An AI algorithm to predict intubations in ICU patients

Predicting intubation to aid medical personnel in the decision to switch from High Flow Nasal Oxygen therapy to intubation

MANON HENDRIKS

APRIL 2023

AN AI ALGORITHM TO PREDICT INTUBATIONS IN ICU PATIENTS

- Predicting intubation to aid medical personnel in the decision to switch from High Flow Nasal Oxygen therapy to intubation -

Hendriks, Manon

Student number : 4530616

April 2023

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Biomechanical Engineering, TUDELFT
May 2022 – April 2023
Supervisor(s):
Dr. ing. , Niehof , Sjoerd
Dr. MD. , den Uil, Corstiaan

Thesis committee members:

Prof. , Reinders, Marcel / TU Delft
Dr. Ing. , Niehof, Sjoerd / Maasstad Ziekenhuis, TU Delft
Dr. MD. , den Uil, Corstiaan / Maasstad Ziekenhuis, Erasmus MC

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Preface

With great pleasure, I present this thesis report for the Master Technical Medicine, called *An AI algorithm to predict intubations in ICU patients*. In May 2022 this thesis project kicked off with a literature review after which the practical research project started.
During this thesis my research qualities were refined. A part of the used data was requested in an official application for the 'Beheercommissie' of Santeon. In this application, I experienced how a committee judges a research proposal and which requirements are necessary. Moreover, my experience in data processing has grown enormously during this thesis project.

In this thesis report a two-sided method that uses two different data sets to develop a Machine Learning (ML) model is described. The ML models are developed to predict intubation in patients that received High Flow Nasal Oxygen (HFNO) therapy.

First, I would like to express my gratitude to Corstiaan den Uil and Sjoerd Niehof, who were my medical and technical supervisors and guided me during my graduation. Corstiaan and Sjoerd supervised me during the whole project and were always available to discuss an approach or help in finding a new path or solution. Still, they gave me the opportunity to explore methods individually. I really enjoyed the collaboration.

During this thesis project I developed experience and skills in different clinical measurements in the ICU. For this, I want to thank Dolf Weller, who is a ventilation practitioner in the ICU of the Maasstad Hospital. Dolf gave me the space and tools to improve my clinical skills. Here I also really enjoyed the discussions and collaboration to figure out how to optimally ventilate ICU patients.

For the data processing part of this thesis Geeke Waverijn and Martijn Kuijper were approached to supervise me in the process of obtaining data and using it to develop ML models. I want to thank Geeke for helping me in the extraction and processing of the data and for her help in working towards the development of an ML model. Martijn has introduced me to joint models and has helped me with processing data that contains repeated measurements. I have learned a lot from Martijn and he was always willing to help if I had run into another R error, for which I am very grateful.
Moreover, for the ML model development Mattia Fornasa, a data scientist at Pacmed, supervised me during the whole thesis. Mattia helped with the structure of my research question, possible methods, and explanation of the results I obtained. I want to thank Mattia for sharing his knowledge of ML models and guidance throughout my thesis project.

In the data application for the 'Beheercommissie' of Santeon, Serena Bruens helped me with the structure of my research proposal and guided me through the paperwork necessary for the application. I want to thank Serena for her help in this process and for the nice meetings.
Without the approval of the 'Beheercommmissie', I would not have had an external validation set to validate the best-performing ML model. For this, I want to thank the 'Beheercommissie'.

Furthermore, I want to thank Marcel Reinders for being part of my exam committee.

Lastly, I want to thank my family and friends for their support and encouragement to finish this thesis project. Especially Daniek and Sanne for all the nice coffee breaks, but also for their insights on this thesis project. Last but definitely not least, I want to thank my boyfriend Luuk for always being there for me and giving me support and motivation.

Manon Hendriks
Delft, April 2023

Summary

Maasstad Hospital is a member of the Santeon hospital group. The ambition of Santeon is to improve healthcare for patients. The project in this internship also aims to improve patients' health, specifically patients in the Intensive Care Unit (ICU).

The treatment of respiratory insufficient patients in the ICU consists of High Flow Nasal Oxygen or Cannula (HFNO or HFNC), among others. There is a substantial uncertainty about the optimal duration of this HFNO therapy and the chance of failure of this therapy. Failure of HFNO therapy will often lead to the progression to mechanical ventilation with intubation. This thesis project researched parameters and predictive models to choose the appropriate treatment, meaning continuing HFNO therapy or escalation to mechanical ventilation by intubating the patient.

In this thesis project, the goal was to develop a Machine Learning (ML) model that can predict intubation at a certain point in time and thereby show that HFNO therapy will not be sufficient. With this eventual model, it would be possible to determine if intubation is necessary on the first day of ICU admission. The proposed model could lead to more elective or early intubation.

The intended ML model was achieved with a two-sided method. Firstly, an aggregated data set was used to compute three different models. These were two tree-based models, a Random Forest (RF) and a Gradient Boosting Model (GBM), and a Logistic Regression Model (LRM). The other parallel method made use of a data set with repeated measurements of vital parameters, such as heart rate. This method resulted in a so-called joint model, which is a combination of Linear Mixed Effects Models and in the second step also an RF, GBM, and LRM.

A nested cross-validation was implemented to test the above-described models, three feature selection methods, and three scaling methods. From the nested cross-validation, the best-performing model was found and tested in the evaluation.

For the evaluation of the models an extra data set was used. This external data set was retrieved via a data request to the Santeon 'Beheercommissie' or data management committee. This data set did contain repeated measurements, but not enough to validate the joint model. Therefore, only the models developed with the aggregated data set could be externally validated.

The two-sided method resulted in an RF with no feature selection and no scaling having the best performance using the aggregated data set, namely an AUC of 0.694 (standard deviation 0.05). The joint model resulted in an RF with no feature selection and Power Transformer scaling with the best performance. It had a performance value of 0.681 (standard deviation 0.07). The external validation of the aggregated data model resulted in an AUC of 0.559. The internal validation of the joint model gave an AUC of 0.699. The precision, recall and f1-score showed that all the models performed better for class 0: the non-intubated patients.

The best-performing aggregated data model shows potential and proves that it is possible to predict intubation using AI, but in its current state is far from implementation. It is therefore advised to train the model with a larger training data set that contains multiple hospitals and perform an external validation with a validation data set that meets the requirements.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
AUC	Area Under Curve
AUROC	Area Under Receiver Operating Curve
BI	Business Intelligence
FiO2	Fraction of inspired Oxygen
GBM	Gradient Boosting Model
HFNO/HFNC	High Flow Nasal Oxygen/Cannula
ICU	Intensive Care Unit
KNN	K-Nearest Neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LMEM	Linear Mixed Effect Model
LRM	Logistic Regression Model
ML	Machine Learning
NN	Neural Network
PCA	Principal Component Analysis
SpO2	Peripheral Oxygen Saturation
RF	Random Forest
RMSE	Root Mean Squared Error
ROX index	Respiratory rate-OXygenation
RR	Respiratory Rate
VAP	Ventilator-associated Pneumonia
VILI	Ventilator-Inflicted Lung Injury

1

Introduction

In this chapter, the introduction of this thesis will be stated. First, the daily practice will be explained, whereafter the aim of this thesis project will be explained. Moreover, the outline of this thesis report will be stated. In Figure 1.1 a graphical overview of what is stated in the introduction of this thesis is shown.

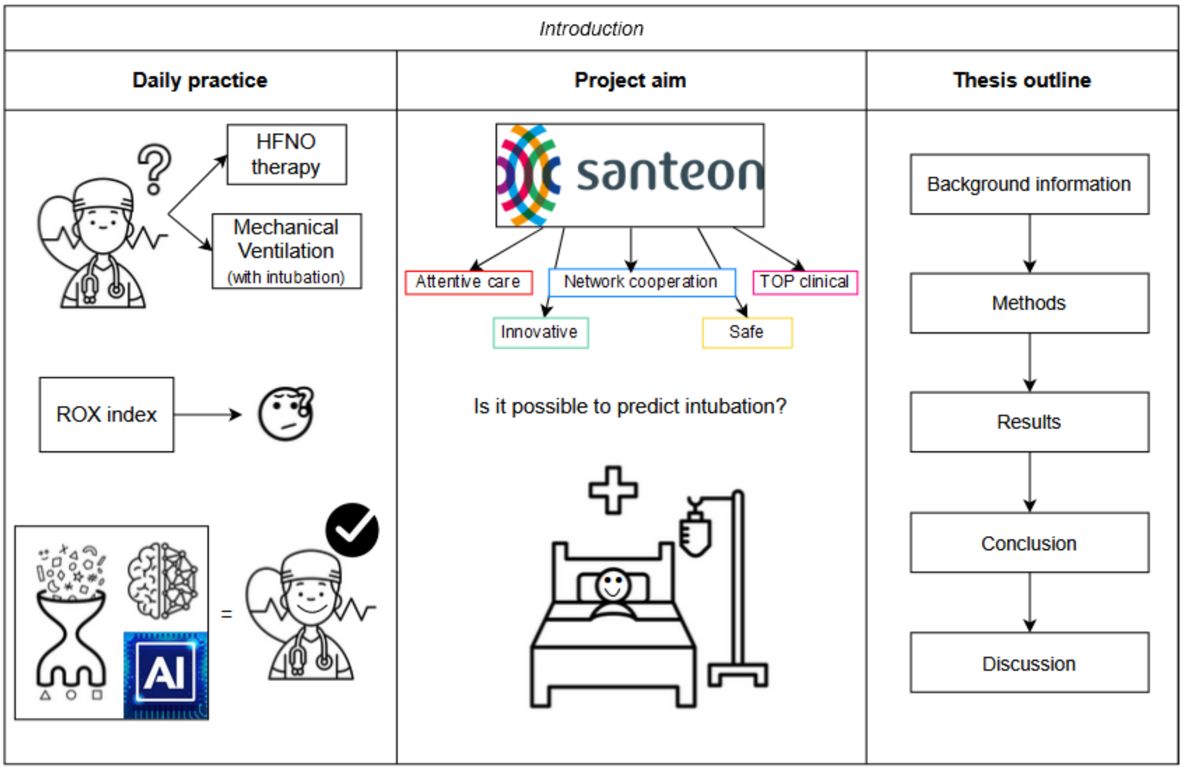


Figure 1.1: Graphical overview of the introduction of this thesis

1.1. Daily practice

The ICU of Maasstad Hospital has 3 units with each 8 beds available for ICU care for patients. During the COVID pandemic, the Maasstad Hospital played a great role in the care of COVID-19 patients. In COVID-19 patients the ventilatory tract was often diseased and therefore needed treatment. The use of High Flow Nasal Oxygen (HFNO) therapy was rising during the pandemic as it could lead to the postponement of invasive ventilation, which was scarce during the pandemic. Even more important for the patient, HFNO therapy is less invasive than mechanical ventilation and is therefore preferred when possible. During the COVID pandemic and also in the current daily practice, it is difficult to determine the duration and efficacy of HFNO therapy.

Specifically, in patients that are showing respiratory deterioration whilst on HFNO therapy it is difficult to determine when the therapy needs to be stopped and mechanical ventilation should be started. HFNO therapy is more commonly known under its brand name Optiflow®. When giving HFNO therapy, two parameters can be altered by the physician, namely the flow in L/min and the fraction of oxygen given through the device (FiO2). The oxygenated air is heated and humidified when given to the patient via a nasal cannula.[1]

The ROX index can be used to predict when an escalation from HFNO therapy to intubation should be considered. The ROX index uses 3 variables to calculate the prediction, namely, peripheral saturation (SpO2), the fraction of admitted oxygen (FiO2), and respiratory rate (RR). The following formula is used to calculate the prediction score.

$$ROXscore = \frac{(SpO_2/FiO_2)}{RR} * 100 \qquad [2]$$

The ROX score can be interpreted as follows. A score of ≥ 4.88 measured at 2, 6 or 12 hours after HFNO therapy is associated with a lower risk for intubation. A ROX index score of < 3.85 indicates that the risk of HFNO therapy failure is high. Therefore, intubation should be discussed. Lastly, if the ROX index score is between 3.85 and 4.88, the scoring could be repeated one or two hours later for further evaluation.[3]

The Area Under the Curve (AUC) value of the ROX index lies around 0.64. [4] As this performance is not sufficient enough to make it trustworthy for physicians, this ROX tool is not applied in the Maasstad Hospital. The performed literature study before this thesis project showed that the AUC value of Machine Learning (ML) models that predict intubation lies around 0.81, which is significantly higher than the AUC value of the ROX index.[5] This higher performance level is probably caused by the larger variation of variables (also called features) that are used in ML models. The literature review in which the AUC value of 0.81 was found is presented in Appendix B.

In an effort to improve the prediction of necessary intubation in patients treated with HFNO therapy, an ML model is developed in this thesis which can be used as a more reliable indicator than the ROX index.

1.2. Project aim

This thesis project will aim to deliver an ML model. The following question will be answered in this project:

Is it possible to predict intubation at a certain point in time and thereby show that HFNO therapy will not be sufficient?

Specifically, it is proposed to develop an ML model that can predict intubation after eight hours in which a patient has received HFNO therapy. This model will be developed with retrospective data that has been collected in Maasstad Hospital before, during, and after COVID-19 and data from the Santeon hospitals during COVID-19. This timeline of data collection is preferred because, in this specific timeline, the use of HFNO therapy was more present in the Maasstad Hospital and other Santeon hospitals. HFNO therapy is a relatively new technique and was not used frequently before COVID-19. This project will use a two-sided method in the development of ML models. Two different data sets will be used: an aggregated data set and a data set that contains repeated measurements.

The eventual models will be compared using the obtained AUC values for the training and test set. The different models with different hyper-parameters can be compared based on their AUC values. The already performed literature review developed a general idea of the performance value for ML models that predict intubation indication.

The literature review gives a clear overview of recent predictive models. Therefore, it is attainable to require a predictive model at the end of the thesis project. A general idea of what model type is preferred and which features are predictive has been formed. The data availability could however create a problem for the attainability. During the project, it took longer than expected to receive permission to use the Santeon data. As a backup plan, data from the Maasstad Hospital was retrieved to start the modelling. This is also beneficial because this will result in a larger database.

The relevance of this study lies in the need to improve patient care. It is of course preferred to give the right treatment at the right time to the patient. ML models can aid medical personnel in the difficult treatment choice between longer HFNO treatment or invasive ventilation which requires intubation.

1.3. Thesis outline

The next chapter of this thesis will be Chapter 2, in which background information is given on the different ways to oxygenate ICU patients and machine learning models. In the next Chapter 3, the methods of this thesis project will be explained. The data extraction method, model development and evaluation process will be stated. Hereafter, Chapter 4 will describe the results of this graduation project. The used data will be described and the different models will be compared on their AUC values. In Chapter 5, the conclusion of this report will be stated. The meaning of the results will be explained in a brief manner. Lastly, in Chapter 6 the relevance of the results, limitations of this research and future recommendations will be stated.

2 Background information

In this chapter, background information will be given on the therapy options for giving oxygen to ICU patients. Additionally, different Machine Learning (ML) models and their characteristics will be explained.

2.1. Oxygenating ICU patients

The different treatment choices that can be made for oxygenating ICU patients are explained in this section. In Figure 2.1 an infographic is shown which shows the different oxygenation treatments and the possible switching between them. The oxygenation treatments are first divided into non-invasive and invasive ventilation. Hereafter, non-invasive ventilation is divided into ”Nasal cannula, oxygen cap” and ”High Flow Nasal Oxygen”. The Invasive ventilation branch is divided into ”Mechanical ventilation with intubation via endotracheal tube” and ”Mechanical ventilation with intubation via tracheostomy”. The treatment choice or path of interest in this specific thesis project is highlighted with a green arrow, namely switching from HFNO therapy to mechanical ventilation with intubation via endotracheal tube. The other smaller black arrows indicate the possible switching between the treatments.

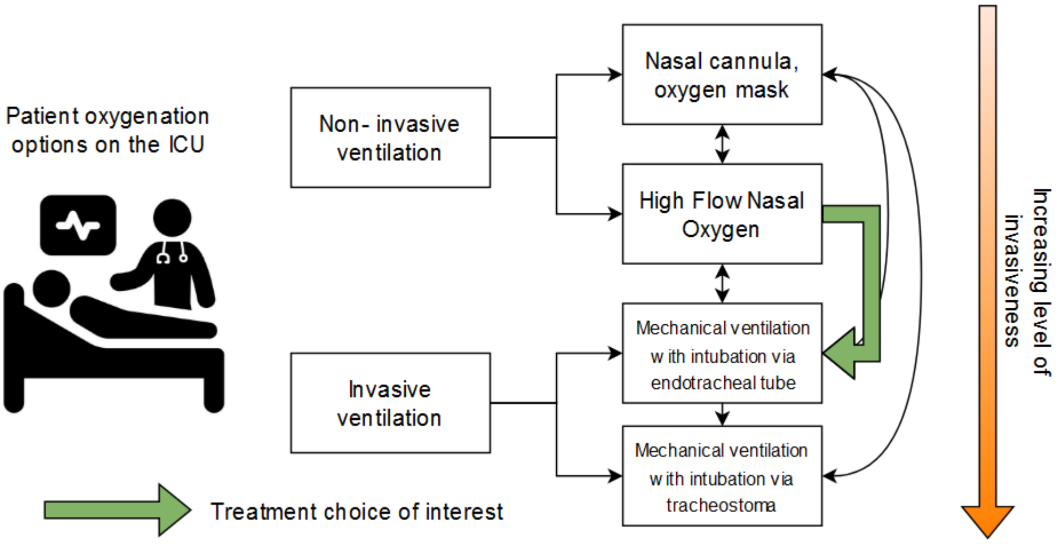

Patient oxygenation options on the ICU
Non- invasive ventilation
Invasive ventilation
Nasal cannula, oxygen mask
High Flow Nasal Oxygen
Mechanical ventilation with intubation via endotracheal tube
Mechanical ventilation with intubation via tracheostoma
Treatment choice of interest
Increasing level of invasiveness

Figure 2.1: Infographic illustrating the different oxygenation treatment options

For patients it is beneficial to prevent the escalation from non-invasive ventilation to invasive ventilation if possible. This is because HFNO therapy does not require intubation, which has a higher risk of complications and simply is more harmful. Complications of intubation include laryngeal injury, infection (ventilator associated pneumonia (VAP)), tearing or puncturing of tissue in the chest cavity that can lead to lung collapse, injury to throat or trachea, damage to dental work or injury to teeth, fluid buildup, aspiration, and ventilator-induced lung injury (VILI).[6, 7, 8] Furthermore, HFNO therapy gives more patient comfort, giving the patient the ability to talk and even eat while receiving HFNO therapy. In contrast, intubated patients are often sedated and will not have the ability to talk because of the tube passing through the vocal cords.

However, in certain situations non-invasive ventilation will not be enough treatment for the patient. In this scenario, the patient will benefit from an escalation to invasive ventilation. Reasons why the HFNO therapy fails could be the severity of the illness of the patient, leading to the need for more sedation and with that the need to take over the ventilation. As stated before, it is difficult to determine this course of the illness in patients. The question is if they will deteriorate and benefit from intubation or if they will improve and have sufficient treatment with HFNO therapy.

The ML model developed in this thesis will give a better understanding of which patients benefit from the escalation to invasive ventilation. In the future, the model can aid in choosing the oxygenation option earlier and therefore create a window in which the intubation can be planned. This is another advantage of knowing to which group the patient will belong. This created window can make sure an emergency intubation can be prevented.

2.2. Different Machine Learning models

For comprehensibility, ML models are explained in two different categories. The categories are based on the data type that is used in the model. The first possibility is to give a tabular-like data set to a model, in which each row contains a different patient. Every column is a feature and one of the columns has the desired outcome or dependent variable (e.g. yes or no for intubation). This type of data set is also called an aggregated data set. Examples of models that work with this kind of data sets are Logistic Regression models and tree-based models, such as a Random Forest and Gradient Boosting model.

The other possibility is to give a more raw data set to the model. For example, the variable heart rate often contains more values than a certain laboratory blood value. In these other types of models, it is possible to use a time series of a certain variable in the model. Models that work with this kind of data sets are deep learning models, such as Neural Networks, or supervised machine learning models, such as a joint model.

In the following subsections, these different models are further explained. The book of Hastie T et al. called *The Elements of Statistical Learning* [9] was used to write the background information of the different machine learning models. When other sources were used, they were cited in the corresponding sections. Different models that make use of the two categories of data sets will be implemented in this thesis project.

Logistic Regression Model

Other than in a linear regression model, a Logistic Regression Model (LRM) gives a true or false outcome. An S-shaped curve classifies which samples are true and which are false. The cut-off is usually made at 50%. In a model that predicts intubation, the outcome *true* would mean the patient is predicted to be intubated and the outcome *false* would mean the patient is predicted to not be intubated. Both continuous and discrete data can be entered to classify samples. How the S-shaped curve is fit through the samples is determined with maximum likelihood of the curve, since the Least Square Method will not work in a logistic regression. Maximum likelihood is a function that calculates the probability of observing the outcome given the input data and the model. This function is optimized to find the set of parameters that result in the largest sum likelihood over the training dataset. The LRM assumes that the relationship between the predictor variable and the predicted outcome is linear. In Figure 2.2 an infographic of the LRM is shown.

Random Forest

A Random Forest (RF) is an ML model that consists of a ”forest” of decision trees. Decision trees alone are often trained too well on the existing data, which makes them inflexible to use on other data sets. Because an RF includes multiple decision trees it is more flexible and accurate on a different data set. When modelling an RF, the first step that is performed is to make a bootstrapped data set. This means that from the existing data set patients are selected randomly. This process will lead to patients randomly not being selected. These out-of-bag samples will become the validation set. With the bootstrapped data set random decision trees are formed. Together, the decision trees become the RF. The accuracy of the RF is then tested using the validation set. A new RF is then built with different variables/features per step. The RF model with the best accuracy and a determined number of variables per step is chosen. In Figure 2.2 an infographic of the RF is shown.

Gradient Boosting Model

A Gradient Boosting Model (GBM) is an ML model that also consists of multiple decision trees. It builds fixed-sized trees based on the previous tree's errors. It starts with a leaf in which the log(odds) is imputed. In this case, the log(odds) would be log(intubated/non-intubated), in which the intubated/non-intubated distribution is extracted from the data. To estimate how well or bad the first prediction of the model is, residuals (the difference between observed and predicted values) are calculated. The calculated residuals are then used to build a new tree. This new tree then predicts the new residuals. With each tree, the residuals become smaller, making the predictions more accurate. New trees are made until the maximum specified number of trees is reached or when adding a tree does not significantly reduce the size of residuals.[10] In Figure 2.2 an infographic of the GBM is shown.

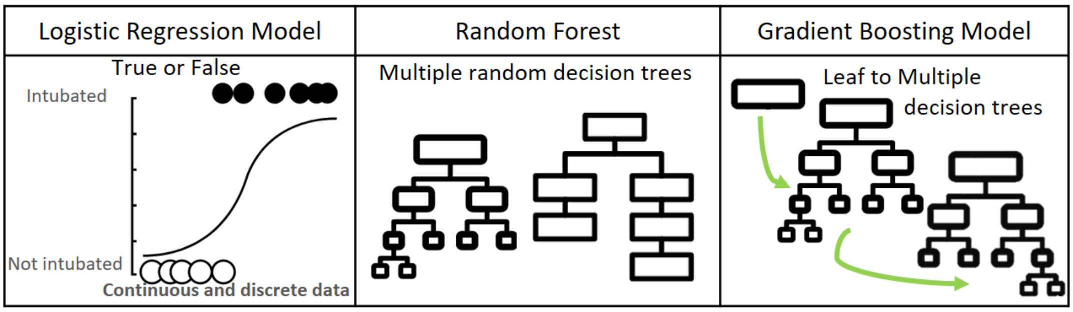


Figure 2.2: Infographic of the different explained ML models that work with an aggregated data set. From left to right: the Logistic Regression Model, Random Forest and Gradient Boosting Model.

Neural Network

A Neural Network (NN) allows multiple inputs and outputs. Between the input and outputs are hidden layers that contain nodes that connect all the layers with the input and next layer or output. The hidden layers contain activation functions that alter the input to form a graph. This graph is eventually used to classify or make a prediction. It is difficult to understand what happens in the hidden layers. In Figure 2.3 an infographic of the NN is shown.

Joint Model

As stated, a Joint Model is also capable of interpreting repeated measurements. A Joint Model is a combination of two models in which the first model calculates the time components of the variable with repeated measures. This first model is often a Linear Mixed Effects Model (LMEM) with a cubic spline through a certain amount of knots based on the data points of the variable. From this first model, different features are collected. These features contain information on the estimated spline graph drawn through the data points. For instance, the intercept of the cubic spline could be a variable. Moreover, the slopes between the intercept and every node are often used as features. These features are then used in an ML model that predicts the event of intubation in this case. The features extracted from the LMEMs are the new data set for the second step of the joint model. This newly formed data set is an aggregated data set. Therefore, it is possible to again apply an LRM, RF, or GBM. In Figure 2.3 an infographic of the Joint Model can be found. In this infographic, an LRM is shown as the second step model but this could be any type of ML model that can handle aggregated data.

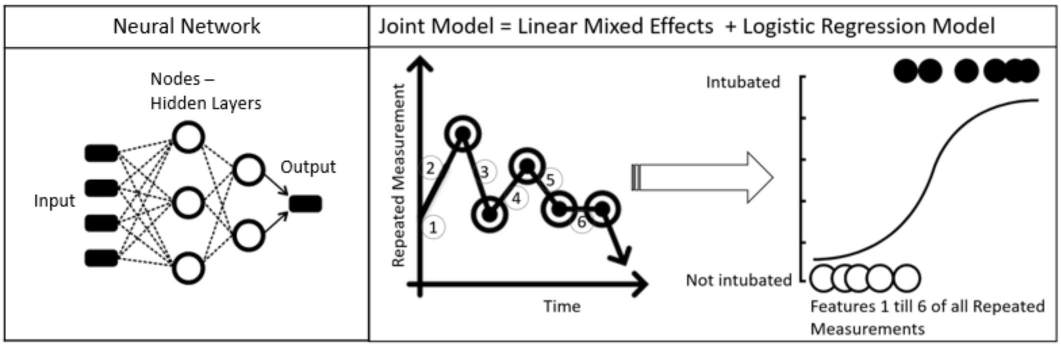

Neural Network
Joint Model = Linear Mixed Effects + Logistic Regression Model
Nodes –
Hidden Layers
Input
Output
Repeated Measurement
Time
1
2
3
4
5
6
Intubated
Not intubated
Features 1 till 6 of all Repeated Measurements

Figure 2.3: Infographic of the different explained ML models that work with a raw data set containing repeated measurements. On the left showing a Neural Network and on the right a Joint Model

3 Methods

In this chapter, the methods of the project will be stated. The method used for data extraction will first be explained. Hereafter, the data processing will be explained for the two different data types, namely the aggregated data set and the repeated measurements data set. The model development phase will also be separately clarified between the two different data sets and models. Lastly, the model validation and evaluation will be defined. In Figure 3.1 a graphical overview of the methods used in this thesis is shown.

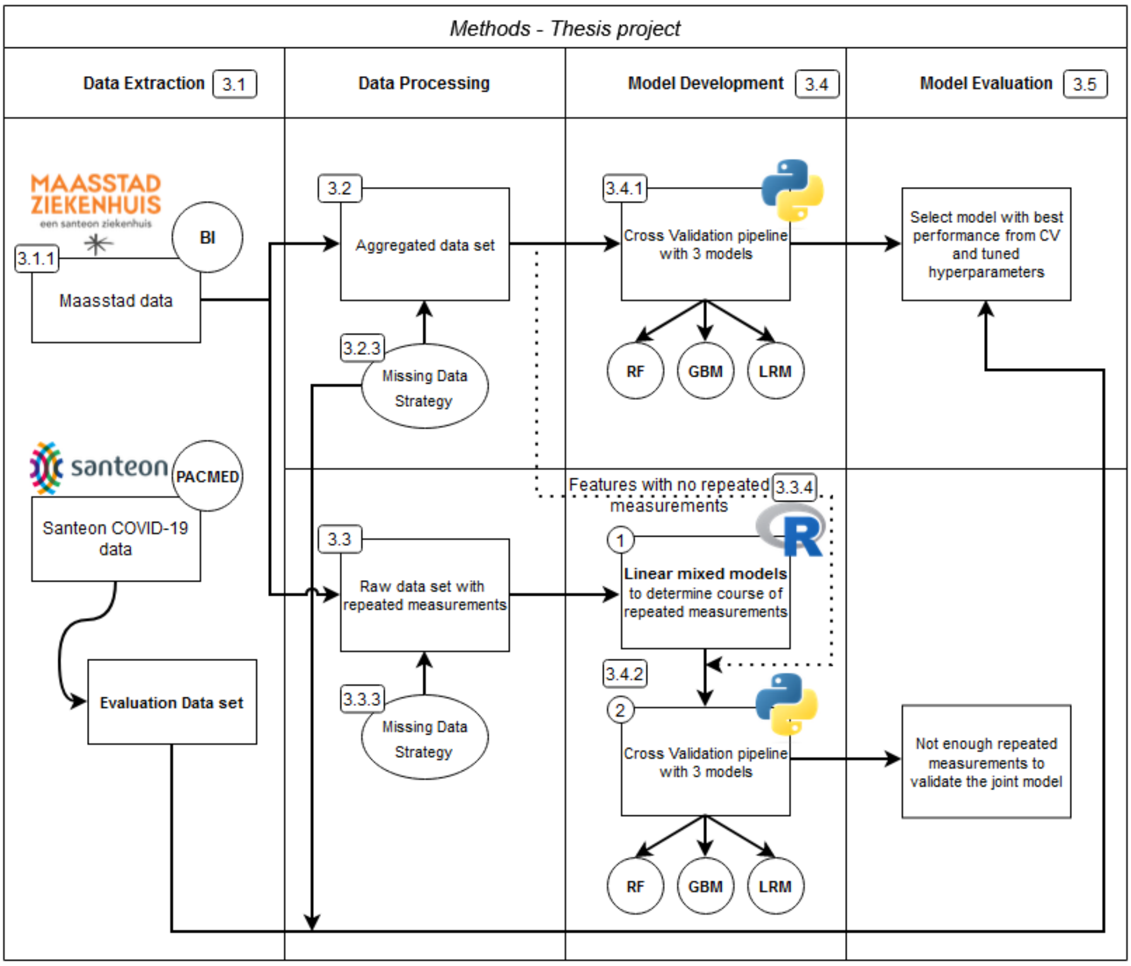

Methods - Thesis project
Data Extraction 3.1
Data Processing
Model Development 3.4
Model Evaluation 3.5
MAASSTAD ZIEKENHUIS
een santeon ziekenhuis
BI
3.1.1
Maasstad data
santeon
PACMED
Santeon COVID-19 data
Evaluation Data set
3.2
Aggregated data set
3.2.3
Missing Data Strategy
3.3
Raw data set with repeated measurements
3.3.3
Missing Data Strategy
3.4.1
Cross Validation pipeline with 3 models
RF
GBM
LRM
Features with no repeated measurements
3.3.4
1
Linear mixed models to determine course of repeated measurements
3.4.2
2
Cross Validation pipeline with 3 models
RF
GBM
LRM
Select model with best performance from CV and tuned hyperparameters
Not enough repeated measurements to validate the joint model

Figure 3.1: Graphical overview of the methods of this thesis, with the section numbers in rectangles corresponding to the sections in which the steps are explained. Abbreviations can be found in the Nomenclature.

3.1. Data extraction

The data extraction process has partly been executed by the Business Intelligence (BI) department of the Maasstad Hospital. Moreover, a data set was used that was distributed by Santeon, a cooperation of 7 hospitals, of which Maasstad Hospital is one. In Appendix A.1 Figure A.1 shows where the data was retrieved and which organizations were involved in this process. Because of the lengthy process of getting permission to use the Santeon data set, the model development was done using the data set provided by the BI department and therefore only contained Maasstad data of the period before, during and after COVID-19. The Santeon data set is used as an external validation data set. In Figure A.1 it can be seen that not all Santeon centres were used in the external validation set. This was due to the timing of receiving the data and the data availability. The Santeon data set consisted purely of COVID-19 data. Because of the different hospitals in the Santeon group, it is beneficial to use this data set as an external validation set. This is due to the fact that it contains patients from a variety of hospitals, making it the ideal data set to test generalizability.

3.1.1. Data extraction by BI

The BI department of the Maasstad Hospital extracted data from electronic patient files. The queries that were used for the data extraction process will not be shared in this thesis project, as they contain sensitive patient data and they are not important for understanding the method of extraction. It will, however, be explained how the data was extracted.

The query used to extract data from the Maasstad Hospital was written to extract data from Metavision, which is the source of the electronic patient files of ICU patients of the Maasstad Hospital. Metavision produces a data set with different tables in Excel format or comma-separated value (CSV), depending on the scale of the data set. The various tables describe the general patient data (age, gender, weight, etc.), measurement of vital functions (heart rate, FiO2 levels, etc.), and lab measurements (bloodwork, electrolytes, etc.). In each of the tables, different columns contain the variable name, value, and timestamp. On every row, a new measurement is given. One of the columns contains the newly generated patient ID.

Within the data set different parameter IDs are given to all the variables. For instance, the variable *Heart Rate* could be represented by parameter ID 43. The parameter IDs that corresponded to the variables implemented in the data set were retrieved from the parameter ID table.

Combining the corresponding parameter IDs for the desired features and the desired patient population, the data were extracted. In Appendix A.2 a figure can be found that provides an overview with examples of the different tables that were used in the data extraction.

The patients included in the data set were required to have at least 8 hours of HFNO therapy. Only these 8 hours of HFNO therapy were considered in the data development. The data therefore only contained measurements of all the features that were measured in these 8 hours of HFNO therapy. In the data extraction process, the following ventilation methods were used to determine when a possible intubation started. When a patient received the following ventilation method it was concluded that the patient was not (yet) intubated: Continuous Positive Airway Pressure (CPAP), Biphasic Positive Airway Pressure (BiPAP), Non-invasive (Niv), manual/spontaneous, and Nasal CPAP. In the case of the remainder following ventilation methods it was concluded that the patient was intubated: Neurally Adjusted Ventilatory Assist (NAVA), Pressure Controlled (PC), Pressure Regulated Volume Control (PRVC), Pressure Support (PS), Volume Controlled (VC), and Volume Support (VS). For the patients that had a ventilation method that indicated intubation, the starting time of the data entry 'ventilation method' was interpreted as the starting time of intubation.

3.2. Data processing for the aggregated data set

In the following subsections, the processing for the aggregated data set is explained. The aggregation of the data, decoding of categorical variables, imputation of missing data, and scaling of the data resulted in the aggregated data set ready to be used in the model development phase. The eventual data set is a table with on every row a new patient and in every column a different variable. Furthermore, one column is the target feature, in this case intubated 'yes' or 'no'. An example of this data set is inserted in Appendix A.3.

The aggregated data set was loaded into Spyder©. This was done on a virtual machine to use Python code to develop models.

3.2.1. Aggregation of the data

In the data set that was aggregated, duplicate patient IDs were removed. These double IDs referred to patients that had multiple admissions during the period from which the data was retrieved. They had to be removed because the joint model could not work with multiple admissions from one patient. To make the resulting models (from the aggregated data set and the repeated measurements data set) more comparable, the double IDs were dropped in both of the data sets.

The aggregated data was aggregated in consultation with the Medical Supervisor (Intensivist, Cardiologist in Maasstad Hospital) of this thesis. His clinical reasoning was used to choose how some of the data needed to be aggregated. For example, for the variable blood pressure and other vital function parameters, the mean, standard deviation and last entered values were kept in the aggregation process. These three different aggregation types were taken into account to retain as much information as possible.

Another example of aggregation is shown in the Fraction of inspired Oxygen (FiO2) variable. The FiO2 value is commonly started at 100% and dialled down when the patient needs less oxygen and is often thus improving. It was therefore chosen to also implement a delta value for the FiO2 value, as this would show how much the oxygen supply has been dialled down. In Appendix A.4 the course of FiO2 values of two different patients are shown with both different deltas.

For the lab measurements there was often only one measurement in the 8 hours of HFNO therapy. In these variables, only the last entered value was taken into account.

The last step of the aggregation was to give a label to the data. Patients were given the label 'intubated' if they had a number higher than zero in the column 'intubation duration'. Furthermore, patients that died during their admission were given the intubation label. This was done because for these patients HFNO therapy was not successful, which means they could have benefited from intubation.

3.2.2. Categorical data decoding

Once the aggregated data set was formed, the categorical data could be decoded to make it workable for ML models. A global search was done to find a decoding strategy. Potdar K et al. 2017 published good results using OneHotEncoder to decode nominal variables.[11] It was therefore chosen to use OneHotEncoder to decode categorical nominal variables, such as gender and origin (the place where the patient was coming from before their ICU admission). OneHotEncoder converts each category value into a new column and assigns a 1 or 0 value to the columns.[11] Gender was decoded with 0 for males and 1 for females. The origin feature contained information from where the patient was admitted to the ICU. It was divided into 4 variables called origin_1 till origin_4. origin_1 means a patient was coming from their home. origin_2 means a patient was admitted from their hospital. origin_3 means a patient was admitted from a different hospital. Lastly, origin_4 means a patient was admitted from the Emergency Room (ER). Next to this, an extra variable was made in which it was determined if a patient was admitted during the day or night. The cut-off was made between 18:00 and 07:00 for night admissions and the other hours would be a day admission. This variable was encoded with 0 for day admissions and 1 for night admission. After the encoding step, the data set only contained numbers in the variables.

3.2.3. Missing data strategy

In the literature review, missing data strategies were researched. Propagate forward was mostly applied when numerical data was missing. For the categorical missing data, an extra variable indicating the missing factor was often implemented.[5] The categorical features did not have missing data in the used data set.

In this specific project it is proposed to implement cross-validation to determine which model with which hyper-parameters show the best performance. When using cross-validation, the missing data can be imputed within the cross-validation or outside of the cross-validation loop. When the missing data is imputed outside the cross-validation loop, it will be imputed before the loop. Literature was searched to find an answer to which option of the two should be chosen. In the article of Jaeger BC et al, 2020, it has been suggested that unsupervised variable selection steps (i.e., steps that ignore the outcome variable) can be applied before conducting cross-validation without incurring bias. As imputing missing data before the cross-validation leads to a reduction of the computational burden, this method is preferred.[12] In a meeting with a statistician from Maasstad Hospital, the option to implement missing data using multiple imputation was opted. This technique imputes a range of multiple values for every

missing data point. It therefore also results in a range of different outcome values. The different outcome values can be compared via their corresponding AUC value.
To get a feeling for the performance of the different ML models, the numerical missing data was first implemented with a mean of the existing data. With this, a data set was formed that can be used in ML models. Moreover, a threshold for a maximum percentage of missing data was set at 60%, meaning that variables that contained more than 60% missing data were dropped out of the data set. After this practice period, missing data was imputed with multiple imputation in the final training data set. The multiple imputation was done before the cross-validation loop. The variable drop threshold remained 60% after the practice period.

3.2.4. Scaling of the data

After the imputation of the missing data, the option to scale the data was researched. This was needed since most machine learning algorithms show greater proficiency using scaled data. To find the best scaling method, the data was first tested on whether it is normally distributed using the Shapiro-Wilk test. The Shapiro-Wilk test is a statistical test used to test the null hypothesis that the data is normally distributed. If the p-value of the test is lower than 0.05, the null hypothesis is rejected and the data is not normally distributed.[13] The data used in the aggregated model was not normally distributed, as all p-values were lower than 0.05. The robust scaling method could therefore be applied. Robust scaling scales the data using centring (subtracting median) and division by the interquartile range (IQR). As the name suggests, robust scaling is robust to the presence of outliers in the data.[14, 15]
To check if Log scaling could be implemented, the data was searched for zeros and negative values, since these can lead to undefined values and errors. Both were present in the data set, which means that Log scaling is not an option. Power Transformation was therefore applied. Power Transforms (PT) are a technique for transforming variables into a uniform distribution, or in other words, stabilising the variance of the distribution.[16] Lastly, before applying these scaling methods, it was checked if the desired ML model types were compatible with this type of scaling.[15, 14]
The article of Ahsan et al. 2021 showed that these scaling methods can be applied in a Logistic Regression Model, Random Forest and Gradient Boosting Model.[15] Decision tree-based models are not sensitive to the scaling of the features. However, scaling can also still be beneficial here because of the reduction of the impact of differences in feature scales. Moreover, it was important to make sure that the data could be descaled after the nested cross-validation pipeline. This step is important to improve the clinical applicability of the model, as medical personnel will understand the model better if the real data is shown.
Both of the found scaling options, Robust Scaling and Power Transformer, and also 'no scaling' were tested in the cross-validation.

3.2.5. Resampling of the data

Resampling of the data was not necessary as the data was quite balanced with 168 of the 348 patients being intubated. This means that in the used data set 48% of the patients were intubated. To give the ML model a balanced data set, this percentage needs to be close to 50%, which applies here. If the data would be imbalanced, many ML models would tend to favour the majority class and ignore the minority class.[17]

3.2.6. Feature selection strategy

Feature selection is a method of reducing the number of input variables in a model by selecting only relevant features and excluding useless features. For the feature selection strategy, the following 4 options were implemented in the ML models: LASSO, PCA, SelectKBest, or no feature selection.
LASSO stands for Least Absolute Shrinkage and Selection Operator and is also called L1 regularization. It gives each variable a certain weight by using a regression analysis. Variables that seem useless will get a high enough weight to shrink the variable to zero and thus cancel it out.[18]
PCA stands for Principal Component Analysis and is a data reduction technique. It uses linear algebra to transform the dataset into a compressed form. With PCA, new features are formed that are a combination of the old features, which thereby reduces the number of features.[19]
SelectKbest is a method that selects features according to the k highest scores. For example, if k=10 the 10 best features will be selected based on their scores. These scores are often determined with a statistical function, such as an ANOVA F-value or chi-squared test.[14, 20] In this project the ANOVA

F-value test was used, since this is the default test and it performed well. ANOVA F-value stands for Analysis of Variance F-value and calculates the variation between sample means divided by the variation within the samples.[21] With the test, features can be found that are independent of the target variable and can thus be removed from the dataset. For the value of k different options were tested, namely k=10 and k=50. Eventually, k=50 gave the best results in the practice period. Therefore, the value of k was set at 50, which means that the 50 best features are selected based on their ANOVA F-value.

Lastly, implementing no feature selection was tested. This was done because all the applied models already have an intrinsic way of selecting features. The tree-based models intrinsically select which features are most predictive for the best outcome. Not all features are used in the trees and therefore a feature selection has taken place. In LRM, an intrinsic hyperparameter that can be defined contains a feature selection method, namely the penalty hyperparameter. This parameter can be set to implement a LASSO selection or Ridge (L2 regularization). In contrary to LASSO, Ridge does not cancel useless features out but gives them a high weight to shrink them close to zero.[18]

3.3. Data processing for the repeated measurements data set

Next to the aggregated data set, a repeated measurements data set was made. In this data set all the repeated measurements that were measured in the 8 hours of HFNO therapy were preserved. This data set has a new measurement on every row, with the patient ID in one of the columns. An example of this repeated measurements data set is inserted in Appendix A.5.

This data set was loaded into RStudio to use R code to develop the first step of the joint model. In the following subsections, the processing of the repeated measurements data set to make it applicable for the first step of the joint model is explained.

3.3.1. Data density decisions

In the first step of the joint model, the longitudinal course of the variables that contained repeated measurements was found using Linear Mixed Effects Models (LMEM). To ensure that the LMEM would make the best estimate of the longitudinal course, as many measurements as possible were taken into account. For many of the variables with repeated measurements, this meant having a value at every minute for every variable (that had repeated measurements) for 8 hours. In the Maasstad Hospital, medical personnel validates the measurements of the patients every hour. The measurements in this data set are therefore not validated. This could mean that the data contains outliers. However, because of the density of the data points, the longitudinal trend of the variables can still be obtained. For all the repeated measurements, the following data was produced with the LMEMs: the value of the variable at every hour or intercept point, the variance around the estimated longitudinal course, and the coefficient or slope of the estimated line between every knot in the longitudinal course.

3.3.2. Processing for timestamp longitudinal course

In the repeated measurements data set, the processing was mainly concerned with making sure that the timestamp of the data was entered in such a way that the joint model could find the longitudinal course of a certain variable. This meant that the timestamp of the *optiflowstart* was used as a starting time for all the repeated measurements. Using this timestamp made sure that all the 8 hours of repeated measurements were started at 0.00 hours for every patient.

3.3.3. Missing data strategy

As mentioned before, in the first step of the joint model an LMEM was used to find the variables that describe the longitudinal course of the repeated measurements. When there were not enough data points at a certain point in the 8 hours of measurements, the LMEM could give a missing data point. The missing data could be in the intercepts, variance, or coefficients. In the Rstudio environment, the missing data were not imputed as the model type that was used to make the LMEM corrected for missing values. This means that the estimation of the longitudinal course was still executed when values are missing. The LMEM then used values and estimation patterns of other patients to predict the course. In the Python environment, the missing data of the LMEM features were imputed using multiple imputations, which was the same strategy that was applied to the aggregated data set. Moreover, the same threshold for missing data was retained, which means that a variable was dropped if it missed more than 60% of its values.

3.3.4. Addition of aggregated variables

Once the values, variance, and coefficients were found for all the variables, the data set with the results of the first step of the joint model was generated. For the second step of the joint model, the variables that did not have repeated measurements were extracted from the aggregated data set and joined together with the data from the LMEMs. Examples of these variables are age, length, weight, origin, and all of the lab measurements. If applicable, the said variables remained encoded and imputed. Combining the two data sets (also called data frames) was done in Python using *merge*. In this function of Python, a parameter can be set to merge the data frames on a specific label. Here, the data frames were merged on *'patientnr'*, which is the newly generated patient ID. Next to the extraction of the aggregated variables, the labels of the patients were also extracted. Due to the fact that the same distribution of intubated patients remained, resampling also was not necessary here.[17]

3.4. Model development

The two different data sets were used in the development of two different predictive models. With the aggregated data set, an LRM, RF, and GBM were developed. With the repeated measurements data set, a joint model was developed that eventually also resulted in an LRM, RF and GBM. The Python codes that were used in the development of the aggregated and joint models can be found via the following link to the GitHub repository in which they are stored.
`https://github.com/manonhendriks/Thesis_Intubation_prediction.git`
For both the aggregated data models and joint models, a graphical overview of the nested cross-validation (also known as cross-validation pipeline) can be found in Appendix A.6 and A.7.

3.4.1. Aggregated data set models

The three models that were developed with the aggregated data set were written in Python using Spyder©, which is a code editor.[22] In the different subsections below, the specific methods for the models are explained.
Before the nested cross-validation could be entered, the data needed to be divided into a training and test set. Because of the availability of an external validation set, the whole data set was used in the nested cross-validation.
Using a nested cross-validation, the three model types were fine-tuned and compared on their performing value. Moreover, the feature selection methods and scaling methods were implemented in the nested cross-validation. As researched, the missing data imputation strategy could be executed before the cross-validation since the training and test data are identically distributed. After this data imputation process, the data were split into k-fold sets, with the *k* chosen based on literature. The hyperparameters of the three models were tuned in the cross-validation. For this inner cross-validation, a fitting *k* was also chosen. In this thesis project, both the inner and outer cross-validation were a 5-fold cross-validation. The splitting of the data into the outer and inner cross-validation folds was done using the function *StratifiedKFold* of sci-kit learn. This function makes sure that the distribution of classes remains the same over all the folds.[14]
The nested cross-validation resulted in the AUC performance value of the three models with different hyperparameters. Moreover, three scaling options were considered: Robustscaling, Power Transformation, or no scaling. Lastly, three feature selection methods were implemented: PCA, SelectKbest, or no feature selection method. For the feature selection methods, LASSO was not implemented in the nested cross-validation as it was already used as a hyperparameter in the LRM. Therefore, the nested cross-validation now resulted in an average AUC value (5 test AUC values for every outer fold) for every model, scaling and selection option. In total, 27 average AUC values were obtained. Using these results, the optimal model with the optimal hyperparameters, (optional) scaling, and (optional) feature selection method can be chosen, which can then be validated using the external validation set.

3.4.2. Hyperparameter tuning in the aggregated data models

In the following subsections, the tuning process of the hyperparameters of the three models will be explained. For every model type, the entered hyperparameters were optimized using the function *GridSearchCV* of sci-kit learn. Before the nested cross-validation was implemented, a general idea was found for the optimal values of the hyperparameters. This suspected optimal value of the hyperparameter was often the middle value of the grid search. For the other two values, entries were chosen that were close to the suspected optimal. In the *GridSearchCV* function, the AUC value was used to determine the best-performing hyperparameters.[14]

Logistic Regression Model
The Logistic Regression Model (LRM) has one hyperparameter that is often tuned, namely the slack parameter (commonly referred to as the C parameter).[14] This hyperparameter is responsible for minimizing over-fitting of the model and determines how much you can 'slack' around objects to improve the classification and prediction of the model. To determine which value for C is preferred, a grid of the following values was tested in the inner cross-validation: 0.1, 1 and 10. Next to the C parameter, the penalty parameter was tested with the entries *'l1'* and *'l2'*. The penalty parameter describes the methods of regularization. As explained before, L1 regularization is also called LASSO and can cancel out features. L2 regularization is also called Ridge and can shrink features to become close to zero. L1 regularization is useful when useless features need to be excluded and L2 regularization is preferred when most variables are useful.[18] Lastly, the solver parameter that determines which regression is applied to shrink the features was set to *'liblinear'* as this solver can handle both the *l1* and *l2* penalty.[14]

Random Forest
The Random Forest has many hyperparameters that can be tuned. In this project, it was chosen to tune the following hyperparameters: *'n_estimators'*, *'max_depth'* and *'min_samples_split'*. For keeping the model computationally efficient, only these three commonly used parameters were tuned in the inner cross-validation.
The *'n_estimators'* parameter determines the number of trees in the random forest. The grid search values were 100, 200 and 500. The *'max_depth'* parameter determines the maximum depth of an individual tree. If this parameter would be *None*, the nodes of the tree would be expanded until all leaves are pure (contain one sample). Here, the *'max_depth'* parameter grid search values were 3, 5 and 10. Another way to make sure that the leaves do not end up containing one sample is with the parameter *'min_samples_split'*. This parameter gives the minimum number of samples required to split a node. The grid search values for this parameter were 5, 10 and 20.
These three hyperparameters are tuned to minimize the over-fitting of the model.[14] If, for instance, the depth of the tree is not specified, the tree can grow into a very specific tree for a selection of the patients, making it not generalizable and too specified. The *'min_samples_split'* parameter is therefore also specified at a higher number than the default, which is 2. This makes sure that single samples cannot end up in their own leaf.

Gradient Boosting Model
Many hyperparameters can be tuned in Gradient Boosting Models. In consultation with a data scientist from Pacmed, the following parameters were looked into: *n_estimators, learning_rate, gamma, max_depth, min_child_weight, subsample, colsample_bytree, reg_lambda, objective*. Eventually, the following three hyperparameters were tuned in the inner cross-validation, again to keep the model computationally efficient. Similar to the Random Forest, the *'n_estimators'* and *'max_depth'* were tuned. The third hyperparameter was the *'learning_rate*. The values for the grid search of the *'n_estimators'* and *'max_depth'* were the same as in the Random Forest, namely 100, 200, and 500 and 3, 5, and 10. The *'learning_rate* hyperparameter shrinks the contribution of each tree by the value of the entered parameter. It determines how fast or slow the model moves towards the optimal weights, with a smaller value indicating a slower learning rate.[14] The following values were entered in the grid search: 0.01, 0.1, and 1.

3.4.3. Repeated measurements data set model

The repeated measurements data set was used in the development of a joint model. In the joint model, a two-stage approach was developed to first determine the longitudinal course of the variables using LMEMs. The newly generated features created with the LMEMs were then used in the second step of the joint model. This second step was again a nested cross-validation with an LRM, RF and GBM. The method of the joint model is explained in the subsections below.

First step of the joint model

As stated before, a joint model uses the information from the longitudinal course of variables, typically variables with repeated measurements. The information of this longitudinal course is extracted from the data using LMEMs. The development of the LMEMs was the first step of the joint model. The models were coded in Rstudio. An existing function *lmer* was used to fit the LMEMs.[23] Cubic splines were used to make an estimation graph of the longitudinal course of the different variables.[24] In these cubic splines, the number of knots could be defined. Between two knots the LMEM fits a cubic spline, which is a summary or prediction of the longitudinal course of the specific variable. In this project, seven knots seemed to fit best as this means every hour had a knot. However, for every repeated measurement the ideal number of knots was individually researched.

From these generated LMEMs different variables were obtained. These variables form the features for the variables that contained repeated measurements in the nested cross-validation of the joint model. These variables contained the slopes (also known as coefficients) between the knots in the graph, the intercept points (values of the parameter at every knot), and the Root Mean Squared Error (RMSE) around the estimated graph. The following formula was used to calculate the RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

In Appendix A.8 an example of an LMEM graph is given for the Systolic Blood Pressure (SBP). In the graph, the three different variables that were obtained for every repeated measurement variable using the LMEM are shown.

Nested cross-validation with repeated measurements data

The second step of the joint model was the development of a predictive model. To make the eventual models comparable, the same three model types, feature selection methods, and scaling methods were again evaluated in a nested cross-validation. Thus, this resulted in the same 27 model options as in the aggregated data model.

Before the model development, the data set was split into a train and test set. 80% of the data was used as a training data set and 20% was used as a test data set. The function *train_test_split()* in Python was used to split the data and the parameter *stratify* was used on the label of the data set (intubated: yes/no). Using this function and parameter, the distribution of intubated patients remained close to the original distribution in the train and test data set.[14]

Moreover, the same grid search values as in the aggregated data model were entered for the hyperparameter tuning.

As stated in the processing section 3.3.4 of the repeated measurement data set, the features that did not contain repeated measurements and therefore were not estimated with an LMEM needed to be added to the joint model data set. Once this data set merge was completed, the nested cross-validation could be executed.

3.5. Model validation and evaluation

The validation of all the models was done by obtaining the AUC values of the models on the test data. The LRM, RF and GBM were validated in a nested cross-validation that included the tuning of the hyperparameters and results of the best-performing model. As stated before, the nested cross-validation of both the aggregated data models and the joint models resulted in 27 average test AUC values, as three models, three scaling methods, and three feature selection methods are tested in a 5-fold outer cross-validation loop. The best-performing model was found by comparing AUC values. In this thesis project an AUC value of 0.7 to 0.8 was considered acceptable, conform the study of Mandrekar et al. 2010.[25]

For the best-performing aggregated data model the Santeon data was used as external validation. This means that for the 27 average test AUC values, the best-performing combination of model type (with tuned hyperparameters), scaling method, and feature selection method was exported to apply an external validation with the Santeon data.

In this external validation, the Maasstad data was not used in the data set as this Maasstad data was already used in the training and internal validation of the model. The number of repeated measurements in the validation set determined its applicability to be used as an external validation set for the joint model. In case of insufficient repeated measurements to externally validate the best-performing joint model, an internal validation set should be generated. In this case, this would contain the 20% test set that was separated from the data before entering the nested cross-validation.

Once the best-performing models for the aggregated data models and joint models were found, these two models with the corresponding hyperparameters that were tuned optimally were trained on all available training data. After this fitting process, the two models predicted new unseen data. This was the external validation data in the case of the best-performing aggregated data model. For the best-performing joint model, this was the external validation data or the separated 20% internal validation data in case of insufficiency.

Of these two models the learning curves were formed when the models were fitted on all available training data. These learning curves show the generalizability of the two models.

Moreover, a classification report was calculated at every outer fold and also in the last validation step. In this classification report the precision, recall, and f1-score were taken into account. After the last validation step, the validation scores were compared to the training scores. Precision is a metric that measures the ratio of true positives to the sum of true positives and false positives. Precision is also called positive predictive value (PPV). Recall is a metric that measures the ratio of true positives to the sum of true positives and false negatives. Recall is also called sensitivity. The f1-score is a metric that combines both precision and recall into a single score, in which higher values indicate better performance.[26] All these three metrics give calculations that are separated based on the class, in which class 0 means non-intubated patients and class 1 means intubated patients. In Appendix A.9 the formulas corresponding to precision, recall, and f1-score are given.[26]

In the validation step of the best-performing models, a Receiver Operating Curve (ROC) was made. This was done when the best-performing model type was trained on all available data. It was not feasible to obtain a ROC in the nested cross-validation, due to computational problems. Therefore, these ROCs had a different distribution of the 5 folds than the 5 folds that were used in the nested cross-validation. However, the same data was used so the results were expected to be in the same range of the test AUC values. In a ROC, the True Positive Rate (recall) is displayed on the y-axis. The False Positive Rate is presented on the x-axis. It thus shows the trade-off between sensitivity (y-axis) and specificity (x-axis). Classifiers that give a curve closer to the top left corner (1.0 sensitivity) indicate better performance. The AUC value is the area under the ROC. Thus, a curve closer to the top left will often lead to a higher AUC value.[27] In Appendix A.9 the formulas corresponding to metrics used in the ROC are stated.

4 Results

In this chapter, the results of this thesis project will be stated. The characteristics of the different data sets and patients will first be explained. Hereafter, the performance of the implemented models will be stated and compared on their used data type. Lastly, the best-performing models are externally or internally evaluated with a validation set.

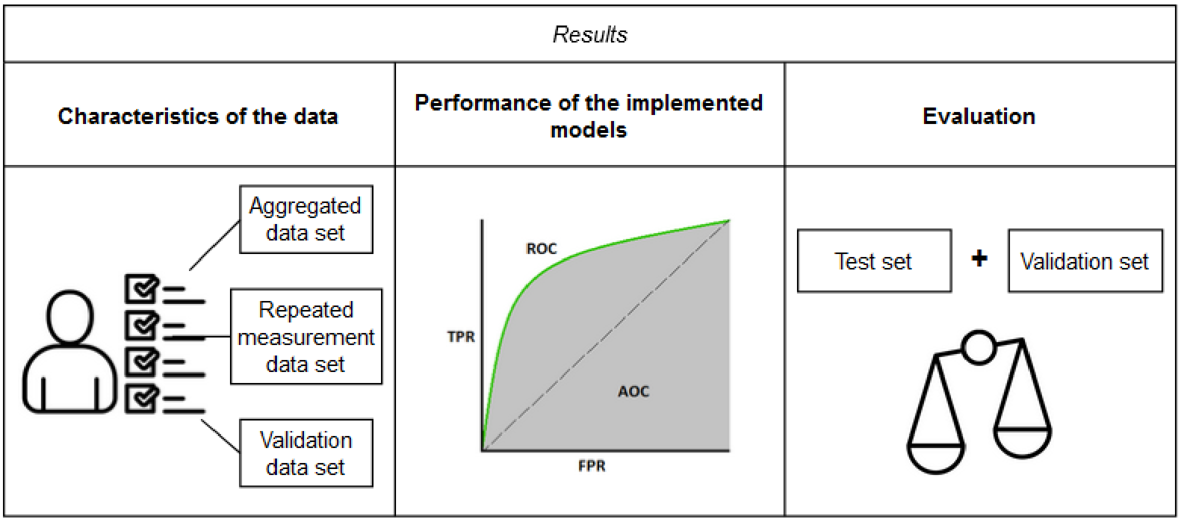

Results
Characteristics of the data
Performance of the implemented models
Evaluation
Aggregated data set
Repeated measurement data set
Validation data set
ROC
TPR
AOC
FPR
Test set
+
Validation set

Figure 4.1: Graphical overview of the results of this thesis

4.1. Description of the data

In this section, a description of the used data sets will be stated, including the aggregated data set, repeated measurements data set, and validation data set. Furthermore, the patient characteristics will be provided. The extraction of the training and test data for both the aggregated data set and the repeated measurements data set, executed by BI, resulted in two data sets with patient admissions to the ICU from January 2018 to February 2022. The validation set obtained through the 'Beheercommissie' of Santeon resulted in a data set with patient admissions to the ICU from March 2020 to February 2021.

4.1.1. Aggregated data set

The aggregated data set contained 350 patient admissions. After removing the duplicate patient IDs, 348 patients remained. Of these 348 patients, 168 patients were given the intubation label. All 348 patients contained at least one missing data point in the features. Of the 116 features, 95 features contained missing data. 75 features remained after excluding features with more than 60% missing values, dropping features that contained useless information, and encoding categorical data. A list of these 75 features with a short description can be found in Appendix A.10.

Some of the vital function parameters did include the mean and last value of the variable but not the standard deviation. For example, the standard deviation of the temperature was not included in the features because of the large number of missing values. The temperature was often measured only once, making it impossible to give a standard deviation.

Multiple imputation was not possible to implement as the missing data strategy because of the computational power needed to apply multiple imputations. The number of iterations for the estimator used to apply multiple imputation needed to be enlarged to get the desired result. The default value is 10. However, from 500 iterations the estimator was able to implement multiple imputation. This, however, gave a memory error. This memory error resulted in the maximum iterations value being set at 150. At this value, the code did result in multiple imputations. Nevertheless, over all of the imputed data frames, the imputed values were the same. For example, patient 2 was missing a lab value, namely CRP. In all the imputed data frames, the value was imputed with 18. Different estimators were tested for the multiple imputation. Eventually, it was chosen to apply a KNearestNeighbour (KNN) multiple imputation with k=3 and the number of imputations also 3. Because the three imputed data frames were identical, only the first was taken into account for the nested cross-validation. The KNN imputation algorithm uses the (in this case) three nearest neighbours to determine the value of the missing data point.[28]

4.1.2. Repeated measurements data set

The number of patients and intubated patients in the repeated measurements data set was equal to the aggregated data set, since the two data sets were linked on the patient IDs. The repeated measurements data set contained 289 features, of which 190 features had missing data. Again, logically, all the patients had at least one missing data point. 235 features remained after excluding useless features and features with more than 60% missing values. A list of the vital function parameters that were different from the aggregated data set is given in Appendix A.11. All the other features corresponded to the aggregated data set.

Once more, in this data set the standard deviation of the temperature was not included because of the large number of missing values. Here too the missing data imputation with KNN k=3 did not result in different imputed values. Therefore, again, only the first imputed data frame was taken into account.

4.1.3. Patient characteristics of the training data

As both the aggregated data set and the repeated measurements data set contained the same patients (with different variables), the patient characteristics will be explained in this subsection for both data sets. The mean age of the patients in the non-intubated group was 61.4 years (±14.3), while in the intubated group it was 62.1 years (±12.3). The non-intubated group consisted of 114 males and 66 females and the intubated group had 117 males and 51 females. The length and weight of the non-intubated group were close to the intubated group being 172 cm (±20) versus 172 cm (±21) and 87.9 kg (±22.7) versus 85 kg (±19.6).

The origin of the non-intubated group was distributed as follows: 164 patients came from their homes, 4 from their hospital, 8 from another hospital and 4 via the ER. For the intubated group, the origin was distributed likewise: 149 patients came from their homes, 7 from their hospital, 9 from another hospital and 3 via the ER. In the non-intubated group, 97 patients had a day admission to the ICU and 83 patients had a night admission. In the intubated group, 105 patients had a day admission and 63 patients had a night admission.

The number of deceased patients was 29 in the non-intubated group and 66 in the intubated group. Patients that died during their admission and were not intubated yet were also given the intubation label. The number of deceased patients was retrieved in the data extraction process. Patients who died after the data extraction have not been considered. This means that the patients who had a date of death in the data either died during their ICU admission or after the admission but within the data extraction period. The seemingly large difference in this characteristic can be attributed to the fact that patients who died during their admission were added to the intubation group. In table 4.1 an overview of the patient characteristics can be found.

Table 4.1: Overview of the patient characteristics of the training data set separated into the non-intubated group and intubated group

	Non-intubated group (n=180)	**Intubated group (n=168)**
Age (y)	61.4 (± 14.3)	62.1 (± 12.3)
Gender	63% M (n=114)	70% M (n=117)
Length (cm)	172 (± 20)	172 (± 21)
Weight (kg)	87.9 (± 22.7)	85 (± 19.6)
Origin	- Home = 164 - Own hospital = 4 - Other hospital = 8 - Emergency Room = 4	- Home = 149 - Own hospital = 7 - Other hospital = 9 - Emergency Room = 3
Day or night admission	- Day admission = 97 - Night admission = 83	- Day admission = 105 - Night admission = 63
Number of deceased patients	29	66

To get a full picture of the course of the patient admission, the two variables *Optiflowstart* and *Intubationstart* were used to determine the number of hours between HFNO therapy and intubation. Logically, only the patients who were intubated had a value in this calculation. The average time of intubation after the start of HFNO therapy was 49 hours, with a standard deviation of 70 hours. Because of the wide range of values, the minimum and maximum values are also shared. In all of these values, the minimal 8 hours of HFNO therapy was also counted. The minimum time of intubation after the start of HFNO therapy was 8.87 hours and the maximum time was 536.48 hours.

4.1.4. Validation data set

The validation data set contained 163 patient admissions. After removing one duplicate patient ID, 162 patients remained. Of these 162 patients, 94 were given the intubation label. The validation data set contained 116 features, of which 101 features had missing values. Again, all the patients had at least 1 missing value.

Unfortunately, the validation set did not contain enough repeated measurements to validate the joint model data set. In Appendix A.12 the number of repeated measurements in both data sets can be found. For example, the parameter Heart Frequency (HF) had 166494 measurements for all the patients in the training data set and only 4889 for all the patients in the validation set. This would mean that there are around 478 measurements per patient in the training set and around 30 in the validation data set. Because of the large difference in the number of repeated measurements, the LMEMs with 7 knots for each hour on HFNO therapy would not be feasible. With more than half of the LMEMs of the training data containing 7 knots, this would result in a great number of missing values in the validation set.

For the encoding of categorical data, the strategy applied to the training data was again implemented. This resulted in gender being encoded as zeros for males and ones for females. The origin parameter again had four options that belonged to similar categories, namely patient admissions from their home, other hospitals, own hospital, or ER.

4.1.5. Patient characteristics of the validation data

The mean age of the patients in the non-intubated group was 60.2 years (±13). In the intubated group it was 66.1 years (±10.3). The non-intubated group consisted of 47 males and 21 females, while the intubated group had 74 males and 20 females. The length and weight of the non-intubated group were close to the intubated group, being 175 cm (±10) versus 176 (±8) and 91.1 kg (±18.8) versus 91.4 kg (±16.5).

The origin of the non-intubated group was distributed as follows: 39 patients came from their homes, 5 from their hospital, 10 from another hospital and 14 via the ER. For the intubated group, the origin was distributed in the following manner: 62 patients came from their homes, 6 from their hospital, 3 from another hospital and 23 via the ER. In the non-intubated group, 36 patients had a day admission to the ICU and 32 patients had a night admission. In the intubated group, 61 patients had a day admission and 33 patients had a night admission.

The number of deceased patients was 3 in the non-intubated group and 38 in the intubated group. Patients that died during their admission and were not intubated yet were again given the intubation label. Once more, deaths after the data extraction have not been considered. The large difference in the number of deceased patients can again be due to the addition of patients that died in their admission to the intubation group. In table 4.2 an overview of the patient characteristics can be found.

Table 4.2: Overview of the patient characteristics of the validation data set separated into the non-intubated group and intubated group

	Non-intubated group (n=68)	**Intubated group (n= 94)**
Age (y)	60.2 (± 13)	66.1 (± 10.3)
Gender	69% M (n=47)	79% M (n=74)
Length (cm)	175 (± 10)	176 (± 8)
Weight (kg)	92.1 (± 18.8)	91.4 (± 16.5)
Origin	- Home = 39 - Own hospital = 5 - Other hospital = 10 - Emergency room = 14	- Home = 62 - Own hospital = 6 - Other hospital = 3 - Emergency room = 23
Day or night admission	- Day admission = 36 - Night admission = 32	- Day admission = 61 - Night admission = 33
Number of deceased patients	3	38

Also in the validation data the variables *Optiflowstart* and *Intubationstart* were used to determine the time between the start of HFNO therapy and (possible) intubation. The average time of intubation after the start of HFNO therapy was 59 hours, with a standard deviation of 69 hours. The minimum time of intubation after the start of HFNO therapy was 8.78 hours and the maximum time was 401.2 hours. Once more, the minimal 8 hours of HFNO therapy were counted in the calculations.

4.2. Performance of the implemented aggregated data set models

In the following subsections, the performance of the implemented aggregated data set models will be stated. This is done by first showing the performance during training and testing of the models and later showing the performance when using an external validation data set.

4.2.1. Training performance of the aggregated data set models

The performance of the aggregated data set models will be shown by first discussing the training and testing performance and hereafter examining the performance of the external validation set.

Training and testing performance of the aggregated data set models

The training AUC values were retrieved from the nested cross-validation. This means that for every inner fold, an AUC score was found. This process was repeated five times for the number of outer folds. This resulted in 25 AUC values for each combination of model, feature selection, and scaling type (27 options). As mentioned before, the model types were LRM, RF, and GBM, the feature selection methods were PCA, SelectKbest, and no feature selection, and the scaling methods were PowerTransformer, Robust Scaling, and no scaling. The inner fold of the nested cross-validation resulted in a best-tuned model. In every outer fold, this best-tuned model was given a test set from which an AUC score resulted. Since there were 5 outer folds, every combination of model, feature selection, and scaling option resulted in 5 test AUC scores. For both the 25 training AUC scores and the 5 test AUC scores, the mean and standard deviation were calculated. The resulting means and standard deviations can be found in Table 4.4. In this table, the green square highlights the model that scored best on the test AUC value.

Thus, with this aggregated data set, it means that the Random Forest with no feature selection and no scaling scores best, with an average test AUC value of 0.694 and a standard deviation of 0.05. In the five test AUC values of this specific model type, the last fold resulted in the highest test AUC value of 0.764. With every training AUC value, a classification report with the corresponding precision, recall, and f1-score was also calculated. In table 4.3 the average precision, recall and f1-score with the calculated standard deviation are shown. Here, it can be seen that the best-performing model performs better for class 0 (the non-intubated patients), because of a higher recall and f1-score. However, for precision, the model performs better in class 1 (the intubated patients).

Table 4.3: Average and standard deviations of the precision, recall, and f1-score of the best-performing aggregated data model: a Random Forest with no feature selection and no scaling.

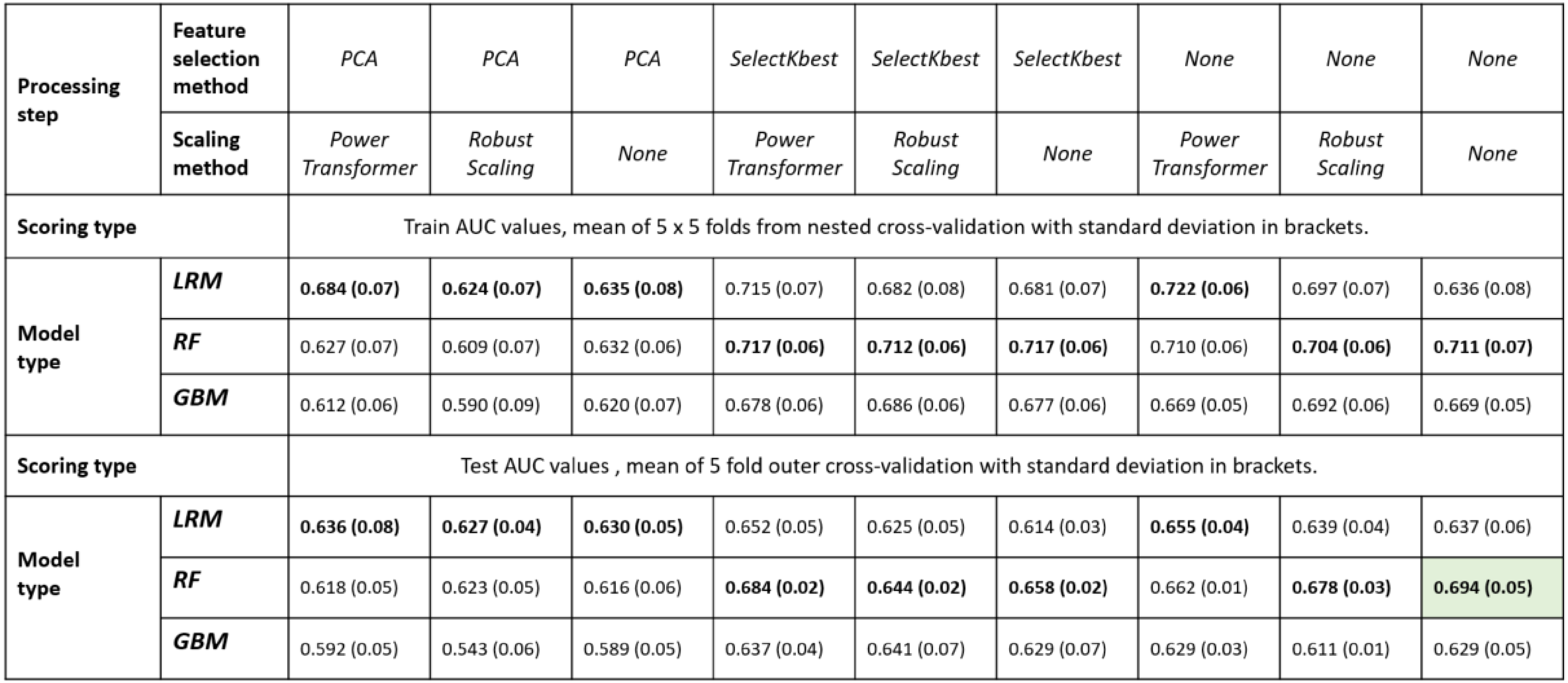
	Average (standard deviation)
Precision class 0	0.683 (0.05)
Precision class 1	0.76 (0.04)
Recall class 0	0.833 (0.03)
Recall class 1	0.583 (0.09)
f1-score class 0	0.747 (0.04)
f1-score class 1	0.657 (0.07)

The model type that scored second best is also an RF. However, in this combination SelectKbest was used as a feature selection method and Power Transformer as a scaling method. This combination resulted in an average test AUC value of 0.684 and a standard deviation of 0.02. In the five test AUC values of this model option, fold 5 again resulted in the highest test AUC value of 0.72. The third-best scoring model is an RF with no feature selection and Robust Scaling, which resulted in an average test AUC of 0.678 and a standard deviation of 0.03. Fold 5 resulted once more in the highest test AUC value of 0.72.

To understand more of what the models use in their prediction, the top 10 predictive features were retrieved at every outer fold in which the test AUC was also obtained. The corresponding tuned hyperparameters are given for the three best-scoring models. The top 10 predictive features and the tuned hyperparameters correspond to the outer fold that resulted in the highest test AUC value, which was fold 5 in these models. These features and hyperparameters can be found below in section 4.2.2. For more information on the different features, a list with a description of the used features can be found in Appendix A.10.

Moreover, the learning curves corresponding to these three model options can be found in Appendix A.13. In these learning curves, it can be seen that for all three model types the training score starts high and grows towards the cross-validation score when the number of training samples is enlarged. This indicates that when using a few samples, the model will be overfitting, giving a high training score, and with more training samples, the training and cross-validation scores become more similar and less overfitted.

Table 4.4: Training and test Area Under the Curve (AUC) values of the different model types made with the aggregated data set with different feature selection and scaling methods. The train AUC values were obtained by the average of the nested cross-validation resulting in 5 x 5 folds of train AUC values and a standard deviation in brackets. The test AUC values were obtained by the average of the 5-fold outer cross-validation, with the standard deviation in brackets. The model type, feature selection method, and scaling method that resulted in the best test AUC score is highlighted in green.



Processing step	**Feature selection method**	*PCA*	*PCA*	*PCA*	*SelectKbest*	*SelectKbest*	*SelectKbest*	*None*	*None*	*None*
	Scaling method	*Power Transformer*	*Robust Scaling*	*None*	*Power Transformer*	*Robust Scaling*	*None*	*Power Transformer*	*Robust Scaling*	*None*
Scoring type		Train AUC values, mean of 5 x 5 folds from nested cross-validation with standard deviation in brackets.								
Model type	***LRM***	**0.684 (0.07)**	**0.624 (0.07)**	**0.635 (0.08)**	0.715 (0.07)	0.682 (0.08)	0.681 (0.07)	**0.722 (0.06)**	0.697 (0.07)	0.636 (0.08)
	RF	0.627 (0.07)	0.609 (0.07)	0.632 (0.06)	**0.717 (0.06)**	**0.712 (0.06)**	**0.717 (0.06)**	0.710 (0.06)	**0.704 (0.06)**	**0.711 (0.07)**
	GBM	0.612 (0.06)	0.590 (0.09)	0.620 (0.07)	0.678 (0.06)	0.686 (0.06)	0.677 (0.06)	0.669 (0.05)	0.692 (0.06)	0.669 (0.05)
Scoring type		Test AUC values , mean of 5 fold outer cross-validation with standard deviation in brackets.								
Model type	***LRM***	**0.636 (0.08)**	**0.627 (0.04)**	**0.630 (0.05)**	0.652 (0.05)	0.625 (0.05)	0.614 (0.03)	**0.655 (0.04)**	0.639 (0.04)	0.637 (0.06)
	RF	0.618 (0.05)	0.623 (0.05)	0.616 (0.06)	**0.684 (0.02)**	**0.644 (0.02)**	**0.658 (0.02)**	0.662 (0.01)	**0.678 (0.03)**	**0.694 (0.05)**
	GBM	0.592 (0.05)	0.543 (0.06)	0.589 (0.05)	0.637 (0.04)	0.641 (0.07)	0.629 (0.07)	0.629 (0.03)	0.611 (0.01)	0.629 (0.05)

4.2.2. Additional information on the best-performing aggregated data models

Below, the optimally tuned hyperparameters and the top 10 predictive features for the three best-performing aggregated data models are stated. These are retrieved from fold 5, which resulted in the highest test AUC for all the three best-performing models.

Hyperparameters and top 10 predictive features of an RF - No feature selection - No scaling

Hyperparameter	Optimized input
'max_depth'	3
'min_samples_split'	20
'n_estimators'	500

1. FiO2 last value
2. SpO2 std dev
3. Spo2 mean
4. Fio2 mean
5. LDH
6. SaO2 mean
7. DBP mean
8. pO2 mean
9. pO2 last value
10. SaO2 last value

Hyperparameters and top 10 predictive features of an RF - SelectKbest - Power Transformer

Hyperparameter	Optimized input
'max_depth'	10
'min_samples_split'	5
'n_estimators'	100

1. FiO2 delta
2. FiO2 mean
3. RR std dev
4. Length
5. Age
6. FiO2 last value
7. PCO2 mean
8. SaO2 mean
9. pH last value
10. PO2 std dev

Hyperparameters and top 10 predictive features of an RF - No feature selection - Robust scaling

Hyperparameter	Optimized input
'max_depth'	3
'min_samples_split'	20
'n_estimators'	200

1. FiO2 last value
2. SpO2 std dev
3. SaO2 mean
4. FiO2 mean
5. SpO2 mean
6. LDH
7. PO2 last value
8. PO2 mean
9. SaO2 last value
10. DBP mean

4.2.3. Validation of the best model made with the aggregated data set

For the validation, the best model found in the training and testing phase was used to examine its performance with the external validation data. The external validation data set was reduced in its number of features using the list of features that were used by the aggregated model, which can also be found in Appendix A.10. After this, the missing data was checked per column. The medication features (*Antibiotics* and *Steroids*) all had no data. Because of the lengthy process of extracting this information from the Santeon data, it was chosen to impute the missing data in these features with zero. Moreover, other features contained a lot of missing data, especially features from the laboratory values. For example, *Serum albumin* had zero data points and was therefore also imputed with zero. Overall, there were few data points of the laboratory values. However, KNN could be used as an imputation method. Again, k=3 was chosen.

To compare the fraction of missing data in the validation data versus the training data, a histogram was made showing the fraction of data present in the aggregated data set next to the fraction of data present in the validation data set. This histogram can be found in Appendix A.14.

After the encoding of categorical data and imputation of missing data, the extracted best model could be fitted with all the training data. Hereafter, a prediction could be made with the validation data. The fitting/training of the best-performing model was done with the tuned hyperparameters as stated above (*max_depth*=3, *min_samples_split*=20, and *n_estimators*=500). When the best model (an RF with no feature selection and no scaling) was fitted on all the training data, a learning curve and the top 10 features were again obtained. The learning curve can be found in Appendix A.15. Also in this learning curve, the training and validation score grow towards each other when the training set size is enlarged. The top 10 predictive features that came from training the best model with all the training data are listed below.

In this validation step, the ROC curve was also obtained. Because of the computational burden of extracting all the ROC curves in the nested cross-validation, only the ROC curve of the best-performing model was made. In Figure 4.2 the ROC curve of the RF with no feature selection and no scaling can be found.

Top 10 predictive features of an RF - no feature selection - no scaling

1. SaO2 std dev
2. SaO2 last value
3. LDH
4. PO2 last value
5. PO2 mean
6. SpO2 mean
7. SaO2 mean
8. FiO2 mean
9. SpO2 std dev
10. FiO2 last value

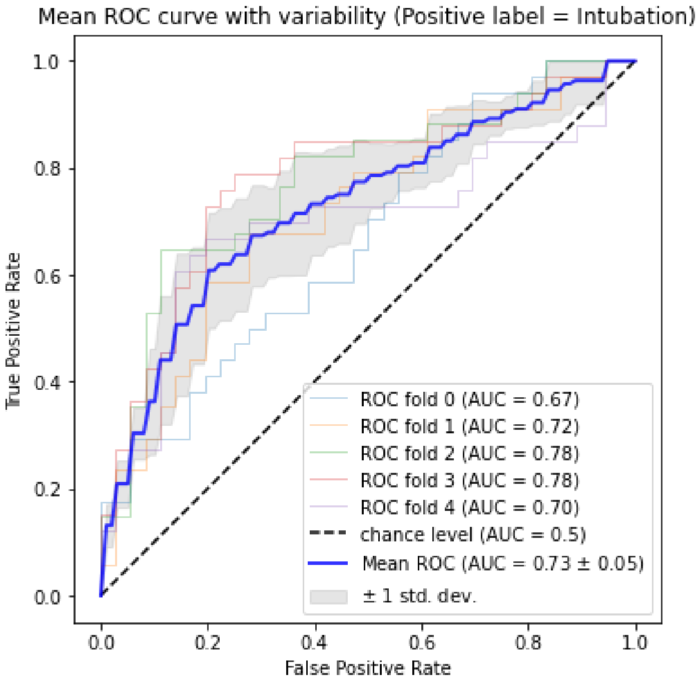


Figure 4.2: ROC curve of the best-performing aggregated model: an RF with no feature selection and no scaling.

The prediction of the fitted best model on the validation data set resulted in an AUC value of 0.559. The precision, recall, and f1-score were also calculated. The precision was higher in class 1 with 0.67 in comparison to 0.46 in class 0. The recall was higher in class 0, being 0.74 in comparison to 0.38 in class 1. The f1-score was also higher in class 0, namely 0.57 compared to 0.49 in class 1. The results can be found in table 4.5 below. Overall, the results are comparable to the previous results since the models seem to perform better on class 0.

Table 4.5: Classification report of the validation of the best-performing aggregated data model.

Classification report	Value
Precision class 0	0.46
Precision class 1	0.67
Recall class 0	0.74
Recall class 1	0.38
f1-score class 0	0.57
f1-score class 1	0.49

4.3. Performance of the implemented joint models

In the following subsections, the Linear Mixed Effects Models results will be discussed, which were generated in the first step of the joint model. Furthermore, the performance of the second step of the joint model will be shown for the different generated models. This is done by first showing the performance during training and testing of the models and later showing the performance when using an internal validation set.

4.3.1. Linear Mixed Effects Models

After the LMEMs were generated, it was checked if there were any notable strange curves. Patients who contained a lot of variation in their measurements were often estimated relatively well. In Appendix A.16 an example is shown of a peripheral oxygen saturation (SpO2) curve in which the possible variation of the measurements is shown. The LMEM still managed to estimate the course of the variable sufficiently. Some notable curves could not be explained by the supervisors in this project. These curves were therefore discussed with a specialised nurse of the ICU who has experience with validating medical data. This was done using the keys to the actual patients in the hospital database, provided by BI. The notable curves could be explained by looking into the patient data and reading what events took place. Eventually, all the LMEMs were determined to be sufficient to continue the model development process. In Appendix A.17 some of the notable curves are shown with an explanation of these courses of the measurements in these patients.

4.3.2. Performance of the second step joint models

The performance of the joint models is shown by discussing the training and testing performance in the following sections.

Training and testing performance of joint models

Comparable to the performance section of the models made with the aggregated data set, the performance of the joint models was also evaluated through the training and test AUC values. The same number of inner and outer folds was applied in the nested cross-validation. This, therefore, also resulted in 25 train AUC values for every combination of model, feature selection, and scaling method. Moreover, 5 test AUC values were obtained for every combination. The mean and standard deviation of the train AUC values and test AUC values can be found in Table 4.7. In this table, the green square again highlights the model that scored best on the test AUC value.

Thus, implementing repeated measurements in this joint model resulted in a Random Forest with no feature selection and Power Transformer scaling with the best performance, with an average test AUC of 0.681 and a standard deviation of 0.07. In the 5 test AUC values of this model type, the third fold scored best with a test AUC of 0.801. With every training AUC value, a classification report with the calculated precision, recall, and f1-score was made. In table 4.6 the average precision, recall, and f1-score with the calculated standard deviation is shown for the best-performing model. The results are similar to the best-performing aggregated model. It can be seen that the model again performs better for class 0, which is the non-intubated group, as seen by a higher recall and f1-score for this class. The precision is slightly better for class 1, which is the intubated group.

Table 4.6: Average and standard deviations of the precision, recall, and f1-score of the best-performing joint model: a Random Forest with no feature selection and Power Transformer scaling.

	Average (standard deviation)
Precision class 0	0.676 (0.07)
Precision class 1	0.698 (0.09)
Recall class 0	0.758 (0.09)
Recall class 1	0.604 (0.12)
f1-score class 0	0.714 (0.07)
f1-score class 1	0.642 (0.09)

The model type that scored second best is an RF with no feature selection and Robust Scaling. For this model, the average test AUC was 0.659 and the standard deviation 0.07. Here, the third fold also had the best test AUC value of 0.713. The third-best scoring model is a Gradient Boost Model with

SelectKbest feature selection and Robust Scaling. This model resulted in an average test AUC of 0.657 and a standard deviation of 0.05. Again, the third fold resulted in the best AUC value of 0.713.
The top 10 predictive features that were used in these three model types to get the highest test AUCs are provided in section 4.3.3, along with the corresponding tuned hyperparameters. For more information about the features, a list with descriptions of all different features used in the joint model compared to the aggregated data model is inserted in Appendix A.11.
Moreover, the learning curves corresponding to these three model options can be found in Appendix A.18. Similar to the learning curves of the best-performing aggregated data models, the training and cross-validation scores grow towards each other when the number of training samples is enlarged.

Table 4.7: Training and test Area Under the Curve (AUC) values of the different model types made with the repeated measurements data set with different feature selection and scaling methods. The train AUC values were obtained by the average of the nested cross-validation resulting in 5 x 5 folds of train AUC values and a standard deviation in brackets. The test AUC values were obtained by the average of the 5-fold outer cross-validation, with the standard deviation in brackets. The model type, feature selection method, and scaling method that resulted in the best test AUC score is highlighted in green.

Processing step	**Feature selection method**	*PCA*	*PCA*	*PCA*	*SelectKbest*	*SelectKbest*	*SelectKbest*	*None*	*None*	*None*
	Scaling method	*Power Transformer*	*Robust Scaling*	*None*	*Power Transformer*	*Robust Scaling*	*None*	*Power Transformer*	*Robust Scaling*	*None*
Scoring type		Train AUC values, mean of 5 x 5 folds from nested cross-validation with standard deviation in brackets.								
Model type	***LRM***	0.669 (0.1)	0.642 (0.08)	0.671 (0.09)	0.747 (0.08)	0.752 (0.08)	0.734 (0.08)	0.689 (0.08)	0.631 (0.09)	0.670 (0.1)
	RF	0.550 (0.07)	0.590 (0.1)	0.589 (0.1)	0.717 (0.08)	0.710 (0.08)	0.708 (0.09)	0.683 (0.09)	0.691 (0.09)	0.685 (0.08)
	GBM	0.606 (0.09)	0.535 (0.09)	0.611 (0.1)	0.677 (0.09)	0.690 (0.08)	0.685 (0.08)	0.623 (0.1)	0.634 (0.1)	0.625 (0.11)
Scoring type		Test AUC values , mean of 5 fold outer cross-validation with standard deviation in brackets.								
Model type	***LRM***	**0.592 (0.07)**	**0.577 (0.04)**	**0.644 (0.03)**	0.642 (0.07)	0.624 (0.06)	**0.650 (0.06)**	0.650 (0.05)	0.603 (0.06)	0.636 (0.06)
	RF	0.552 (0.06)	0.563 (0.07)	0.543 (0.08)	**0.655 (0.08)**	0.647 (0.08)	0.644 (0.09)	**0.681 (0.07)**	**0.659 (0.04)**	**0.652 (0.07)**
	GBM	0.511 (0.04)	0.515 (0.03)	0.472 (0.07)	0.633 (0.07)	**0.657 (0.05)**	0.642 (0.08)	0.642 (0.06)	0.615 (0.08)	0.627 (0.05)

4.3.3. Additional information on the best-performing joint models

Below, the optimally tuned hyperparameters and the top 10 predictive features for the three best-performing joint models are stated. These are retrieved from fold 3, which resulted in the highest test AUC for all the three best-performing models.

Hyperparameters and top 10 predictive features of an RF - No feature selection - Power Transformer

Hyperparameter	Optimized input
'max_depth'	10
'min_samples_split'	5
'n_estimators'	200

1. FiO2_int7
2. FiO2_int6
3. FiO2_int5
4. FiO2_int2
5. FiO2_int3
6. FiO2_int0
7. SpO2_int7
8. FiO2_int4
9. FiO2_slope6
10. FiO2_int1

Hyperparameters and top 10 predictive features of an RF - No feature selection - Robust Scaling

Hyperparameter	Optimized input
'max_depth'	3
'min_samples_split'	5
'n_estimators'	200

1. FiO2_int7
2. FiO2_int5
3. FiO2_int6
4. FiO2_int1
5. FiO2_int4
6. FiO2_int2
7. FiO2_int3
8. PO2_int4
9. SaO2_int4
10. FiO2_slope6

Hyperparameters and top 10 predictive features of a GBM - SelectKbest feature selection - Robust Scaling

Hyperparameter	Optimized input
'max_depth'	5
'learning_rate'	1
'n_estimators'	200

1. CK
2. Ureum
3. Origin_3
4. Chloride
5. DBP_slope0
6. DBP_int1
7. Admission during day or night
8. MCV
9. Hemoglobin
10. DBP_int7

4.3.4. Validation of the best joint model

As stated before, the external validation set distributed by Santeon did not contain enough repeated measurements to validate the best joint model. It was therefore chosen to split the data into a train and test set before the nested cross-validation. For the validation, the 20% test data was used. The extracted best-performing model was trained on all the training data (the 80% of other data) using the best-tuned hyperparameters (*max_depth*=10, *min_samples_split*=5, and *n_estimators*=200). Hereafter, the fitted model was used to predict the remaining 20% test data. Once the best model (an RF with no feature selection and Power Transformer) was fitted on the whole training data set, a learning curve and the top 10 predictive features could be obtained. This learning curve can be found in Appendix A.19. In this learning curve, the training and validation score do not grow towards each other, which could be a sign of overfitting. The top 10 features from training the best model with the whole training data set are listed below.

In this validation step of training the best-performing model on all available data, the ROC curve was obtained. In Figure 4.3 the ROC curve of the RF with no feature selection and Power Transformer scaling can be seen.

Top 10 predictive features of the best-performing joint model trained on all available data (RF - no feature selection - Power Transformer)

1. SpO2_rmse1
2. SpO2_int5
3. DBP_slope0
4. PO2_int0
5. FiO2_int2
6. LDH
7. SpO2_int4
8. FiO2_int5
9. FiO2_int4
10. FiO2_int6

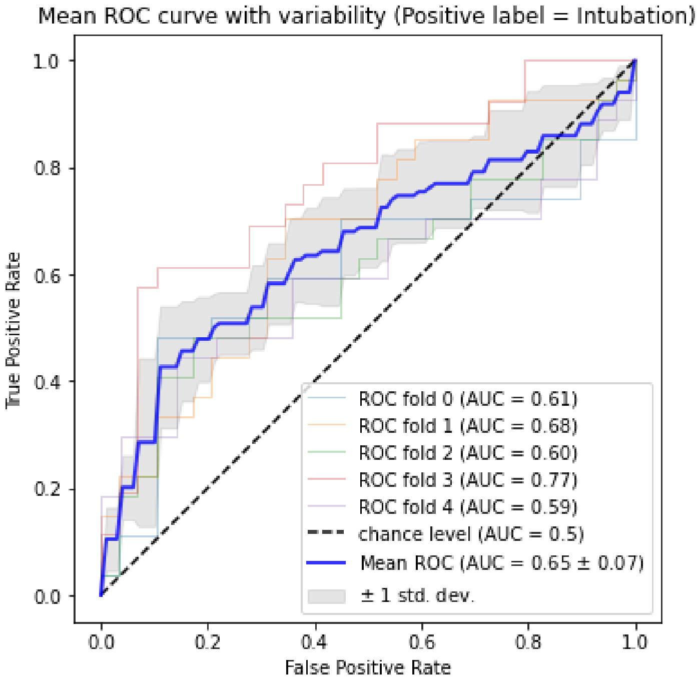


Figure 4.3: ROC curve of the best-performing joint model: an RF with no feature selection and Power Transformer scaling.

The prediction of the fitted best model on the 20% test set resulted in an AUC value of 0.699. The precision, recall, and f1-score were also calculated. The precision was equal in class 0 (non-intubated) and class 1 (intubated), namely 0.70. The recall was higher in class 0, namely 0.72 compared to 0.68 in class 1. The f1-score was also higher in class 0, being 0.71 in comparison to 0.69 in class 1. The results can be found in table 4.8 below. These results are comparable to the training results because the model performs better on class 0.

Table 4.8: Classification report of the validation of the best-performing joint model.

Classification report	Value
Precision class 0	0.70
Precision class 1	0.70
Recall class 0	0.72
Recall class 1	0.68
f1-score class 0	0.71
f1-score class 1	0.69

5 Conclusion

In this thesis project, two different data types were used to develop ML models. Aggregated data and repeated measurements data were used to develop a Logistic Regression Model (LRM), Random Forest (RF) and Gradient Boosting Model (GBM). A nested cross-validation was implemented to test 27 combinations of these three models, three feature selection methods, and three scaling methods. The aggregated data models outperformed the joint models (repeated measurements data). The addition of repeated measurements to the training data does, therefore, not seem to be of added value.

The best-performing aggregated data model was an RF with no feature selection and no scaling, which had a performance value of 0.694 (standard deviation 0.05). The external validation of this model with Santeon data resulted in an AUC of 0.559. The precision, recall and f1-score showed that the model had a better performance on predicting class 0: the non-intubated patients.

With the test AUCs close to the required 0.7 in the aggregated models, it can be concluded that it is possible to predict intubation with data of patients that received HFNO therapy for at least 8 hours.[25] However, with these models, the performance is not sufficient enough to conclude that intubation should be implemented and HFNO therapy will not be sufficient. The precision, recall and f1-score are too poor to rely on the model's prediction to intubate a patient or not. Specifically, the recall of class 1 is 0.583, meaning 41.7% would wrongfully not be intubated. Moreover, the precision of class 1 is 0.76, which means 24% would wrongfully be intubated. Both recall and precision need to be improved to limit the number of false negatives and false positives.

To conclude, the best-performing aggregated data model shows potential and proves that it is possible to predict intubation using AI, but in its current state is far from implementation.

6 Discussion

In this chapter, the results of this thesis project will be discussed. First, the results will be interpreted. Hereafter, the relevance of the results will be stated. Moreover, the limitations of the results will be discussed. Lastly, future recommendations will be given.

6.1. Interpretation of the results

The two-sided method of this thesis project resulted in two models that predicted intubation using different data. The best-performing aggregated data model had a test AUC performance value of 0.694 (standard deviation 0.05) and was a Random Forest with no feature selection and no scaling. The external validation with Santeon data showed a decline in the AUC value, from 0.694 to 0.559. The learning curves from the training fold and the final fitting of the best-performing model on all available training data were showing a similar pattern. The precision, recall, and f1-score were also comparable, with the model having a better performance on class 0 (non-intubated patients).

The decline in AUC value when applying the external validation data can be explained by several reasons. Firstly, the validation data contained a large amount of missing data. In Appendix A.14 a histogram is shown in which the amount of present data in the training data set is compared to the validation data set. The variables that were present in the top 10 predictive features would ideally be significantly present in the validation data set. The arterial saturation measurements (*SaO2 mean/std dev/last value*) were often present in the top 10 predictive features but had almost no data in the validation data set. This also applied to the laboratory features. *LDH* frequently appeared in the most predictive features. Nevertheless, this feature did not have many values in the validation data.

Secondly, an explanation for the decline is the different distribution of the target variable in the external validation data. The percentage of intubated patients was 58% in the external validation data compared to 48% in the training data. This could be an explanation for the decline in AUC since the model showed a better performance for recall and f1-score on class 0, which is the non-intubated class. It therefore could be the case that the model is better trained on detecting patients that are not intubated and therefore scores worse on a data set in which the majority of patients were intubated.[17]

Thirdly, the size of the data set could explain the lower AUC. The external validation set was smaller than the training data set, with 162 patients in comparison to 348 patients. This could lead to a decline in the model's performance as the data set does not capture the full range of variation in the patient population. Preferably, the external validation set would have the same amount of patients or more.[29]

Fourthly, sampling bias could have played a role. Sampling bias can occur when the patient population in the training data set may not be representative of the patient population in the external validation data set. The external validation data set was sampled from a different population, namely different hospitals. This could lead to a sample being produced that performs well on training data but does not generalize well to new data, in this case the external validation set.[21]

Lastly, overfitting could be a reason for a decline in AUC. Overfitting occurs when the model is too complex relative to the size and variability of the training data set.[30] This can result in a model that is highly tuned to the training data but does not generalize well to new data. The learning curves do not directly show overfitting. However, this reason cannot be excluded.

The average AUC found in the literature review was 0.810 and thus significantly higher than the AUC value reached in this model, namely 0.694.[5] This difference in AUC value can be explained by the same reasons that were described above to explain the decline in AUC when the external validation set was applied. The developed model did outperform the ROX index, which had an AUC of 0.64.[4] As described in the introduction, there are three parameters used to calculate the ROX index, namely *SpO2*, *FiO2* and *RR*. Both *SpO2* and *FiO2* were variables that were often present in the top 10 features. The respiratory rate did not come up in the top 10 features. An explanation for the enhanced performance compared to the ROX index could be the implementation of more significant features.
Going further on the top 10 features that were most predictive, as mentioned, *FiO2* and *SpO2* were widely represented in the top 10 features. Remarkably, *LDH* was also often present in the top 10. LDH stands for lactate dehydrogenase and is an enzyme that appears in all body cells. It is extracted to diagnose cell and tissue damage in patients. LDH can diagnose liver illness, myocardial infarcts, haemolysis, muscle damage, and lung function.[31] The predictive value of LDH can thus be due to its wide variety of describing the patient's status. Moreover, the *DBP mean* and *PO2* were often represented in the top 10. Diastolic blood pressure (DBP) is a measure of the patient's circulation. PO2 stands for the partial pressure of oxygen and is a measure of blood oxygenation. It is understandable that the model uses these two variables in the prediction as they provide insight into both circulation and the amount of oxygen in the blood.

The best-performing joint model had a test AUC performance value of 0.681 (standard deviation 0.07) and was a Random Forest with no feature selection and Power Transformer scaling. The internal validation with a 20% separated test set resulted in an enlarged AUC value of 0.699. The learning curves of the training fold and the final fitting of the best-performing model on all available training data showed different patterns. The last learning curve had a flat line at 1 for the training score, which is an indication of overfitting. It means that the model is perfectly fitting the training data set.[21] The precision, recall, and f1-score showed that overall the model performs best on class 0: the patients that were not intubated.
The increased AUC when performing the internal validation with the 20% left out data set can be explained by the possible overfitting of the model, which is also indicated by the last learning curve. Similar to the aggregated model, it is expected that an external validation data set would lead to a (small) decrease in performance. The overfitting could mean that the model has memorized the training data set, rather than learning the underlying patterns that are present in the data.
The following article by Ying et al. 2019 describes a technique to minimize overfitting in Random Forest models. First of all, it is important to monitor the validation score during training and to stop training when the validation score stops improving or starts to decline. In this case, the algorithm stops improving because it is learning the noise of the data, e.g. outliers of the data.[30] In the best-performing joint model, the second-best, and third-best model, the highest AUC values were all reached in cross-validation fold 3 of the 5 folds in total. It is possible that the overfitting occurred when the model continued training on the 4th and 5th fold. In Random Forest models, early stopping can be implemented by setting a threshold for the number of trees. With the hyperparameter *n_estimators* the number of trees is defined. In this model, a grid search was performed to find the optimal number of trees with the following entries: 100, 200 and 500. The optimally tuned hyperparameters resulted in the number of trees being 200 in the joint model. It could have been that the optimal number of trees lies between 100 and 200. This is a disadvantage of using grid search to tune hyperparameters. It is also an option to tune hyperparameters using a randomized search. With a randomized search, it is more likely to find an optimal value of a parameter, since there is a smaller chance of the optimal value being just between two points. In Appendix A.20 an image is shown in which the difference between grid search and randomized search for tuning hyperparameters is explained.[32] Next to setting this *n_estimators* parameter, an actual early stop argument can also be coded. This argument analyzes the validation scores and automatically stops when the scores are not improving or are declining.
Secondly, it is known that training the model with more data could also lead to a reduction in overfitting because the model then sees more different patients and cannot train perfectly for a subset of patients.[21]
The top 10 predictive features of the best-performing joint model only contains *FiO2* variables and one *SpO2* value. This list is thus quite different from the best-performing aggregated data model. The presence of 9 different FiO2 values (mainly intercept values, and one slope) shows that information on the

FiO2 value for every hour has a predictive value. Nevertheless, having a smaller variance of different features can mean that the model gives a less reliable prediction.

When the results for both of the model types are taken into account, it can be concluded that the addition of repeated measurements does not seem to have a benefit for the performance of the model in predicting intubation. However, ideally, the joint model should be externally validated to make it a fair comparison to the aggregated model. Moreover, a larger external validation set that contains more data on the important features and is balanced likewise to the training data is desired to improve the external validation.

With the test AUCs close to the required 0.7 in the aggregated models it can be concluded that it is possible to predict intubation with data of patients that received HFNO therapy for at least 8 hours.[25] However, with these models, the performance is not sufficient enough to conclude that intubation should be implemented and HFNO therapy will not be sufficient. To predict intubation with this performance value poses an unacceptable risk, especially considering the average recall of this model for class 1 is 0.583. This means that 58.3% of the patients were correctly classified as class 1 and thus as being intubated. The other 41.7% of patients were thus misclassified as non-intubated patients.[26] These patients who would most likely benefit from intubation should not be missed. It is therefore preferred to get a higher recall.

On the other hand, precision should not be ignored. It is important to consider both precision and recall to balance the trade-off between false positives and false negatives. A precision of 0.76 in class 1 means that 76% of the patients were correctly classified as class 1, and thus 24% were misclassified as intubated patients.[26] To make sure that patients do not get intubated if it is not necessary, high precision is desired. As mentioned, a balance between precision and recall is relevant.

To visualize the trade-off between sensitivity and specificity, the ROC is obtained. In Figure 4.2 the ROC of the best-performing aggregated model is shown. In the ROC, it can be seen that fold 2 and fold 3 (of fold 0 to fold 4) result in the best AUC value, namely 0.78. These two ROCs indicated in red and green are also plotted closer to the top left corner, which indicates higher performance. Furthermore, it can be seen that the model results in a higher sensitivity at the beginning with fewer false positives. Eventually, the curve flattens and the sensitivity becomes more equal to the false positive rate. Overall, these findings suggest that the model performed well in discriminating between positive and negative cases, particularly in fold 2 and fold 3.

6.2. Relevance of the results

The model developed in this thesis was developed to aid medical personnel in the treatment option from HFNO therapy to mechanical ventilation and thus intubation. The ROX index with an AUC of 0.64 is not trustworthy enough to determine if escalation from HFNO therapy to intubation should be done.[4] The results of the developed model in this thesis project did not reach the AUC performance of the studies that were found in the literature research.[5] However, the best-performing models with both datatypes did outperform the ROX index and thus show potential.

Ideally, the developed model would aid medical personnel in the treatment choice of HFNO therapy or intubation. Intubation is in certain patients the best treatment option. However, it also comes with risks and complications and thus should not be applied unnecessarily. HFNO therapy is sufficient in certain patients and has benefits for patient comfort compared to intubation. Nevertheless, for more severely ill patients it will not give enough treatment.

Next to the relevance of aiding medical personnel in this difficult treatment choice, it is also relevant to embrace innovations, such as ML models, as a hospital and see opportunities in these new techniques. With this thesis project that was executed in the Maasstad Hospital, the model was developed inside the hospital, which makes the knowledge obtained with the research valuable for the hospital. This is an advantage compared to a situation in which an outside company is hired to develop a model, due to the fact that with this project the medical personnel influenced the development and knowledge was shared with them directly.

6.3. Limitations of the research

There were several limitations to the research performed in this thesis project. Firstly, the completeness of the used data was a limitation. The data that was used to train and validate the model was a limited amount of all data available from ICU patients. For the development of the models as much data as possible was extracted. In the aggregation of the data, certain choices were made that limited the amount of data, such as the features that were taken into account for medication. In the used data set, it was captured whether a patient got antibiotics or steroids during their admission. It was not captured which kind of antibiotics or steroids were given, in which dose, and how many times. Also in the aggregation of vital functions, it was chosen to limit the number of features by only taking into account the mean, standard deviation and last value. Moreover, in the features that described the HFNO therapy, only the amount of inspired oxygen was taken into account, not the flow rate of the HFNO therapy.

It is inevitable that certain data will not be taken into account in the development of the model. This incompleteness of data could, however, mean that the population characteristics are not well defined in the data used to train the model. Nevertheless, with this strategy of aggregating the data with clinical reasoning, the features that are meaningful for physicians in their daily practice are preserved. The possibility of the model finding an irrelevant pattern is hereby reduced. Moreover, the top 10 predictive features over all the models are quite similar, which suggests these features were enough information for the model to make a prediction. However, it is not possible to know if the addition of other data would have resulted in a different performance.

Secondly, the data that was used in the development of the model in this thesis project mainly consisted of COVID-19 patients. Especially in this patient population, medical personnel were experiencing difficult situations in which patients deteriorated fast and the moment of intubation or continuing HFNO therapy was difficult to determine. It is therefore questionable if this developed model is applicable in the current medical setting in which almost no patients get admitted to the ICU because of COVID-19. Brinkman et al. 2022 researched the differences between COVID-19 patients and patients with viral pneumonia in the ICU.[33] They found that mechanical ventilation at ICU admission was more prevalent in the viral pneumonia group and mechanical ventilation in the first 24 hours of ICU admission was distributed comparably among the COVID-19 and viral pneumonia groups. The developed model may therefore apply to viral pneumonia patients that are admitted to the ICU.

Thirdly, a limitation of the used data is that it was retrospective. Thus, the decision to intubate a patient was made by the physician and it is not known if this was the best treatment option. For instance, it could be the case that a patient was intubated while this was not the best option and HFNO therapy would have been sufficient. The ML model then interpreted this patient as rightfully being intubated. This makes it more difficult to find the right treatment option based on this data.

Lastly, a limitation of the developed model is that it was only trained with the Maasstad Hospital data and could therefore be difficult to implement in other hospitals. As seen with the implementation of the external validation data, the AUC declined. As described above, this could be due to the validation data containing significantly different patients. Afterwards, it would have been better to fit the model on the Santeon data set that contained different hospitals. Then, the Maasstad data could be used to externally validate the model. However, because of the lengthy process of receiving the Santeon data, this method could not be implemented.

6.4. Future recommendations

In this thesis project, the implementation of the developed model in daily practice for ICU personnel has not been taken into account, due to the premature state of the developed model. However, for future versions of the model, the implementation should be taken into account. In the following section, different aspects of the implementation of the model are discussed.

First of all, before the design of a model, it should be carefully considered which model type is preferred. The article by Sidney-Gibson et al. 2019 describes the trade-off between complexity and interpretability in machine learning models.[34] On the one hand, a more complex model works better for complex data, but is hardly interpretable due to it being a black box. On the other hand, 'auditable algorithms' cannot handle complex data, but are better interpretable.

In this thesis project it was chosen to develop a Logistic Regression Model, Random Forest and Gradient Boosting Model. These model types are all closer to being 'auditable algorithms' than black boxes. However, it is still difficult to visualize how the model makes its predictions. For example, a classic 'auditable algorithm' is a decision tree in which it is seen which features are used for the prediction and even which values of these features are taken as a classification point. In this thesis project, the best-performing model was a Random Forest that consisted of 200 decision trees. It is imaginable that it is difficult to visualize how classifications are formed in the forest of all these trees.

Moreover, it is important to involve all the stakeholders in the development phases of the model. In this thesis project, the medical supervisor was involved in the development of the model. With his reasoning, the aggregation of the data was performed. In the implementation trajectory, more stakeholders should be involved, as the medical personnel will eventually need to use the model and trust its predictions. It is advised to involve medical personnel from all layers in the development process: from physicians to nurses.

During the medical activities in the ICU department of Maasstad Hospital, it was researched what kind of questions arise from the medical personnel when talking about implementing an ML model that predicts intubation. Interesting subjects and questions were brought up by the medical personnel. For instance: ”Who is responsible for the well-being of a patient?” If it occurs that a patient dies because the model advised to stay on HFNO therapy but intubation was necessary, who is responsible? Or the other way around: the model advises intubation but the medical personnel thinks this is not necessary. Later, it turns out that intubation should have been performed. Did the medical personnel do wrong by not listening to the model? In these cases, it is important to consider that the model would not replace the medical personnel. Instead, it would be an addition to the team. Eventually, the physician will always have the final decision and with that the responsibility. It is however interesting how this dynamic between humans and machines may change in the future.

This possible shift in dynamic comes with worries. For instance, the nurses were worried that their clinical view of the patient would be replaced by the model. Moreover, it is difficult to know who is correct because, as of now, there is no golden standard to which the model can measure its performance. The model was trained on retrospective data in which medical personnel decided to intubate the patient or not. There was no information on this treatment choice being the best choice with the best outcome for the patient. However, as this is the current situation and the aim is to implement the model in the current situation, it was not wrong to take the decision made by the physician as the golden standard. All of the implementation factors mentioned above should be taken into account when implementing the developed model.

For a second future recommendation, certain parameters to monitor the patient outcome could be taken into account. For instance, mortality, the rate of comorbidities, and quality of life for the patients that survived their ICU admission could be additional parameters. When these parameters are collected, it is also possible to perform a clinical performance of the model. For instance, does the implementation of the model lower mortality in the ICU because patients that need to be intubated are discovered earlier in their admission? Furthermore, are these patients that are at risk for intubation found more often? This clinical performance can be found when the model has been implemented for a certain period, so it would be an a posteriori impact measurement.

Thirdly, it is recommended to train and develop a model without COVID-19 data to capture the characteristics of the current population in the ICU. With this recommendation, it can be researched if the developed model can be applied to patients with viral pneumonia admitted to the ICU. Moreover, it is advised to implement data from multiple hospitals in this new data set to capture the characteristics of the current ICU population.

Fourthly, a recommendation would be to research if measurements that are not standardly measured in the ICU could have a beneficial effect on the performance of the model. During the clinical activities in the thesis project, it became clear that many techniques are used in the Maasstad Hospital to determine the status of (ventilated) patients in the ICU. For instance, Electro Impedance Tomography (EIT) can be used to determine the distribution of ventilation and perfusion in the lungs.[35] This non-invasive technique could be applied to get more data on the ventilation and perfusion status of patients on HFNO therapy. Moreover, lung ultrasounds could be implemented to get more information on the status of the lungs and possible fluid or infectious areas in the lungs.[36, 37] It would be very interesting to take these kinds of measurements into account in the data used to develop the ML model.

A last recommendation is to advocate for a universally applicable and available database for all hospitals. During this thesis project, it became clear that it was almost impossible to request data from an existing database of ICU data of the Santeon hospitals. Moreover, when the data was obtained, another difficulty was the different structure of the data between the different Santeon hospitals. The cooperation between Santeon hospitals is a unique opportunity to generate a database that contains enough patients to develop and validate ML models. In a future with more AI applications, data is very valuable and should thus be safely shared to get as much out of it as possible.

Bibliography

[1] S. M. Raboni, V. C. Neves, R. M. Silva, G. L. Breda, A. C. Ceregato, T. P. Broza, G. de Oliveira, L. L. Melo-Diaz, C. B. Braga, C. F. Carraro, N. C. Arroyo, R. F. Bardy, G. F. Devetak, C. M. Ozawa, M. E. Graf, V. L. Dias, M. A. Ducroquet, D. P. Nunes, C. S. Sokoloski, and R. R. Petterle, "High-Flow Nasal Cannula Therapy in Patients With COVID-19: Predictive Response Factors," *Respiratory care*, vol. 67, pp. 1443–1451, 11 2022.

[2] O. Roca, J. Messika, B. Caralt, M. García-de Acilu, B. Sztrymf, J. D. Ricard, and J. R. Masclans, "Predicting success of high-flow nasal cannula in pneumonia patients with hypoxemic respiratory failure: The utility of the ROX index," *Journal of Critical Care*, vol. 35, pp. 200–205, 10 2016.

[3] O. Roca, B. Caralt, J. Messika, M. Samper, B. Sztrymf, G. Hernández, M. García-De-Acilu, J. P. Frat, J. R. Masclans, and J. D. Ricard, "An index combining respiratory rate and oxygenation to predict outcome of nasal high-flow therapy," *American Journal of Respiratory and Critical Care Medicine*, vol. 199, no. 11, pp. 1368–1376, 2019.

[4] V. Arvind, J. S. Kim, B. H. Cho, E. Geng, and S. K. Cho, "Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19," *Journal of Critical Care*, vol. 62, pp. 25–30, 4 2021.

[5] Hendriks Manon, "Predicting intubation in ICU patients using artificial intelligence: a systematic review with meta-analysis," pp. 1–19, 8 2022.

[6] N. R. Macintyre, "Ventilator-Associated Pneumonia: The Role of Ventilator Management Strategies Introduction Reducing Ventilator-Induced Lung Injury Lung-Protective Mechanical Ventilatory Strategies Require Tradeoffs Hypercapnic Respiratory Acidosis Sedation Atelectasis Weaning Delays Summary," tech. rep., 2005.

[7] "Intubation: What is it, types, procedure, side effects, and pictures."

[8] "Complications of the endotracheal tube following initial placement: Prevention and management in adult intensive care unit patients - UpToDate."

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY: Springer New York, 2009.

[10] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, pp. 367–378, 2 2002.

[11] K. Potdar, T. S., and C. D., "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *International Journal of Computer Applications*, vol. 175, pp. 7–9, 10 2017.

[12] B. C. Jaeger, N. J. Tierney, and N. R. Simon, "When to Impute? Imputation before and during cross-validation," tech. rep.

[13] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: A guide for non-statisticians," *International Journal of Endocrinology and Metabolism*, vol. 10, no. 2, pp. 486–489, 2012.

[14] F. Pedregosa, V. Michel, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, and f. Duchesnay, "Scikit-learn: Machine Learning in Python," Tech. Rep. 85, 2011.

[15] M. Ahsan, M. Mahmud, P. Saha, K. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, p. 52, 7 2021.

[16] “9 Feature Transformation & Scaling Techniques| Boost Model Performance.”

[17] F. Alahmari, “A Comparison of Resampling Techniques for Medical Data Using Machine Learning,” *Journal of Information and Knowledge Management*, vol. 19, 3 2020.

[18] R. Muthukrishnan and R. Rohini, “LASSO: A feature selection technique in predictive modeling for machine learning,” in *2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016*, pp. 18–20, Institute of Electrical and Electronics Engineers Inc., 3 2017.

[19] S. B. Kotsiantis, “Feature selection for machine learning classification problems: a recent overview,” *Artif Intell Rev*.

[20] “DataTechNotes: SelectKBest Feature Selection Example in Python.”

[21] R. H. Riffenburg, *Statistics in Medicine, Third Edition*. Elsevier, 1 2012.

[22] “Home — Spyder IDE.”

[23] D. Bates, M. Mächler, E. Zurich, B. M. Bolker, and S. C. Walker, “Fitting Linear Mixed-Effects Models Using lme4,”

[24] D. Rizopoulos, “Introduction to the Joint Modeling of Longitudinal and Survival Data, with Applications in R,” tech. rep.

[25] J. N. Mandrekar, “Receiver operating characteristic curve in diagnostic test assessment,” *Journal of Thoracic Oncology*, vol. 5, pp. 1315–1316, 9 2010.

[26] D. M. W. Powers and Ailab, “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION,” tech. rep.

[27] K. Hajian-Tilaki, “Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation,” 2013.

[28] L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms: A critical evaluation,” *BMC Medical Informatics and Decision Making*, vol. 16, p. 74, 7 2016.

[29] R. D. Riley, T. P. A. Debray, G. S. Collins, L. Archer, J. Ensor, M. Smeden, and K. I. E. Snell, “Minimum sample size for external validation of a clinical prediction model with a binary outcome,” *Statistics in Medicine*, vol. 40, pp. 4230–4251, 8 2021.

[30] X. Ying, “An Overview of Overfitting and its Solutions,” in *Journal of Physics: Conference Series*, vol. 1168, Institute of Physics Publishing, 3 2019.

[31] “LD iso-enzymen | NVKC.”

[32] K. E. S. Pilario, Y. Cao, and M. Shafiee, “A Kernel Design Approach to Improve Kernel Subspace Identification,” *IEEE Transactions on Industrial Electronics*, vol. 68, pp. 6171–6180, 7 2021.

[33] S. Brinkman, F. Termorshuizen, D. A. Dongelmans, F. Bakhshi-Raiez, M. S. Arbous, D. W. de Lange, N. F. de Keizer, and et al., “Comparison of outcome and characteristics between 6343 COVID-19 patients and 2256 other community-acquired viral pneumonia patients admitted to Dutch ICUs,” *Journal of Critical Care*, vol. 68, pp. 76–82, 4 2022.

[34] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction,” *BMC Medical Research Methodology*, vol. 19, p. 64, 3 2019.

[35] S. Mansouri, Y. Alharbi, F. Haddad, S. Chabcoub, A. Alshrouf, and A. A. Abd-Elghany, “Electrical Impedance tomography – Recent applications and developments,” 2021.

[36] M. E. Haaksma, J. M. Smit, M. L. Heldeweg, J. S. Nooitgedacht, H. J. De Grooth, A. H. Jonkman, A. R. Girbes, L. Heunks, and P. R. Tuinman, “Extended Lung Ultrasound to Differentiate between Pneumonia and Atelectasis in Critically Ill Patients: A Diagnostic Accuracy Study,” *Critical Care Medicine*, vol. 50, pp. 750–759, 5 2022.

[37] D. A. Lichtenstein, “Lung ultrasound in the critically ill,” 1 2014.

[38] “Medical Dictionary.”

A
Supplementary figures and tables

A.1. Data extraction overview

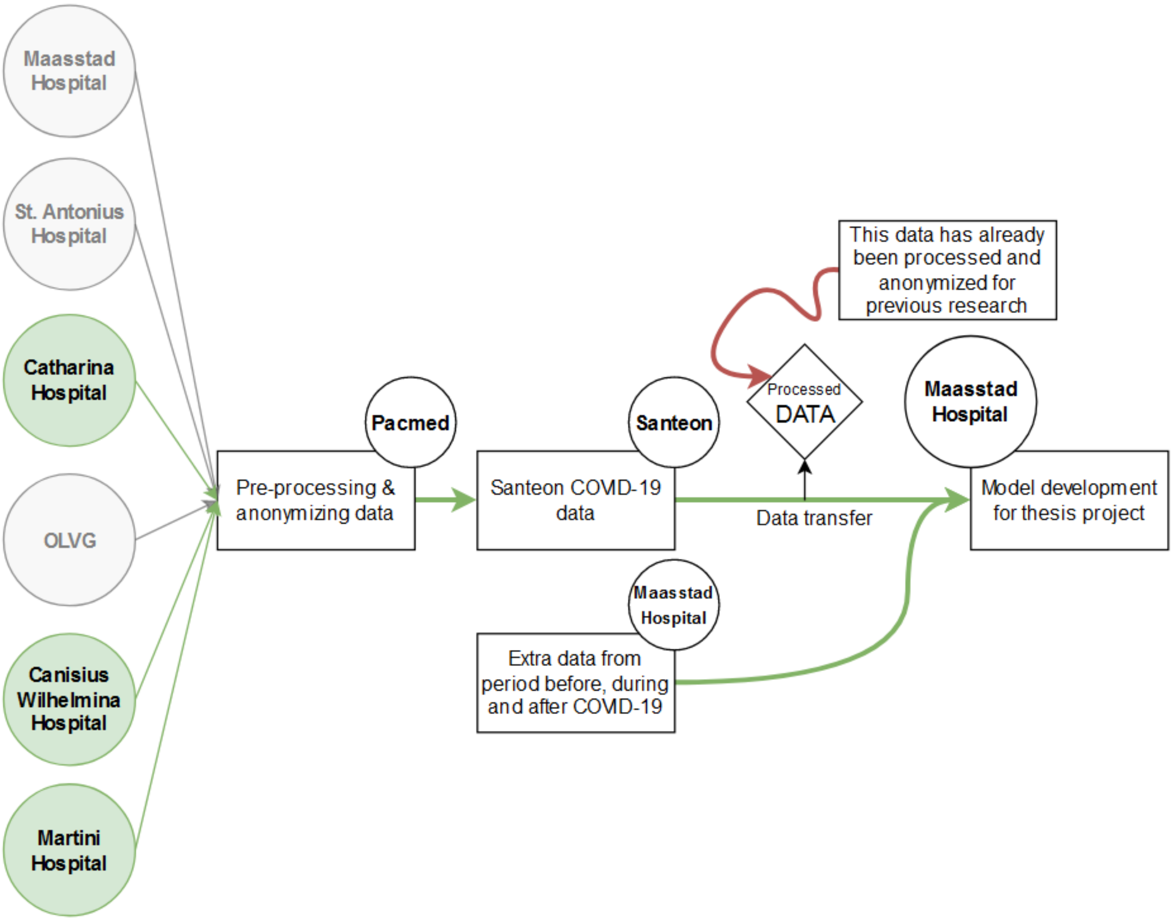
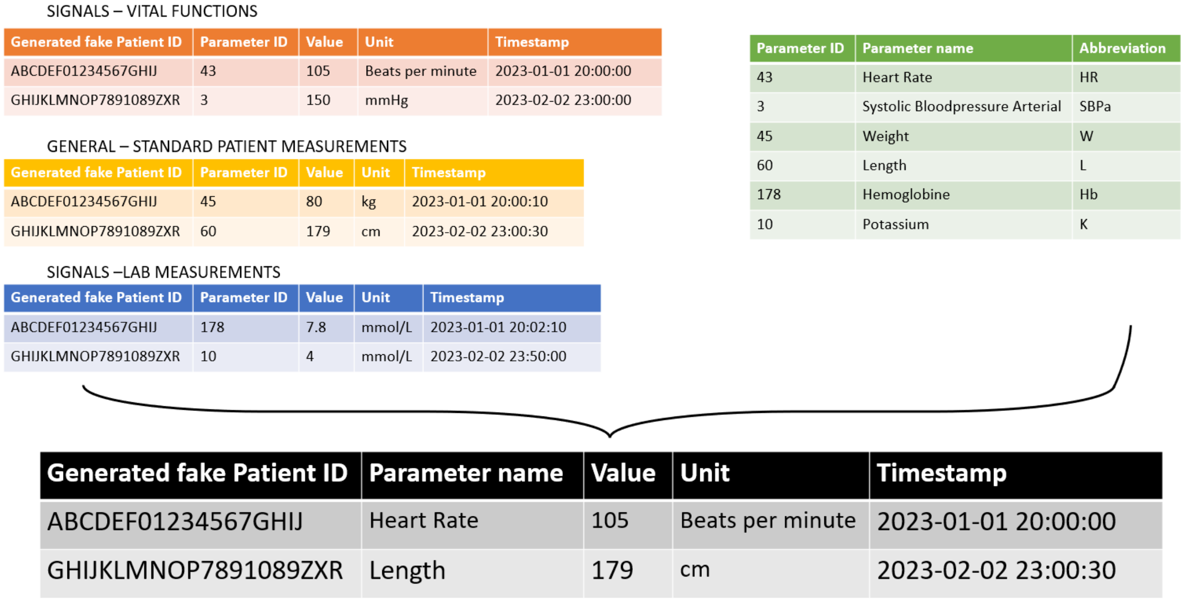

Maasstad Hospital
St. Antonius Hospital
Catharina Hospital
OLVG
Canisius Wilhelmina Hospital
Martini Hospital
Pacmed
Pre-processing & anonymizing data
Santeon
Santeon COVID-19 data
This data has already been processed and anonymized for previous research
Processed DATA
Data transfer
Maasstad Hospital
Model development for thesis project
Maasstad Hospital
Extra data from period before, during and after COVID-19

Figure A.1: Infographic of the data extraction process for all the data, including the Santeon data set and data from the Maasstad Hospital. From the Santeon hospitals, the data of the hospitals that are marked green were used in the external validation data.

A.2. Data extraction by BI



SIGNALS – VITAL FUNCTIONS

Generated fake Patient ID	Parameter ID	Value	Unit	Timestamp
ABCDEF01234567GHIJ	43	105	Beats per minute	2023-01-01 20:00:00
GHIJKLMNOP7891089ZXR	3	150	mmHg	2023-02-02 23:00:00

GENERAL – STANDARD PATIENT MEASUREMENTS

Generated fake Patient ID	Parameter ID	Value	Unit	Timestamp
ABCDEF01234567GHIJ	45	80	kg	2023-01-01 20:00:10
GHIJKLMNOP7891089ZXR	60	179	cm	2023-02-02 23:00:30

SIGNALS –LAB MEASUREMENTS

Generated fake Patient ID	Parameter ID	Value	Unit	Timestamp
ABCDEF01234567GHIJ	178	7.8	mmol/L	2023-01-01 20:02:10
GHIJKLMNOP7891089ZXR	10	4	mmol/L	2023-02-02 23:50:00

Parameter ID	Parameter name	Abbreviation
43	Heart Rate	HR
3	Systolic Bloodpressure Arterial	SBPa
45	Weight	W
60	Length	L
178	Hemoglobine	Hb
10	Potassium	K

Generated fake Patient ID	Parameter name	Value	Unit	Timestamp
ABCDEF01234567GHIJ	Heart Rate	105	Beats per minute	2023-01-01 20:00:00
GHIJKLMNOP7891089ZXR	Length	179	cm	2023-02-02 23:00:30

Figure A.2: This figure shows which tables were used in the data extraction process to generate the data set. In this figure, fake data is used to give an insight into using the different tables.

A.3. Aggregated data example data set

Table A.1: Example of the aggregated data set with fictional patient data. For an explanation of the different features, see Appendix A.10 for the list of features that were used.

# row	Age	Length	...	FiO2 mean	FiO2 std dev	FiO2 last value	...	CRP	ALAT	AF	Intubated
0	64	180	...	68.3	2.83	60	...	18	20	100	0
1	70	175		83	1.75	60		24	30	95	1
...											
349	78	165	...	45	4.56	80	...	3	14	80	1
350	59	178		75	3.23	40		23	10	110	0

A.4. FiO2 example patients

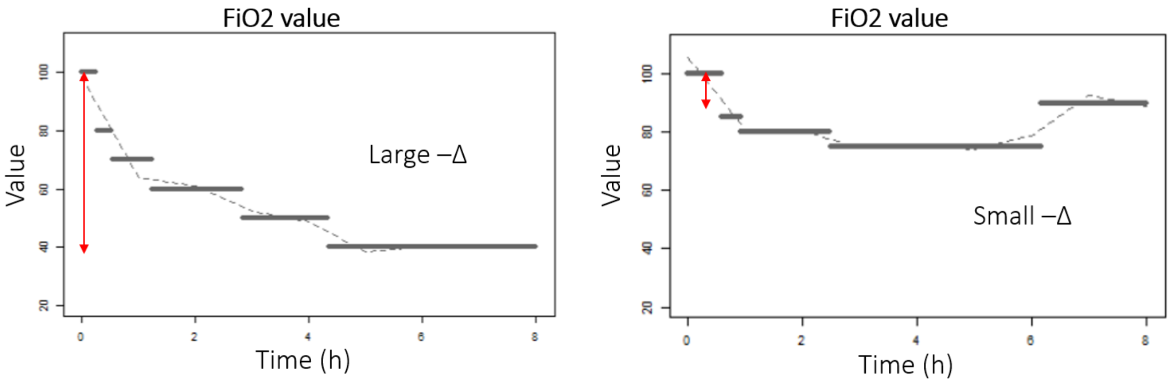


Figure A.3: Two example patients and their course of FiO2 values showing different deltas.

A.5. Repeated measurements example data set

Table A.2: Example of the repeated measurements data set, filled with fictional data. Appendix A.10 and A.11 can be consulted for an explanation of the used features.

Patient ID	Parameter ID	Parameter name	Value	Calender time	Tijd
1	4	HR	109	2021-09-08 14:50:00	0.000
1	4	HR	112	2021-09-08 14:51:00	0.01666667
1	4	HR	110	2021-09-08 14:52:00	0.03333333
1	8	FiO2	40	2021-09-08 14:50:00	0.000
...	...	...	...	...	...
2	10	SBP Mean	100	2020-10-10 22:06:00	0.000

A.6. Graphical overview of the nested cross-validation for the aggregated models

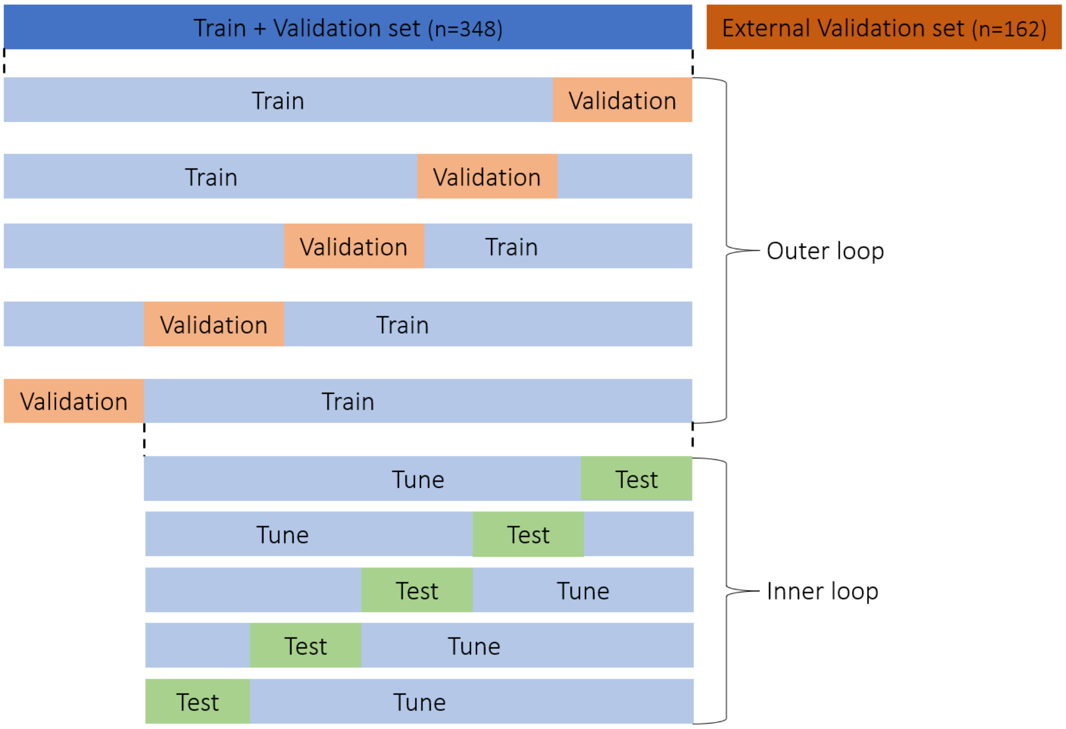


Figure A.4: Graphical overview of the nested cross-validation for the aggregated models with *k=5* for the outer cross-validation and *k=5* for the inner cross-validation.

A.7. Graphical overview of the nested cross-validation for the joint models

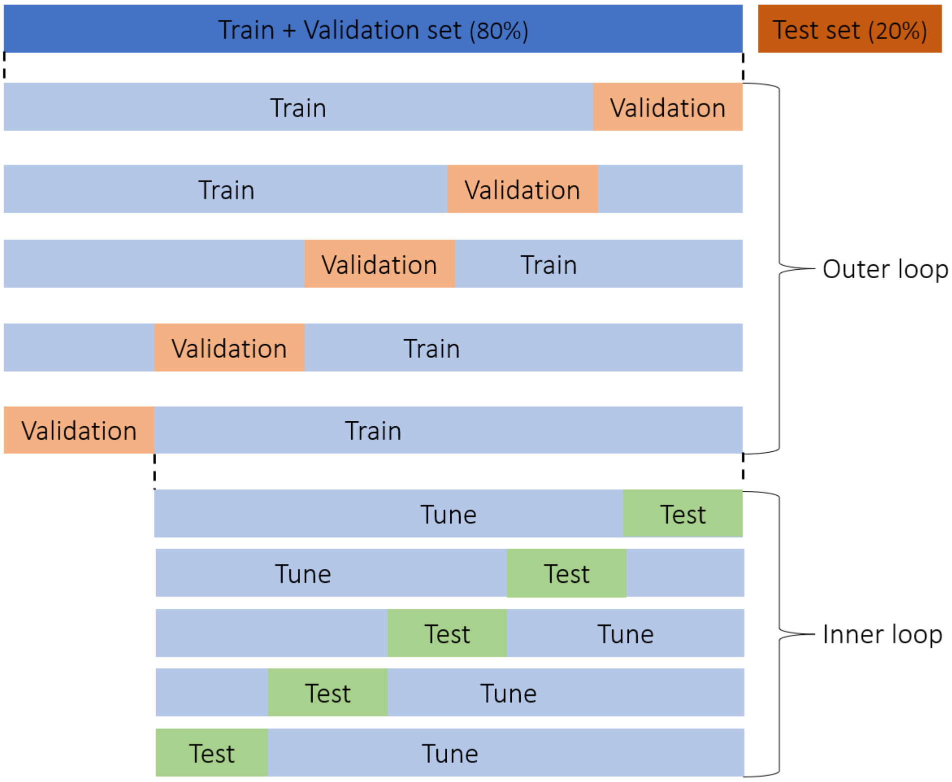


Figure A.5: Graphical overview of the nested cross-validation for the joint models with *k=5* for the outer cross-validation and *k=5* for the inner cross-validation.

A.8. Example LMEM showing 3 variables

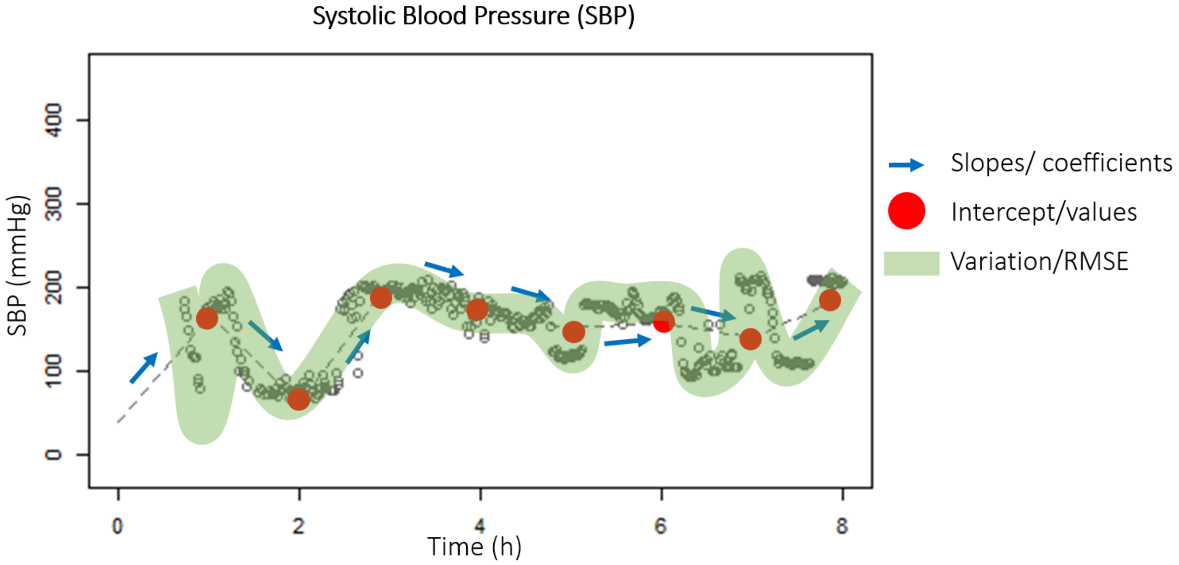


Figure A.6: Example of an LMEM of a Systolic Blood Pressure course for the three variables: slopes, intercepts, and RMSE.

A.9. Precision, recall, f1-score, and related formulas

In this section, the formulas of precision, recall, and f1-score are given.[26, 27]

$$Precision, PPV = \frac{TP}{TP + FP}$$

$$Recall, Sensitivity, TPR = \frac{TP}{TP + FN}$$

$$Specificity, TNR = \frac{TN}{TN + FP}$$

$$f1 - score = \frac{2 * precision * recall}{precision + recall}$$

$$FNR, 1 - Sensitivity = \frac{FN}{TP + FN}$$

$$FPR, 1 - Specificity = \frac{FP}{TN + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

For these formulas, TP means true positive, TN means true negative, FP means false positive, and FN means false negative. TPR means true positive rate, TNR means true negative rate, FNR means false negative rate, FPR means false positive rate, PPV means positive predicted value, and NPV means negative predicted value.

A.10. Feature list used in aggregated data model

A.10.1. General patient information

10 features contain general patient information. Some of these features contained categorical data and were therefore encoded. What the newly generate codes mean is explained per feature.

- **Age**: Age of the patient as entered in the electronic patient file, given in years.
- **Length**: Length of the patient as entered in the electronic patient file, given in centimetres.
- **Weight**: Weight of the patient as entered in the electronic patient file, given in kilograms.
- **Origin_1 - Origin_4**: The location from where the patient was admitted to the ICU, from 1 to 4 meaning admission from their home, their hospital, a different hospital and the emergency room (ER).
- **Gender**: Gender of the patient encoded to males stated with a 0 and females with a 1.
- **Admission during day or night**: Patients with night admissions (between 18:00 and 07:00) were given a 1 and day admissions were encoded with a 0.
- **Intubation**: The target label of the models. Patients who had an intubation duration that was higher than zero were given the intubation label (1). Moreover, patients who died during their admission and were not intubated were given the label.

A.10.2. Vital function parameters

21 features contain vital function parameters. Below is explained what the features measure.[38]

- **FiO2 mean**: Fraction of inspired oxygen. This feature represents the average of the FiO2 for 8 hours of HFNO therapy.
- **FiO2 std dev**: Fraction of inspired oxygen, which is the standard deviation of the average.
- **FiO2 last value**: Fraction of inspired oxygen, which is the last entered value.
- **FiO2 delta**: Fraction of inspired oxygen. The delta is the difference between the first and the last entered value.
- **SpO2 mean**: Saturation of arterial blood with oxygen as measured by pulse oximetry and therefore called peripheral oxygen saturation. This feature represents the average saturation.
- **SpO2 std dev**: Saturation, which is the standard deviation of the average.
- **SpO2 last value**: Saturation, which is the last entered value.
- **RR mean**: Respiratory rate is the frequency of breathing, recorded in the number of breaths per minute. This feature represents the average respiratory rate.
- **RR std dev**: Respiratory rate, which is the standard deviation of the average.
- **RR last value**: Respiratory rate, which is the last entered value.
- **SBP mean**: Systolic blood pressure. Blood pressure during contraction of the ventricles is measured in mmHg. This feature represents the average systolic blood pressure.
- **SBP std dev**: Systolic blood pressure, which is the standard deviation of the average.
- **SBP last value** Systolic blood pressure, which is the last entered value.
- **DBP mean**: Diastolic blood pressure. Blood pressure during relaxation of the ventricles is measured in mmHg. This feature represents the average diastolic blood pressure.
- **DBP std dev**: Diastolic blood pressure, which is the standard deviation of the average.
- **DBP last value**: Diastolic blood pressure, which is the last entered value.
- **HR mean**: Heart rate. The number of contractions of the heart is measured in beats per minute. This feature represents the average heart rate.
- **HR std dev**: Heart rate, which is the standard deviation of the average.
- **HR last value**: Heart rate, which is the last entered value.
- **Temperature mean**: Temperature measured axillary (in the armpit) given in Fahrenheit.
- **Temperature last value**: Temperature measured axillary, which is the last entered value.

A.10.3. Blood gas analysis

In the blood gas analysis, 16 features are measured. The different features are explained below.[38]

- **SaO2 mean**: Saturation of arterial blood with oxygen measured with Co-oximeter and therefore called arterial oxygen saturation. This feature represents the average arterial saturation.
- **SaO2 std dev**: Arterial saturation, which is the standard deviation of the average.
- **SaO2 last value**: Arterial saturation, which is the last entered value.

- **HCO3- mean**: Bicarbonate. An important parameter of the acid-base balance that is measured in blood gas analysis. This feature represents the average bicarbonate.
- **HCO3- std dev**L Bicarbonate, which is the standard deviation of the average.
- **HCO3- last value**: Bicarbonate, which is the last entered value
- **PCO2 mean**: Partial pressure of carbon dioxide. This feature represents the average partial pressure of carbon dioxide.
- **PCO2 std dev**: Partial pressure of carbon dioxide, which is the standard deviation of the average.
- **PCO2 last value**: Partial pressure of carbon dioxide, which is the last entered value.
- **PO2 mean**: Partial pressure of oxygen. This feature represents the average partial pressure of oxygen.
- **PO2 std dev**: Partial pressure of oxygen, which is the standard deviation of the average.
- **PO2 last value**: Partial pressure of oxygen, which is the last entered value.
- **pH mean**: Measure of acidity of the patient. This feature represents the average pH.
- **pH std dev**: pH, which is the standard deviation of the average.
- **pH last value**: pH, which is the last entered value.
- **Haemoglobin blood gas**: Protein in red blood cells that transports oxygen. This feature represents the haemoglobin in a blood gas analysis.

A.10.4. Laboratory parameters

Within the laboratory parameters, 24 features are measured. Below is explained what the features mean and what is measured.[38]

- **CRP**: C-reactive protein, which is a marker for inflammation, infection, or after injury.
- **ALAT**: Alanine aminotransferase, which is a liver enzyme.
- **ASAT**: Aspartate aminotransferase, which is a liver enzyme.
- **GGT**: Gamma glutamyltranspeptidase, which is a liver enzyme.
- **AF**: Alkaline phosphatase, which is a liver enzyme.
- **Bilirubin**: Bilirubin is waste material that is released when breaking down red blood cells. It is used as a marker for liver function.
- **LDH**: Lactacte dehydrogenase, which is an enzyme that appears in all body cells but is especially abundant in kidney, skeletal muscle, liver, and myocardium.
- **K**: Potassium, which is one of the important electrolytes. Among others, it is responsible for normal heart rhythm, fluid balance, and conduction of nerve impulses.
- **Cl**: Chloride, which is also an electrolyte responsible for fluid balance.
- **Plasma albumin**: Albumin is a protein in the blood that is made in the liver. This feature represents the albumin in blood plasma.
- **Mg**: Magnesium, which is a mineral important for skeletal muscles and formation of bones.
- **P**: Phosphate, which is a mineral important for bone mineralisation, energy storing, and cellular processes.
- **Creatinin**: Creatinin is a waste product of breaking down muscle cells. It is an indication of kidney function.
- **CK**: Creatine Kinase is an enzyme that appears in skeletal muscles, heart, and brain. It is a marker of muscle damage.
- **Urea**: Urea occurs in urine and other body fluids and is a product of protein metabolism.
- **Blood glucose**: Blood glucose is the level of glucose in the blood.
- **Lactate**: Lactate acid is a substance made by muscle tissue and red blood cells and may be an indication of a lack of oxygen.
- **Serum albumin**: Albumin is a protein in the blood that is made in the liver. This feature represents the albumin in blood serum.
- **Haemoglobin**: Haemoglobin is a protein in red blood cells that transports oxygen.
- **MCV**: Mean corpuscular volume is a blood test that measures the average size of the red blood cells.

- **Thrombocytes**: Platelets are disk-shaped structures in the blood that promote clotting.
- **APTT**: Activated partial thromboplastin time is a test in which the clotting of blood is timed.
- **Leukocytes**: White blood cells. The number of leukocytes gives information on inflammation or infection.
- **CK-MB** Creatine kinase cytosol isoenzym. Likewise to CK, it is a marker of muscle damage.

A.10.5. Medication parameters

There were two features taken into account that considered medication parameters. The two features are listed below.

- **Antibiotics**: Patients who received antibiotic treatment during the 8 hours of HFNO treatment were encoded with a 1 in this feature
- **Steroids**: Patients who received a steroid during the 8 hours of HFNO treatment were encoded with a 1 in this feature.

A.10.6. Cannula parameters

There were two features taken into account that considered cannula parameters. The two features are listed below.

- **Urine catheter**: Patients who had a urine catheter during the 8 hours of HFNO treatment were encoded with a 1 in this feature.
- **Gastric cannula**: Patients who had a gastric cannula during the 8 hours of HFNO treatment were encoded with a 1 in this feature.

A.11. Feature list used in joint data model

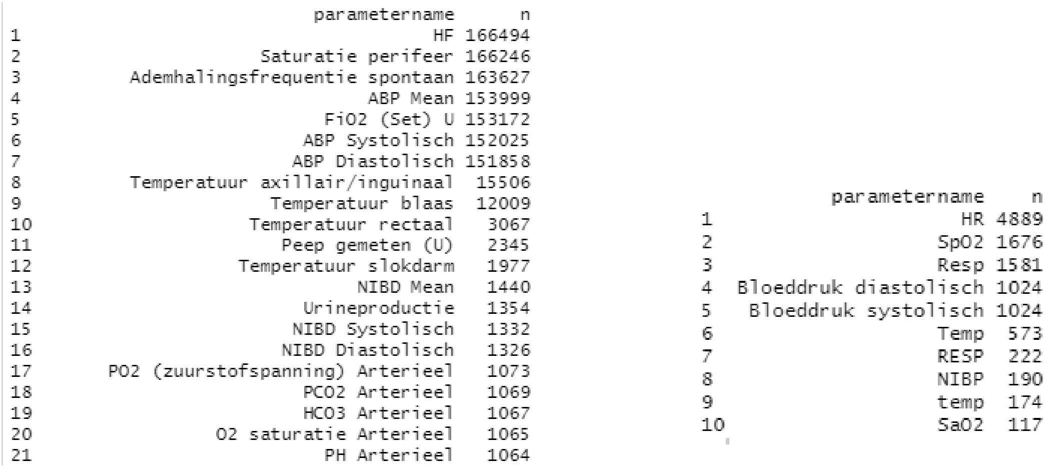
A.11.1. Vital functions parameters

The following vital functions parameters were used in the development of the joint model. These parameters replaced the vital functions and blood gas analysis parameters of the aggregated data. In total, 196 parameters were added from the LMEM.

- **DBP_int0 - DBP_int7**: Diastolic blood pressure (DBP) value at every intercept. There are 7 knots and the 8th intercept is the endpoint.
- **DBP_slope0 - DBP_slope7**: Slope of the DBP between every knot, from the starting point to the first knot, and the last knot to the endpoint.
- **DBP_rmse0 - DBP_rmse7**: RMSE of the DBP, which is the variation of data points around every slope.
- **FiO2_int0 - FiO2_int7**: Fraction of inspired Oxygen (FiO2) value at every intercept. There are 7 knots and the 8th intercept is the endpoint.
- **FiO2_slope0 - FiO2_slope7**: Slope of the FiO2 between every knot, from the starting point to the first knot, and the last knot to the endpoint.
- **FiO2_rmse0 - FiO2_rmse7**: RMSE of the FiO2, which is the variation of data points around every slope.
- **HCO3art_int0 & HCO3art_int4**: Bicarbonate value at the intercept and endpoint. There is one knot at hour 4.
- **HCO3art_slope0 & HCO3art_slope4**: Slope of the HCO3- before the knot and after the knot.
- **HCO3art_rmse0 & HCO3art_rmse4**: RMSE of the HCO3-, which is the variation of data points around every slope.
- **HF_int0 - HF_int7**: Hart frequency (HF) value at every intercept. There are 7 knots and the 8th intercept is the endpoint.
- **HF_slope0 - HF_slope7**: Slope of the HF between every knot, from the starting point to the first knot, and the last knot to the endpoint.
- **HF_rmse0 - HF_rmse7**: RMSE of the HF, which is the variation of data points around every slope.

- **SaO2_int0 & SaO2_int4**: Arterial saturation value at the intercept and endpoint. There is one knot at hour 4.
- **SaO2_slope0 & SaO2_slope4**: Slope of the SaO2 before the knot and after the knot.
- **SaO2_rmse0 & SaO2_rmse4**: RMSE of the SaO2, which is the variation of data points around every slope.
- **PCO2_int0 & PCO2_int4**: Partial pressure of carbon dioxide value at the intercept and endpoint. There is one knot at hour 4.
- **PCO2_slope0 & PCO2_slope4**: Slope of the PCO2 before the knot and after the knot.
- **PCO2_rmse0 & PCO2_rmse4**: RMSE of the PCO2, which is the variation of data points around every slope.
- **pH_int0 & pH_int4**: pH (a measure of acidity) value at the intercept and endpoint. There is one knot at hour 4.
- **pH_slope0 & pH_slope4**: Slope of the pH before the knot and after the knot.
- **pH_rmse0 & pH_rmse4**: RMSE of the pH, which is the variation of data points around every slope.
- **PO2_int0 & PO2_int4**: Partial pressure of oxygen value at the intercept and endpoint. There is one knot at hour 4.
- **PO2_slope0 & PO2_slope4**: Slope of the PO2 before the knot and after the knot.
- **PO2_rmse0 & PO2_rmse4**: RMSE of the PO2, which is the variation of data points around every slope.
- **SpO2_int0 - SpO2_int7**: Peripheral saturation (SpO2) value at every intercept. There are 7 knots and the 8th intercept is the endpoint.
- **SpO2_slope0 - SpO2_slope7**: Slope of the SpO2 between every knot, from the starting point to the first knot, and the last knot to the endpoint.
- **SpO2_rmse0- SpO2_rmse7**: RMSE of the SpO2, which is the variation of data points around every slope.
- **RR_int0 - RR_int7**: Respiratory rate (RR) value at every intercept. There are 7 knots and the 8th intercept is the endpoint.
- **RR_slope0 - RR_slope7**: Slope of the RR between every knot, from the starting point to the first knot, and the last knot to the endpoint.
- **RR_rmse0 - RR_rmse7**: RMSE of the RR, which is the variation of data points around every slope.
- **SBP_int0 - SBP_int7**: Systolic blood pressure (SBP) value at every intercept. There are 7 knots and the 8th intercept is the endpoint.
- **SBP_slope0 - SBP_slope7**: Slope of the SBP between every knot, from the starting point to the first knot, and the last knot to the endpoint.
- **SBP_rmse0 - SBP_rmse7**: RMSE of the SBP, which is the variation of data points around every slope.
- **Temperature_int0 - Temperature_int7**: Temperature value at every intercept. There are 7 knots and the 8th intercept is the endpoint.
- **Temperature_slope0 - Temperature_slope7**: Slope of the temperature between every knot, from the starting point to the first knot, and the last knot to the endpoint.
- **Urine production_int0 & Urine production_int4**: Urine production value at the intercept and endpoint. There is one knot at hour 4.
- **Urine production_slope0 & Urine production_slope4**: Slope of the urine production before the knot and after the knot.
- **Urine production_rmse0 & Urine production_rmse4**: RMSE of the urine production, which is the variation of data points around every slope.

A.12. Number of repeated measurements training set versus validation set



```
                          parametername      n
1                                    HF 166494
2                    Saturatie perifeer 166246
3      Ademhalingsfrequentie spontaan 163627
4                              ABP Mean 153999
5                          FiO2 (Set) U 153172
6                        ABP Systolisch 152025
7                       ABP Diastolisch 151858
8         Temperatuur axillair/inguinaal  15506
9                      Temperatuur blaas  12009
10                   Temperatuur rectaal   3067
11                      Peep gemeten (U)   2345
12                   Temperatuur slokdarm   1977
13                             NIBD Mean   1440
14                          Urineproductie   1354
15                        NIBD Systolisch   1332
16                       NIBD Diastolisch   1326
17        PO2 (zuurstofspanning) Arterieel   1073
18                         PCO2 Arterieel   1069
19                         HCO3 Arterieel   1067
20                   O2 saturatie Arterieel   1065
21                           PH Arterieel   1064
```

```
           parametername    n
1                     HR 4889
2                   SpO2 1676
3                   Resp 1581
4   Bloeddruk diastolisch 1024
5    Bloeddruk systolisch 1024
6                   Temp  573
7                   RESP  222
8                   NIBP  190
9                   temp  174
10                  SaO2  117
```

Figure A.7: List of repeated measurements in the training data set and validation data set

A.13. Learning curves of the best-performing aggregated data set models

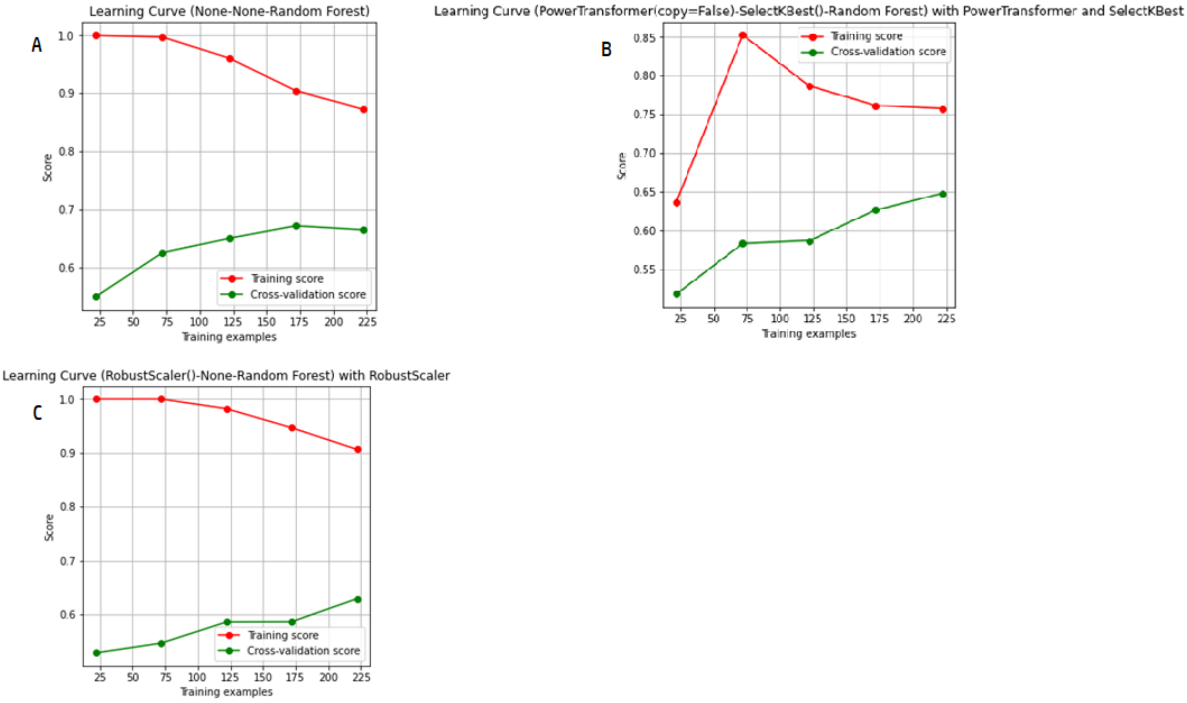


Figure A.8: Learning curves of the best-performing aggregated data set models. Graph **A** is the learning curve of a Random Forest (RF) with no feature selection and no scaling. Graph **B** is the second best-performing model, which is an RF with SelectKbest feature selection and Power Transformer scaling. Graph **C** is the learning curve of the third best-performing model, namely an RF with no feature selection and Robust Scaling.

A.14. Histogram of data distribution of aggregated data set vs. validation data set

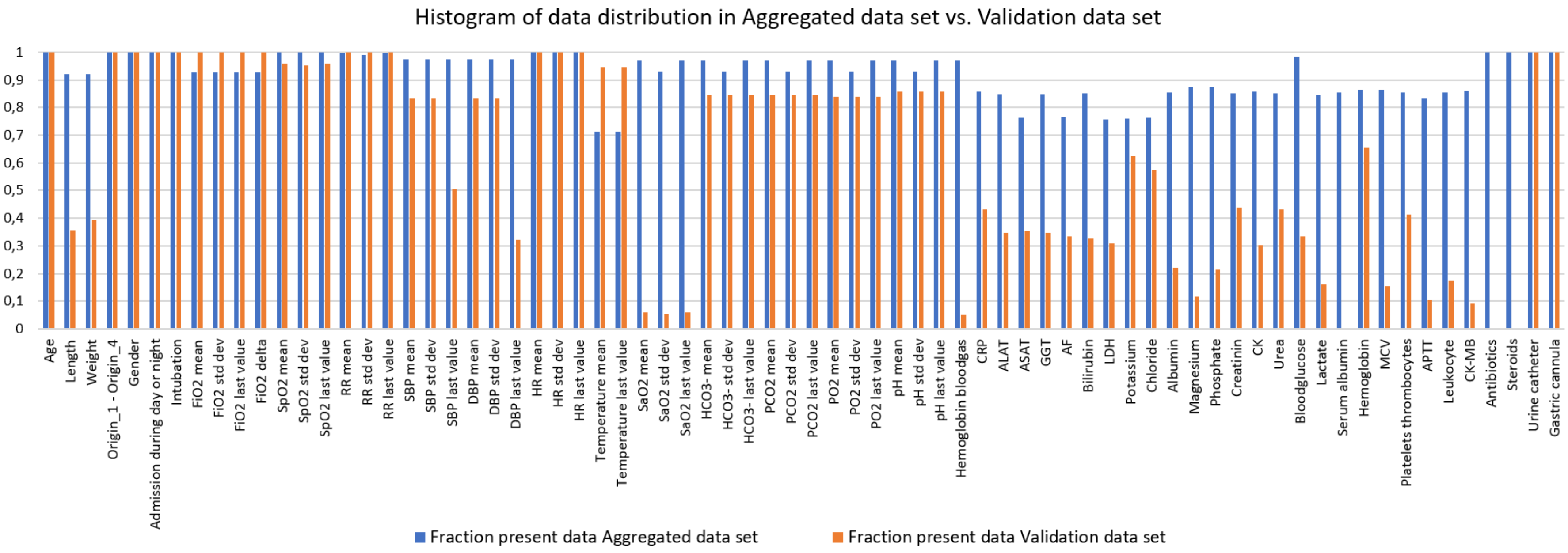


Figure A.9: Histogram showing the fraction of present aggregated data vs. the fraction of present validation data

A.15. Learning curve of training the best-performing aggregated data model on the whole available data set

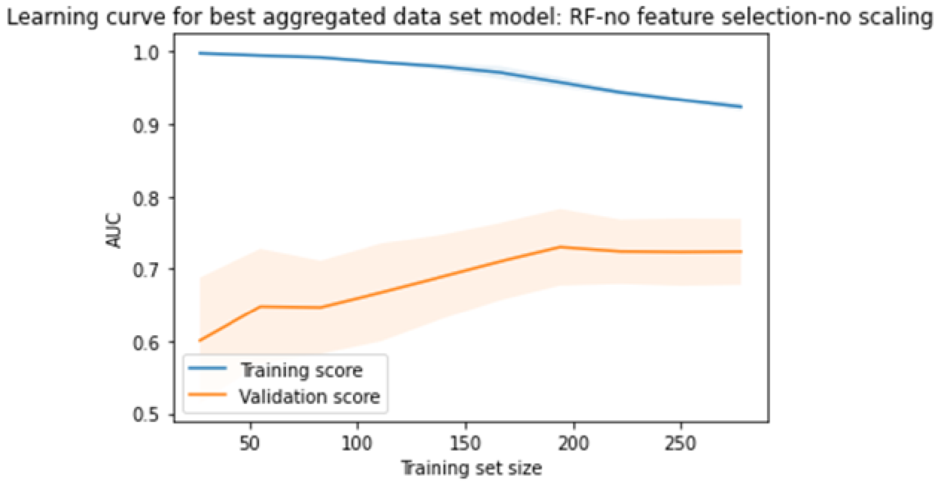


Figure A.10: Learning curve obtained when training the best-performing aggregated data model (combination of RF, no feature selection, and no scaling) on all available training data.

A.16. Example of LMEM in which variation of measurements can be seen

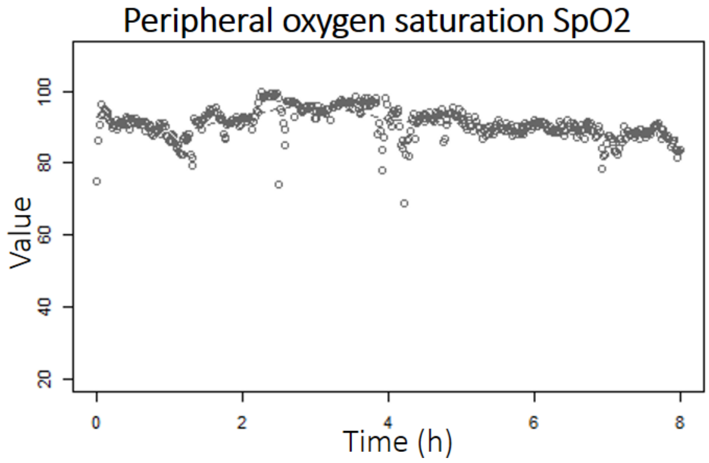


Figure A.11: Example LMEM curve of the peripheral oxygen saturation (SpO2) in which variation around the estimated curve is seen. It can be observed that the curve is minimally influenced by the variance.

A.17. Notable LMEM curves with an explanation for the shown course of the variable

In Figure A.12 a Heart Frequency (HF) signal is shown. The notable thing about this specific LMEM of the HF is the significant variation in the first 4 hours of the graph, contrary to the last 4 hours, which show almost no variation. The patient files explained this phenomenon. In this case, the starting time of HFNO therapy was 20:00 and around 24:00 this patient was given Zopiclon, which is a strong sleep medication. It can be assumed that the patient was in deep sleep after 4 hours of HFNO therapy, which made the HF steadier.

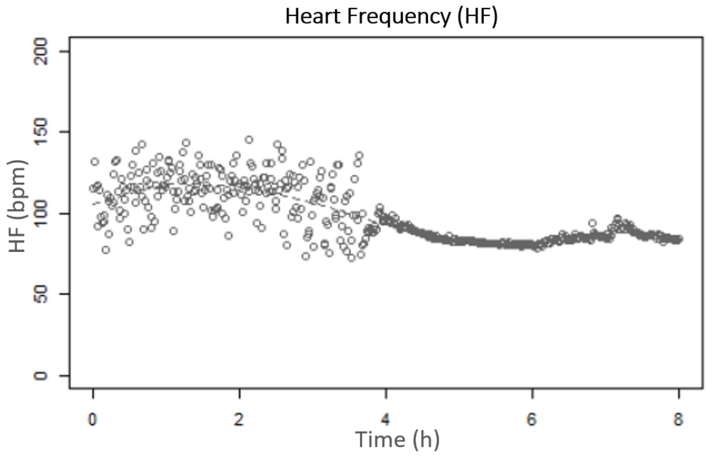


Figure A.12: Heart Frequency of a patient, with the first 4 hours significant variation and the last 4 hours almost no variation.

In Figure A.13 four different blood pressure curves are shown for two patients. Both the Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) are given. In the first patient shown in graphs **A** and **C**, the first 2 hours of HFNO therapy do not have measurements for both SBP and DBP. It turned out that this patient immediately started with HFNO therapy after being admitted to the ICU, which left a gap of 2 hours before the arterial line (that measures the blood pressure) could be punctured.
For the second patient shown in graphs **B** and **D**, the last 6 hours of HFNO therapy do not have measurements for both SBP and DBP. Here, the reason was that the arterial line was broken and not repunctured. The blood pressure was measured non-invasively during these hours.

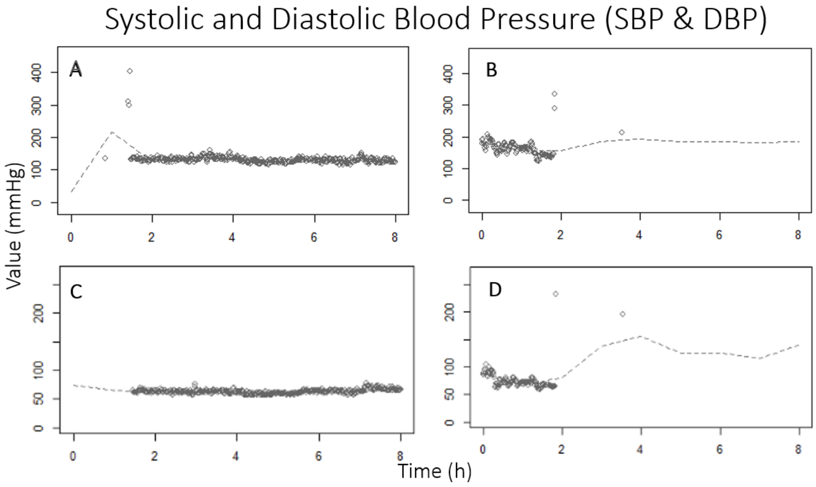


Figure A.13: Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) of two different patients, within both graphs a period without measurements. **A** SBP of a patient that has no values in the first 2 hours of HFNO therapy, **B** SBP of a patient that has no values in the last 6 hours of HNFO therapy, **C** DBP of the same patient as **A** and **D** DBP of the same patient as **B**.

A.18. Learning curves of best-performing joint models

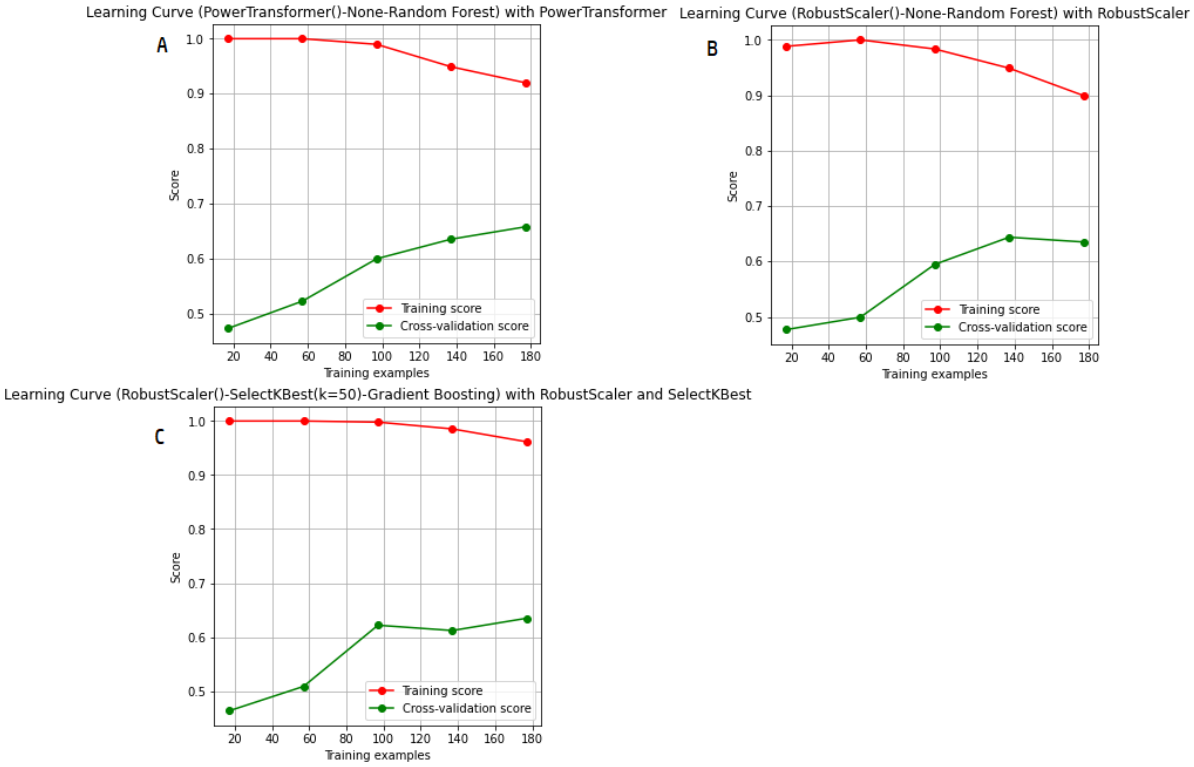


Figure A.14: Learning curves of the best-performing joint models. Graph **A** is the learning curve of a Random Forest with no feature selection and Power Transformer scaling. Graph **B** is the learning curve of the second best-performing model, which is an RF with no feature selection and Robust Scaling. Graph **C** shows the learning curve of the third best-performing model, namely a Gradient Boosting Model with SelectKbest feature selection and Robust Scaling.

A.19. Learning curve of training the best-performing joint model on the whole available data set

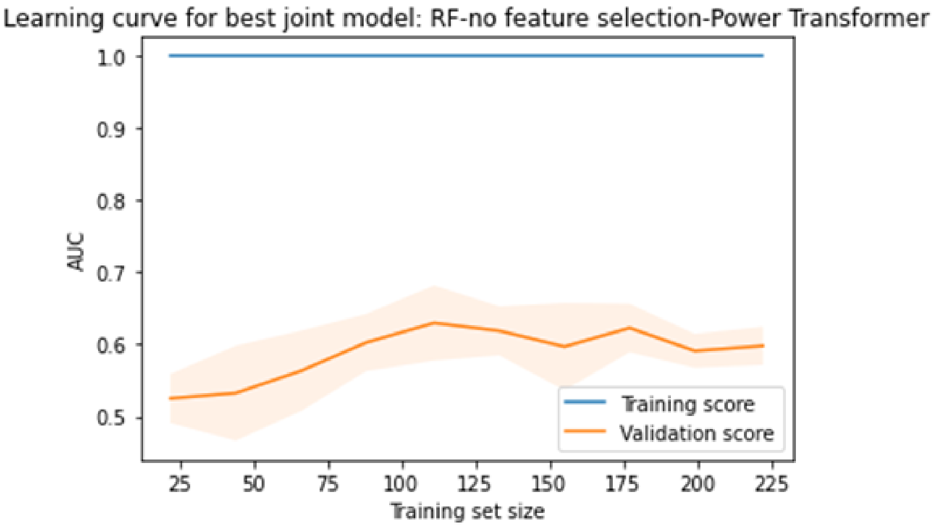


Figure A.15: Learning curve obtained when training the best-performing joint model (combination of RF, no feature selection, and Power Transformer scaling) on all available training data.

A.20. Grid search versus randomized search for hyperparameter tuning

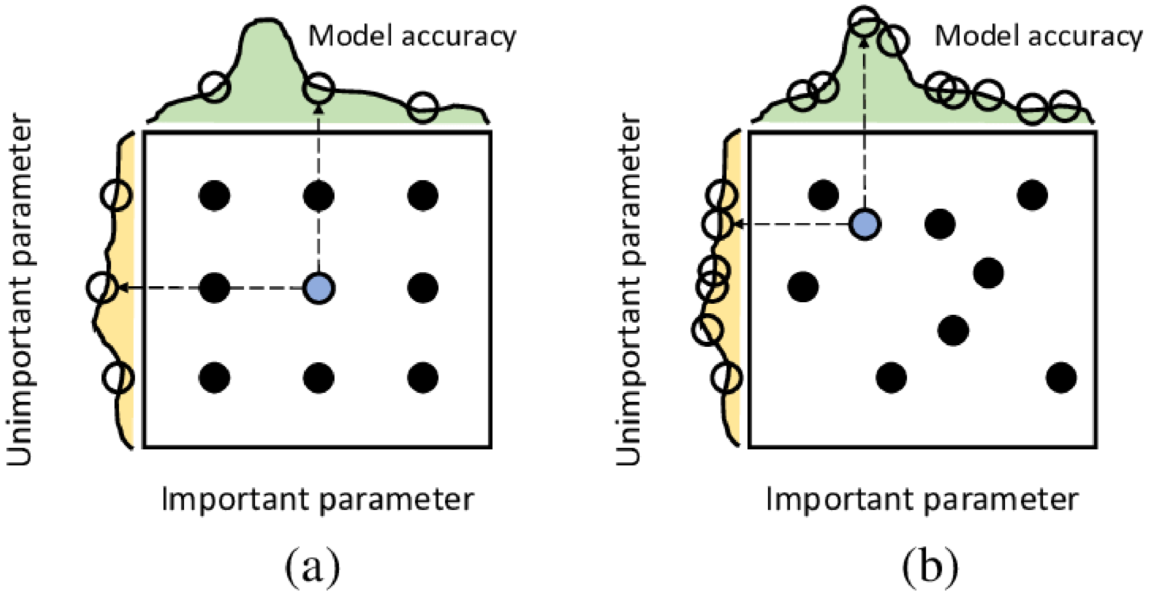


Figure A.16: Image showing the difference in *a* grid search and *b* randomized search for hyperparameter tuning. It is shown that in grid search, the peak of the important parameter is not found, while in randomized search it is found.[32]

B

Literature review

In this Appendix B the literature review that has been conducted before this thesis project has been inserted. The literature review is a comprehensive overview of the currently used models to predict intubation in ICU patients. The literature review resulted in the model type choice of this thesis project.

Predicting intubation in ICU patients using artificial intelligence: a systematic review with meta-analysis

Manon Hendriks

1. Educational program Technical Medicine; Leiden University Medical Center, Delft University of Technology & Erasmus University Medical Center Rotterdam.

2. Intensive Care Unit, Maasstad Hospital, Rotterdam, The Netherlands

August 2022

Abstract

Introduction In this systematic review recent literature on predictive models in ICU patients that predict intubation is reviewed. The possibility to predict intubation will be of great value to medical personnel in their treatment choice, as it could lead to fewer unnecessary and fewer emergency intubations.
Methods Pubmed was used to retrieve recent literature concerning machine learning models that predict intubation. Using a search string Pubmed was systematically searched for the concerning articles. Numerous in- and exclusion criteria were used to make a selection of the articles.
Results The Pubmed search resulted in 111 articles, 16 articles were included in the systematic review. An adjusted Newcastle-Ottawa Quality Assessment Score (NOS) was used to test the quality of the articles. A forest plot of the Area Under the Curve (AUC) values of the different articles is given in the result section. The meta-analysis of the AUCs of the included articles gave a Mean AUC value of 0.810 with a 95% Confidence Interval (CI) of (0.687, 0.933). The best scoring models were respectively two feed-forward Neural Networks, a Random Forest and a Gradient Boosting Model.
Conclusion This systematic review gives a comprehensive overview of the current models that have been developed to predict intubation among ICU patients. Several options for a considered model choice can be taken from this review.

Keywords Artificial Intelligence, Prediction, Endotracheal intubation.

1 Introduction

The decision to intubate a patient in the ICU that suffers from a form of respiratory failure continues to be challenging for hospital personnel. It is thought that earlier intubation results in shorter hospital stays and less deterioration of patients, however on the other hand several articles (citation) are unable to prove these advantages.[1, 2] Moreover, intubation can be dangerous for patients and lead to additional damage such as ventilator-induced lung injury (VILI) or ventilator-associated pneumonia (VAP).[3]

A prediction model based on Machine Learning (ML) could be useful to predict which patients should be intubated and which not, based on their clinical parameters. The most recent literature is reviewed in this article to give a summary of the existing predictive models and their performance value.
The goal of this review is on the one hand to find out what kind of models already exist that can predict intubation to use as background information for the eventual thesis project and to get an idea of features that are used in these models. On the other hand, this review is used to get a feeling of the performing value

of these different ML models, and thereby a feeling for the range of performing value the model of the thesis project should have. The selection criteria of the articles in this review will depend on these two goals. For the first goal, it does not matter that much what kind of patient population is used in the articles, as the model will probably not differ that much. Except for the articles in which neonate patients are included, neonate ventilation medicine differs a lot from adult ventilation medicine. For the performance value the patient population does however matter, as it can influence this value greatly. For instance a patient population with mainly young patients will probably have a lower rate of intubation and thus a different prediction for intubation compared to a patient population of mainly elderly.

2 Methods

2.1 Search methods

Pubmed was searched to find articles concerning machine learning models that predict intubation. The following search string was entered in Pubmed on the 13th of May 2022.
((((((”Decision Support Systems, Clinical”[Mesh]) OR (”decision-making”[TIAB])) OR (”prediction”[TIAB])) OR (”predict”[TIAB])) OR (”predicting”[TIAB])) AND (((”Intubation, Intratracheal”[Mesh]) OR (”Intubation”[TIAB])) OR (”Mechanical ventilator”[TIAB]) OR (”extubation”[TIAB]))) AND (((((”Artificial Intelligence”[Mesh]) OR (”Artificial Intelligence”[TIAB])) OR (”Artificial Intelligent”[TIAB])) OR (”machine learning”[MESH])) OR (”machine learning”[TIAB]))

2.2 Study selection

Titles and abstracts were screened to filter out the articles that needed screening for full text. Moreover, via cross-reference additional articles could be included, that were not found with the Pubmed search string.
The following in- and exclusion criteria were taken into account while screening the titles and abstracts. The first criterion was that the article should be about a predictive model that predicts intubation in at least one of the outcomes. Secondly, it was required that the article contained humans only in their population. Articles with neonates were excluded, and articles with children and adults were included. Articles that were about a predictive model that predicts extubation or intubation failure were excluded, High Flow Nasal Cannula (HFNC) or High Flow Nasal Oxygen (HFNO) failure articles were included as a failure of these treatments often leads to intubation. Moreover, articles that developed a predictive model to predict favourable intubation locations such as the ideal depth and predictions of endotracheal tube locations on X-ray images were excluded. Articles that build their predictive model solely on imaging data were also excluded.

2.2.1 Quality Assessment

After title and abstract screening, a selection of the articles were assessed for their quality. This Quality Assessment (QA) was performed based on a well-known QA scale, namely the Newcastle Ottawa Scale (NOS).[4] There are two versions of the NOS, in this review the NOS for cohort studies was used. The NOS was adjusted to fit better with the articles included in the study. Specifically, the fourth question of the NOS was removed, namely the question: ”Demonstration that outcome of interest was not present at the start of study”. This question would have been answered with ”No” for every included article as all the articles had an outcome of interest at the beginning of the study. Next to the removal of this question, two additional questions were added. These two questions were about Machine Learning specific topics. Namely, it was questioned whether the train and test set during training and validation of the model were kept completely separate and also which kind of performance measure was used to interpret results. The adjusted NOS can be found in Appendix A. In the adjusted NOS a total of 10 points can be scored, in this review articles that score below 6 points will be excluded.

2.3 Statistical analysis

For the statistical analysis, the AUC values or AUROC were compared between all the included articles. The program OpenMeta[analyst][5] was used to perform a meta-analysis with all the AUC values. This specific program uses the size of the population group, AUC value and standard deviation of the AUC value. Not all the included articles provided the standard deviation of the AUC value. The following formula was therefore applied to calculate the standard deviation of the AUC values. This formula could only be implemented for articles of which the 95% Confidence Intervals (CI) were known.[6]

$$SD = \sqrt{N} * (UpperCI - LowerCI)/(t_{alpha,df} * 2)$$

In which SD is the standard deviation, N is the number of subjects, $UpperCI$ is the upper bound of the CI, $LowerCI$ is the lower bound of the CI and $t_{alpha,df}$ the factor that considers the probability and degrees of freedom. In this article 0.05 was used as $alpha$ and $N-1$ for df.[6] The T.INV.2T function is used in excel to calculate the $t_{alpha,df}$ factor.
The heterogeneity value of the meta-analysis was not given a specific cut-off value, when a very large heterogeneity (e.g more than 90%) occurred a second meta-analysis would be performed. In this extra meta-analysis, the most heterogeneous study should be excluded. Which is the study that lies furthest below the

mean AUC value.
The weights that will be given to each study in the meta-analysis will be compared, as a higher weight means a higher precision it will be analyzed which study receives the highest weight.

3 Results

3.1 Included literature

After title abstract analysis three additional articles were found via cross-reference inclusion. Of the 114 articles, 94 were excluded based on their title and abstract. Reasons for exclusion were different article types such as reviews, letters to the editor and symposium abstracts. Other reasons for exclusion were, that the article was not about intubation, the article was about extubation, the subject of the article was intubation location, the article's model was based on imaging or histological data, the study was in animals or neonates and no full text was available. For articles that seemed useful in the title/abstract screening all full text was retrieved or bought.
20 articles were included to be read full text. Hereafter, 2 articles were excluded. One article did not predict intubation and the other article included neonates as patients. 18 articles were included in the Quality Assessment Scale, namely the previously mentioned adapted NOS. After the NOS assessment, two additional articles were excluded based on a lower score than 6 on the adjusted NOS. Namely the articles Varzaneh ZA et al. 2022 and Lundon DJ et al 2020.[7, 8] Eventually the study selection resulted in 16 articles that were taken into account in this systematic review. In Figure 2 a flowchart that demonstrates the study selection process can be found. In Appendix B the results of the adjusted NOS are shown, and the answers to every question from the adjusted NOS are listed per article that was included in the Quality Assessment.

3.2 Study Characteristics of included articles

During the full-text assessment of the included articles, different information parameters were collected. The study type, population size, best performing algorithm type, top 5 predictive features, applied missing data strategy, follow-up time, patient/population & intervention & comparison & outcomes (PICO), country of origin of data, population age, gender and quality assessment score of every article is listed in a snapshot of the study characteristics table in Figure 4.

3.2.1 Population characteristics

Almost every article was a cohort study, most of them a retrospective cohort study. Except for Veermani A et al. 2022 who performed a retrospective case-control study [9] and Burdick H et al. 2020 that performed a multicenter clinical trial validation study [10]. In all the included studies the intervention group, the intubated patients, was always smaller than the control group. Often the intubation population would be about a third of the control group population. It seems that is comparable to the true patient population in the ICU as all the articles from many different countries have the same distribution. The age of the population in the included articles is often around 60 years, except for the articles of Pappy G et al. 2022 [11] and Im D et al. 2022[12]. These articles included children in their study. In these articles the ages were around 3 and 9 years.

3.2.2 Machine Learning characteristics

There are roughly two kind of Machine Learning models. The difference between the two kind lies in the type of data that is used in the model. It is a possibility to give a tabular-like data set to a model, in which each row contains a different patient, every column is a feature and one of the columns contains the desired outcome (e.g. yes or no for intubation). Models that work with this kind of data sets are Logistic Regression model and tree-based models such as Random Forest and Gradient Boosting models. In Appendix C more background information about these model types is given.
In this review the following articles made use of this kind of model; Veermani A et al 2022 [9], Aljouie AF et al. 2021 [13], Bolourani S et al. 2021 [14], Campbell TW et al. 2021 [15], Mauer E et al. 2021 [16], Arvind V et al. 2020 [17], Burdick H et al. 2020 [10], Siu BMK et al. 2020 [18] and Ren O et al. 2018 [19].
The other possibility is to give a more raw data set to the model, for example a data set of a heart rate often contains more values than a data set of a certain laboratory blood value. In these other models, it is possible to use time series in the model. Models that work with this kind of data sets are deep learning models such as Neural Networks. Again in Appendix C more background information is given for these model types. The following articles use this kind of model; Boussen S et al. 2022 [20], Im D et al. 2022 [12], Pappy G et al. 2022 [11], Shashikumar SP et al. 2021 [21], Wanyan T et al. 2021 [22], Catling FJR et al. 2020 [23] and Suresh H et al. 2017 [24].
As stated the included articles in this review contain both of the mentioned model types. Namely, 9 included articles used tabular-like data in their model and the other 7 included articles made use of raw data.

3.2.2.1 Applied missing data strategy

The included articles used different strategies to handle missing data. In some studies, the missing drug or intervention measurements were imputed with zero and the physiological or lab measurements were propagated forward.[12, 11, 22, 17] Other studies did not apply a missing data strategy or excluded patients with missing data.[20, 15, 19] Another commonly used technique was to implement an extra indicator for missing values that resulted in a column or feature.[16, 10, 23, 24]. Moreover, a K Nearest Neighbour Algorithm was used to impute missing data by some studies.[9, 14] Aljouie AF et al. 2021 used random undersampling and random downsampling to impute missing data [13], Shashikumar SP et al. 2021 used mean imputation [21] and Siu BMK et al. 2020 used autoencoder [18].

3.2.2.2 Most predictive features

From the top 5 predictive features listed for every study in the study characteristics table the most common predictive features were retrieved. This resulted in the following top 5 predictive features with the most occurred predicted feature at the top.

1. Respiratory Rate (8 times)
2. Oxygen Saturation (SpO2) (5 times)
3. Temperature (4 times)
4. Fraction of inspired Oxygen (FiO2) (3 times)
5. Heart Rate (3 times)

3.3 Meta-analysis

More than half of the included studies did not report a standard deviation or 95% CI to support their found AUC. For this reason these studies could not be analyzed using OpenMeta[analyst][5] the AUC values are therefore listed in Table 1 below. Veermani A et al. 2022 did report a brier score of 0.05 which says that the provided AUC is a reliable estimation as a brier score of 0 means a perfect prediction.[9] Moreover, Aljouie AF 2021 reported a balanced accuracy of 0.79.[13] P Campbell TW et al. 2021 did not report an AUC at all, they gave the precision, recall and F1 scores for 4 different risk groups.[15] In Appendix D the scores are shown for the intubation sub-research. Suresh H et al. 2017 mentioned that the AUCs had a difference of 0.12 but this was between the developed models and not a difference or variance of the AUC value itself.[24]

Table 1: Overview of AUC values of the studies that were not included in the Forest Plot

Study name	AUC
Boussen S et al. 2022 [20]	0.94
Pappy G et al. 2022 [11]	0.78
Veermani A et al. 2022 [9]	0.737
Aljouie AF et al. 2021 [13]	0.82
Campbell TW et al. 2021 [15]	No AUC
Mauer E et al. 2021 [16]	0.891-0.934
Arvind V et al. 2021 [17]	0.84
Burdick H et al. 2020 [10]	0.866
Suresh H et al. 2017 [24]	0.75

In the following Table 2, the weights that are given to each study in the meta-analysis are shown per study. The weights are evenly divided which means that the studies are evenly divided precision-wise, moreover larger population size often leads to a heavier weight. However, in this meta-analysis the relatively small population size of Im D et al. 2022 [12] is cancelled out by the good precision of this study.

Table 2: Overview of weights within Forest Plot for every included article

Study name	Weight
Im D et al. 2022 [12]	14.288%
Bolourani S et al. 2021 [14]	14.301%
Shashikumar SP et al. 2021 [21]	14.289%
Wanyan T et al. 2021 [22]	14.302%
Catling FJR et al. 2020 [23]	14.284%
Siu BMK et al. 2020 [18]	14.288%
Ren O et al. 2018 [19]	14.248%

An image of the Forest Plot (FP) is shown in Figure 3. In this FP it can be seen that the mean AUC value for all the included studies in the meta-analysis is an AUC of 0.810 with a 95% CI of (0.687, 0.933). Also shown is that the studies Shashikumar SP et al. 2021 [21], Catling FJR et al. 2020 [23], Siu BMK et al. 2020 [18] and Ren O et al. 2018 [19] have a higher AUC value than the mean AUC value. Moreover, their 95% CI bands do not cross the mean AUC value line which means that their AUC values are significantly higher than the mean AUC value. Im D et al. 2022 have their square exactly on the mean AUC value as their AUC value is also 0.810.[12] Bolourani S et al. 2021 [14] and Wanyan T et al. 2021 [22] score significantly lower than the mean AUC value.

The significant high heterogeneity of this FP I^2=99.97% means that the pooled effect estimate is not shown as the studies are too heterogeneous.[25, 26] The exclusion of the two significantly lower scoring studies Bolourani S et al. and Wanyan T et al. resulted in a heterogeneity of I^2=97.62% which is still very high. Which is actually expected as predictive studies usually differ in design and execution, therefore variation between their results is unlikely to occur only by chance. The chosen model to perform the meta-analysis was therefore a Continuous Random-Effects Model.[27]

4 Discussion

This systematic review has resulted in a comprehensive overview of the current articles about models that have been developed to predict intubation in ICU patients. Specifically, the meta-analysis has shown that the following 4 articles developed the models with the best AUC values. Namely, Shashikumar SP et al. who developed a feed-forward Neural Network with an AUC of 0.886 (0.876-0.896) [21], Catling FJR et al. also developed a feed-forward Neural Network with an AUC value of 0.896 (0.885-0.907) [23], Siu BMK et al. developed a Random Forest model with an AUC value of 0.860 (0.850-0.870) [18] and Ren O et al. developed a Gradient Boosting Model with an AUC value of 0.890 (0.870-0.910) [19].

From this analysis it seems that Neural Network models perform best in predicting intubation, these are the models that make use of raw data to give a prediction. In the following thesis project, these types of models are not preferred because the clinical reasoning and explainability toward medical personnel are difficult with Neural Networks.[28] The models such as Random Forest and Gradient Boosting Model are more easily understood and therefore more likely to be used by medical personnel. Moreover, the AUC values of the studies that use Random Forest and Gradient Boosting Model do not differ that much.

It is therefore decided that in the thesis project the type of models that use a tabular-like data set with an outcome column will be developed first.

The results of this systematic review matter because it gives an overview of the currently available models and the corresponding performance value. This will be of great value in the development of the model in the thesis project. Eventually, it will be very valuable for the Maasstad Hospital to implement a model that can predict whether a patient needs to be intubated or not. With a predicted score the medical personnel can make an advised decision, which will result in fewer unnecessary intubations and fewer unexpected/unplanned intubations. This will result in improved patient care.

Limitations

Some of the included studies did not develop a model to predict intubation but developed a model that predicts HFNC or HFNO therapy failure. As the failure of this therapy often ultimately leads to intubation (or death) these studies were still considered in the review. Specifically, Im D et al. predicted Bilevel Positive Airway Pressure (BIPAP) failure and Pappy G et al. predicted HFNC failure. As Im D et al. was included in the meta-analysis this could have introduced a form of bias as the outcomes were not completely the same. However, there was still data available on how many of these study participants were intubated.

Many studies did not report a CI or standard deviation and therefore could not be included in the meta-analysis, some of the AUC values of these studies were quite high when compared to the studies that were included in the meta-analysis. As 5 of the 9 studies that were not included in the meta-analysis scored above the obtained mean AUC value the inclusion of these articles could have given a different mean AUC value. Moreover, more included studies and data would have resulted in a more variant database and therefore a more reliable result.

This systematic review and its meta-analysis were performed by one author, this may have resulted in bias as only one author performed the inclusion process and Quality Assessment of the included studies.

Future research

For future research, it would be interesting to perform a systematic review in which the method for defining when intubation has taken place is compared. The studies included in this study give different methods for determining when a patient is intubated. Some studies do not describe how they determined this. A technique that is commonly used is to use the FiO2 values. A sudden elevation in FiO2 could indicate that a patient has been intubated. Other studies make use of the Positive End Expiratory Pressure (PEEP) value that can be entered into a mechanical ventilator. In this future research, it could be investigated which method should be used to determine how long a patient stays intubated and if the patient is extubated in the meantime.

5 Conclusion

This systematic review gives a comprehensive overview of the current models that have been developed to predict intubation among ICU patients. Several options for different model types are researched and their performance value can easily be compared. To make sure that an eventual model is explainable and clinically valuable it should be considered to use a tabular-like database and a model that uses this kind of data such as a Logistic Regression, Random Forest or other tree-based models such as Gradient Boosting Models.

Funding

No funding was obtained to undertake this systematic review and meta-analysis.

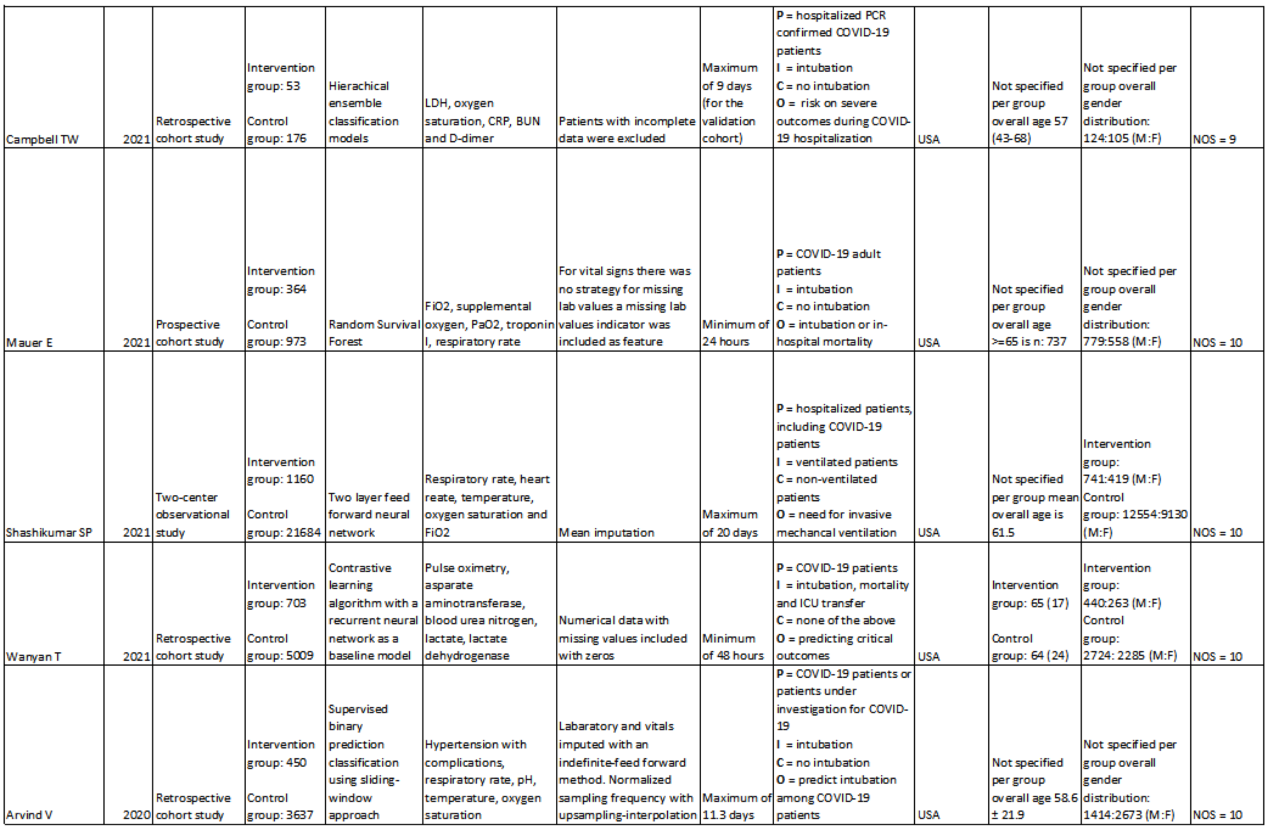
Availability of data and materials

This review is solely based on secondary data reported by published research studies.

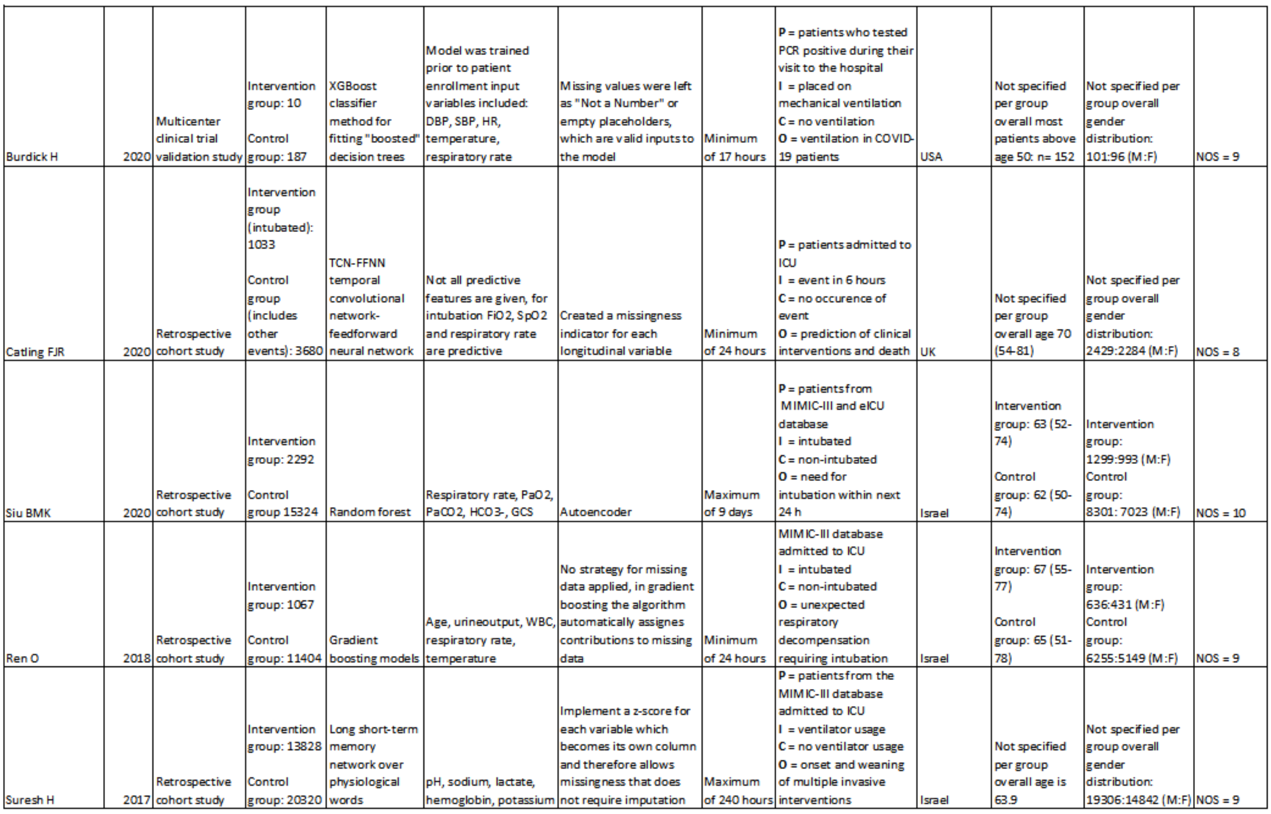
Competing interests

The author declares that she has no competing interest.

First author	Year	Study type	Population size	Best performing algorithm type	Top 5 predicting features	Applied missing data strategy	Follow-up time	PICO	Country of data	Population Age (in years)	Gender	Quality assessment score
Boussen s	2022	Retrospective observational cohort study	Intervention group: 111 Control group: 168	Unsupervised machine learning model	Combination of parameters of model 4 have the best performance: SpO2M, SpO2min, SpO2-90, BFM	Only two parameters that were recorded for all patients	Maximum of 20 days	**P** = patients admitted to the ICU **I** = intubation **C** = no intubation **O** = predict an extended stay on ICU	France	Not specified per group overall age 63 [56-70]	Not specified per group overall gender distribution: 186:93 (M:F)	NOS = 8
Im D	2022	Retrospective cohort study	Intervention group: 175 Control group: 455	Long short-term memory recurrent neural network	Not stated which of the 301 input variables had the most predictive value	Missing drug or intervention measurements imputed with zero and physiological or lab measurement were propagated forward	Minimum of 28 days	**P** = pediatric patients that received BIPAP **I** = BIPAP failure **C** = BIPAP non-failure **O** = BIPAP failure	USA	Intervention group: 9.0 (4.3-13.7) Control group: 10.7 (5.0-14.5)	Intervention group: 91:88 (M:F) Control group: 243:212 (M:F)	NOS = 9
Pappy G	2022	Single-center retrospective cohort study	Divided data in training, validation and test set. Overall n= 834	Multi ensemble of long short-term memory with 3 times input perseveration and transfer learning	Heart rate, mean arterial pressure, pulse oximetry, respiratory rate, systolic blood pressure	Missing drug or intervention measurements imputed with zero and physiological or lab measurement were propagated forward	Minimum of 24 hours	**P** = pediatric patients admitted to ICU in which HFNC was used **I** = HFNC failure **C** = HFNC success **O** = HFNC failure	USA	Overall 3.1 years (± 4.4)	Not mentioned	NOS = 8
Veermani A	2022	Retrospective case control study	Intervention group: 238 Control group: 53864	Gradient Boosting algorithm	Disseminated cancer, dialysis, weight loss, old age, elevated operation time	Patients with categorial variables missing were excluded, quantative variables were imputed with nearest neighbour algorithm	Maximum of 6 days	**P** = patients who underwent ACDF surgery **I** = unplanned intubation **C** = no unplanned intubation **O** = unplanned intubation	USA	Intervention group: 54.68 Control group: 63.37	Intervention group: 82:156 (M:F) Control group: 26660:27204 (M:F)	NOS = 6
Aljouie AF	2021	Retrospective cohort study	Intervention group: 184 Control group: 1324 (both no ventilation and non-invasive ventilation)	Logistic regression model with random undersampling and ReliefF for feature selection	CXR zone 11, CXR zone 12, age, CXR zone 5, gender	Random udersampling and random downsampling , also tested SMOTE and ADASYN for other models	Maximum of 30 days	**P** = hospitalized confirmed COVID-19 patients **I** = mechanical ventilation **C** = no/non-invasive ventilation **O** = mortality and ventilation requirement classification	Saudi Arabia	Intervention group: 61.69 (±14) Non-invasive ventilation: 60.23 (± 15.76) No ventilation: 52.52 (± 16.94)	Intervention group: 140:44 (M:F) Control group: 717:607 (M:F)	NOS = 10
Bolourani S	2021	Retrospective observational cohort study	Intervention group: 933 Control group: 10592	XGBoost model	Invasive oxygen supply via nonrebreather mask, ESI values of 1 and 3, max respiratory rate, max oxygen saturation, black race	Imputed numerical missing data with a weighted k-nearest neighbours algorithm, added category "missing" for categorial variables	Minimum of 48 hours	**P** = hospitalized PCR confirmed COVID-19 patients **I** = intubated **C** = not intubated **O** = intubation and MV within 48h of admission	USA	Intervention group: 66 (56-75) Control group: 65 (54-77)	Intervention group: 606:372 (M:F) Control group: 6062:4530 (M:F)	NOS = 10



Campbell TW	2021	Retrospective cohort study	Intervention group: 53 Control group: 176	Hierachical ensemble classification models	LDH, oxygen saturation, CRP, BUN and D-dimer	Patients with incomplete data were excluded	Maximum of 9 days (for the validation cohort)	**P** = hospitalized PCR confirmed COVID-19 patients **I** = intubation **C** = no intubation **O** = risk on severe outcomes during COVID-19 hospitalization	USA	Not specified per group overall age 57 (43-68)	Not specified per group overall gender distribution: 124:105 (M:F)	NOS = 9
Mauer E	2021	Prospective cohort study	Intervention group: 364 Control group: 973	Random Survival Forest	FiO2, supplemental oxygen, PaO2, troponin I, respiratory rate	For vital signs there was no strategy for missing lab values a missing lab values indicator was included as feature	Minimum of 24 hours	**P** = COVID-19 adult patients **I** = intubation **C** = no intubation **O** = intubation or in-hospital mortality	USA	Not specified per group overall age >=65 is n: 737	Not specified per group overall gender distribution: 779:558 (M:F)	NOS = 10
Shashikumar SP	2021	Two-center observational study	Intervention group: 1160 Control group: 21684	Two layer feed forward neural network	Respiratory rate, heart reate, temperature, oxygen saturation and FiO2	Mean imputation	Maximum of 20 days	**P** = hospitalized patients, including COVID-19 patients **I** = ventilated patients **C** = non-ventilated patients **O** = need for invasive mechancal ventilation	USA	Not specified per group mean overall age is 61.5	Intervention group: 741:419 (M:F) Control group: 12554:9130 (M:F)	NOS = 10
Wanyan T	2021	Retrospective cohort study	Intervention group: 703 Control group: 5009	Contrastive learning algorithm with a recurrent neural network as a baseline model	Pulse oximetry, asparate aminotransferase, blood urea nitrogen, lactate, lactate dehydrogenase	Numerical data with missing values included with zeros	Minimum of 48 hours	**P** = COVID-19 patients **I** = intubation, mortality and ICU transfer **C** = none of the above **O** = predicting critical outcomes	USA	Intervention group: 65 (17) Control group: 64 (24)	Intervention group: 440:263 (M:F) Control group: 2724: 2285 (M:F)	NOS = 10
Arvind V	2020	Retrospective cohort study	Intervention group: 450 Control group: 3637	Supervised binary prediction classification using sliding-window approach	Hypertension with complications, respiratory rate, pH, temperature, oxygen saturation	Labaratory and vitals imputed with an indefinite-feed forward method. Normalized sampling frequency with upsampling-interpolation	Maximum of 11.3 days	**P** = COVID-19 patients or patients under investigation for COVID-19 **I** = intubation **C** = no intubation **O** = predict intubation among COVID-19 patients	USA	Not specified per group overall age 58.6 ± 21.9	Not specified per group overall gender distribution: 1414:2673 (M:F)	NOS = 10



Burdick H	2020	Multicenter clinical trial validation study	Intervention group: 10 Control group: 187	XGBoost classifier method for fitting "boosted" decision trees	Model was trained prior to patient enrollment input variables included: DBP, SBP, HR, temperature, respiratory rate	Missing values were left as "Not a Number" or empty placeholders, which are valid inputs to the model	Minimum of 17 hours	**P** = patients who tested PCR positive during their visit to the hospital **I** = placed on mechanical ventilation **C** = no ventilation **O** = ventilation in COVID-19 patients	USA	Not specified per group overall most patients above age 50: n= 152	Not specified per group overall gender distribution: 101:96 (M:F)	NOS = 9
Catling FJR	2020	Retrospective cohort study	Intervention group (intubated): 1033 Control group (includes other events): 3680	TCN-FFNN temporal convolutional network-feedforward neural network	Not all predictive features are given, for intubation FiO2, SpO2 and respiratory rate are predictive	Created a missingness indicator for each longitudinal variable	Minimum of 24 hours	**P** = patients admitted to ICU **I** = event in 6 hours **C** = no occurence of event **O** = prediction of clinical interventions and death	UK	Not specified per group overall age 70 (54-81)	Not specified per group overall gender distribution: 2429:2284 (M:F)	NOS = 8
Siu BMK	2020	Retrospective cohort study	Intervention group: 2292 Control group 15324	Random forest	Respiratory rate, PaO2, PaCO2, HCO3-, GCS	Autoencoder	Maximum of 9 days	**P** = patients from MIMIC-III and eICU database **I** = intubated **C** = non-intubated **O** = need for intubation within next 24 h	Israel	Intervention group: 63 (52-74) Control group: 62 (50-74)	Intervention group: 1299:993 (M:F) Control group: 8301: 7023 (M:F)	NOS = 10
Ren O	2018	Retrospective cohort study	Intervention group: 1067 Control group: 11404	Gradient boosting models	Age, urineoutput, WBC, respiratory rate, temperature	No strategy for missing data applied, in gradient boosting the algorithm automatically assignes contributions to missing data	Minimum of 24 hours	MIMIC-III database admitted to ICU **I** = intubated **C** = non-intubated **O** = unexpected respiratory decompensation requiring intubation	Israel	Intervention group: 67 (55-77) Control group: 65 (51-78)	Intervention group: 636:431 (M:F) Control group: 6255:5149 (M:F)	NOS = 9
Suresh H	2017	Retrospective cohort study	Intervention group: 13828 Control group: 20320	Long short-term memory network over physiological words	pH, sodium, lactate, hemoglobin, potassium	Implement a z-score for each variable which becomes its own column and therefore allows missingness that does not require imputation	Maximum of 240 hours	**P** = patients from the MIMIC-III database admitted to ICU **I** = ventilator usage **C** = no ventilator usage **O** = onset and weaning of multiple invasive interventions	Israel	Not specified per group overall age is 63.9	Not specified per group overall gender distribution: 19306:14842 (M:F)	NOS = 9

Figure 1: Overview of the study characteristics shown per study. List of abbreviations in alphabetical order: ACDF= Anterior Cervical Discectomy and Fusion, ADYSYN= Adaptive Synthetic, BFM= Mean value of Breathing Frequency, BIPAP= Bilevel Positive Airway Pressure, BUN= Blood Urea Nitrogen, CRP= C-reactive protein, CXR= chest X-ray, ESI= Emergency Severity Index, FiO2= Fraction of inspired Oxygen, GCS= Glasgow Coma Scale, GU= Genito-Urinary, HCO3-= Bicarbonate, LDH= Lactate dehydrogenase, MV= Mechanical Ventilation, NOS= New Ottawa Scale, PaCO2= Partial pressure of Carbon Dioxide, PaO2= Partial pressure of Oxygen, PCR= Polymerase Chain Reaction, SMOTE= Synthetic Minority Over-Sampling Technique, SpO2-90= Percentage of time under 90% of oxygen saturation, SpO2M= Mean value of oxygen saturation, SpO2min= Minimum value of oxygen saturation, WBC= White Blood Cell

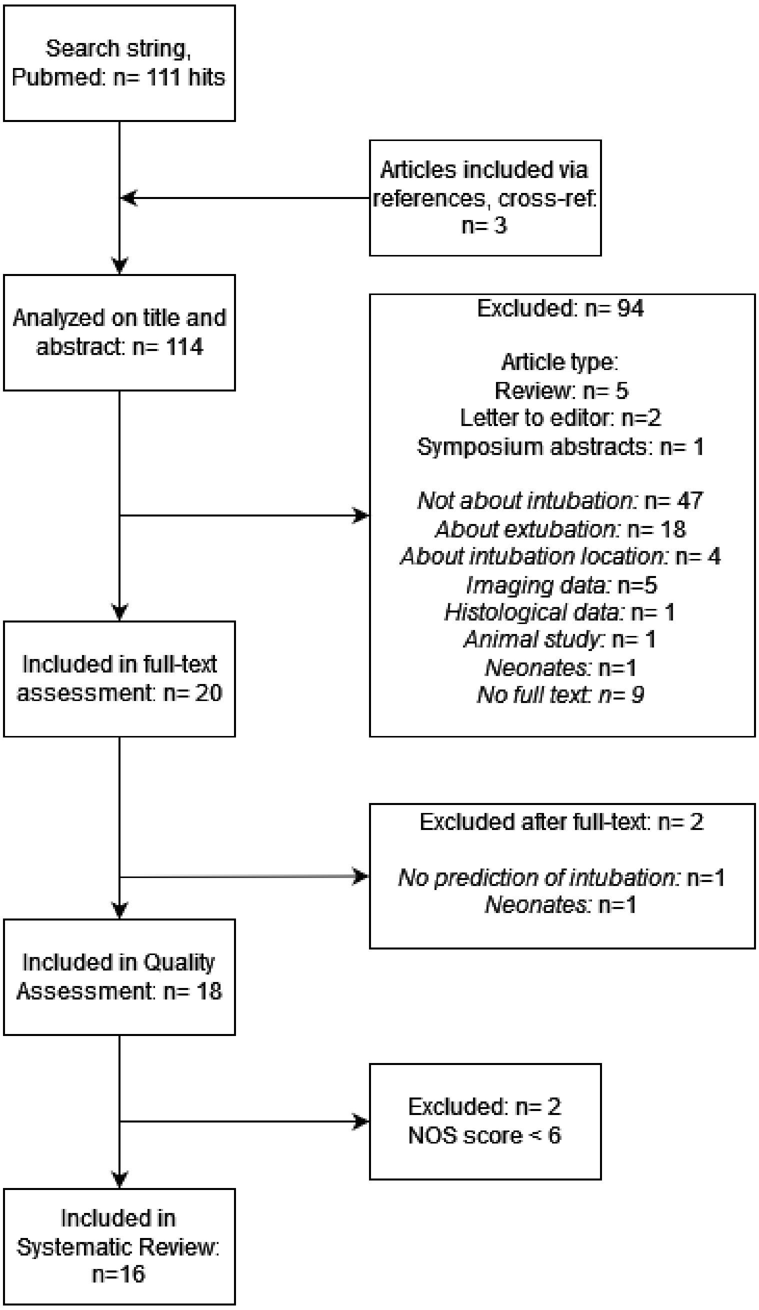

Search string,
Pubmed: n= 111 hits
Articles included via references, cross-ref: n= 3
Analyzed on title and abstract: n= 114
Excluded: n= 94
Article type:
Review: n= 5
Letter to editor: n=2
Symposium abstracts: n= 1
Not about intubation: n= 47
About extubation: n= 18
About intubation location: n= 4
Imaging data: n=5
Histological data: n= 1
Animal study: n= 1
Neonates: n=1
No full text: n= 9
Included in full-text assessment: n= 20
Excluded after full-text: n= 2
No prediction of intubation: n=1
Neonates: n=1
Included in Quality Assessment: n= 18
Excluded: n= 2
NOS score < 6
Included in Systematic Review: n=16

Figure 2: Flowchart demonstrating the study selection process.

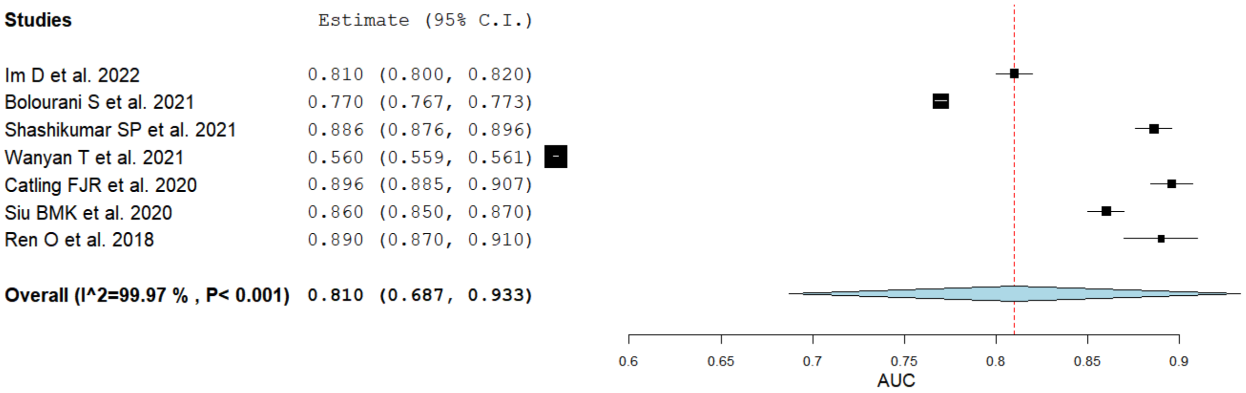


Figure 3: Forest plot showing the AUCs of all the included articles that reported an AUC and also a supporting standard deviation or 95% CI

References

[1] P. Chopra, K. Sodhi, A. Shrivastava, S. Tandon, and R. K. Joia, "Impact of early versus late tracheostomy on patient outcomes in a tertiary care multispeciality ICU," *Journal of Anaesthesiology Clinical Pharmacology*, vol. 37, pp. 458–463, 7 2021.

[2] I. I. Siempos, E. Xourgia, T. K. Ntaidou, D. Zervakis, E. E. Magira, A. Kotanidou, C. Routsi, and S. G. Zakynthinos, "Effect of Early vs. Delayed or No Intubation on Clinical Outcomes of Patients With COVID-19: An Observational Study," *Frontiers in Medicine*, vol. 7, p. 1040, 12 2020.

[3] N. R. Macintyre, "Ventilator-Associated Pneumonia: The Role of Ventilator Management Strategies Introduction Reducing Ventilator-Induced Lung Injury Lung-Protective Mechanical Ventilatory Strategies Require Tradeoffs Hypercapnic Respiratory Acidosis Sedation Atelectasis Weaning Delays Summary," tech. rep., 2005.

[4] "Ottawa Hospital Research Institute."

[5] B. C. Wallace, I. J. Dahabreh, T. A. Trikalinos, J. Lau, P. Trow, and C. H. Schmid, "Closing the gap between methodologists and end-users: R as a computational back-end," *Journal of Statistical Software*, vol. 49, pp. 1–15, 6 2012.

[6] J. P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch, *Cochrane Handbook for Systematic Reviews of Interventions.* Wiley, 9 2019.

[7] Z. A. Varzaneh, A. Orooji, L. Erfannia, and M. Shanbehzadeh, "A new COVID-19 intubation prediction strategy using an intelligent feature selection and K-NN method," *Informatics in Medicine Unlocked*, vol. 28, 1 2022.

[8] D. J. Lundon, B. D. Kelly, D. Shukla, D. M. Bolton, P. Wiklund, and A. Tewari, "A decision aide for the risk stratification of gu cancer patients at risk of SARS-CoV-2 infection, COVID-19 related hospitalization, intubation, and mortality," *Journal of Clinical Medicine*, vol. 9, pp. 1–9, 9 2020.

[9] A. Veeramani, A. S. Zhang, A. Z. Blackburn, C. M. Etzel, K. J. DiSilvestro, C. L. McDonald, and A. H. Daniels, "An Artificial Intelligence Approach to Predicting Unplanned Intubation Following Anterior Cervical Discectomy and Fusion," *GSJ-Original Research Global Spine Journal*, vol. 2022, no. 0, pp. 1–7.

[10] H. Burdick, C. Lam, S. Mataraso, A. Siefkas, G. Braden, R. P. Dellinger, A. McCoy, J. L. Vincent, A. Green-Saxena, G. Barnes, J. Hoffman, J. Calvert, E. Pellegrini, and R. Das, "Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial," *Computers in Biology and Medicine*, vol. 124, p. 103949, 9 2020.

[11] G. Pappy, M. Aczon, R. Wetzel, and D. Ledbetter, "Predicting High Flow Nasal Cannula Failure in an Intensive Care Unit Using a Recurrent Neural Network With Transfer Learning and Input Data Perseveration: Retrospective Analysis," *JMIR Medical Informatics*, vol. 10, 3 2022.

[12] D. D. Im, E. Laksana, D. R. Ledbetter, M. D. Aczon, R. G. Khemani, and R. C. Wetzel, "Development of a deep learning model that predicts Bi-level positive airway pressure failure," *Scientific Reports*, vol. 12, pp. 1–9, 12 2022.

[13] A. F. Aljouie, A. Almazroa, Y. Bokhari, M. Alawad, E. Mahmoud, E. Alawad, A. Alsehawi, M. Rashid, L. Alomair, S. Almozaai, B. Albesher, H. Alomaish, R. Daghistani, N. K. Alharbi, M. Alaamery, M. Bosaeed, and H. Alshaalan, "Early prediction of COVID-19 ventilation requirement and mortality from routinely collected baseline chest radiographs, laboratory, and clinical data with machine learning," *Journal of Multidisciplinary Healthcare*, vol. 14, pp. 2017–2033, 2021.

[14] S. Bolourani, M. Brenner, P. Wang, T. McGinn, J. S. Hirsch, D. Barnaby, T. P. Zanos, M. Barish, S. L. Cohen, K. Coppa, K. W. Davidson, S. Debnath, L. Lau, T. J. Levy, A. Makhnevich, M. D. Paradis, and V. Tóth, "A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: Model development and validation," *Journal of Medical Internet Research*, vol. 23, 2 2021.

[15] T. W. Campbell, M. P. Wilson, H. Roder, S. MaWhinney, R. W. Georgantas, L. K. Maguire, J. Roder, and K. M. Erlandson, "Predicting prognosis in COVID-19 patients using machine learning and readily available clinical data," *International Journal of Medical Informatics*, vol. 155, 11 2021.

[16] E. Mauer, J. Lee, J. Choi, H. Zhang, K. L. Hoffman, I. J. Easthausen, M. Rajan, M. G. Weiner, R. Kaushal, M. M. Safford, P. A. Steel, and S. Banerjee, “A predictive model of clinical deterioration among hospitalized COVID-19 patients by harnessing hospital course trajectories,” *Journal of Biomedical Informatics*, vol. 118, 6 2021.

[17] V. Arvind, J. S. Kim, B. H. Cho, E. Geng, and S. K. Cho, “Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19,” *Journal of Critical Care*, vol. 62, pp. 25–30, 4 2021.

[18] B. M. K. Siu, G. H. Kwak, L. Ling, and P. Hui, “Predicting the need for intubation in the first 24 h after critical care admission using machine learning approaches,” *Scientific Reports*, vol. 10, pp. 1–8, 12 2020.

[19] O. Ren, A. E. Johnson, E. P. Lehman, M. Komorowski, J. Aboab, F. Tang, Z. Shahn, D. Sow, R. Mark, and L. W. Lehman, “Predicting and understanding unexpected respiratory decompensation in critical care using sparse and heterogeneous clinical data,” in *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, pp. 144–151, Institute of Electrical and Electronics Engineers Inc., 7 2018.

[20] S. Boussen, P. Y. Cordier, A. Malet, P. Simeone, S. Cataldi, C. Vaisse, X. Roche, A. Castelli, M. Assal, G. Pepin, K. Cot, J. B. Denis, T. Morales, L. Velly, and N. Bruder, “Triage and monitoring of COVID-19 patients in intensive care using unsupervised machine learning,” *Computers in Biology and Medicine*, vol. 142, 3 2022.

[21] S. P. Shashikumar, G. Wardi, P. Paul, M. Carlile, L. N. Brenner, K. A. Hibbert, C. M. North, S. S. Mukerji, G. K. Robbins, Y. P. Shao, M. B. Westover, S. Nemati, and A. Malhotra, “Development and Prospective Validation of a Deep Learning Algorithm for Predicting Need for Mechanical Ventilation,” *Chest*, vol. 159, pp. 2264–2273, 6 2021.

[22] T. Wanyan, H. Honarvar, S. K. Jaladanki, C. Zang, N. Naik, S. Somani, J. K. De Freitas, I. Paranjpe, A. Vaid, J. Zhang, R. Miotto, Z. Wang, G. N. Nadkarni, M. Zitnik, A. Azad, F. Wang, Y. Ding, and B. S. Glicksberg, “Contrastive learning improves critical event prediction in COVID-19 patients,” *Patterns*, vol. 2, p. 100389, 12 2021.

[23] F. J. Catling and A. H. Wolff, “Temporal convolutional networks allow early prediction of events in critical care,” *Journal of the American Medical Informatics Association*, vol. 27, pp. 355–365, 3 2020.

[24] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, “Clinical Intervention Prediction and Understanding using Deep Networks,” 5 2017.

[25] T. P. A. Debray, J. A. A. G. Damen, K. I. E. Snell, J. Ensor, L. Hooft, J. B. Reitsma, R. D. Riley, and K. G. M. Moons, “A guide to systematic review and meta-analysis of prediction model performance,” *BMJ*, vol. 356, p. i6460, 1 2017.

[26] S. A. Glasmacher and W. Stones, “Anion gap as a prognostic tool for risk stratification in critically ill patients - a systematic review and meta-analysis,” *BMC Anesthesiology*, vol. 16, 8 2016.

[27] T. B. Huedo-Medina, J. Sánchez-Meca, F. Marín-Martínez, and J. Botella, “Assessing heterogeneity in meta-analysis: Q statistic or I 2 Index?,” *Psychological Methods*, vol. 11, pp. 193–206, 6 2006.

[28] Y. Ge, Y. Xiao, Z. Xu, M. Zheng, S. Karanam, T. Chen, L. Itti, and Z. Wu, “A Peek Into the Reasoning of Neural Networks: Interpreting with Structural Visual Concepts,” tech. rep.

[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* Springer Series in Statistics, New York, NY: Springer New York, 2009.

[30] J. H. Friedman, “Stochastic gradient boosting,” *Computational Statistics and Data Analysis*, vol. 38, pp. 367–378, 2 2002.

Appendices

A Adjusted NOS

On the following page the adjusted NOS[4] is given.

NEWCASTLE - OTTAWA QUALITY ASSESSMENT SCALE
COHORT STUDIES – ADJUSTED VERSION

Note: A study can be awarded a maximum of one star for each numbered item within the Selection, Outcome and Machine learning specific categories. A maximum of two stars can be given for Comparability

Selection

1) Representativeness of the exposed cohort
   a) truly representative of the average patients in the ICU community *
   b) somewhat representative of the average patients in the ICU community *
   c) selected group of users eg nurses, volunteers
   d) no description of the derivation of the cohort

2) Selection of the non exposed cohort
   a) drawn from the same community as the exposed cohort *
   b) drawn from a different source
   c) no description of the derivation of the non exposed cohort

3) Ascertainment of exposure
   a) secure record (eg surgical records) *
   b) structured interview *
   c) written self report
   d) no description

Comparability

1) Comparability of cohorts on the basis of the design or analysis
   a) study controls for age and gender *
   b) study controls for any additional factor * (This criteria could be modified to indicate specific control for a second important factor.) e.g. comorbidities

Outcome

1) Assessment of outcome
   a) independent blind assessment *
   b) record linkage *
   c) self report
   d) no description

2) Was follow-up long enough for outcomes to occur
   a) yes (select an adequate follow up period for outcome of interest) *
   b) no

3) Adequacy of follow up of cohorts
   a) complete follow up - all subjects accounted for *
   b) subjects lost to follow up unlikely to introduce bias - number lost <= 20 % , or description suggested no different from those followed *
   c) follow up rate less than 80% and no description of those lost
   d) no statement

Machine Learning specific

1) Seperation of train and test set in training and validation of the model

   a) yes *

   b) no

   c) not stated

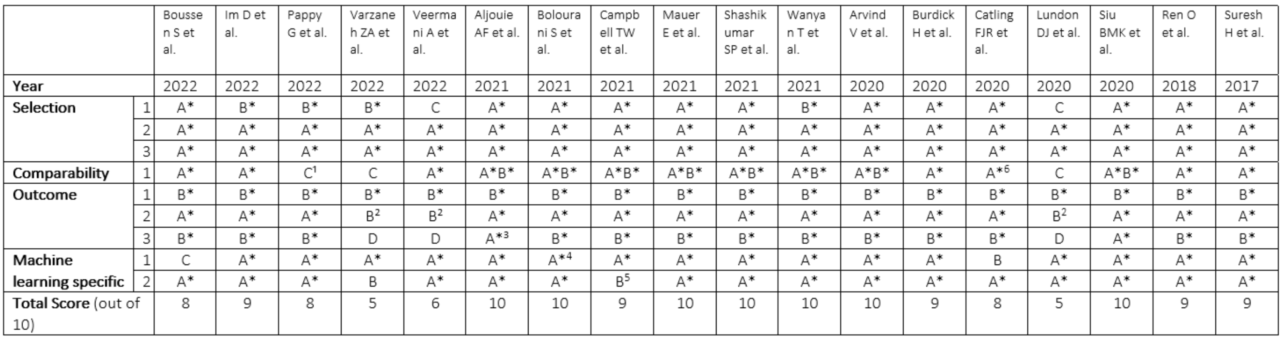
2) Performance measurement scale or statistical measure used to interpret results

   a) AUC or AUROC *

   b) Accuracy, predictive positive value or any other

The maximum score of this adjusted NOS is 10 stars.

B Results adjusted NOS



		Boussen S et al.	Im D et al.	Pappy G et al.	Varzaneh ZA et al.	Veermani A et al.	Aljouie AF et al.	Bolourani S et al.	Campbell TW et al.	Mauer E et al.	Shashikumar SP et al.	Wanyan T et al.	Arvind V et al.	Burdick H et al.	Catling FJR et al.	Lundon DJ et al.	Siu BMK et al.	Ren O et al.	Suresh H et al.
Year		2022	2022	2022	2022	2022	2021	2021	2021	2021	2021	2021	2020	2020	2020	2020	2020	2018	2017
Selection	1	A*	B*	B*	B*	C	A*	A*	A*	A*	A*	B*	A*	A*	A*	C	A*	A*	A*
	2	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*
	3	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*
Comparability	1	A*	A*	C[1]	C	A*	A*B*	A*B*	A*B*	A*B*	A*B*	A*B*	A*B*	A*	A*[6]	C	A*B*	A*	A*
Outcome	1	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*	B*
	2	A*	A*	A*	B[2]	B[2]	A*	A*	A*	A*	A*	A*	A*	A*	A*	B[2]	A*	A*	A*
	3	B*	B*	B*	D	D	A*[3]	B*	B*	B*	B*	B*	B*	B*	B*	D	A*	B*	B*
Machine learning specific	1	C	A*	A*	A*	A*	A*	A*[4]	A*	A*	A*	A*	A*	A*	B	A*	A*	A*	A*
	2	A*	A*	A*	B	A*	A*	A*	B[5]	A*	A*	A*	A*	A*	A*	A*	A*	A*	A*
Total Score (out of 10)		8	9	8	5	6	10	10	9	10	10	10	10	9	8	5	10	9	9

Figure 4: Results of the adjusted NOS for the included studies, the adjusted NOS has a maximum score of 10.

[1] Study is in children no control for gender and comorbities

[2] Not stated how long follow-up was

[3] MV and NIV added together as MV group but also separately documented

[4] Used 1 hospital as test set in every fold, but kept test and training separate within fold

[5] Used PPV, sensitivity and F1 score instead

[6] Groups were not compared, no comorbidities

C Background information on Machine Learning Model types

In this section more background information is given for the different model types.[29] In the first category of models data is given to the model in a table with on every row a different patient or measurement point, every column a different measurement type and one column that contains the desired outcome.

- Logistic Regression models
  Other than in a linear regression model, a Logistic Regression Model (LRM) gives a true or false as an outcome. An S-shaped curve classifies which samples are true and which are false. The cut-off is usually made at 50%. In a model that predicts intubation true would be intubated and false would be not-intubated. Both continuous and discrete data can be entered to classify samples. How the S-shaped curve is fit through the samples is determined with maximum likelyhood. The LRM assumes that the relationship between the predictor variable and the predicted outcome is linear. In case of repeated measures, which is the case in most medical data sets a mixed effects logistic regression model should be implemented. In Figure 5 an info-graphic of the LRM is shown.

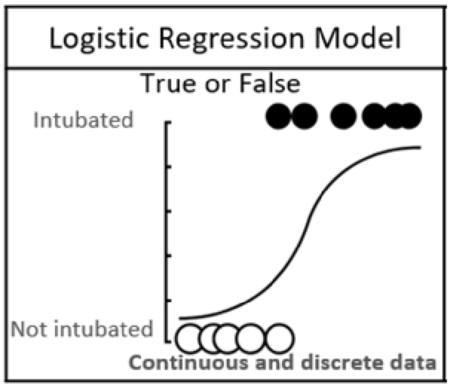


Figure 5: Info-graphic of the Logistic Regression Model

- Random Forest
  A Random Forest (RF) is a ML model that consists out of a ”forest” of decision trees. Decision trees alone are often trained to good on the existing data, which makes them inflexible to use on other data sets. Because an RF includes multiple decision trees it is more flexible and accurate on a different data set. When modeling an RF, the first step that is taken is to make a bootstrapped data set. This means that from the existing data set patients are selected randomly, this process will lead to patients randomly not being selected. These out-of-bag samples will become the validation set. With the bootstrapped data set random decision trees are formed, together the decision trees become the RF. The accuracy of the RF is then tested using the validation set. A new RF is then build with different variables/features used per step. The RF model with the best accuracy and a determined number of variables per step is chosen. In Figure 6 an info-graphic of the RF is shown.

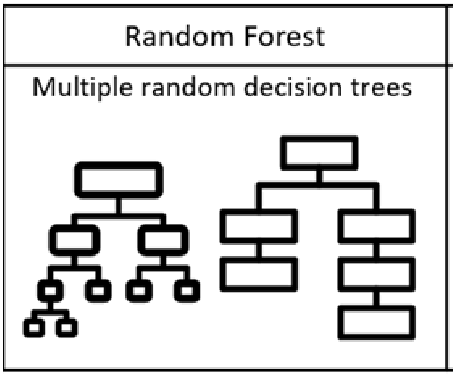


Figure 6: Info-graphic of the Random Forest

- Gradient Boosting Model
  A Gradient Boosting model (GBM) is a ML model that also consist out of multiple decision trees. It builds fixed sized trees based on the previous tree's errors. It is started with a leaf in which the log(odds) is imputed, in this case the log(odds) would be log(intubated/not intubated). Residuals (difference between observed and predicted values) are then used to build a new tree. This new tree than predicts the new residuals. With each tree the residuals become larger or smaller making the predictions more accurate. New trees are made till the maximum specified number of trees is reached or if adding a tree does not significantly reduce the size of residuals.[30] In Figure 7 an info-graphic of the GBM is shown.

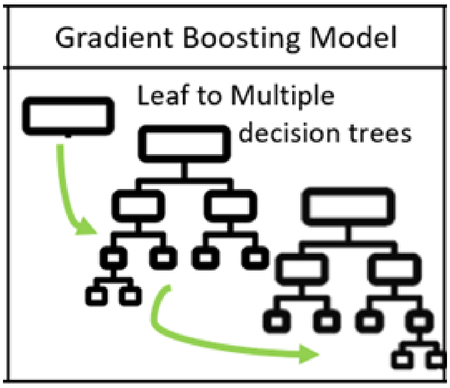


Figure 7: Info-graphic of the Gradient Boosting Model

The second category contains models that work with a more raw data set, in which it is possible to feed time series to the model. These models are also called deep learning models.

- Neural Networks
  A Neural Network (NN) allows multiple inputs and outputs. Between the input and outputs are hidden layers that contain nodes that connect all the layers with the input and next layer or output. The hidden layers contain activation functions that alter the input to form a graph. It is difficult to understand what happens in the hidden layers. In Figure 8 an info-graphic of the NN is shown.

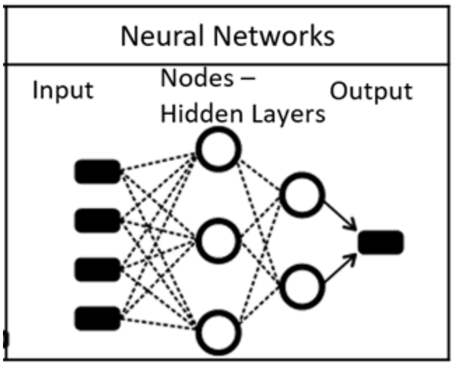
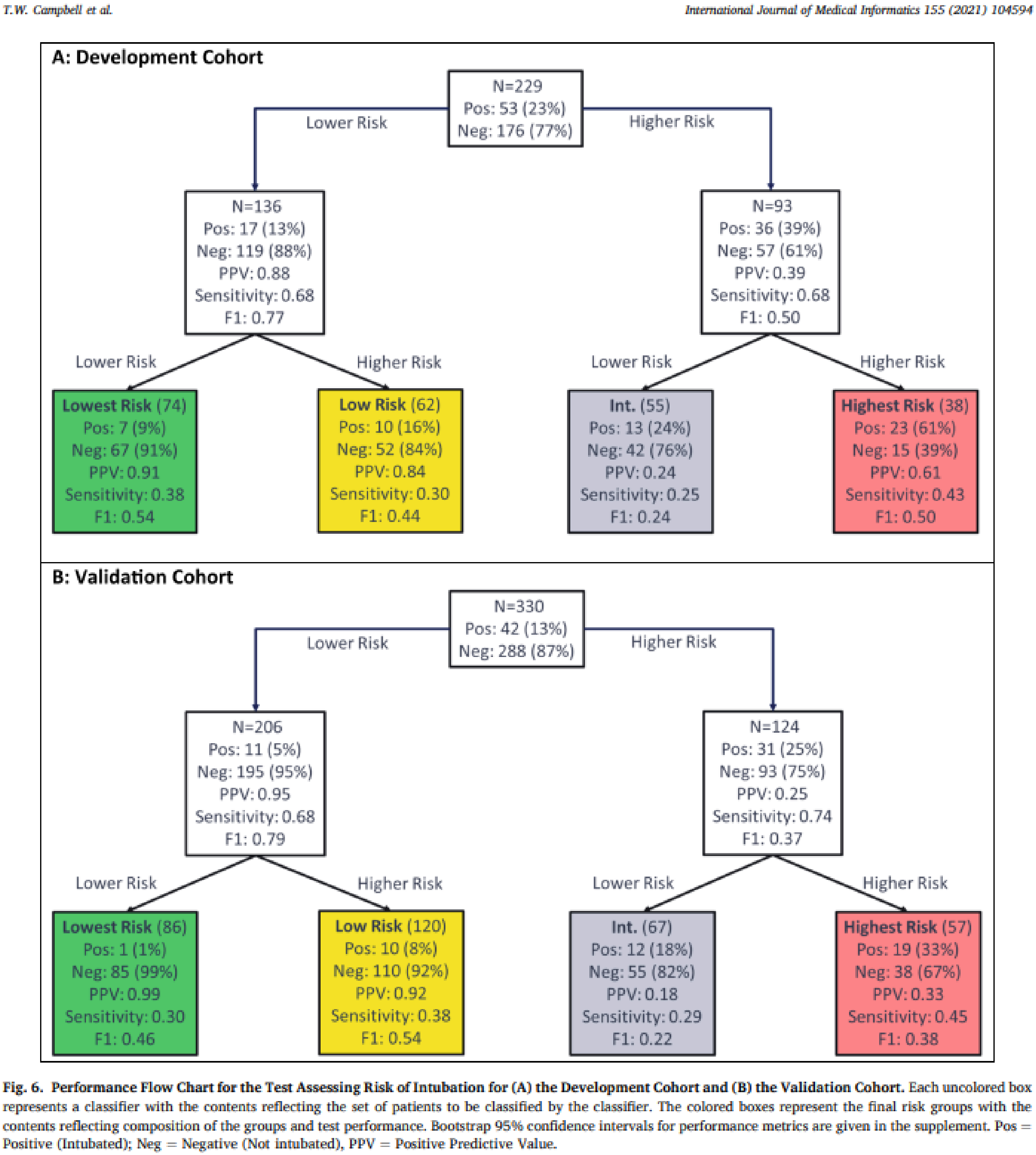


Figure 8: Caption

D Result section of Campbell TW et al. 2021 [15]



T.W. Campbell et al. International Journal of Medical Informatics 155 (2021) 104594

Fig. 6. Performance Flow Chart for the Test Assessing Risk of Intubation for (A) the Development Cohort and (B) the Validation Cohort. Each uncolored box represents a classifier with the contents reflecting the set of patients to be classified by the classifier. The colored boxes represent the final risk groups with the contents reflecting composition of the groups and test performance. Bootstrap 95% confidence intervals for performance metrics are given in the supplement. Pos = Positive (Intubated); Neg = Negative (Not intubated), PPV = Positive Predictive Value.

Figure 9: Snapshot from the article of Campbell TW et al. 2021 including the results for the intubation sub-research.[15]