		92.84%	66.08%	51.05%
	Recall		94.31%	81.14%	19.88%
	Accuracy	83.92%			
SVM	Precision		97.39%	64.70%	56.93%
	Recall		92.60%	86.92%	28.09%

Although the machine learning models are quite capable of detecting observations with a medium LoS, as the recall is higher than 81% for all models, the relatively low precision values of the medium LoS show that a large part of the observations classified as having a medium LoS actually do not have a medium LoS. Together with the relatively low recall values of the long LoS class, this indicates that a large part of the observations with a long LoS is 'underestimated'. Notable is the fact that the RNN performs poorly relative to the other machine learning models. What is particularly striking is the low recall value for the long LoS class. This shows that the RNN is unable to detect admissions with a long LoS, as a result of which the overall accuracy is logically lower. It should be noted that the RNN considered in this research has a relatively simple structure. Using a more complex RNN may increase the overall accuracy.

It is striking that the machine learning models RF, XGBoost, and SVM only perform slightly better than the mutinomial logit model. Although the interpretation of the coefficients in a multinomial logit model is not completely straightforward, it is possible to deduce based on the log-odds ratio how a longer LoS could be caused. However, the estimated effects are always relative with respect to another class, as a result of which it is not possible to derive how much longer the LoS will be if certain factors apply to a particular patient. Nevertheless, ChipSoft and/or hospitals might prefer the multinomial logit model over the machine learning models, because the estimated coefficients are somewhat interpretable and the classification accuracy is not substantially lower than that of the machine learning models.

5.2.2 Regression

Table 9 shows the performance results of the different models for predicting the LoS. This table shows the MAE and WAPE that are obtained by the different models. As the MAE and WAPE are performance measures representing the deviation of the predicted values from the actual values, lower values for these measures indicate that a model is better able to accurately predict the LoS of nonsurgical patients. SVM achieves the lowest values for the MAE and WAPE.

In case of the statistical models, the linear regression model using the log LoS as dependent variable obtains the lowest value for both the MAE and the WAPE. However, the MAE and WAPE values achieved by the duration model using the log-logistic distribution and the GLMs assuming the Poisson and negative binomial distribution do not differ substantially from the values of the linear regression model using the log LoS as dependent variable. The standard linear regression model is the worst at predicting the LoS, because it achieves the highest values for the performance measures. The difference in performance measures for the standard linear regression model and the linear regression model with a log-transformed dependent variable shows that transforming the dependent variable is a valuable technique to achieve a better predictive performance. This transformation ensures that the predicted LoS cannot be negative, which in any case reduces the deviation from the true value for the admissions with a short LoS.

Noticeable is that all statistical models are relatively well able to predict the LoS of the medium and long LoS classes. The percent deviation of the true value is relatively small, as becomes clear from the WAPE values. However, it should be noted that the absolute deviation from the true value is larger for these classes than for admissions with a short LoS. The fact that the standard linear regression model can predict the LoS of the long LoS class relatively

accurate, shows that this model is sensitive to observations with large LoS values. The GLM with the Poisson distribution is best able to predict the LoS of the long LoS class. Remarkable is the fact that the linear regression model with the log LoS as dependent variable is relatively less able to predict the LoS for the long LoS class, while overall this model performs best among the statistical models.

The machine learning models are better able to predict the LoS of nonsurgical patients than the most accurate statistical model, namely, linear regression using the log LoS as dependent variable. The machine learning models RF, XGBoost, and SVM are about equally capable of predicting the LoS, because the MAE and WAPE are around 20.5 and 62.5%, respectively. It is noticeable that the MAE and WAPE of the RNN are a bit higher than those of the other machine learning models. SVM is the most accurate model, and achieves a MAE of 20.38 and a WAPE of 62.17%. This means that the average absolute deviation of the prediction from the true value is around 20 hours, and the average percentage deviation of the predicted value weighted relatively to the true value is 62.17%. The MAE value of SVM is substantially smaller than the MAE value of the most accurate statistical model. SVM is able to create predictions that are on average almost 2 hours closer to the true value than the predictions of the linear regression model using the log LoS. SVM is less accurate in predicting the LoS for patients with a short or medium LoS than XGBoost and RF. However, the difference in performance of SVM compared to these two models is relatively small. SVM is in absolute terms more accurate for patients with a long LoS than XGBoost and RF. This result shows that SVM is able to construct a relatively accurate LoS prediction for almost all admitted nonsurgical patients.

Table 9: Regression performance of the different predictive models, where the length of stay (LoS) classes are the actual classes.

Model	Performance measure	Total	LoS class		
			Short	Medium	Long
Statistical models					
Linear regression	MAE	27.67	13.55	31.75	98.09
	WAPE	84.41%	375.16%	77.87%	54.72%
Linear regression (log LoS)	MAE	22.26	3.73	19.52	133.20
	WAPE	67.91%	103.11%	47.88%	74.31%
GLM Poisson	MAE	22.95	6.70	31.91	94.22
	WAPE	69.99%	185.47%	78.26%	52.56%
GLM negative binomial	MAE	23.24	6.79	30.92	98.53
	WAPE	70.87%	187.97%	75.82%	54.97%
Duration log-logistic	MAE	22.86	6.25	25.65	110.25
	WAPE	69.72%	172.98%	62.91%	61.51%
Machine learning models					
RF	MAE	20.76	2.92	20.03	123.13
	WAPE	63.30%	80.78%	49.12%	68.69%
XGBoost	MAE	20.65	2.93	20.78	120.39
	WAPE	62.97%	81.08%	50.96%	67.16%
RNN	MAE	21.82	3.24	20.14	130.58
	WAPE	66.55%	89.56%	49.39%	72.84%
SVM	MAE	20.38	3.40	22.43	111.58
	WAPE	62.17%	94.18%	55.02%	62.25%

Due to the fact that SVM is able to construct the most accurate predictions over the entire nonsurgical patient flow, and because the difference in performance compared to the interpretable linear regression model is substantial, SVM could be preferred for predicting the LoS of nonsurgical patients. However, it should be noted that the differences in performance with respect to RF and XGBoost are relatively small. It takes a lot of time train SVM, as the time complexity of this model is between $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ (Bottou & Lin, 2007; List & Simon, 2009), where n denotes the number of observations in the training data. The time complexity for training RF is $\mathcal{O}(kpn\log(n))$ (Louppe, 2014), where k is the number of decision trees, p the dimensionality of the data, and n the number of observations in the training data. XGBoost has a training time complexity of $\mathcal{O}(kd||x||_0\log(n))$ (Chen & Guestrin, 2016), where k is the number of trees, d is the depth of the tree, $||x||_0$ represents the number of non-missing entries in the training data, and n is the number of observations in the training data. If the number of variables and the number of decision trees in RF and XGBoost are small compared to the number of observations, the time complexity of these two algorithms is much lower than that of SVM. Therefore, as it takes less time to train RF and XGBoost, and because the values for the performance measures are not substantially higher than those of SVM, the models RF and XGBoost may be better suited to be used as a predictive model if ChipSoft chooses to implement a prediction framework in their software HiX. The fact that XGBoost offers more flexibility than RF by having multiple hyperparameters that can be tuned (see also Appendix B), ensures that XGBoost is the preferred and recommended model for a practical implementation in a prediction framework in HiX.

Table 10 shows the number of times the different predictive models under- and overestimate the LoS. From this table it becomes clear that the statistical models, with the exception of the linear regression model using the log LoS as dependent variable, 'overfit' the long LoS, because the LoS for patients belonging to the short or medium LoS class is more often overestimated than underestimated. Especially, the GLMs overfit to the observations with a long LoS. These GLMs overestimate the LoS of the short and medium LoS class relatively more often than the other statistical models. In addition, the GLMs underestimate the LoS for the long LoS class less often. For all statistical models, the LoS for patients with a long LoS is much more often underestimated than overestimated. This could be due to the fact that there are relatively few patients with a long LoS and because the LoS of these patients can differ relatively much compared to the LoS of patients belonging to the short or medium LoS class. This result suggests that is not necessarily possible to derive from the included explanatory variables how a longer LoS is caused.

The machine learning models seem to less overfit the patients with a long LoS, because these model do not necessarily more often overestimate the LoS of patients belonging to the short and medium LoS class. Also the machine learning models almost always underestimate the LoS of the long LoS class. Together with the results in Table 9, which showed that the deviation of the predicted LoS from the actual LoS is very large in absolute terms in case of the long LoS class, this indicates that it is practically impossible to predict the LoS of nonsurgical patients who stay in the hospital for a longer period of time. It seems that the considered models cannot find patterns in the data that cause a long LoS. However, this could be due to the fact that

the factors that can cause a long LoS are also important for patients with a shorter LoS, such that the models do not consider these factors as having a strong influence on a longer LoS. In addition, there could be factors that only become known during the admission that can explain why certain patients have had a longer LoS.

Table 10: Number of times the different predictive models under- and overestimate the length of stay (LoS).

	Total		LoS class					
			Short		Medium		Long	
Model	−	+	−	+	−	+	−	+
Statistical models								
Linear regression	17497	24059	11195	15076	1691	8941	4611	42
Linear regression (log LoS)	19968	21588	10525	15746	4797	5835	4646	7
GLM Poisson	15877	25679	9422	16849	2051	8581	4404	249
GLM negative binomial	15325	26231	8822	17449	1964	8668	4539	114
Duration log-logistic	12080	29476	4742	21529	2737	7895	4601	52
Machine learning models								
RF	20018	21538	10631	15640	4756	5876	4631	22
XGBoost	20062	21494	10808	15463	4644	5988	4610	43
RNN	19938	21618	10315	15956	4998	5634	4625	28
SVM	20187	21369	11236	15035	4497	6135	4454	199

Note. The columns with the minus sign (−) represent the number of times the LoS is underestimated, and the columns with the plus sign (+) show how often the LoS is overestimated.

5.3 Prediction diagnostics

Table 9 in Section 5.2.2 presents the performance measures with respect to the actual classes. That table shows how well the different predictive models are able to predict the LoS for the different classes. However, this does not reflect the prediction deviation when a particular class is predicted. For a capacity manager in the hospital it is important to have idea of the possible deviation when a certain LoS has been predicted. Therefore, in Table 11 it is chosen to present the MAE and WAPE for the different models with respect to the predicted class. In this case, the predicted class corresponds to the class the predictions of the models that are used for regression would fall into.

Table 11 shows that the prediction deviation of all models, with the exception of the standard linear regression model, is very small when the predicted value belongs to the short LoS class. If the predicted value belongs to the short LoS class, the MAE is smaller than 2, which means that the prediction deviates less than 2 hours from the actual LoS on average. The MAE and WAPE for the standard linear regression model are a bit higher due to the fact this model also predicts negative values. This model is also relatively strongly influenced by admissions with a longer LoS, which also became clear from Table 10 in Section 5.2.2. What directly stands out is the fact that the MAE and WAPE are a bit higher for the predicted medium LoS class than the MAE and WAPE for the actual medium LoS as presented in Table 9 in Section 5.2.2.

If a short LoS is predicted, the absolute deviation from the true value is smaller for the RF model than for SVM, as shown by the MAE values. However, SVM is in absolute terms more accurate for the predicted medium and long LoS class. Remarkable is the fact that this result

is not necessarily presented by the WAPE values of SVM. The WAPE values of the different predicted classes are higher for SVM than the WAPE values of RF and XGBoost. However, this does not necessarily cause the overall WAPE to be higher for SVM. It turned out from Table 9 in Section 5.2.2 that the overall WAPE of SVM is lower than that of RF. This can be explained by the fact that the numerator in the WAPE consists of the sum of the absolute prediction deviations from the true values. These absolute deviations are in general smaller for SVM than for RF. In case the overall WAPE is determined, the numerator is the same, regardless of the applied model. The numerator is not equal for both RF and SVM when the WAPE is derived with respect to the different predicted classes. As a result, it is possible that the values for the WAPE with respect to the predicted class are lower for RF than those of SVM, while the overall WAPE of SVM is lower than the one of RF.

Table 11: Regression performance of the different predictive models, where the length of stay (LoS) classes are the predicted classes.

Model	Performance measure	LoS class		
		Short	Medium	Long
Statistical models				
Linear regression	MAE	6.89	44.77	74.03
	WAPE	199.21%	81.96%	53.45%
Linear regression (log LoS)	MAE	1.25	50.51	84.56
	WAPE	34.07%	70.56%	50.96%
GLM Poisson	MAE	6.70	31.91	94.22
	WAPE	185.47%	78.26%	52.56%
GLM negative binomial	MAE	6.79	30.92	98.53
	WAPE	187.97%	75.82%	54.97%
Duration log-logistic	MAE	1.65	50.15	76.01
	WAPE	45.41%	72.98%	52.77%
Machine learning models				
RF	MAE	1.16	47.59	78.11
	WAPE	29.61%	66.79%	50.73%
XGBoost	MAE	1.74	47.65	77.08
	WAPE	38.78%	66.32%	51.77%
RNN	MAE	1.23	49.90	81.25
	WAPE	31.58%	69.94%	48.78%
SVM	MAE	1.75	44.94	71.37
	WAPE	39.06%	66.80%	54.19%

In order to get a more in-depth idea of the reliability of the LoS predictions, Table 12 presents the percentage of predictions by the linear regression model using the log LoS as dependent variable and XGBoost for which the absolute prediction deviation is smaller than a certain value. It is chosen to only present these percentages for the best statistical model and the machine learning model that is recommended for a practical implementation in HiX, which are the linear regression model with the log LoS as dependent variable and XGBoost, respectively. The results in Table 12 are easier to interpret than the values for the performance measures. The use of percentages for which the absolute prediction deviation is smaller than a certain value is therefore probably the way in which the reliability of the predictions would be presented in HiX. Despite the fact that it was already clear from the results in Section 5.2.2 that the machine

learning models RF, XGBoost, and SVM can predict the LoS more accurately than the best statistical model, it is decided to also present the performance of the best statistical model, so that is once again possible to compare the performance to that of a machine learning model. This has been done so that, based on interpretable results, it is possible to emphasize once again that a machine learning model, in this case XGBoost, can make more accurate LoS predictions than the best statistical model. The machine learning models RF and SVM perform similarly to XGBoost, as a result of which presenting the prediction diagnostics for RF and SVM has no added value, because comparing the results of these models with both the linear regression model using the log LoS as dependent variable and XGBoost will not lead to new insights. For reference, the prediction diagnostics of RF and SVM are presented in Appendix C.

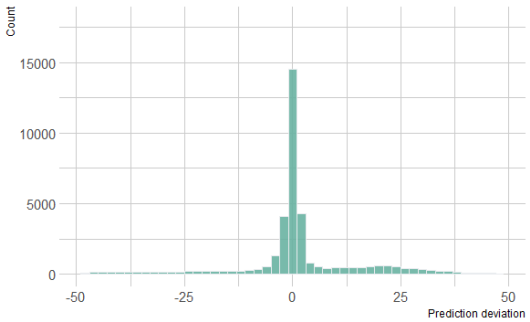
Table 12 shows that XGBoost can more accurately predict the LoS for a large part of the nonsurgical patient flow compared to the linear regression model using the log LoS as dependent variable, because the percentages for which the absolute prediction deviation is smaller than 2, 6, and 12 hours are substantially higher for XGBoost. The results in Table 12 show that more than half of the predicted values by XGBoost does not deviate more than 2 hours from the actual LoS. Almost 80% of the predicted values has an absolute deviation that is less than a day, i.e. 24 hours. 87.39% of the predicted values deviates 2 days or less in absolute terms from the actual LoS. These results show that XGBoost is capable of accurately predicting the LoS for the largest part of the nonsurgical patient flow. Especially, the fact that 53.02% of the predicted values has an absolute deviation that is smaller than 2 hours shows that XGBoost is extremely suitable for predicting the LoS for a large part of the nonsurgical admissions. There is only a small percentage of admissions for which XGBoost is unable to predict the LoS, because the percentage for which the absolute prediction deviation is more than a week, i.e. 7 days or 168 hours, is only 2.54%.

Table 12: Percentages of predictions by the linear regression model using the log LoS as dependent variable and eXtreme Gradient Boosting (XGBoost) for which the absolute prediction deviation is smaller than a certain amount of time.

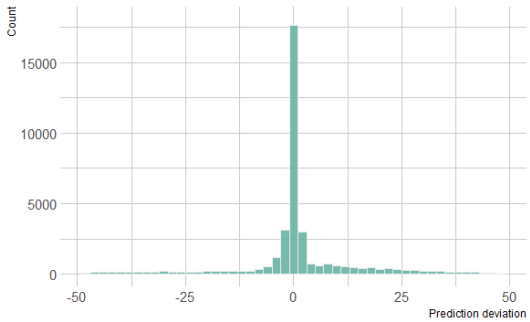
	Absolute prediction deviation (in hours)						
Model	≤ 2	≤ 6	≤ 12	≤ 24	≤ 48	≤ 96	≤ 168
Linear regression (log LoS)	48.48%	61.60%	67.02%	77.39%	87.50%	92.92%	97.14%
XGBoost	53.02%	63.19%	69.85%	78.22%	87.39%	93.63%	97.46%

Figure 2 shows the histogram of the deviation from the actual value for the predictions by the linear regression model using the log LoS as dependent variable and XGBoost. This figure does not include the admissions for which the absolute prediction deviation is more than 48 hours, such that the figure is not stretched. A positive prediction deviation means that the predicted value is larger than the actual value, and a negative prediction deviation represents a predicted value that is smaller than the actual value. The bin-size is equal to 2 hours and the histogram is centered around 0. Therefore, Figure 2b shows that more than 17500 of the 41556 predictions made by XGBoost have a prediction deviation of less than an hour. The number of predictions by the linear regression model using the log LoS as dependent variable for which the prediction deviation is less than an hour is substantially lower, as Figure 2a shows that less

than 15000 predictions have a prediction deviation smaller than 1 hour. Also from Figure 2 it becomes clear that the prediction deviation is really small for a large part of the nonsurgical patients, which is in line with the results of Table 12.



(a) Histogram of the prediction deviations linear regression (log LoS).



(b) Histogram of the prediction deviations XGBoost.

Figure 2: Histograms of the deviations from the actual values for the predictions by the linear regression model using the log length of stay (LoS) as dependent variable and eXtreme Gradient Boosting (XGBoost).

In order to gain insight into which variables are the most important to the predictions of XGBoost, the method of SHAP is used. In Figure 3 the means of the absolute SHAP values for the different variables in the predictions by XGBoost are presented. A variable can have a negative or a positive impact on the prediction. The magnitude of a SHAP value represents how big the impact is. Using the absolute value for the SHAP values ensures that the SHAP values do not balance out, if for example a certain variable has a negative impact on one prediction and a positive impact on another. By using the means of the absolute SHAP values, it is possible to deduce which variables have had the biggest impact on the considered predictions, which means that a global explanation for the considered predictions is provided.

It should be noted that the means shown in Figure 3 are not the means over the whole test set. The SHAP values are determined for only 0.25% of the test set. For the calculation of the explanations 0.25% of the training data is used. This 0.25% for both the training and the test set is randomly sampled and takes into account the proportions of the short, medium, and long LoS classes in the original data. Although this 0.25% represents only a small part of the data, by taking into account the proportions, this should be a representative sample for the whole dataset. It is highly possible that the means of the absolute SHAP values deviate when the SHAP values are determined for all predictions for the test set. However, it was not possible to derive the SHAP values for all predictions. This was due to the fact that determining the SHAP values was computationally expensive and required too much memory. As the SHAP values are not derived for all predictions, it is not possible to conclude which variables have had the biggest influence on the predictions in general. It is only possible to decide which variables, on average, were the most influential for the considered predictions.

From Figure 3 it becomes clear that the admission type, in particular the clinical admission type, is important for constructing the predictions. Also certain specialisms have a relatively large mean absolute SHAP value. Age, admissions outside office hours, and emergency admis-

sions are the other factors that belong to the 8 variables with the largest mean absolute SHAP value. The remaining variables have a relatively small mean absolute SHAP value in comparison to these 8 variables. Based on this result it is not possible to conclude that the variables with the largest mean absolute SHAP values are the most important for constructing the predictions. However, this result indicates that, in particular, the **admission type** and the assigned **specialism** have a large impact on the predictions. It is not surprising that the admission type and the assigned specialism emerge as important factors for making LoS predictions. The admission type and specialism offer an indication of the possible diagnosis and/or treatment of the patient. In Sections 5.1.1 and 5.1.2, it also became clear that these variables strongly cohere with the LoS and can therefore be considered determining factors of the LoS.

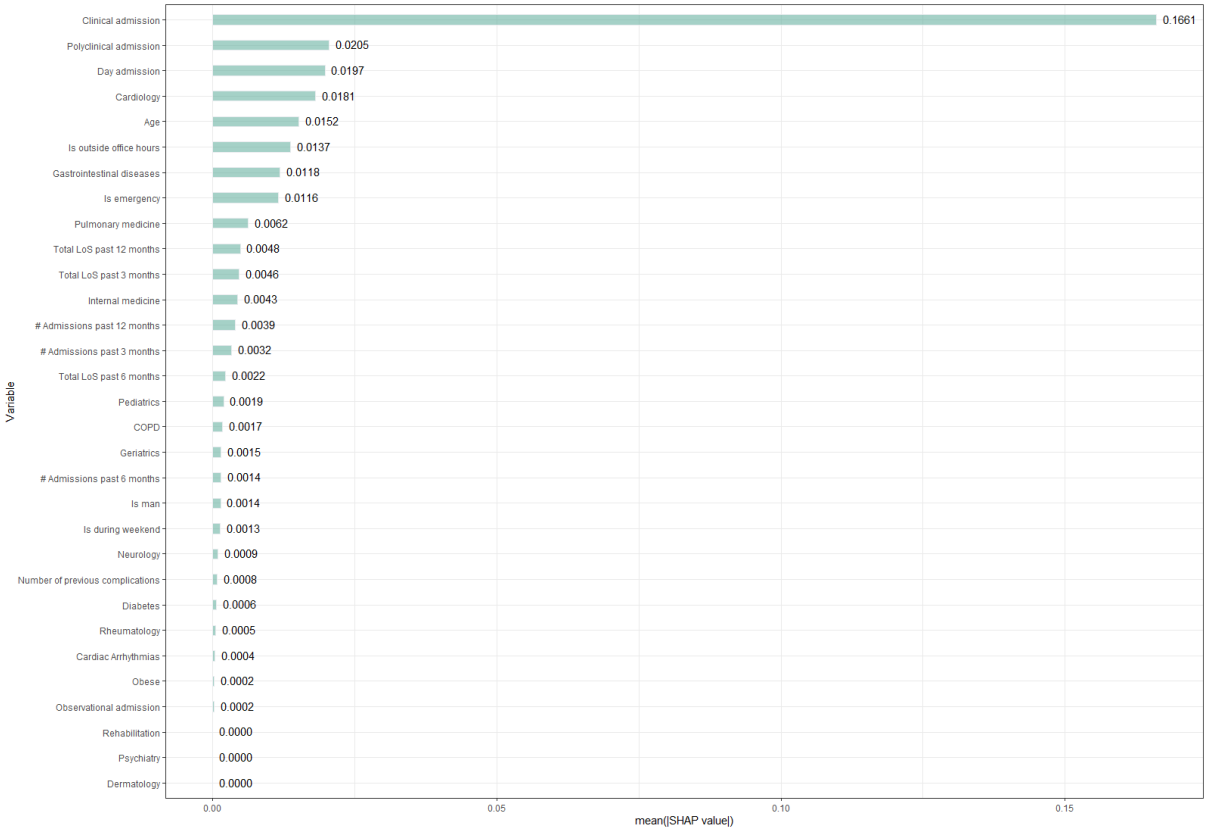


Figure 3: Means of the absolute SHapley Additive exPlanations (SHAP) values for the different variables in the predictions by eXtreme Gradient Boosting (XGBoost).

5.4 Possible explanations prediction deviations

It is not immediately possible to derive from the data why it is not possible to accurately predict the LoS for certain patients with a long LoS. However, the medical record in HiX describes in words what the patient's condition has been and what possible peculiarities may have occurred during the admission. It is difficult to contain this information in an explanatory variable for a predictive model, because it is cumbersome to translate the written information into a variable and the information generally only becomes known during the admission. Nevertheless, this information can probably explain why a patient has had a longer LoS and why a possible large prediction deviation may have arisen. The possible explanations of the prediction deviations can provide an insight why patients could spend a longer period of time in the hospital. However,

it is too short-sighted to conclude what generally causes a longer LoS based on a few possible explanations.

What stands out is the fact that patients for whom the prediction deviation is large have a rather severe diagnosis, such as a cerebral hemorrhage or a metastatic tumor. Sometimes the patients with such severe diagnoses have not been admitted in the past year and do not have a condition considered as a risk factor. As a result, the models logically predict that these patients have had a much shorter LoS, which leads to a large prediction deviation. It is not possible to include the diagnosis as a variable in the predictive models, because there are too many possible diagnoses and the diagnosis is often only determined during the admission. It is also impractical to predict the LoS from the moment the diagnosis is determined, because this moment can differ per patient and the exact moments could not be derived from the data. If it had been possible to derive the moment the diagnosis is determined, it would eventually be interesting to predict the remaining LoS, so that capacity managers are still provided with an estimate when the bed could become available again. It could then be possible to do a simulation study where an initial prediction for the LoS is made at time of admission and this prediction is adjusted when the diagnosis becomes known. When such an approach is applied, the diagnoses need to be grouped, because, as mentioned earlier, it is not possible to include every possible diagnosis as a variable.

It also stands out that other (additional) risk factors are present for patients for whom the prediction deviation is large. It can occur that the patient is a smoker. However, it was not possible to derive for each patient whether this patient is a smoker, because the questionnaire in which the patients must indicate whether they smoke differs per department in the hospital and therefore this information is stored in different places in the database. In general, these questionnaires are also only administered during the admission. There are also patients with a high blood pressure. The blood pressure is measured during the admission. In addition, not every patient takes all tests, and the test forms can differ per department, treatment, and/or patient condition. The risk factors contained in this research were determined in a previous admission as a (sub)diagnosis. This made it possible to retrieve these risk factors from the database, and it could be determined for each patient whether they have these risk factors. For example, for being a smoker and/or having high blood pressure, this was not possible, because these risk factor are never established as a (sub)diagnosis.

6 Conclusion

This research derives the most determining factors for a nonsurgical patient's LoS and evaluates whether it is possible to predict the LoS of these nonsurgical admissions. The results of this research provide ChipSoft insight into which variables should be included in a predictive model when they will choose to implement a prediction framework for the nonsurgical patient flow in their software HiX. In addition, this research gives an advice which considered predictive model is the most suitable for a practical implementation in HiX. Implementing a prediction framework for the nonsurgical patient flow in HiX can provide the capacity managers with insight into the expected LoS, as a result of which hospitals can better plan and manage their bed occupancy.

The central research question this research aimed to answer is: *"Is it possible to predict a nonsurgical patient's length of stay, and which of the considered predictive models is the most*

accurate?". In order to answer this central research question, several statistical- and machine learning models are considered, and their predictive performance is evaluated based on different performance measures. First, an initial analysis using linear regression models and a duration model is performed to derive the most determining factors for the LoS of nonsurgical patients. After that, it is evaluated whether the considered models are able to accurately classify the admissions in short, medium, and long LoS. As this does only provide a 'rough' estimate of the LoS, predicting the LoS is also examined as a regression problem. For the predictive model that is most suitable for a practical implementation in HiX diagnostics about the predictions are presented, and it is derived by applying the method of SHAP which explanatory were the most important for constructing the predictions of this model.

The estimation results of the linear regression models and the duration model show that the **admission type** and the assigned **specialism** can explain a large part of the LoS of nonsurgical patients. In addition, the **number of mutations**, which represents the number of changes during an admission, has an impact on the LoS. Therefore, it is concluded that the admission type, the assigned specialism, and the number of mutations are determining factors for the LoS of nonsurgical patients. The total number of mutations that have occurred is only known after the patient has been discharged. As a result, it was not possible to include this factor as an explanatory variable in the predictive models.

Based on the results it is concluded that it is possible to predict the LoS for a large part of the nonsurgical patient flow. However, for a small part of the nonsurgical admissions it is not possible to accurately predict the LoS. The classification results showed that the models were incapable of detecting the admissions with a long LoS. From the regression results it became clear that the prediction deviation in absolute terms was the largest for the long LoS class. These results indicated that there are probably certain factors that are not included as explanatory variable in the predictive models, which makes a longer LoS generally unpredictable.

The machine learning models RF, XGBoost, and SVM are substantially better able to predict the LoS than the considered statistical models. Despite the fact that the statistical models generally provide more interpretation, these machine learning models should be preferred due to their predictive power. The XGBoost model is considered to be the most suitable for a practical implementation in HiX. This model was the most accurate in classifying the LoS. SVM was a little more accurate than XGBoost in regression. However, the fact that XGBoost requires less computation time than SVM and offers more flexibility by having multiple hyperparameters that can be tuned, has led this research to recommend XGBoost to ChipSoft for a practical implementation in HiX.

Despite the strong aspects of this research, this research also has some limitations. The grid search that is used for tuning the hyperparameters of the machine learning models is relatively limited. A more extensive grid search could probably lead to better hyperparameters, which can improve the overall performance of the models. The fact that the bed type is not consistently indicated for each patient in HiX, necessitated the assumption that a patient would only have been assigned to a bed if they had a LoS of at least one hour. As a result, it cannot be stated with certainty that all patients considered in this research were actually assigned to a bed. It was not possible to include the urgency of the admission and/or a possible first impression of

the severity of the patient's condition as an explanatory variable in the predictive models. Most of the information contained in the medical record of a patient in HiX is described in words. This is beneficial for the communication between the doctors and nurses, because it can provide additional clarity regarding the admission. However, this makes it challenging to incorporate certain information that may be of importance for explaining a longer LoS, and that might have been already known at time of admission, into a predictive model as explanatory variable.

For further research, it could be interesting to consider a simulation model where the LoS prediction is adjusted when a mutation, i.e. a change in the admission, occurs. This way, it could be possible to include additional information that becomes known during the admission in the prediction framework. In addition, this allows for a LoS prediction that is conditional on the LoS of the current admission so far. As the considered models are incapable of predicting the LoS of patients that stay for a longer period of time in the hospital, it could be of interest to evaluate the effect of oversampling. However, this is probably accompanied by a worse overall performance, and this causes the data to no longer be representative of the original sample.

References

Aalbers, R. & Roos, A. (2022). Zorguitgaven, ons een zorg? *CPB*. Retrieved from `https://www.cpb.nl/sites/default/files/omnidownload/CPB-Publicatie-Zorguitgaven-ons-een-zorg.pdf`.

Aghajani, S. & Kargari, M. (2016). Determining factors influencing length of stay and predicting length of stay using data mining in the general surgery department. *Hospital Practices and Research*, *1*(2), 53–58.

Alsinglawi, B., Alshari, O., Alorjani, M., Mubin, O., Alnajjar, F., Novoa, M. & Darwish, O. (2022). An explainable machine learning framework for lung cancer hospital length of stay prediction. *Scientific Reports*, *12*(1), 607-610.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, *25*(6), 954–961.

Austin, P. C., Rothwell, D. M. & Tu, J. V. (2002). A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. *Health Services and Outcomes Research Methodology*, *3*(2), 107–133.

Azari, A., Janeja, V. P. & Mohseni, A. (2012). Predicting hospital length of stay (PHLOS): A multi-tiered data mining approach. In *Proceedings of the IEEE 12th International Conference on Data Mining Workshops* (pp. 17–24).

Bennett, S. (1983). Log-logistic regression models for survival data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *32*(2), 165–171.

Bottou, L. & Lin, C.-J. (2007). Support vector machine solvers. *Large-Scale Kernel Machines*, *3*(1), 301-320.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, *43*(1), 45–57.

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, *17*(31), 334–355.

Breusch, T. S. & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287–1294.

Caetano, N., Laureano, R. M. & 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* (pp. 407–414).

CBS. (2023a). In 2021 meer ziekenhuisopnamen dan in 2020. Retrieved from `https://www.cbs.nl/nl-nl/nieuws/2023/24/in-2021-meer-ziekenhuisopnamen-dan-in-2020`.

CBS. (2023b). Zorguitgaven stegen in 2022 met 1,2 procent. Retrieved from `https://www.cbs.nl/nl-nl/nieuws/2023/27/zorguitgaven-stegen-in-2022-met-1-2-procent`.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(2), 248–265.

Daghistani, T. A., Elshawi, R., Sakr, S., Ahmed, A. M., Al-Thwayee, A. & Al-Mallah, M. H. (2019). Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *International Journal of Cardiology*, *288*(1), 140–147.

Earnest, A., Chen, M. I. & Seow, E. (2006). Exploring if day and time of admission is associated with average length of stay among inpatients from a tertiary hospital in Singapore: An analytic study based on routine admission data. *BMC Health Services Research*, *6*(6), 1–8.

Escobar, G. J., Greene, J. D., Gardner, M. N., Marelich, G. P., Quick, B. & Kipnis, P. (2011). Intra-hospital transfers to a higher level of care: Contribution to total hospital and intensive care unit (ICU) mortality and length of stay (LOS). *Journal of Hospital Medicine*, *6*(2), 74–80.

Escobar, G. J., Turk, B. J., Ragins, A., Ha, J., Hoberman, B., LeVine, S. M., ... Kipnis, P. (2016). Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals. *Journal of Hospital Medicine*, *11*(1), 18–24.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, *46*(6), 1293–1301.

Habehh, H. & Gohel, S. (2021). Machine learning in healthcare. *Current Genomics*, *22*(4), 291-300.

Harerimana, G., Kim, J. W. & Jang, B. (2021). A deep attention model to forecast the length of stay and the in-hospital mortality right on admission from ICD codes and demographic data. *Journal of Biomedical Informatics*, *118*(1), 103778.

Hautz, W. E., Kämmer, J. E., Hautz, S. C., Sauter, T. C., Zwaan, L., Exadaktylos, A. K., ...

Schauber, S. K. (2019). Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *27*(54), 1–12.

Heij, C. (2004). *Econometric methods with applications in business and economics*. Oxford: Oxford University Press.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hosmer, D. W. & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. New York: John Wiley & Sons.

Hulshof, P. J., Kortbeek, N., Boucherie, R. J., Hans, E. W. & Bakker, P. J. (2012). Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems*, *1*(2), 129–175.

Huntley, D. A., Cho, D. W., Christman, J. & Csernansky, J. G. (1998). Predicting length of stay in an acute psychiatric hospital. *Psychiatric Services*, *49*(8), 1049–1053.

Jarque, C. M. & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259.

Jiang, R. & Murthy, D. (2011). A study of Weibull shape parameter: Properties and significance. *Reliability Engineering & System Safety*, *96*(12), 1619–1626.

Kalbfleisch, J. D. & Prentice, R. L. (2011). *The statistical analysis of failure time data*. New York: John Wiley & Sons.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*(282), 457–481.

Kiefer, N. M. (1988). Economic duration data and hazard functions. *Journal of Economic Literature*, *26*(2), 646–679.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, J. P., Moeschberger, M. L. et al. (2003). *Survival analysis: Techniques for censored and truncated data* (Vol. 1230). Berlin: Springer.

Kurniasari, D., Widyarini, R. & Antonio, Y. (2019). Characteristics of hazard rate functions of log-normal distributions. In *Proceedings of the 2nd International Conference on Applied Sciences, Mathematics, and Informatics* (pp. 1–6).

Launay, C., Rivière, H., Kabeshova, A. & Beauchet, O. (2015). Predicting prolonged length of hospital stay in older emergency department users: Use of a novel analysis method, the artificial neural network. *European Journal of Internal Medicine*, *26*(7), 478–482.

Liew, D., Liew, D. & Kennedy, M. P. (2003). Emergency department length of stay independently predicts excess inpatient length of stay. *Medical Journal of Australia*, *179*(10), 524–526.

List, N. & Simon, H. U. (2009). SVM-optimization and steepest-descent line search. In *Proceedings of the 22nd Annual Conference on Computational Learning Theory* (pp. 1–11).

Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Advances in Neural Information Processing Systems, *30*(1), 1–10.

Ma, F., Yu, L., Ye, L., Yao, D. D. & Zhuang, W. (2020). Length-of-stay prediction for pediatric patients with respiratory diseases using decision tree methods. *IEEE Journal of Biomedical and Health Informatics*, *24*(9), 2651–2662.

McAleese, P. & Odling-Smee, W. (1994). The effect of complications on length of stay. *Annals of Surgery*, *220*(6), 740–744.

McCullagh, P. (2019). *Generalized linear models*. Abingdon-on-Thames: Routledge.

Medisch Contact. (2018). ChipSoft marktleider van ziekenhuis-epd's. Retrieved from `https://www.medischcontact.nl/actueel/laatste-nieuws/artikel/chipsoft-marktleider-van-ziekenhuis-epds`.

M&I/Partners. (2021). EPD-marktinventarisatie ziekenhuizen 2021: Consolidatie EPD-markt zet door. Retrieved from `https://mxi.nl/kennis/541/epd-marktinventarisatie-ziekenhuizen-2021-consolidatie-epd-markt-zet-door`.

Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A. & Jha, N. K. (2014). Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE Journal of Biomedical and Health Informatics*, *19*(6), 1893–1905.

Nippak, P., Isaac, W., Ikeda-Douglas, C., Marion, A. & VandenBroek, M. (2014). Is there a relation between emergency department and inpatient lengths of stay. *Can J Rural Med*, *19*(1), 12–20.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*(1), 1–10.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Rathi, S. (2019). Generating counterfactual and contrastive explanations using SHAP. *arXiv preprint arXiv:1906.09293*.

Ravangard, R., Arab, M., Rashidian, A., Akbari, S. A., Zare, A. & Zeraati, H. (2011). Comparison of the results of Cox proportional hazards model and parametric models in the study of length of stay in a tertiary teaching hospital in Tehran, Iran. *Acta MedicaIranica*, *49*(10), 650-658.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Rijksoverheid. (2023). Plannen voor zorg en gezondheid. Retrieved from `https://www.rijksoverheid.nl/onderwerpen/prinsjesdag/zorg-en-gezondheid`.

Ryan, K., Levit, K. & Davis, P. H. (2010). Characteristics of weekday and weekend hospital admissions. *Agency for Healthcare Research and Quality*. Retrieved from `https://www.ncbi.nlm.nih.gov/books/NBK53602/`.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241.

Steele, R. J. & Thompson, B. (2019). Data mining for generalizable pre-admission prediction of elective length of stay. In *Proceedings of the IEEE 9th Annual Computing and Communication Workshop and Conference* (pp. 127–133).

Sykora, D., Traub, S. J., Buras, M. R., Hodgson, N. R. & Geyer, H. L. (2020). Increased inpatient length of stay after early unplanned transfer to higher levels of care. *Critical Care Explorations*, *2*(4), 1–5.

Turgeman, L., May, J. H. & Sciulli, R. (2017). Insights from a machine learning model for predicting the hospital length of stay (LOS) at the time of admission. *Expert Systems with Applications*, *78*(1), 376–385.

Vapnik, V. (1995). *The nature of statistical learning theory.* Berlin: Springer.

Verburg, I. W., de Keizer, N. F., de Jonge, E. & Peek, N. (2014). Comparison of regression methods for modeling intensive care length of stay. *PLOS One*, *9*(10), 1–11.

Wang, Y., Stavem, K., Dahl, F. A., Humerfelt, S. & Haugen, T. (2014). Factors associated with a prolonged length of stay after acute exacerbation of chronic obstructive pulmonary disease (AECOPD). *International Journal of Chronic Obstructive Pulmonary Disease*, *9*(1), 99–105.

Weerts, H. J., Mueller, A. C. & Vanschoren, J. (2020). Importance of tuning hyperparameters of machine learning algorithms. *arXiv preprint arXiv:2007.07588*.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, *1*(4), 339–356.

Yoon, P., Steiner, I. & Reinhardt, G. (2003). Analysis of factors influencing length of stay in the emergency department. *Canadian Journal of Emergency Medicine*, *5*(3), 155–161.

A Estimated coefficients Cox's proportional hazards model

As the assumption of Cox's PH model does not hold, the estimated coefficients by this model are incorrect. As a result, the estimated coefficients by Cox's PH model are not presented in the main research and Cox's PH model is also not used to construct LoS predictions. Table 13 shows both the coefficients estimated by the parametric model using the log-logistic distribution and the estimated coefficients of Cox's PH model. The coefficients of the duration model using the log-logistic distribution were already presented in Table 7 of Section 5.1.2 in the main research. By also showing the estimated coefficients of the duration model using the log-logistic distribution in Table 13, it is possible to directly deduce how the estimated coefficients by Cox's PH model relate to the coefficients of the duration model using the log-logistic distribution. It directly stands out in Table 13 that the coefficients estimated by Cox's PH model and the parametric model using the log-logistic distribution have opposite signs. This can be explained by the fact that the coefficients have a different interpretation in these models. In Section 4.1 of the main research, it is mentioned that the exponent of an estimated coefficient by Cox's PH model represents the effect of an increase in an explanatory variable on the hazard rate. In case of the duration model using the log-logistic distribution, the estimated coefficients represent the effect of the explanatory variables on the log of the scale parameter. As a result, a positive coefficient leads to a decrease in the hazard rate for the duration model using the log-logistic distribution, while in case of Cox's PH a positive coefficient represents an increase in the hazard rate. Note that Cox's PH model does not contain an intercept, because the baseline hazard rate takes the place of it. In case of Cox's PH model, the baseline hazard includes all elements of the hazard that do not depend on the explanatory variables, including any intercept term, which remains constant for all observations by definition.

Table 13: Estimated coefficients by a duration model assuming the log-logistic distribution and Cox's proportional hazards (PH) model.

	Model	
Variable	Log-logistic	Cox's PH
Intercept	0.502 (0.009)***	
Known at time of admission		
Age	0.005 (0.000)***	-0.007 (0.000)***
Is man	0.006 (0.003)*	-0.028 (0.004)***
Admission type		
Day admission		
Clinical admission	1.996 (0.006)***	-2.639 (0.011)***
Observational admission	0.851 (0.009)***	-1.057 (0.013)***
Polyclinical admission	-0.835 (0.005)***	2.547 (0.009)***
Specialism		
Cardiology (CAR)		
Dermatology (DER)	0.496 (0.329)	-0.294 (0.378)
Geriatrics (GER)	1.128 (0.014)***	-1.320 (0.019)***
Internal medicine (INT)	0.549 (0.006)***	-0.471 (0.008)***

Pediatrics (KIN)	0.831 (0.009)***	-0.962 (0.014)***
Pulmonary medicine (LON)	0.281 (0.006)***	-0.101 (0.009)***
Gastrointestinal diseases (MDL)	0.190 (0.006)***	0.197 (0.008)***
Neurology (NEU)	0.297 (0.007)***	-0.189 (0.010)***
Psychiatry (PSY)	-1.559 (0.026)***	-1.107 (0.044)***
Rheumatology (REU)	0.293 (0.011)***	0.070 (0.018)***
Rehabilitation (REV)	0.719 (0.078)***	-0.857 (0.137)***
Admission information		
Is emergency	0.007 (0.006)	-0.005 (0.008)
Is during weekend	0.026 (0.003)***	-0.049 (0.005)***
Is outside office hours	0.173 (0.004)***	-0.161 (0.006)***
Risk factors		
Cardiac arrhythmias	-0.024 (0.004)***	0.014 (0.007)*
COPD	0.069 (0.004)***	-0.076 (0.007)***
Obese	0.011 (0.016)	-0.099 (0.024)***
Diabetes	0.027 (0.006)	-0.014 (0.009)
Previous admissions		
Number of previous complications	-0.002 (0.000)***	0.004 (0.001)***
# Admissions past 3 months	-0.010 (0.002)***	0.001 (0.003)***
# Admissions past 6 months	-0.003 (0.002)*	0.016 (0.003)***
# Admissions past 12 months	-0.002 (0.001)*	0.005 (0.001)***
Total LoS past 3 months	0.001 (0.000)***	-0.001 (0.000)***
Total LoS past 6 months	0.000 (0.000)	-0.000 (0.000)***
Total LoS past 12 months	0.000 (0.000)***	-0.000 (0.000)***
Becomes known during admission		
# Surgeries	0.203 (0.006)***	-0.245 (0.008)
Complication occurred	0.009 (0.004)*	0.012 (0.006)*
# Mutations	0.474 (0.002)***	-0.424 (0.003)***
Mutation to ICU	-0.188 (0.041)***	-0.206 (0.051)
# Diagnosis changes	0.053 (0.010)***	-0.041 (0.013)**

Note. Standard errors are in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

B Hyperparameter tuning

In Section 4.2.3 of the main research it is stated that the hyperparameters for the machine learning models are tuned by performing a grid search using stratified k-fold cross-validation. This section presents the grids that are used for tuning each model for both classification and regression. For each model, it is discussed which hyperparameters are tuned and it is presented which values for these hyperparameters are found as the best. If there are multiple hyperparameters tuned in a model, the evaluated values and a short description for each hyperparameter are shown in a table. In these tables, the best found hyperparameters in classification are presented in bold, and the best hyperparameters in regression are underlined.

Random Forest. As became clear earlier in Section 4.2.2 of the main research, RF uses a bagging scheme to select a random subset of variables for each bootstrap sample in order to ensure that the trees do not become correlated. The hyperparameter that is tuned for RF is the number of variables that is included in the random subset of variables. Including the admission types and specialisms, the total number of explanatory variables is equal to 31. For classification, it is evaluated which value of the set {1, 2, 3, 5, 10, 15, 20, 25, 30, 31} is the best number of variables to be included in the random subset by RF. In case of regression, RF requires more memory and computation time compared to classification. This is due to the fact that RF needs to store and compute the average of the target values at each leaf node in case of regression, while the algorithm only needs to store the majority vote counts at each leaf node of the trees when it is used for classification. Hence, to reduce the computation time of tuning the hyperparameters for RF in case of regression, a smaller set of values for the number of variables to be included in the random subset is considered. For regression, it is examined which value of the set {1, 2, 3, 5, 10, 15} is the best. For both classification and regression, the number of trees is held constant at 500. In case of classification, the best found number of variables to be included in the random subset is equal to 5. Also for regression the best value for this hyperparameter is 5.

eXtreme Gradient Boosting. Table 14 shows the tuned hyperparameters for XGBoost. The maximum depth of the tree, *max_depth*, determines how many splits a decision tree in XGBoost can have. Higher values of max_depth make the XGBoost model more complex and more likely to overfit. The learning rate is controlled by *eta*, as it scales the contribution of each tree when it is added to the current approximation. Lower values of eta make XGBoost more robust to overfitting, but this is accompanied by more computation time. In addition, eta controls the number of boosting iterations to be performed by XGBoost, where lower eta values imply more boosting iterations. The hyperparameter *gamma* represents the minimum loss reduction needed to create an additional split on a leaf node in the decision tree. A higher gamma value makes the XGBoost algorithm more cautious about making new splits. The subsample ratio of the training data is represented by *subsample*. Lower values for subsample prevent overfitting and lead to shorter computation times.

Table 14: Short description and evaluated values for each tuned hyperparameter in eXtreme Gradient Boosting (XGBoost).

Hyperparameter	Description	Evaluated values
max_depth	Maximum tree depth	3, **6**, 9, 12
eta	Shrinkage	0.01, **0.05**, 0.1, 0.3
gamma	Minimum loss reduction	0, **0.1**, 0.2
colsample_bytree	Subsample ratio of columns	1
min_child_weight	Minimum sum of instance weight	1
subsample	Subsample percentage	**0.5**, 1

Note. The best found hyperparameters for XGBoost in classification are presented in bold, and the best hyperparameters in regression are underlined.

Recurrent Neural Network. The tuned hyperparameters for the RNN are shown in Table 15. The hyperparameter *units* represents the number of neurons in each recurrent layer. A higher

value for the number of units increases the model its capacity to learn complex relationships, but leads to more computation time and increases the risk of overfitting. The *dropout_rate* represents the fraction of input units to drop. Properly tuning this hyperparameter prevents the model from becoming too reliant on specific neurons. The *learning_rate* determines the step size of the optimizer in each iteration while moving towards a minimum for the loss function. Lower values for the learning rate lead to slow convergence, while higher values can cause the model to converge too quickly to a suboptimal solution or even diverge.

Table 15: Short description and evaluated values for each tuned hyperparameter in the Recurrent Neural Network (RNN).

Hyperparameter	Description	Evaluated values
units	# Neurons in each recurrent layer	50, 100, **150**
dropout_rate	Fraction of the input units to drop	**0**, 0.1, 0.2, 0.4
learning_rate	Step size of the optimizer	0.001, **0.01**, 0.1

Note. The best found hyperparameters for the RNN in classification are presented in bold, and the best hyperparameters in regression are underlined.

Support Vector Machine. As mentioned earlier in Section 4.2.2 of the main research, the Gaussian radial basis function is used as kernel in SVM. The hyperparameter *sigma* represents the inverse kernel width. Lower values of sigma make the kernel more narrow, which causes the model to be more sensitive to individual data points. This leads to a more complex decision boundary that can capture more detailed relationships between the variables. However, this increases the risk of overfitting, because the model could become too sensitive to noisy data points. Likewise, a higher value of sigma leads to a wider kernel function, and makes the model less sensitive to individual data points. As a result, the model is more focused on capturing the overall trend of the data. Using a wider kernel reduces the risk of overfitting, but it can also lead to underfitting, as the model could become too simple to capture the most important relations in the data. An interval of possible values for sigma is estimated based on the complete training set. Any value between the two boundary values should produce good results, and therefore the middle value between the two boundary values is selected as the sigma value. The selected value of sigma, in both classification and regression, is equal to 0.02191.

In case of the regression problem, SVM contains an additional hyperparameter, *epsilon*, which represents the maximum deviation of the fitted function from the actual target values. Larger values of epsilon make SVM less sensitive to small errors, which leads to a simpler model that could possibly generalize better to unseen data. For smaller epsilon values, the model tries to fit the training data more precisely. This leads to a more complex model that fits the training data very well, but this causes an increased risk of overfitting and performing poorly on new data. In this research, the value for epsilon is held constant at 0.1.

The hyperparameter C is the cost regularization parameter, which controls the smoothness of the fitted function. A lower value for C leads to a smoother decision boundary, allowing some misclassifications. This makes the model more robust to outliers, but it could also lead to underfitting if the value for C is too low. Higher values of C put more emphasis on minimizing the training error, as a result of which the fit to the training data is 'tighter'. This means that fewer margin violations are allowed, potentially leading to a more complex model. While

this can reduce training error, it increases the risk of overfitting, especially if the training data contains noise. For classification, it is evaluated which value of the set {0.125, 0.25, 0.5, 1, 2} is the best value for C. As stated earlier in Section 4.2.2 of the main research, the resulting QP problem in case of regression contains additional variables and constraints compared to the classification problem, which makes the model more complex. As a result, SVM requires more computation time compared to classification. Therefore, in order to reduce the computation time of tuning the hyperparameters for SVM in case of regression, the set of evaluated values for C is smaller. For the regression problem, it is examined which value of the set {1, 2} is the best value for C. The best value for C in case of classification is equal to 2, and for regression the best value for C is also 2.

C Prediction diagnostics Random Forest and Support Vector Machine

As already stated in Section 5.2.2 of the main research, the XGBoost model is recommended to ChipSoft for a practical implementation in HiX. In Section 5.3 of the main research, it was chosen to present only the prediction diagnostics of the most accurate statistical model and the machine learning model that is best suited for a practical implementation, namely the linear regression model using the log LoS as dependent variable and XGBoost, respectively. SVM performed better than XGBoost in regression, and RF achieved a performance similar to XGBoost. Hence, for reference, the prediction diagnostics of RF and SVM are shown here, because it could still be of interest for both ChipSoft and possible other research purposes.

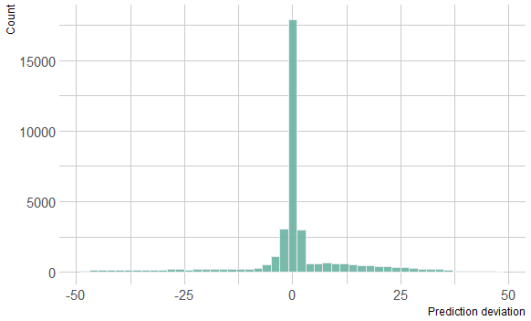
Table 16 shows the percentages of predictions by XGBoost, RF, and SVM for which the absolute prediction deviation is smaller than a certain value. It becomes clear that the percentages of RF and SVM are comparable to those of XGBoost. The percentages of predictions where the absolute prediction deviation is 2 hours or smaller is a little lower for SVM than the respective percentages of XGBoost and RF. However, the percentage of predictions that have an extreme prediction deviation is lower for SVM, because the percentage of predictions where the absolute prediction deviation is 168 hours or smaller is 97.71% for SVM, which is higher than the respective percentages of XGBoost and RF.

Table 16: Percentages of predictions by eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machine (SVM) for which the absolute prediction deviation is smaller than a certain value.

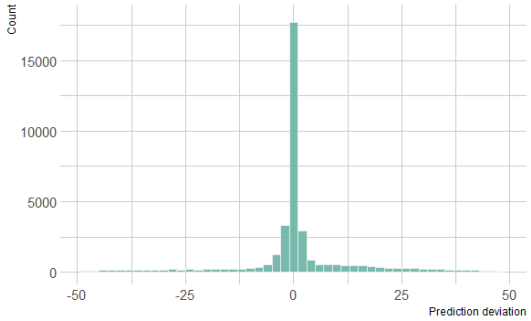
	Absolute prediction deviation (in hours)						
Model	≤ 2	≤ 6	≤ 12	≤ 24	≤ 48	≤ 96	≤ 168
XGBoost	53.02%	63.19%	69.85%	78.22%	87.39%	93.63%	97.46%
RF	53.20%	63.08%	69.58%	78.58%	87.46%	93.49%	97.36%
SVM	52.68%	64.09%	69.93%	78.20%	86.99%	93.75%	97.71%

Figure 4 shows the histograms of the deviation from the actual values for the predictions by RF and SVM. These histograms have a similar shape to the histogram of the deviation from the actual values for the predictions by XGBoost presented by Figure 2 in Section 5.3 of the main research. This is not unexpected, because the percentages of predictions by RF and SVM for

which the absolute prediction deviation is smaller than a certain value, as shown in Table 16, are also fairly similar to the respective percentages of XGBoost.



(a) Histogram of the prediction deviations RF.



(b) Histogram of the prediction deviations SVM.

Figure 4: Histograms of the deviations from the actual values for the predictions by Random Forest (RF) and Support Vector Machine (SVM).

D Switched hyperparameters

In this research, the hyperparameters are tuned by using stratified 5-fold cross-validation over the training set. The best hyperparameters over the five folds are used as hyperparameters for testing. It is not necessarily the case that the best found hyperparameters over the training set are also the best for the test set (Weerts et al., 2020). In addition, it is possible that the best hyperparameters for classification and regression are not the same. Therefore, it is evaluated whether the performance of the machine learning models can change when the best hyperparameters for regression are used in classification, and vice versa. For both RF and SVM the best found hyperparameters for classification were the same as for regression. Hence, only for XGBoost and the RNN the hyperparameters for classification and regression are switched.

Table 17 shows the classification performance of XGBoost and the RNN when the best found hyperparameters for regression are used. From this table it becomes clear that the XGBoost achieves a lower overall accuracy than when the best found hyperparameters for classification are used. The RNN obtains the same overall accuracy as when using the best found hyperparameters for classification. The RNN is able to detect more observations belonging to the medium LoS class. However, this is accompanied by the fact that fewer observations are correctly classified as having a short or long LoS.

Table 17: Classification performance of eXtreme Gradient Boosting (XGBoost) and the Recurrent Neural Network (RNN) where the best found hyperparameters for regression are used.

Model	Performance measure	Total	LoS class		
			Short	Medium	Long
XGBoost	Accuracy	83.22%			
	Precision		98.56%	61.77%	59.57%
	Recall		91.20%	92.08%	17.92%
RNN	Accuracy	82.61%			
	Precision		97.96%	62.30%	48.96%
	Recall		90.74%	90.46%	18.76%

Table 18 presents the regression performance of XGBoost and the RNN where the best found hyperparameters for classification are used. XGBoost performs worse than when the best found hyperparameters for regression are used. It is striking that the RNN performs better using the best found hyperparameters for classification. Although the difference in performance with respect to using the best found hyperparameters for regression are not substantial, it is still notable. This result emphasizes that it is not necessarily the case that the best found hyperparameters over the training set are also the best for the test set. It is not necessarily unexpected that the best found hyperparameters for the RNN in classification achieve similar results as when the best found hyperparameters for regression are used. The combination of hyperparameters that was best for classification achieved an average MAE over the five folds that was only slightly higher than the MAE of the hyperparameter combination that was selected as the best for regression.

Table 18: Regression performance of eXtreme Gradient Boosting (XGBoost) and the Recurrent Neural Network (RNN) where the best found hyperparameters for classification are used.

Model	Performance measure	Total	LoS class		
			Short	Medium	Long
XGBoost	MAE	20.69	2.97	20.90	120.26
	WAPE	63.10%	82.17%	51.25%	67.09%
RNN	MAE	21.51	3.41	21.26	124.27
	WAPE	65.60%	94.32%	52.15%	69.32%

Note. The length of stay (LoS) classes are the actual classes.