DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH

A COMPARISON BETWEEN EXISTING MORTALITY RISK ALGORITHMS AND MACHINE LEARNING TECHNIQUES:

A Retrospective Analysis of Covid-19 Patient Mortality Risk for a Large South African Private Hospital Group

by

Assignment presented in partial fulfilment of the requirements for the degree of Master's of Commerce at the University of Stellenbosch.

Supervisor: Prof. Rulof Burger & Riani Retief

December 2022

Declaration

I, the undersigned, hereby declare that:

(i) the work contained in this assignment is my own work; and

(ii) in the instance where my research assignment is based on previously submitted work, I have provided detailed information:

(a) regarding the nature, substance and origin of the overlap in the space below and throughout my research assignment (using standard referencing conventions or footnotes),

(b) regarding content that has been added to the previous submission,

(c) and that I understand that the evaluation of my research assignment will be primarily based on the new work.

By submitting this thesis electronically, I declare that the entirety of the work contained herein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: …….....

Date: ……24 October 2022.....

A Comparison Between Existing Mortality Risk Algorithms and Machine Learning Techniques - A Retrospective Analysis of Covid-19 Patient Mortality Risk for a Large South African Private Hospital Group

Jenny Swart[a]

[a]*Stellenbosch University, South Africa*

Abstract

This thesis assesses the feasibility and benefits of using the patient data of a large private South African hospital group to estimate a model of mortality risk using flexible machine learning techniques. Specifically, I investigate whether such a model would have been able to outperform a commonly used medical scoring system, SAPS 3, in predicting mortality during the second half of the Covid-19 pandemic. A LightGBM machine learning model is shown to be much more accurate in predicting mortality (76.15% accuracy, compared to 56.58% for SAPS 3) for the Covid-19 positive sample. Roughly half of this gain in predictive accuracy is obtained from using the most recent and relevant data to train the model, while the remaining lift is attributable to allowing the model to find patient symptoms and attributes that are measured but ignored by SAPS 3. Interestingly, the flexible functional form of the machine learning models, which allow the predictors to affect mortality through non-linearities and interactions, has a negligible effect on predictive accuracy. The same method is also found to produce more accurate forecasts for patients who tested negative for Covid-19, but this improvement is smaller than for Covid-19 positive sample. The results of this thesis illustrate that machine learning methods are valuable tools to predict patient outcomes, particularly when there are unexpected shifts in the relationship between patient features and patient outcomes. Large hospital groups can obtain more accurate forecasts from a dynamic scoring system which is frequently retrained on their own patient data.

Table of Contents

List of Figures

1. Introduction

Coronavirus disease (Covid-19) is an infectious disease caused by the SARS-CoV-2 virus and was first reported in December 2019. Since then, it has caused an estimated 559 infections and 6.3 million deaths. In response to its spread governments implemented restrictive lockdowns which disrupted international and local travel, supply chains, and economic activity. In a pandemic, or any kind of health crisis, patient triage is required. Ideally, this triage would maximise the benefits gained from limited resources through the quick and effective allocation of patients to appropriate risk groups. However, humans are often emotional, distractible, and prone to cognitive biases and this is particularly true during the uncertainty and stress of a health crisis or pandemic. Due to this, triage would ideally be implemented using an algorithm that reflects the relevant objectives and based on appropriate data.

Such algorithms exist as medical scoring systems. These systems are particularly useful during pandemics, when the availability of intensive care unit (ICU) beds and ventilators are a binding constraint on medical care, and human decision makers are overburdened. Unfortunately, it is exactly during such episodes that the data used to calibrate these scoring rules may be less relevant. For instance, the medical scoring systems applied during the Covid-19 pandemic in South Africa (e.g., SAPS 3), suffer from several shortcomings: i) they utilise very simple formulas for easy application, which can reduce their accuracy, ii) they are estimated on pre-Covid data, and therefore do not reflect pandemic-era health risks, iii) they are estimated on non-SA data and therefore may not be optimally informative about the risk faced by South African patients. It may be preferable to develop a dynamic scoring model which periodically re-estimates mortality risk using the most recent and relevant data. Large hospital groups may create sufficient sample sizes to do this on their own data, without any need for data from other countries which may be less relevant for the mortality risk of its own patients. If technology facilitates ease of application (e.g., by nurses simply entering symptoms on a tablet), then perhaps an oversimplified algorithm is not necessary.

This thesis will assess the feasibility and potential benefits of using the patient data of a large South African hospital group to assess mortality risk using recent data and sophisticated and flexible statistical techniques. This will be achieved by comparing the predictive accuracy of a model estimated on the hospital's own patient data relative to the predictions obtained from the most commonly used conventional medical scoring system: SAPS 3 (Simplified Acute Physiology Score 3). The patient sample is restricted to patients admitted to the ICU and who also tested positive for Covid-19, since these are the patients for whom rapid and accurate clinical decisions are most impactful. The empirical analysis will proceed in three steps, in

order to assess the accuracy costs of each of the three shortcomings of the SAPS 3 system listed above. First, a linear additive prediction model is estimated on the hospital's own data and using the same variables included in SAPS 3. This will reveal how much predictive accuracy can be gained from using the most relevant data to predict mortality risk. Secondly, flexible machine learning methods will be estimated on the same patient data. These methods allow for non-linear and interaction effects of risk factors on mortality risk, so will indicate how much predictive accuracy can be gained from allowing more flexible functional forms than SAPS 3. Thirdly, the same models will be estimated using patient attributes and clinical indicators that are available but not used in calculating the SAPS 3 score. This will reveal the benefit of using an approach that allows more factors to determine mortality risk.

The Covid-19 pandemic is an example of a rare global medical emergency during which patient triage decisions were exceptionally important and difficult. This provides an interesting backdrop to the question of whether dynamically updated algorithms could have improved in-hospital decision-making. However, the anomalous nature of this event may call into question the external validity of this study's conclusion: would these models be similarly useful in normal times when the relationship between patient attributes and mortality risk is more stable? In order to answer this question, the models are also estimated on patients who tested negative for Covid-19.

The results of this thesis illustrate that machine learning methods are valuable tools to predict patient outcomes, and this is true when there are unexpected shifts in the relationship between patient features and patient outcomes but also under normal circumstances because machine learning techniques are particularly good at identifying interesting interactions and explaining model errors to improve prediction accuracy. To conduct the analysis, a gradient boosting framework implementation, LightGBM, was utilised particularly because of its ability to accelerate the training process up to 20 times while maintaining similar accuracy, in contrast to other implementations. Since LightGBM provides both efficiency and accuracy, the results of this thesis can be extremely powerful if hospitals were to implement it by developing a dynamic scoring model to calculate mortality risk for triage decisions. Furthermore, to the best of my knowledge, this thesis is the first to provide a mortality risk prediction model for ICU patients infected with Covid-19 in South Africa.

The sections are outlined as follows: Section 2 provides a literature overview of triage rules and algorithms, machine learning techniques which have been implemented to predict hospital mortality, and the SAPS 3 scoring system. Section 3 discusses the data provided by a large South African hospital for the purpose of the thesis. Section 4 and 5 provides the methodology

and empirical analysis conducted in this thesis, and Section 6 concludes.

2. Literature Review

During the Covid-19 pandemic the sudden influx of infected patients posed challenges to health-care services, overwhelming hospitals and intensive care units. Medical professionals were required to allocate critical but limited medical resources, such as ventilators and intensive care beds, to patients who would benefit most from treatment (Emanuel et al., 2020). Patient triage is required during a pandemic or a health crisis and, should ideally be quick and effective to allocate patients to the appropriate risk group to ensure that maximal benefits are gained from limited resources. In reality, patient triage occurred while doctors, nurses, hospital administrators and policy makers had incomplete information regarding the health effects and risk factors of the pandemic, and limited cognitive bandwidth to make important decisions. It is well documented that humans are emotional, distractible, and prone to cognitive biases and this is particularly true during the emotional distress of a pandemic (Tversky & Kahneman, 1974). Triage decisions, including withdrawing ventilators from one patient to allocate it to another, might result in emotional distress for medical professionals, which could hamper the quality of decision-making. Due to this, triage should ideally occur using algorithms that are based on appropriate data and reflect relevant objectives.

There is a substantial literature on the socio-ethic ramifications of these triage rules, and what they should attempt to achieve, and this debate has been reignited during the Covid-19 pandemic. Under normal circumstances, when medical resources are abundant, the odds of survival after treatment and disease severity are the primary determinants of resource allocation in triage (O'Laughlin & Hick, 2008). However, during a pandemic resources are constrained and often allocated to benefit the greatest number of patients in terms of lives saved or life-years gained (Fiest et al., 2020). Scarce resources would be allocated to patients who are more likely to survive, and live a long and healthy life subsequently. White & Lo (2020) and Basu (2021) have recommended that triage frameworks should be developed with priority scores determined by multiple criteria, including the likelihood of survival until discharge, likelihood of long-term survival based on comorbidities, and the patient's current life cycle stage. There is a broad consensus that the benefits gained from limited resources should be maximised in the above-mentioned regard, while giving priority to the worst-off and ensuring non-discrimination. However, how exactly this is achieved remains a point of contention (Emanuel et al., 2020; Basu, 2021). Random selection and first-come, first-served approaches have been proposed by some (Emanuel et al., 2020) while others emphasise social justice and non-discrimination ideals as

means to uphold equality (Reid, 2020), with the exception of instrumental value advocated by saving those who can save others, such as nurses and other health care workers (DB. et al., 2009). Other recommendations include that triage guidelines should differ by intervention and should be flexible to new scientific evidence (Emanuel et al., 2020). The objectives of these triage rules fall outside of the scope of this study, which will focus only on modelling mortality risk, which is an important input into any sensible triage rule.

Psychologists have identified several cognitive biases that can affect human judgements. Such biases are one of the reasons that motivated the establishment of triage committees who would formulate general triage algorithms to remove the responsibility from bedside clinicians (Truog, Mitchell & Daley, 2020) As mentioned previously, these should ideally depend on algorithms which reflect clear objectives and are based on relevant data and medical evidence. These algorithms exist as medical scoring systems. Most hospitals use one of a variety of medical scoring systems to allocate patients to ICU beds to ensure effective and consistent treatment. These scoring systems use easily measured data in statistical algorithms to provide informative single scores to enable medical professionals to assess patient conditions, to estimate and stratify risk, to predict outcomes and to diagnose diseases accurately. Specifically, scoring systems are used to predict mortality risk in hospitals and intensive care units to inform decisions for patient triage and treatment.

Although medical scoring systems are a crucial input into patient triage decisions, such scoring systems usually suffer from several shortcomings. Many scoring systems (including SAPS 3, which is discussed in more depth in section 2.2 below) make use of very simple linear additive formulas, with coefficients being restricted to integers. This simplified functional form ensures convenience of use and allows nurses to implement without the need for computers or calculators. However, if the most accurate mortality forecasting model is a non-linear function of patient attributes and symptoms, then a linear approximation necessarily implies sacrificing model fit and by implication forecasting accuracy. This would imply that the predicted mortality risk is not maximally accurate, and some patients may be incorrectly prioritised over others. Furthermore, these scoring systems were developed prior to the pandemic, which means that they cannot capture new interactions and risks that are relevant in a Covid-19 disease environment. Finally, these scoring systems were estimated on non-SA data which implies that it might not be optimally relevant for triage decisions in a South African environment. SAPS 3 has been successfully implemented in intensive care units worldwide and is among the most used modern scoring systems. However, these potential drawbacks emphasise the need for an alternative scoring system during the Covid-19 pandemic and in a South African context to aid decision-making in patient triage through accurate and effective mortality risk prediction.

Given all these shortcomings, it may be useful for hospitals to develop a dynamic scoring model which periodically uses the most recent and relevant data to calibrate this algorithm, instead of using a static scoring model. Large hospital groups may create sufficient sample sizes to do this with their own patient data. If modern technology can facilitate ease of application – by nurses simply entering symptoms on a tablet, which then calculates the mortality risk, for example - then an oversimplified algorithm is unnecessary. These prediction models for mortality risk could then be used to make more effective triage choices, and to inform policy decisions.

This need for a prediction model that could account for potential non-linear and interaction effects of patient features and symptoms on mortality risk, and that is applicable during Covid-19, has been recognised by several researchers who accordingly set out to model these complexities using various machine learning techniques (Banoei et al., 2021; Ottenhoff et al., 2021; Pourhomayoun & Shakibi, 2021). There is a related literature that uses machine learning techniques to to predict the number of new Covid-19 cases to inform policymaking (Al-qaness, Ewees, Fan & Abd ElAziz, 2020; Al-qaness, Ewees, Fan, Abualigah, et al., 2020; Alsayed et al., 2020; ArunKumar et al., 2021). The most common machine learning techniques that are used in these studies are Support Vector Machines (SVM), Artificial Neural Networks, Random Forests, Logistic Regressions, K-Nearest Neighbours (KNN), Tree-based Gradient Boosting (XGB), and shrinkage methods such as LASSO (Least Absolute Shrinkage and Selection Operator1. The empirical analyses conducted by these studies utilised different data sources with different patient characteristics and used different variables to predict mortality risk. Despite the differences in data and variables used, a consensus has emerged regarding the risk features that significantly influence mortality among Covid-19 patients: age, hypertension, asthma, respiratory rate, oxygen saturation, diabetes, systolic blood pressure and the Glasgow Coma Scale (GCS Score).

Most of this literature has estimated models that predict mortality risk for all hospitalised patients, whereas this study will focus only on critically ill patients admitted to intensive care units. Critically ill patients are uniquely characterised by uncertainty and their swift deterioration, which makes a model that can predict mortality risk for critically ill patients essential. In this respect, this study is similar to Zhai et al. (2020) which also restricted its sample to patients admitted to emergency department intensive care units in Chinese hospitals, and used XGBoost, SVM and a Logistic Regression to predict mortality risk. However, the patient sample used in this study preceded the Covid-19 pandemic so the estimated models could not assess the benefits of using this approach during the pandemic. It is also doubtful that the predictions of models estimated on patients from one country could be accurately extrapolated to the patients of a different country. South Africa is a country uniquely characterised by

a high incidence of diseases such as HIV and Tuberculosis which could potentially alter the relationship between patient features and mortality risk in the presence of Covid-19. SAPS 3 is currently the only scoring system used in intensive care units in South African hospitals to predict mortality and, to the best of my knowledge, no research exists that have predicted mortality of intensive care patients infected with Covid-19 in South Africa, further emphasising the need and urgency of this study.

2.1. SAPS 3

The Acute Physiology and Chronic Health Evaluation Score (APACHE) and the Simplified Acute Physiology Score (SAPS) exist among several scoring systems developed to classify the severity of disease of patients admitted to ICU. The original APACHE consisted of two parts: the Acute Physiological Score (APS) and the Chronic Health Evaluation (CHE) which represents the severity of acute illness and the physiological status of the patient prior to the illness, respectively. SAPS was mainly designed to overcome the complexity of APACHE. Approximately ten years after the first publication of the APACHE and SAPS came the newer generation of these instruments, SAPS II and APACHE 3, which performed significantly better than their forerunners since they were developed using sophisticated statistical techniques and larger multinational data (Metnitz et al., 2005). These risk adjustment models significantly facilitated research in critical care. However, over time more studies emerged assessing the performance of these risk adjustment systems, exposing their insufficient prognostic performance (Apolone et al., 1996; Poole et al., 2009; Saleh et al., 2015). The poor prognostic performance was evident in the lack of calibration for patient subgroups and an overestimation of mortality for high-risk patients and an underestimation of mortality for low-risk patients. Since the collection of SAPS II's database the prevalent major diseases have changed as well as the available and regularly used diagnostic methods which have resulted in poor calibration. Moreover, this database was developed on European and North American patients which implies that the sample is potentially not representative of medical practices in ICUs worldwide. Given that outcome is likely related ICU practices, the results of the model have limited generalisability. Researchers accordingly aimed to improve these systems through recalibration on the equation for mortality prediction, and recalibration on the variable weighting within the model. As the problems seemed to be intrinsic to the models, because the population baseline characteristics likely changed over time and important prognostic variables were excluded, recalibration was unable to solve them. Since recalibration was ineffective to improve the prognostic performance of the models, a new model had to be developed which included variables shown to be influential on the outcome. SAPS 3 was thus developed in 2002 with the objective to provide an

improved risk-adjustment model for critically ill patients available free of charge to the scientific community (Metnitz et al., 2005).

SAPS 3 was published in 2005 as a significant improvement over SAPS II and SAPS to predict mortality. The main difference observed in the SAPS 3 model is the usage of patient features available within the first hour of ICU admission, rather than the first 24 hours of admission and contrary to other commonly used scores, the major driving force for its predictive power is drawn from patient features known prior to admission. The data describes prior chronic conditions and diseases, circumstances related to and physiological derangement at ICU admission. A total of 16 784 patients consecutively admitted to 303 ICUs across America, Europe, Mediterranean and Australasia during a 2-month period in 2002 comprises the SAPS 3 Hosptial Outcome Cohort (Metnitz et al., 2005). The data were collected on days 1, 2, 3 and on the last day of ICU stay. Data collected on day 1 of admission (one hour before or after admission) were categorised into three levels to describe the patient's condition before ICU admission to indicate any chronic illnesses or medical conditions, the reason for admission and any infections or surgery done at time of admission, and the patient's physiological derangement at admission. Days 2 and 3, and the last day of ICU stay information was collected to describe the severity of illness, organ dysfunction information, length of stay and vital status at discharge. Five-fold cross-validation was employed to build and validate the model. Thereafter, a logistic regression was used to simplify the model and a multilevel logistic regression was implemented to estimate the regression coefficients. Lastly, bootstrapping methods were used to check the variable selection and their weighting. The final model included 20 different variables and major geographic areas have their own customised equations (Moreno et al., 2005).

The final selected variables are contained in three respective boxes or classifications. This is because SAPS 3 distinguishes between evaluation of the individual patient from evaluation of the ICU. For individual assessment, the patient factors are therefore captured in the information that is available prior to ICU admission, i.e., Box I, which interestingly provides half of the model's predictive power. Five variables are used for evaluating Box I: age, co-morbidities, use of vasoactive drugs before ICU admission, intrahospital location before ICU admission, and length of stay in the hospital before ICU admission. The prognostic performance of the model can further be enhanced by including the data relevant to the circumstances at ICU admission, i.e., Box II, which includes reason(s) for ICU admission, planned/unplanned ICU admission, surgical status at ICU admission, anatomical site of surgery, and presence of infection at ICU admission and place acquired. Another 27.5% can be won by incorporating the physiological variables contained in Box III. Box III consists of lowest estimated Glasgow coma scale, highest heart rate, lowest systolic blood pressure, highest bilirubine, highest body temperature, highest

creatinine, highest leukocytes, lowest platelets, lowest hydrogen ion concentration (pH), and ventilatory support and oxygenation (Moreno et al., 2005).

Mortality is a binary variable, and the risk of mortality is necessarily between 0 and 1. A common functional form that produces outcomes in the unit interval is the logistic function. The logistic function is an S-shaped curve given by the following equation:

$$f(x) = \frac{L}{1 + \exp^{-k(x-x_0)}}$$

with

x_0, the midpoint of the curve;

L, the maximum value of the curve;

k, the steepness of the curve.

This function takes as input a linear additive function of observable variables (which we can think of as an underlying propensity towards mortality), and transforms it to probabilities.

These variables are assigned respective scores indicating their contribution to mortality risk. The scores of all 20 variables are then summed to form a total score, referred to as the SAPS 3 admission score, which relates to the vital status at hospital admission given by the equation:

$$logit = -32.6659 + log(SAPS3score + 20.5958) * 7.3068 \tag{2.1}$$

The likelihood of mortality relates to the vital status at hospital discharge given by the equation:

$$Probabilityofdeath = \frac{e^{logit}}{1 + e^{logit}} \tag{2.2}$$

Customised models for large geographic areas were calculated to allow the intercept and the gradient of the curve to differ across regions. The logit equation utilised by Mediclinic South Africa is given by:

$$logit = -27.38054 + log(SAPS3score + 5.5077) * 6.2746 \tag{2.3}$$

Figure 2.1 below illustrates the probability of death for ICU patients based on the standard logit equation and the customised logit equation utilised by Mediclinic South Africa. There is a downward adjustment of the probability of death for Mediclinic South Africa especially for mid-range SAPS 3 scores.

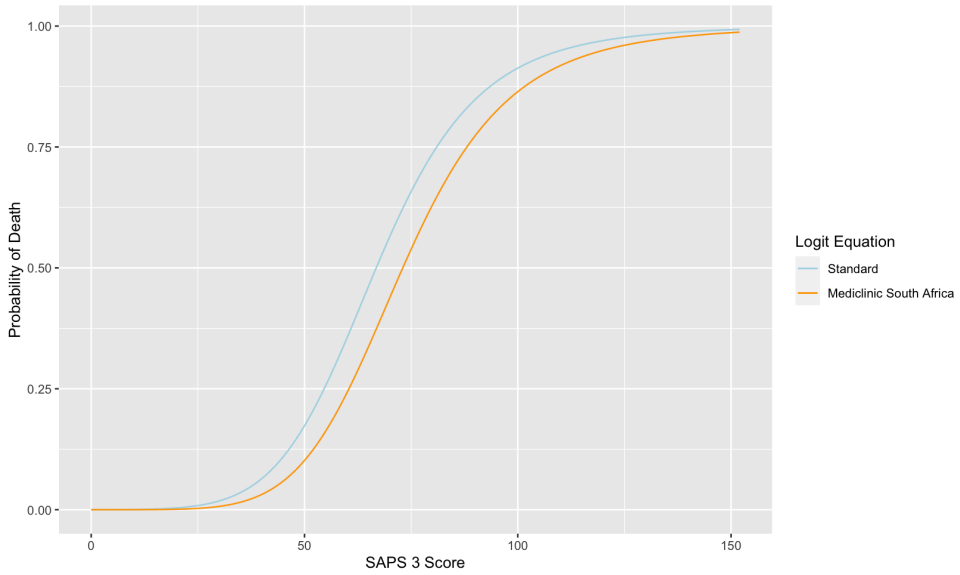

Figure 2.1: Probability of Death based on SAPS 3 Score.

Although SAPS 3 was published as a significant improvement over SAPS II to predict mortality, there are still some potential drawbacks as mentioned in section 2.1. First, SAPS 3 was developed prior to the Covid-19 pandemic. This implies that mortality risk might be differently influenced by patient features due to the virus. In other words, the model might be predicting mortality risk without accounting for how Covid-19 could have changed the way patient features now influence the probability of death. It must be considered that there might exist new interactions among patient features that would influence the effect on mortality risk. Re-

sultingly, a model developed with data prior to the Covid-19 outbreak may not be particularly useful for mortality prediction of Covid-19 infected individuals.

Secondly, SAPS 3 makes use of a very simple linear additive formula, with the variables' coefficients being restricted to integers. The benefit of this simplistic functional form is that the formula could be easily implemented by nurses and other health care workers. However, these coefficients are most likely not chosen because they optimally describe the relationship between patient features and mortality risk, but specifically because they are easily implemented and simplistic to use and calculate. In addition, mortality risk might perhaps be a more complicated function of patient features - it might even be non-linear. To illustrate a non-linear effect, the danger of a high heart rate affecting old patients differently than young patients can be considered. Consequently, a linear function designed for convenience rather than describing the exact relationship, might not be maximally accurate. Therefore, there might exist some other functional form, as opposed to a linear additive form, or some other coefficients, that would better describe the relationship between patient features and mortality risk. The interactions between the variables might be more complex than depicted by a simple linear function. This is a valid concern in the context of Covid-19, since there are different interactions among variables with patients infected with Covid-19, but it might even be a valid concern in the absence of Covid-19. Even prior to Covid-19 there exists a possibility that non-linear effects are present when estimating mortality risk. However, speculating about the form of non-linearities is not only difficult, but almost trivial due to countless possibilities and consequences run high when the incorrect functional form is chosen. This provides an opportunity for alternative modelling techniques. If there are a multitude of variables that might be relevant, but uncertainty exists regarding which variables are important, or if an interesting functional form exists with non-linear interaction, but there exists uncertainty as to the functional form of the non-linearity, machine learning techniques offer the advantage of seeking out those non-linearities and interactions. Therefore, this study utilises machine learning techniques in order to uncover the functional form that describes the effect of patient features on mortality risk.

Therefore, a model that accurately predicts the mortality of Covid-19 patients should use data captured during the pandemic, rather than data prior to the pandemic, to estimate the model in order to identify the key variables that are significant in the presence of Covid-19. Additionally, the model should be able to capture the potential non-linear effects that patient features might have on mortality risk and the new interesting interactions that might exist between the variables.

3. Data

Mediclinic Southern Africa operates a range of multi-disciplinary acute care private hospitals in South Africa and Namibia and utilises the commonly used ICU scoring system, SAPS 3, to predict mortality risk for patients. As part of its normal operations, Mediclinic hospitals collect all the patient information that is required to calculate the SAPS 3 admission score. This data consists of patient characteristics prior to admission (e.g., age, gender, arrival type, doctor speciality, length of stay before ICU admission), co-morbidities, various physiological variables (e.g., reason for ICU admission, acute infection and surgical status at ICU admission, heart rate, oxygenation, systolic blood pressure and temperature) and whether the patients suffer from diseases such as HIV and Tuberculosis.

Mediclinic has agreed to share this data, under the provision of ethical approval obtained (Project ID 22525; Ethics Reference Number S21/07/010_COVID-19). This retrospective study used clinical data from 34,292 patients admitted to ICU in 43 different Mediclinic hospitals across all nine provinces in South Africa, and three provinces in Namibia. The data is collected from the respective hospitals across South Africa and Namibia and stored in a central database. The data includes all ICU patients over the age of 16. Of the 34,292 observations, or patients we have in the data set, only 392 (approximately 1.14%) were omitted which had either no SAPS 3 admission score or a score of zero due to missing values in the individual variables used to calculate the score. Figure 3.1 provides information regarding the population distribution within the data utilised in the study. The population has slightly fewer females than males and a slightly higher median age for female patients. The death rate for females is also higher compared to males. Furthermore, as one would expect, the death rate for Covid-19 positive patients is significantly higher compared to Covid-19 negative patients.

Variables	Population (%)	Count	Age (Median)	Mortality Count	Death (%)
Total Patients	100.00	34,292	62	9,241	26.95
Females	42.42	14,547	63	4,135	28.43
Males	57.58	19,745	62	5,106	25.86
Covid Negative	74.48	25,541	64	4,250	16.64
Covid Positive	25.52	8,751	59	4,991	57.03

Figure 3.1: Descriptive Statistics

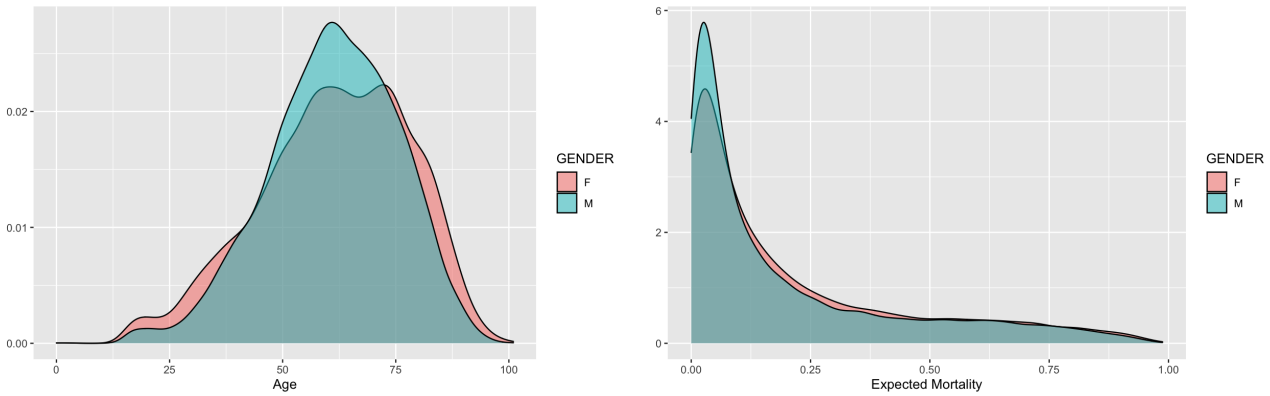

Figure 3.2: Distribution of Age and Expected Mortality

According to the Figure 3.2 above we see that female age has a bimodal distribution and there is a greater proportion of females below the age of 30 and above the age of 75 in ICU compared to males. Furthermore, the distribution for expected mortality of males and females is similarly shaped, although there exists a higher proportion of females with an expected mortality between 12.5% and 50%, and a higher proportion of males with expected mortality close to zero. Figure 3.3 below shows that the average expected mortality is higher for patients infected with Covid-19 compared to the Covid-19 negative patients for all age groups. Covid-19 clearly influences the distribution of expected mortality across age and this new relationship in the presence of the Covid-19 disease should be captured in any model that aims to predict mortality during the Covid-19 pandemic.

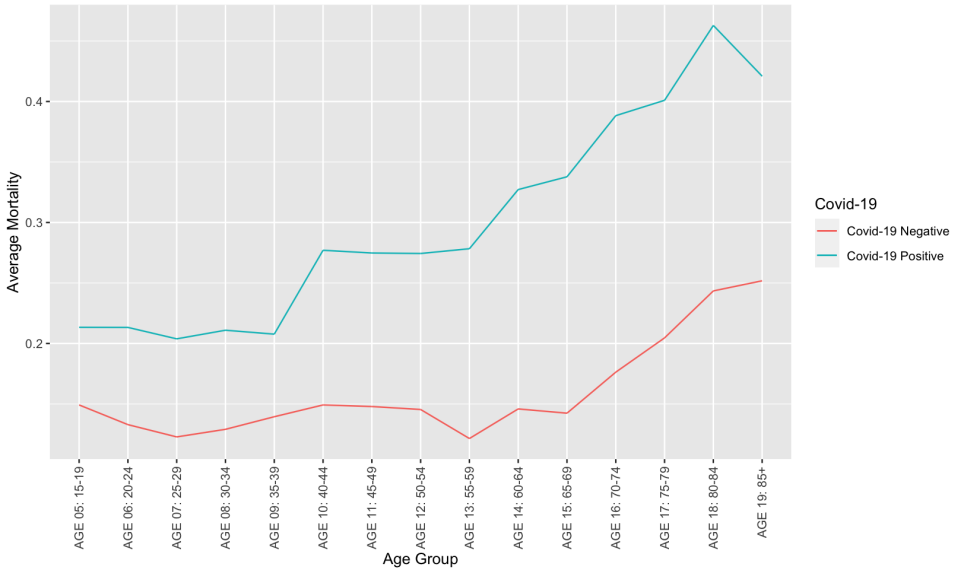

Figure 3.3: Average Mortality Per Age Group

From the data set obtained there are additional variables available not included in the SAPS 3 calculation that is also utilised in the study to see whether other variables available at the time of admission could potentially benefit the mortality prediction model by improving the model accuracy. For instance, Figure 3.4 below shows the relationship between BMI category and the average mortality for the Covid-19 positive sample compared to the Covid-19 negative sample. The graph indicates that the BMI category differently impacts the mortality risk based on Covid-19 infection, suggesting that the inclusion of BMI as an additional variable in the model, among others, could potentially further explain mortality risk during the pandemic.

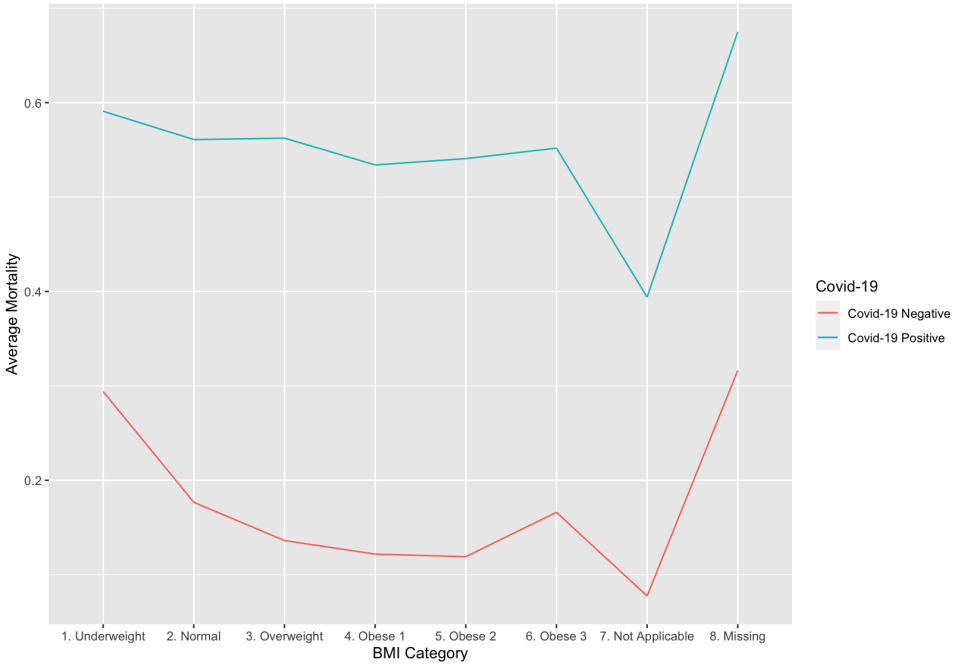

Figure 3.4: Average Mortality based on BMI group

Furthermore, the data obtained from Mediclinic which informs the SAPS 3 score are provided in score format based on the model's severity index for each variable. The drawback to the scored data is that the models will not have the advantage of being trained and modelled on raw data, but only on pre-categorised data. This reduces the models' potential to identify features and patterns for prediction, and relationships between mortality and patient characteristics since the variation within each variable is limited.

4. Methodology

Modelling the effect of patient features on mortality risk requires that we specify some functional form for the underlying data generating process. However, some problems may potentially arise if interaction effects or non-linear effects are not accounted for, or when irrelevant controls are excluded or even relevant controls excluded, for instance. If the underlying data generating process is incorrectly specified, the estimations of mortality risk will most likely be inaccurate and not particularly useful. Therefore, the closer we can get to the true data generating process, the more useful the model will be. There are typically two approaches to statistical modelling. The first assumes the likely functional form for the population model and then estimates the parameters accordingly. These are typically interested in the exact relationship between a variable and the outcome. The second approach assumes that the functional form is unknown and thus seeks to employ flexible techniques to uncover this underlying data generating process. The research question at hand is a predictive problem by nature since it is concerned with identifying patients that are at high risk for mortality based on various patient features. Therefore, the main concern is not to understand and estimate the parameters to explain the effect of a specific patient feature on mortality risk. Rather, the study aims to uncover the unknown data generating process that produces a certain mortality risk from various patient feature inputs. The study will therefore emphasise an approach to statistical modelling of the second kind. Since prediction accuracy is our concern, machine learning will be well-suited for our problem.

Machine learning relies on regularisation and empirical tuning which involves the process of introducing restrictions through hyperparameters and finding the optimal hyperparameter values for the best test set accuracy performance, respectively. Regularisation involves choosing a function class which consists of a set of hyperparameters that determines how complex the model will be. Our empirical analysis will focus on three different function classes to evaluate how accurately the different functional forms and model complexities can predict mortality. The study will consider a logistic regression, classification and regression tree, and a gradient boosting decision tree implementation called Light Gradient Boosting Machine (LightGBM). Once the function class has been chosen, the optimal hyperparameter values that produce the best out-of-sample fit must be determined. This is accomplished by identifying potential values for the hyperparameters, obtaining estimates for the out-of-sample performance of the models with different values of the hyperparameters and then choosing the hyperparameter value that yield the best out of sample fit. A simple grid-search was utilised to evaluate the performance of the different hyperparameter values except for the logistic regression which does not have critical hyperparameters to tune. Once the final model has been estimated with the optimal hyperparameter value, the prediction accuracy must be measured on the test

set. The prediction accuracy must be assessed on a sample that has not been utilised in the training or the tuning of the hyperparameters. Accordingly, three data sets are required that provide data for training, validating (the tuning of hyperparameters) and the testing for the prediction accuracy of the final model. This requires some sample splitting. There are diverse ways to split the sample depending on the sample size and the model objectives. In this study, the observations are ordered according to date of admission and then split 75:25 between a training set (for training and validation) and a testing set (for assessing prediction accuracy of the final model). In other words, the data from the first 75% of admissions are used to predict for the last 25% of admissions. There are a few difficulties faced when choosing the ideal sizes for the sample splitting as a small training set could produce imprecise coefficient estimates which could negatively affect the choice of the hyperparameters, but a small validation set can likewise cause a sample-dependent fit of the out-of-sample estimates even when coefficients are precise, resulting in an inferior choice of the optimal hyperparameter. Moreover, a small test set can result in erroneous prediction accuracy estimates. These difficulties can be solved by using k-fold cross-validation in training and tuning the parameters. This process involves removing the test set from the randomly shuffled dataset and then splitting the remaining set into k smaller sets. For each individual k fold (or subset) the kth-fold is taken as the validation set and the remaining k-1 folds is taken as the training set. The model is then trained on the training set and evaluated on the validation set and the process is repeated for every unique k-fold. The average performance over all the k validation sets is then computed by

$$CV_k = \frac{1}{k} \sum_{i=1}^{k} MSE_i \tag{4.1}$$

This allows the study to obtain an improved selection of hyperparameters which in return produces more accurate predictions.

As briefly mentioned, the study will consider three different functional forms (class functions) to predict the mortality risk for ICU patients infected with Covid-19. The first is a logistic regression that linearly models the probability of a certain outcomes occurring. In our case, the logistic regression will linearly model the probability of a patient in ICU not surviving. The logistic regression relaxes many of the assumptions made for linear regressions as it does not require a linear relationship between the outcome variable and the explanatory variables, it does not require the normal distribution of the error terms, and homoscedasticity is not needed.

However, logistic regression assumes that observations are independent and that there is little or no multicollinearity among explanatory variables. Each variables gets a coefficient and there are no interactions. Important to note here is that variables are allowed to differ in importance, although non-linear effects are not yet allowed for (like allowing the effect of high heart rate to differ for old vs. young patients). The benefit of this functional form therefore it that it allows for the relevance of variables to differ which SAPS 3 does not.

The relationship $p(X) = Pr(Y = 1|X)$ must be modelled without allowing predicted probabilities above one or below zero. Hence, a function that provides values only between zero and one given any input value must be used. Accordingly, the logistic function is used for logistic regression

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

The model is fitted with the maximum likelihood technique and will generate an S-shaped function between zero and one. After some manipulation, the above formula becomes,

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}$$

Taking the logs of both sides produces,

$$\log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

The second functional form this study considers is a classification tree. The study uses the R package rpart (Recursive Partitioning) that utilises many of the concepts found in CART (Classification and Regression Trees). The rpart programme builds general structure classification or regression models using a two-step process: first, the variable is identified which best splits the data into two groups. Once the data is split into two groups, the process is repeated on each subgroup, respectively. This process is repeated until improvement is impossible or until the subgroups reach a minimum size; secondly, since the model is often too intricate and complex, it requires pruning that is accomplished by cross-validation. The nature of the research question in this study requires a classification tree which provides a qualitative outcome rather than a regression tree that provides a quantitative outcome (James et al., 2013). In other words, the classification tree predicts the class to which each observation is most likely to

belong to, so whether an individual is more likely to belong to the survived or the not survived category. The tree is grown through recursive binary splitting according to the Gini index (Therneau & Atkinson, 1997). All the values would go through different splitting points and be tested according to a cost function - which in this case is the Gini index. Decision trees are generally referred to as greedy algorithms since they opt to make the optimal decision at each step of the procedure. However, a potential pitfall of a greedy algorithm is that it tends to overfit the data - it becomes exceptionally good at explaining the sample data, but not so great to explain the out-of-sample. To avoid this sort of complexity in the tree we restrict the growth of the tree by restricting the complexity parameter value. One of the advantages of a classification tree is that it is easy to interpret, and it can be displayed graphically. The tree can model basic non-linear effects and it allows for interaction between variables, both of which SAPS 3 does not.

The third method the study considers, light gradient boosting machine (LightGBM), is a gradient boosting framework that is based on decision tree algorithms. The boosting framework simply uses weak learners, such as decision trees, as foundational units to generate more powerful prediction models. The approach combines multiple decision trees in an ensemble to improve the accuracy of prediction. In contrast to bagging where separately grown trees are combined in an ensemble, boosting is a sequential tree-growing approach combining many different decision trees that all grow using information from previously grown trees. Since individual trees often suffer from high bias or high variance, combining these weak learners reaches a better variance-bias trade-off which improves the performance of the model. The objective of boosting is to allow a procedure that learns slowly by avoiding the fitting of one large decision tree over the data that might be subject to overfitting. Essentially, the first trees (or models) fit the data very simplistically and thereafter the residuals are analysed to update the following model. The sequential decision trees are fitted to the residuals of the model and those trees are then included to update the residuals. The aim is therefore to solve for the remaining error from the previous tree. The adding of these small trees to the residuals slowly improves the function specifically in the areas where it tends to perform poorly by increasing the weight of an input that was misclassified in the previous decision. Fundamental to gradient boosting is the use of weak learners, sequential tree-growing, and the minimisation of some loss function. Thus, as the model grows sequentially on weak learners, the model error reduces.

Gradient boosting decision tree is a very popular and prevalent algorithm in machine learning specifically because it is known to be efficient, accurate and interpretable. Although there are various implementations of the algorithm, such as XGBoost, scikit-learn and pGBRT, that are widely used, recent developments in data availability have brought to light new hindrances in

obtaining accuracy with efficiency. Scalability is more challenging as computational intricacies become more complex with an increase in the number of features and observations available with big data. Particularly because the gradient boosting decision tree algorithm (GBDT) necessarily assesses the estimated information gain of all potential split points by scanning all the data instances for all available features. Accordingly, the size of the data and the number of features is directly proportionate to the complexity of the computations resulting in time intensive implementations of the algorithm. A sensible notion would be to reduce data size and number of features to solve the inefficiencies that arise because of computational complexities. Unfortunately, what seems straightforward in theory is not always equally simple in practice. Turns out that data sampling for gradient boosting decision tree is quite tricky since the commonly used weighted sampling technique cannot be imitated with this algorithm because there exists no sample weight (Ke et al., 2017). Accordingly, Ke et al. (2017) proposed two novel algorithms that constituted the new implementation of GBDT called Light Gradient Boosting Machine (LightGBM).

The first technique, Gradient-based One-Side Sampling (GOSS), aims to reduce the number of data instances. Lui et al. recognised that gradients of data instances play a significant role in computing information gain since the greater the gradient, the more it contributes to information gain. Accordingly, the number of data instances can be reduced whilst retaining accuracy by keeping those instances with larger gradients and randomly dropping those with smaller gradients. This method was proven to produce increased estimation gain accuracy compared to random sampling that is uniform in nature (Ke et al., 2017). The second technique, Exclusive Feature Bundling (EFB), reduces the number of features with practically no loss. The technique involves the design of an algorithm that simplifies the optimal bundling problem, i.e., how to optimally bundle exclusive features. Real-world applications typically have sparse feature spaces, regardless of the number of features, which implies that most features never take nonzero values concurrently. In other words, most features are nearly exclusive which allows them to be bundled safely. Accordingly, an algorithm was designed to simplify the optimal bundling of exclusive features by means of a graph colouring approach which makes vertices from features and edges from two non-mutually exclusive features. The graph colouring problem is then solved with a greedy algorithm and accordingly the number of effective features is reduced with nearly no loss. Therefore, the complex bundling problem is proven to be solved by a greedy algorithm that produces a good approximation ratio which allows the effective reduction of the number of features while retaining the accuracy of determining split points. The combination of these two novel algorithms produces the implementation of the GBDT algorithm called LightGBM. Various experiments on public data have indicated that LightGBM accelerates the training process of common GBDT implementations up to 20 times while maintaining similar

accuracy.

Each model was evaluated using various metric scores that provide insights for how well the model performed. These include sensitivity and specificity, accuracy and area under the curve (AUC). Sensitivity refers to the ability of the model to correctly identify patients that will not survive. In other words, if the model predicts that an individual does not survive, the individual also does not survive in reality. Whereas specificity refers to the ability of the model to correctly identify patients that will survive. Accuracy is a metric score that indicates the proportion of correct predictions. This metric uses a threshold to determine the predicted mortality score. If an individual received a mortality risk above the 50% threshold, they are given a binary mortality score of one and those below the 50% threshold are given a mortality score of zero. In other words, if a patient has a 60% probability of dying, they are above the threshold and therefore predicted to have a mortality score of 1 implying that they are predicted to not survive. Accuracy is then determined by showing the proportion of individuals with the correct binary prediction – that is to say, if they received a mortality score of one, they actually did not survive. There is a downside to this metric since no differentiation is made between individuals with an expected mortality of 51% and 90% since both are equivalently assigned with a mortality score of one, indicating that they are predicted to not survive. However, it is obvious that these patients are clearly different. Therefore, an additional metric, area under the curve (AUC), is included to provide a measure of performance of the ranking of the model, or stated differently, the relative performance of the model. Essentially, it shows how well the model can distinguish between classes. Thus, an AUC score of 80% would imply that the model can with 80% accuracy take two individuals who respectively survived and did not survive and rank them correctly. Furthermore, the 95% confidence interval for the accuracy metric is also provided.

5. Empirical Analysis

The primary objective of this study is to determine how accurately mortality risk can be predicted for Covid-19 positive patients admitted to ICU. A secondary objective is to understand how much of this accuracy gain, if any, is due to estimating the model on the most appropriate sample of patients, allowing for this probability to be a non-linear function of the same variables utilised in SAPS 3, and from including additional variables available at the time of admission in the model. Figure 5.1 below contains a list of candidate variables that are collected by Mediclinic but are not used in SAPS 3. These questions will be answered by estimating three models on the Covid-19 positive sample: a logistic model (which assumes a linear additive functional

form, which is then transformed using a link function), a recursive partitioning (rpart) model (which allows for non-linear and interactive effects, although in a highly restrictive way) and a LightGBM model (which is highly flexible). These three models will be first be estimated using only the variables utilised in SAPS 3, and then re-estimated using a longer list of candidate explanatory variables.

Patient Admission Categories	Hypertension
Gender	Asthma
Doctor Speciality	Chronic obstructive pulmonary disease
Arrival Type (ambulance, ER, etc.)	Postpartum
Covid Case	Pregnancy
Body Mass Index	Mechanical ventilator
HIV	Room oxygen aid
Tuberculosis	ICU Days
Immunosuppressive disease	Blood FFP
Diabetes	Dialysis

Figure 5.1: Additional Variables

If we find that a more flexible functional form or additional variables could improve the accuracy predictions for mortality risk for Covid-19 positive sample, this raises the question of whether the same applies to the Covid-19 negative sample. If so, then the benefits of an alternative modelling approach to improve SAPS 3 could be generalised to non-pandemic periods. To investigate these questions, the empirical analysis described above will be repeated on the Covid-19 negative sample.

Tables 1 and 2 from Figure 5.4 provide the model evaluations for the Covid-19 positive sample. Table 1 reports the various goodness-of-fit metrics for the four alternative models (SAPS 3, logistic, rpart and LightGBM) using only the variables included in SAPS 3, whereas Table 2 shows the same metrics across the same models now estimated using the additional variables as well. From Table 1 it is evident that the performance of the SAPS 3 model is improved by allowing alternative functional forms. The logistic model, which has a similar functional form as the SAPS 3 model, achieves more than a 11-percentage point improvement in accurately predicting mortality. The confidence bands reveal that this difference is highly statistically significant. This may be attributable either to the fact that the logistic model does not restrict coefficients to non-negative integers or because it allows the relative importance of the variables to be determined by the sample of Covid-19 positive South African patients. Figure 8.1 in the Appendix confirms the latter. It provides a comparison between the logistic and SAPS 3 coefficients which indicates that the accuracy improvement of the logistic model is the result of more appropriate data being utilised since the relative magnitude of the coefficients differ significantly. The rpart model also outperforms the SAPS 3 model, although its performance is notably worse than that of the logistic model. The LigthGBM model likewise outperforms

the SAPS 3 model, but its out-of-sample performance is the same as that of the logistic model. Notably, the accuracy confidence intervals for all three alternative models do not even include the point estimate of the SAPS 3 model, indicating that the accuracy improvement is significant.

Table 2 also shows how much additional predictive accuracy can be gained over SAPS 3 by including additional variables in the models. The results of the logistic regression model estimated on the SAPS 3 and the additional variables are displayed below in Figure 5.2 with coefficients limited to those that are statistically significant. General admittance, arrival type of patients, mechanical ventilation, TB and dialysis all increase the risk of mortality. Allowing these additional non-SAPS variables to affect mortality raises the forecast accuracy by nearly 20 percentage points during the Covid-19 pandemic and more than 2 percentage points during normal circumstances. Presumably, the logistic model's performance could be further improved by obtaining and utilising the values of the variables used to determine the individual SAPS 3 points rather than the point scores. The highly flexible functional form of the LightGBM further improves this performance in the Covid negative sample, although only by .23 percentage points.

Characteristic	log(OR)[1]	95% CI[1]	p-value
AGE_ADM	0.04	0.03, 0.05	**<0.001**
Acc_catGeneral.Admit	0.86	0.34, 1.4	**0.001**
Arrival_TypeSplit.Account	0.70	0.33, 1.1	**<0.001**
TB	0.70	0.13, 1.3	**0.016**
DIALYSIS	0.44	0.21, 0.66	**<0.001**
MECHANICALVENTY	2.3	2.1, 2.4	**<0.001**
EXPRSSN0.11	0.71	0.23, 1.2	**0.004**
Points_Creatinine.7	0.43	0.15, 0.72	**0.003**
Points_GCS.2	0.44	0.21, 0.67	**<0.001**
Points_GCS.15	0.36	0.16, 0.55	**<0.001**
Points_HeartRate.5	0.32	0.17, 0.48	**<0.001**
Points_Oxygenation.7	0.50	0.30, 0.69	**<0.001**
Points_Oxygenation.11	0.47	0.31, 0.63	**<0.001**
Points_Respitory.5	0.26	0.07, 0.45	**0.006**
Points_SystolicBP.11	1.2	0.67, 1.7	**<0.001**

[1] OR = Odds Ratio, CI = Confidence Interval

Figure 5.2: Logistic Regression

As was the case when not including additional variables, the rpart model with additional variables is outperformed by both the logistic and LightGBM models. Nevertheless, this model provides interesting insights into the variables that are most important in predicting mortality for Covid-19 positive patients. The results of the rpart estimated on SAPS 3 and additional

variables are displayed in Figure 5.3 below. Oxygenation and age, which are both included in SAPS 3, are the most important variables in determining mortality risk. Mechanical ventilation, dialysis and BMI category "missing" also play a key role. Concerns might arise for the practicality of including a variable such as BMI in the model due to the potential difficulty of measuring body weight and height of patients during ICU admission. However, additional variables are informative and important, even when submitted with missing values, since they provide valuable information on patient status and characteristics to inform mortality prediction.

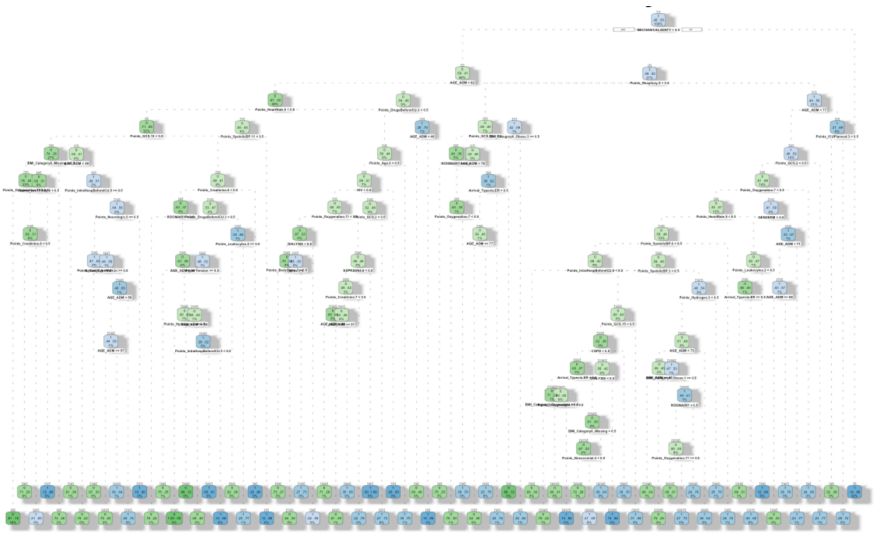

Figure 5.3: Rpart Tree

We can conclude from Tables 1 and 2 that, towards the end of the pandemic, it would have been possible to predict mortality risk of Covid-19 positive patients much more accurately with a model trained on Mediclinic's own data compared to using SAPS 3. Using the non-linear LightGBM model and including additional variables not used in SAPS 3 resulted in a substantial accuracy improvement of more than 19 percentage points relative to SAPS 3. The most important contribution of this new model would have been that it used the most relevant data to identify the relative importance of the SAPS 3 variables: this would have increased predictive accuracy by 10.88 percentage points. Using a highly flexible estimator which allows for non-linear and interactive effects of patient attributes on mortality risk, and additional symptoms would have increased predictive accuracy by an additional 8.69 percentage points.

RESULTS: Covid-19 Positive Sample

Table 1

Estimated on SAPS 3 Variables Only:

Metrics	SAPS 3	Logistic	RPART	LightGBM
Sensitivity	88.12	53.61	39.73	46.09
Specificity	31.13	76.26	80.24	81.81
AUC	59.62	64.94	59.99	63.95
Accuracy	56.58	67.46	64.42	67.85
CI: (95%)	(55.32, 57.82)	(65.62, 69.26)	(62.54, 66.25)	(66.02, 69.64)

Table 2

Estimated on Additional Variables:

Metrics	SAPS	Logistic	RPART	LightGBM
Sensitivity	88.12	76.07	69.67	77.01
Specificity	31.13	76.22	76.47	75.60
AUC	59.62	76.15	73.08	76.31
Accuracy	56.58	76.16	73.81	76.15
CI: (95%)	(55.32, 57.82)	(74.47, 77.79)	(72.09, 75.49)	(74.47, 77.77)

Figure 5.4: Covid Positive Results

Table 3 and 4 from Figure 5.5 below show the results of the Covid-19 negative samples. The AUC metric for Table 3 shows that SAPS 3 performed the best compared to the alternative models. The AUCs of the logistic regression, rpart and LightGBM actually deteriorated slightly, whereas the accuracy metric improved slightly. However, the results that include the additional variables as displayed in Table 4 shows the AUCs improving moderately, as well as the accuracy. The accuracy confidence intervals of all three alternative functional forms do not even include the point estimate of the SAPS 3 accuracy, once again implying that the accuracy improvement from allowing additional variables and flexible functional forms is significant. These tables imply that an alternative functional form is not necessarily better for the Covid-19 negative sample and that perhaps the sample size is too small for machine learning to be useful, but that the inclusion of additional variables could improve prediction accuracy.

RESULTS: Covid-19 Negative Sample

Table 3

Estimated on SAPS 3 Variables Only:

Metrics	SAPS	Logistic	RPART	LightGBM
Sensitivity	95.63	96.81	96.95	97.09
Specificity	43.31	41.80	33.30	36.77
AUC	69.47	69.31	65.13	66.93
Accuracy	86.93	87.76	86.49	87.18
CI: (95%)	(86.51, 87.34)	(86.89, 88.58)	(85.59, 87.35)	(86.30, 88.02)

Table 4

Estimated on Additional Variables:

Metrics	SAPS	Logistic	RPART	LightGBM
Sensitivity	95.63	97.03	97.49	97.39
Specificity	43.31	48.85	41.57	48.83
AUC	69.47	72.94	69.54	73.11
Accuracy	86.93	89.18	88.30	89.41
CI: (95%)	(86.51, 87.34)	(88.36, 89.97)	(87.46, 89.11)	(88.60, 90.18)

Figure 5.5: Covid Negative Results

In summary, the results show that machine learning methods improve prediction accuracy for mortality risk, particularly when there is uncertainty regarding the factors influencing mortality, such as with the Covid-19 infection, but also during normal circumstances since the Covid-19 negative sample also indicated an improvement in prediction accuracy. Accuracy gains could further be improved by including other available patient features, especially for Covid-19 positive patients. For the Covid-19 positive sample the machine learning improved the accuracy by nearly 12 percentage points, whereas it improved accuracy for the Covid-19 negative sample by 0.25 percentage points. The effect of using additional patient feature variables resulted in substantial accuracy gains, particularly in the Covid-19 positive sample of nearly 20 percentage points. Large hospitals, such as Mediclinic, that relied on SAPS 3 or similar models which have functional forms designed for ease of application could potentially have been better prepared for the Covid-19 pandemic if advances in statistical techniques and technology were incorporated to assist in triage decisions through mortality risk predictions. Firstly, to improve mortality risk predictions, custom parameters could have been estimated with the in-house data available at large hospitals. Additional variables available at the time of admission could also have been incorporated into these models to improve prediction accuracy. Moreover, this thesis provides evidence that flexible estimation methods such as machine learning are incredibly powerful

tools for prediction accuracy during times of uncertainty, but also during normal circumstances and that this accuracy can be achieved while maintaining efficiency.

6. Conclusion

This thesis illustrates that Mediclinic ICU data provides useful information to outperform SAPS 3 mortality predictions and that machine learning methods are valuable tools to predict patient outcomes. This is particularly true when sudden shifts in the relationship occur between patient features and patient outcomes, for example in the presence of Covid-19 infection. Using the most relevant data to predict mortality risk increased prediction accuracy with 10.88 percentage points, allowing a flexible functional form gained nearly 12 percentage points, and including other patient attributes and clinical indicators that are available but not used in the SAPS 3 score resulted in accuracy gains of more than 19 percentage points. This implies that large hospitals could potentially have been better prepared for the pandemic to make more effective triage choices by utilising tools that are readily available. For instance, SAPS 3 parameters could have been customised or recalibrated with in-house hospital data, additional variables already captured within these hospitals could have been utilised, and more advanced flexible statistical techniques which are better suited for times of uncertainty could have been implemented to improve the prediction accuracy of mortality. Although the contributions made by this thesis are important during the pandemic, they are also valuable, true and relevant during normal circumstances as this thesis illustrates that machine learning techniques are equally powerful tools to identify interesting interactions and explain model errors to improve prediction accuracy in the absence of Covid-19 infection. A useful endeavour for hospitals could be to develop a dynamic scoring model which periodically uses the most recent and relevant data obtained from their own patients to calibrate this algorithm, rather than using a static scoring model. Ease of application could be facilitated by modern technology which allows nurses to capture symptoms on a tablet which then calculates mortality risk. This would eradicate the need for an oversimplified algorithm and result in more effective triage choices and policy decisions.

7. References

Al-qaness, MAA., Ewees, AA., Fan, H., Abualigah, L. & Abd Elaziz, M. 2020. Marine predators algorithm for forecasting confirmed cases of COVID-19 in italy, USA, iran and korea. *International Journal of Environmental Research and Public Health*. 17(10):3520.

Al-qaness, MAA., Ewees, AA., Fan, H. & Abd ElAziz, M. 2020. Optimization method for forecasting confirmed cases of COVID-19 in china. *Journal of Clinical Medicine*. 9(3):674.

Alsayed, A., Sadir, H., Kamil, R. & Sari, H. 2020. Prediction of epidemic peak and infected cases for COVID-19 disease in malaysia. *International journal of environmental research and public health*. 17(11):4074.

Apolone, G., Bertolini, G., D'Amico, G. and C., R.and Iapichino & Melotti, R.M. 1996. The performance of SAPS II in a cohort of patients admitted to 99 italian ICUs: Results from GiViTI. *Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva. Intensive care medicine*. 22(12).

ArunKumar, K.E., Kalaga, D.V., Kumar, C., Kawaji, M. & Brenza, T.M. 2021. Forecasting of COVID-19 using deep layer recurrent neural networks (RNNs) with gated recurrent units (GRUs) and long short-term memory (LSTM) cells. *Chaos, solitons, and fractals*. 146(1).

Banoei, M.M., Dinparastisaleh, R., Zadeh, A.V. & Mirsaeidi, M. 2021. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. *Critical care (London, England)*. 25(1):328.

Basu, S. 2021. Approaches to critical care resource allocation and triage during the COVID-19 pandemic: An examination from a developing world perspective. *Journal of medical ethics and history of medicine*. 14(5).

DB., W., MH., K., JM., L. & B., L. 2009. Who should receive life support during a public health emergency? Using ethical principles to improve allocation decisions. *Ann Intern Med*. 150(2).

Emanuel, E.J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickamn, A., Zhang, C., Boyle, C., et al. 2020. Fair allocation of scarce medical resources in the time of covid-19. *The New England journal of medicine*. 382(21):2049–2055.

Fiest, K.M., Krewulak, K.D., Plotnikoff, K.M. & al., et. 2020. Allocation of intensive care resources during an infectious disease outbreak: A rapid review to inform practice. *BMC Medicine*. 18(1).

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. *Introduction to statistical learning with applications in r*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *NIPS*.

Metnitz, P.G., Moreno, R.P., Almeida, E., Jordan, B., Bauer, P., Campos, R.A., Iapichino, G., Edbrooke, D., et al. 2005. SAPS 3: From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. *Intensive care medicine*. 31(10):1336–1344.

Moreno, R.P., Metnitz, P.G., Almeida, E., Jordan, B., Bauer, P., Campos, R.A., Iapichino, G., Edbrooke, D., et al. 2005. SAPS 3: From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive care medicine*. 31(10).

O'Laughlin, D.T. & Hick, J.L. 2008. Ethical issues in resource triage. *Respiratory care*. 53(2):190–200.

Ottenhoff, M.C., Ramos, L.A., Potters, W., Janssen, M., Hubers, D., Hu, S., Fridgeirsson, E.A., Piña-Fuentes, D., et al. 2021. Predicting mortality of individual patients with COVID-19: A multicentre dutch cohort. *BMJ open*. 11(7).

Poole, D., Rossi, C., Anghileri, A., Giardino, N., M.and Latronico, Radrizzani, D., Langer, M. & Bertolini, G. 2009. External validation of the simplified acute physiology score (SAPS) 3 in a cohort of 28,357 patients from 147 italian intensive care units. *Intensive care medicine*. 35(11).

Pourhomayoun, M. & Shakibi, M. 2021. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health Volume*. 20(1).

Reid, L. 2020. Triage of critical care resources in COVID-19: A stronger role for justice. *Journal of medical ethics*. 46(8):526–530.

Saleh, A., Ahmed, M., Sultan, I. & Abdel-lateif, A. 2015. Comparison of the mortality prediction of different ICU scoring systems (APACHE II and III, SAPS II, and SOFA) in a

single-center ICU subpopulation with acute respiratory distress syndrome. *Egyptian Journal of Chest Diseases and Tuberculosis*. 64(4).

Therneau, T. & Atkinson, E. 1997. An introduction to recursive partitioning using the RPART routines. *Mayo Clinic*. 61.

Truog, R.D., Mitchell, C. & Daley, G.Q. 2020. The toughest triage - allocating ventilators in a pandemic. *The New England journal of medicine*. 382(21):1973–1975.

Tversky, A. & Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science*. 185(4157).

White, D.B. & Lo, B. 2020. A framework for rationing ventilators and critical care beds during the COVID-19 pandemic. *JAMA*. 323(18):1773–1774.

Zhai, Q., Lin, Z., Ge, H., Liang, Y., Li, N., Ma, Q. & Ye, C. 2020. Using machine learning tools to predict outcomes for emergency department intensive care unit patients. *Scientific reports*. 10(1).

8. Appendix

Characteristic	Logit	SAPS 3
Age, years		
40-59	0,47	0,54
60-69	0,69	0,90
70-74	1,00	1,31
75-79	1,20	1,47
≥80	1,60	1,79
Total bilirubin, mg/dL (µmol/L)		
2-5.9 mg/dL	-0,02	0,44
≥6 mg/dL	0,12	0,46
Body temperature, °C		
<35 °C	0,16	0,68
Comorbidities		
Cancer therapy	0,69	0,32
Metastatic cancer	0,66	1,07
Chronic Heart Failure	0,54	0,62
Haematological cancer	0,24	0,59
Cirrhosis	0,62	0,76
AIDS	0,89	0,77
Reasons for ICU Admission		
Cardiovascular: rhythm disturbances	0,01	-0,49
Cardiovascular: hypovolemic hemorrhagic shock, hypovolemic non-hemorrhagic shock	0,57	0,28
Cardiovascular: septic shock	0,44	0,48
Neurologic: seizures	0,32	-0,43
Neurologic: coma, stupor, obtunded patient, vigilance disturbances, confusion, agitation,	0,57	0,40
Neurologic: focal neurologic deficit	0,47	0,71
Neurologic: intracranial mass effect	1,3	1,00
Digestive: acute abdomen, other	0,15	0,35
Digestive: severe pancreatitis	1	0,92
Hepatic: liver failure	0,75	0,59
Creatinine, mg/dL (µmol/L)		
1.2-1.9 mg/dL	0,04	0,20
2-3.4 mg/dL	0,64	0,67
≥3.5 mg/dL	0,66	0,80
Use of major therapeutic options before ICU admission		
Vasoactive Drugs	0,36	0,30
Glascow Coma Scale/Score		
7-12	1	0,22
6	0,86	0,71
5	1,4	1,03
3-4	1,3	1,48
Heart rate, beats/min		
120-159	0,46	0,45
>160	0,64	0,66
Planned or unplanned ICU admission		
Planned	0,17	0,35
Length of stay before ICU admission, days		
14-27	0,75	0,55
>28	1,1	0,69
Intrahospital location before ICU admission		
Emergency room	0,6	0,49
Other ICU	0,65	0,69
Other ward	0,64	0,76
Leukocytes, G/L		
>15	0,22	0,15
Oxygenation		
PaO_2<60 and no MV	0,73	0,54
PaO_2/FiO_2≥100 and MV	0,51	0,74
PaO_2/FiO_2<100 and MV	0,84	1,07
Platelets, G/L		
50-99	0,17	0,49
20-49	0,68	0,76
<20	0,59	1,33
Acute infection at ICU admission		
Nosocomial	-0,03	0,4137
Respiratory	0,68	0,5403
Surgical status at ICU admission		
No surgery	0,42	0,49
Emergency surgery	0,24	0,63
Systolic blood pressure, mm Hg		
70-119	0,38	0,32
40-69	1	0,83
<40	1,4	1,11

Figure 8.1: Coefficient Comparison between SAPS 3 and Logistic Regression