Predicting postoperative complications after esophagectomy using machine learning algorithms

Marieke Reincke
Master Thesis

Supervisors:

Chairman:	Prof. Dr. Ir. H.J. Hermens
Medical supervisor:	dr. E.A. Kouwenhoven
Technical supervisor:	dr. M. Poel
Technical supervisor institution:	ing. J. Geerdink
External supervisor:	R.F.M. van Doremalen MSc
Process supervisor:	N.C. Cramer-Bornemann MSc

UNIVERSITY OF TWENTE.

zgt
Topzorg voor uw levenskwaliteit

Abstract

Background: Currently, surgical resection of the esophagus is the only curative treatment for a patient with non-metastatic esophagus cancer. Despite a substantial improvement in the survival of these patients, esophagectomy is burdened with high procedure-related morbidity. The most common and severe postoperative complications after esophagectomy are pneumonia and anastomotic leakage. To assist clinicians in the early detection of postoperative complications, machine learning models could support in detecting novel predictors and patterns of postoperative deterioration.

Objective: This research aimed to explore the ability of machine learning algorithms to predict major complications in patients who underwent esophagectomy by using structured and unstructured postoperative data.

Methods: Postoperative structured and unstructured data of patients who underwent esophageal resection for cancer were extracted from the electronic health record. These patients were divided into two groups, one reference group, group 0 and a group with patients who suffered from either pneumonia or anastomotic leakage, group 1. The structured postoperative data contained vital signs and laboratory tests. The unstructured data consisted of nursing assessments reports, which we converted to text features by using a bag of words model. Both the structured and unstructured data was used to predict postoperative complications, specifically anastomotic leakage and pneumonia, using logistic regression, support vector machines, decision trees and random forest.

Results: We identified 164 patients of which 112 patients belong to group 0 and the other 52 in group 1. When using structured data alone we predicted postoperative complications using random forest with an area under the receiver operating curve of 0.88, a sensitivity of 44% and a specificity of 94%. After the addition of text features, the AUC improved to 0.90 and the specificity increased to 97%, while the sensitivity decreased to 12%. The overall performance of all of our models did not improve when adding text features to the models.

Conclusions: This study revealed that machine learning models have an overall fair prediction of postoperative complications after surgery when using postoperative data, both structured and unstructured. Within these models, C-reactive protein and temperature are important predictors of anastomotic leakage and pneumonia. Furthermore, the potential of text features needs to be further explored to improve the prediction of postoperative complications after esophagectomy.

List of abbreviations

CRP:	C-reactive protein
ZGT:	Ziekenhuisgroep Twente
ML:	Machine Learning
EHR:	Electronic Health Record
SVM:	Support Vector Machines
BP:	Blood Pressure
HR:	Heartrate
RR:	Respiratory Rate
ICU:	Intensive Care Unit
***k*-NN:**	*k*-Nearest Neighbor
BoW:	Bag of Words
TF-IDF:	Term Frequency – Inverse Document Frequency
ROC:	Receiving Operating Curve
AUC:	Area Under Curve
LDA:	Latent Dirichlet Allocation

Table of contents

Chapter 1 – Introduction

Esophageal cancer is a prominent cause of cancer-related mortality across the world. The two predominant histological subtypes of esophageal cancer are squamous cell carcinoma and adenocarcinoma. In western countries, there is a preponderance of the adenocarcinoma subtype, which is related to an increase in the incidence of gastro-esophageal reflux disease, Barrett's esophagus and obesity. Other risk factors of adenocarcinoma are smoking, the male gender and the white race [1]. In the Netherlands, the incidence of patients diagnosed with esophageal cancer has increased over the past decades, from 814 patients in 1990 to 2.521 in 2019 [2]. Currently, surgical resection of the esophagus is the only curative treatment for a patient with non-metastatic esophagus cancer [3]. Traditionally, esophagectomy was performed using an open surgical procedure, but several randomized controlled trials have shown that minimally invasive esophagectomy decreases postoperative complications and increases the quality of life [4]–[6]. Esophagectomy is followed by reconstruction surgery to restore intestinal continuity. The stomach's vascularity is very rich and is therefore suitable as a conduit for esophageal reconstruction. There are two approaches to accomplish digestive tract reconstruction, namely the gastric tube and the whole-stomach approach, of which the gastric-tube approach is superior [7]. Furthermore, there are various techniques devised to design an anastomosis of the esophageal remnant with the stomach e.g. end-to-side, side-to-side, hand-sewn or mechanically-sewn. However, there is insufficient scientific evidence to point out the pre-eminence of certain techniques [8].

Esophagectomy is a high-risk procedure and despite a substantial improvement of the survival due to multimodality therapy and centralization of care, postoperative complication rates remain high, around 60%, even in renowned centers of expertise [3]. The most common and severe postoperative complications are pneumonia and anastomotic leakage [9]. The pathophysiology of postoperative pneumonia after esophagectomy is associated with the patient's age and comorbidities, postoperative pain, atelectasis, aspiration and postoperative ventilatory requirements [10]–[12]. The most important predisposing factors for anastomotic leaks are ischemia of the gastric conduit, impairment of oxygen delivery and errors in surgical techniques. During esophageal reconstruction, both venous drainage as arterial supply is sacrificed, which has a negative effect on the healing of the anastomoses and could result in a leak [13]. These postoperative complications not only exert an ongoing negative impact on the quality of life but are also related to unplanned readmission [14], [15]. The onset of postoperative complications is not well understood and the recognition of these complications can be challenging due to variations in patient's response to a complication and varying levels of a clinician's experience. As a result, postoperative complications also occur post-discharge provoking unplanned readmissions and re-operations [15]. The 30-days readmissions rate is reported in several studies and lies between 11% and 18% [15]–[18]. Hospital readmissions are associated with worse long term survival and a substantial increase in healthcare costs [17]. Therefore, early detection of postoperative complications is important, not only to prevent or manage a complication but also to avoid premature discharge after esophagectomy. The utility of predictors of anastomotic leakage or pneumonia, such as C-reactive protein (CRP) [19]–[24], the neutrophil/lymphocyte ratio [24], [25] or drain amylase levels [26]–[28] can be helpful to early recognize postoperative complications. In the Ziekenhuisgroep Twente (ZGT) in Almelo the Netherlands, surgeons already utilize such laboratory measurements together with vital signs and the overall condition of the patient to assess whether the patient is suffering from an anastomotic leakage or pneumonia. In addition to these predictors, the use of machine learning (ML) algorithms could enable precise prognostication and risk stratification of patients who underwent esophagectomy. Furthermore, ML techniques could identify patterns in data that are yet unknown by combining such predictive values [29]. Ultimately, a ML model could assist in identifying patients at risk for postoperative complications, which will support clinicians in making actionable decisions to diagnose or manage a postoperative complication.

To follow up on the current research within this field, we want to use postoperative parameters together with ML techniques to early predict postoperative complications. Examples of such postoperative parameters are the number of leukocytes or heart rate, which are measured on daily basis. In earlier exploratory research these postoperative features showed to be of high value to predict major postoperative complications. In addition to these postoperative parameters, unstructured data, i.e. free-text data, could improve ML models, since it contains valuable information, nuances and context. Most of the collected data in healthcare are unstructured, e.g. medical prescriptions or radiological assessments. Extraction and analysis of unstructured data are more challenging compared to structured data because it comes from different sources and is more variable and heterogeneous [30]. However, unstructured data can play a key role in the prediction of postoperative complications, especially in a clinical setting where the patient's well-being is not easily captured in structured data.

The aim of this research is to explore the ability of machine learning algorithms to predict major complications in patients who underwent esophagectomy by using structured and unstructured postoperative data from electronic health records (EHR). In this research, we focus on the occurrence of anastomotic leakage and pneumonia, because these are the most common severe postoperative complications for esophageal cancer patients.

Chapter 2 – Related work

Predicting postoperative complications is difficult, not only after esophagectomy but in all surgical areas. Talmor et al. [31] summarized the scoring systems that have been researched for predicting postoperative complications. They have discovered that most predictive scores and models are developed from administrative databases and have at best moderate discriminatory value. Furthermore the heterogeneity of the study population makes it challenging to utilize such models to specific surgical subgroups. However, they have found the following risk factors of postoperative morbidity: increasing age, frailty, poor cardiorespiratory reserve and chronic renal failure. Recently, ML algorithms are increasingly used to predict postoperative complications after surgery in general or after different types of surgery [29], [32]–[35]. Some of these studies used clinical data from electronic health records (EHR), while others relied on national quality registrations. The ML models that were often utilized are logistic regression [29], [34], support vector machines (SVM) [29], decision trees [32] and random forest [29]. ML is gaining interest in healthcare for its ability to learn from large datasets without being explicitly programmed. These algorithms are able to recognize statistical patterns from large sets of data by combining different features, which is an impossible performance for humans. Furthermore, ML models can learn patterns from patient data incredibly fast, which could help physicians draw information from the experience of such models [36]. An example is the study of Bronsert et al. [35], who developed a ML model for the surveillance of patients who underwent surgery by using EHR data. They constructed a model that correctly classified 83% of the patients with a postoperative complication. However, the generalizability of this model is limited as they used data from one hospital. ML techniques have not been used yet to predict postoperative complications after esophagectomy. Only one study, performed by Bolourani et al [37], used ML algorithms to predict early readmission after esophagectomy. Although Bolourani et al showed promising results, the ML model they constructed was based on a national readmission database, which does not contain granular information about the patient's daily, postoperative condition. For example, this national readmission database records the diagnosis of sepsis but does not capture the levels of leukocytes or core temperature of a patient. Consequently, this model is not yet ready to be adopted in clinical practice, since it is not a head-to-head comparison against the judgment of an experienced physician [38].

Postoperative complications, especially post discharge, drive unplanned readmissions after esophagectomy [15]. As a result, surgeons are reluctant to discharge patients after esophagectomy. Safe discharge criteria after esophagectomy are therefore of high value. Müller et al [39] aimed to achieve international consensus on safe hospital discharge criteria after esophagectomy using the Delphi methodology. An international expert panel found agreement on nine criteria to determine a 'fit-for-discharge' status [39]. The general domains of the criteria were: vital signs, laboratory tests, wound status, drains and catheters, pain control, recovery of respiratory function, restoring bowel movement, upper gastrointestinal symptoms, tolerance on nutrition and mobilization and selfcare. These criteria need to be validated and are not implemented in clinical practice yet, but could assist in the decision-making regarding a patient's discharge. Furthermore these criteria could also be used as indicators to rule out postoperative complications. Other studies have already shown the importance of laboratory values [19]–[28], physical activity [40] and vital signs [41] as prognostic indicators of severe postoperative complications after esophagectomy.

Another emerging data source to predict outcomes after surgery is unstructured data. Other studies already proved the significance of text data from the EHR, e.g. clinical notes and radiologist reports, to discover patterns and topics to support data-driven decision-making [42], [43]. Particularly in a clinical setting where the symptoms of deterioration are not easily captured in structured data, text data could be of added value. For example, Barber et al [44] used unstructured data in a retrospective study to predict postoperative complications after ovarian cancer surgery. They used natural language processing to add preoperative CT scans, which led to an increase of 20–25% in the ability to predict postoperative complications. They utilized a BoW model to convert free written text into features.

Another approach to use clinical text data is shown by Goh et al [43]. They utilized Latent Dirichlet Allocation, which is a topic model that generates topics, based on patterns of word frequency from a set of documents.

Based on the related work discussed in this chapter, we expect that ML technique have great potential to predict postoperative complications after esophagectomy. Models that are most often utilized in this field are logistic regression, SVM, decision trees and random forest. Features that could play an important role in such models are laboratory test, especially CRP levels, and vital signs. Moreover, unstructured data could add value to the prediction of postoperative complication as it could explain more about the general patient's wellbeing.

Chapter 3 – Description of dataset

In this chapter we describe the contents and characteristics of our dataset. In section 3.1 we explain how we included and labelled our patient population. Thereafter, in section 3.2, we describe the data we extracted from the EHR of these patients, both the structured data and unstructured data.

3.1 Study population

In our institution, laparo-thoracoscopic minimally invasive esophagectomy was introduced in 2010 and became the standard surgical procedure for esophagectomy. For that reason, we only included patients who underwent surgery from 2010 in this research. In total, 412 patients were selected for this study. All of these patients underwent esophagostomy for esophageal cancer from December 2010 to December 2020 in ZGT Almelo. Only one patient was excluded due to the patient's death on the day of surgery. After patient selection, we divided the patients into a reference group (group 0) and a complication group (group 1). Cases that could not be grouped into any of these categories were excluded.

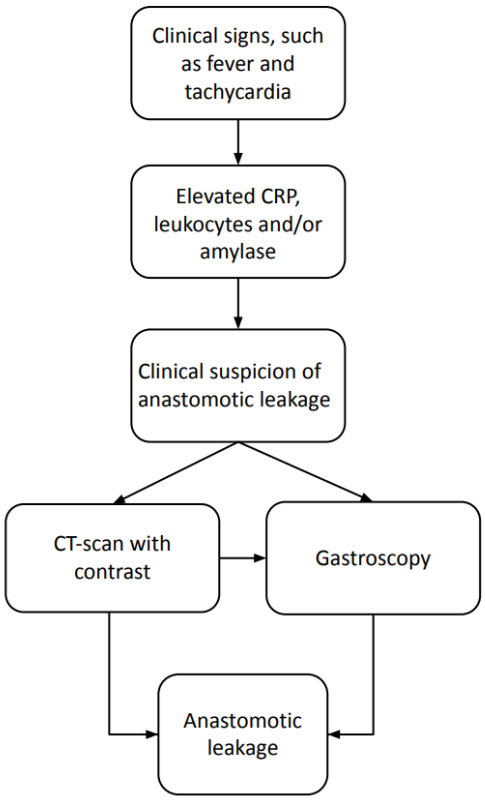


***Figure 1** – Flowchart of most common diagnostic approach of anastomotic leakage in patients who underwent esophagectomy.*

Patients with a major complication were appointed to group 1. In this study, we focused on the most common and severe complications, namely anastomotic leakage and pneumonia. We combined the two complications in one group as these are both inflammatory diseases. Of our patient population, 40% was diagnosed with either anastomotic leakage (83) or pneumonia (117). These complications were manually registered by an experienced clinician, who used the Utrecht pneumonia scoring system to record patients diagnosed with pneumonia. To ascertain the diagnosis of anastomotic leakage the clinician used the diagnostic pathway depicted in figure 1, which shows the most common ways to diagnose anastomotic leakage. Both the diagnosis and the date of diagnosis of these complications in these patients were documented.

The reference group consisted of patients who were not diagnosed with major complications, such as pneumonia or atrial fibrillation. We did not register each possible postoperative complications in our data acquisition, since that is a highly time-consuming task. Therefore, we could not guarantee that the patients in group 0 experienced no complications at all during hospitalization, e.g. a wound infection could still have occurred in patients from this group. Therefore, we only included patients whose admission duration was according to our postoperative recovery protocol. In 2017 a fast-track protocol after esophagectomy was implemented in our hospital and since then patients with a normal postoperative recovery were generally discharged after eight days. Before November 2017, patients without complications were usually hospitalized for ten days. Accordingly, we selected patients with an admission duration of maximal ten days if hospitalized before November 2017 and patients with an admission of eight days or less after November 2017.

3.2 Data extraction and preparation

From the patients in this study, both structured and unstructured clinical data were collected from the electronic health record, from the first day after surgery (day 1) until two weeks after surgery (day 14).

3.2.1 Structured data

The structured data consisted of the vital signs, i.e. heart rate (HR), systolic blood pressure (systolic BP), respiratory rate (RR) and temperature and laboratory values, namely leukocyte count, CRP and amylase levels. After esophagectomy, patients are initially admitted to the intensive care unit (ICU), where vital signs are mostly monitored continuously. When a patient is transferred to the surgical ward, vital signs are manually measured several times per day depending on the patient's clinical well-being. In this study, we combined these measurements, besides their difference in acquisition. For example, at the ICU the HR is measured using electrocardiographic electrodes, while at the ward a pulse-oximeter is used. We chose to combine the measurements to reduce the number of missing values and because the clinical interpretation and application is similar. A bulk extraction from the EHR was done of each day in our postoperative window, for both the continuous vital data and the manual vital data. Since the vital signs are measured continuously at the ICU, we would receive an enormous amount of data if we would extract all the data available. To avoid this, we chose to collect one measurement per hour of each day, resulting in a maximum of 24 measurements per day for each vital sign. For the manual vital data, we extracted a maximum of 20 measurement per day, since vital signs are measured usually two to three times per day at the ward. Ultimately, after we concatenated the continuous vital data with the manual data, our dataset consisted of a maximum of 44 measurements per day for each vital sign. We filtered these measurements to remove physiologically unrealistic values. In table 1 the cut-off values of each vital parameter that we used to filter the vital sign measurements are shown. We decided to not use all these measurements, because we would have too many missing values. We retained three measurements per day in different time windows. We divided each postoperative day into three time intervals: 1) from midnight to 8 am, 2) from 8 am to 4 pm and 3) from 4 pm to midnight. These windows are based on the daily shifts of the nurses in which they often measure the vital signs. In each interval, we preserved the value that occurred at the earliest time within this interval. In this way we expected roughly the same amount of time between the three values of each day. For the laboratory measurements, we extracted a maximum of 3 measurements per day and preserved the most abnormal value of each day in our dataset. For leukocytes, CRP and amylase, the highest value is the most deviating value. This resulted in one measurement per day for each laboratory measure.

Table 1 – Cut-off values of vital parameters

Heartrate	30 – 220 beats/min
Systolic blood pressure	40 – 200 mmHg
Respiratory rate	6 – 40 breaths/min
Temperature	34 – 42 ^{0}C

3.2.2 Unstructured data

The unstructured data consisted of daily reports of the nurses from the ICU and the surgical ward, written in Dutch. In our institution, the nursing assessment of the patient is documented according to Gordon's functional health patterns. Two patterns of this system were used in this study, namely health perception and activity. In addition, the general information and observations reported by nurses were also included. We chose these in particular because we expected to find indications of clinical deterioration, specifically signs of anastomotic leakage or pneumonia, in these patterns and reports. We extracted all the available text data from these particular forms from the first day after surgery until two weeks after surgery. When we found multiple reports of the same form on the same day for one patient, we concatenated the text together. By doing this, we preserved one text document of each form on each day for every patient.

In table 2, an overview is shown of the entire dataset used in this research. This table shows the number of measurements we collected from the EHR and which we preserved in our dataset.

Furthermore we added the percentage of missing values per parameter, which we further explain in the next section.

Table 2 *– Overview of the dataset used in this study. For each parameter the number of measurements we extracted from the EHR are shown. Furthermore the number of datapoints we selected from the extraction into our dataset and the method of selection is given. The percentage of missing values is also shown for each parameter in our dataset. For the vital parameters we selected one unit for each of the three predefined time windows from the extracted data. For the laboratory parameters we selected the most deviating value for our dataset from the three measurements we extracted. In regards to the unstructured data, we extracted all the available data from the EHR and concatenated the reports if more than one report of one day for a particular patient was found.*

	Maximum number of datapoints extracted	**Selection method**	**Number of datapoints selected for dataset**	**Total of missing values**
	per day		per day	%
Structured data:				
HR	Continuous: 24 Manual: 20	Time window*	3	42.4
Systolic BP	Continuous: 24 Manual: 20	Time window*	3	45.2
RR	Continuous: 24 Manual: 20	Time window*	3	68.8
Temperature	Continuous: 24 Manual: 20	Time window*	3	43.8
Amylase	3	Most deviating value	1	50.6
Leukocytes	3	Most deviating value	1	34.6
CRP	3	Most deviating value	1	34.9
Unstructured data:				
Activity	All available	Concatenation	1	49.2
General information	All available	Concatenation	1	48.4
Health Perception	All available	Concatenation	1	58.6
Observations	All available	Concatenation	1	64.4

** We divided each postoperative day into three time windows: 1) from midnight to 8 am, 2) from 8 am to 4 pm and 3) from 4 pm to midnight and selected one measurement per time window.*

3.2.3 Missing values

Missing values are inevitable and ubiquitous in clinical research. During postoperative recovery, clinical parameters are measured more frequently during the first days after surgery, especially at the ICU, compared to the end of admission. As a result, our dataset contains missing values, of which the extend grows towards discharge. In table 2, the percentage of missing values of each variable in the dataset is given. Note that there is no postoperative data available after discharge, so when a patient is hospitalized for less than two weeks, which is often the case, there are many missing values. As a result the percentage given in table 2 shows a distorted view on the actual missing data during admission. Therefore we created missing plots to give more insight in the occurrence of missing values for every

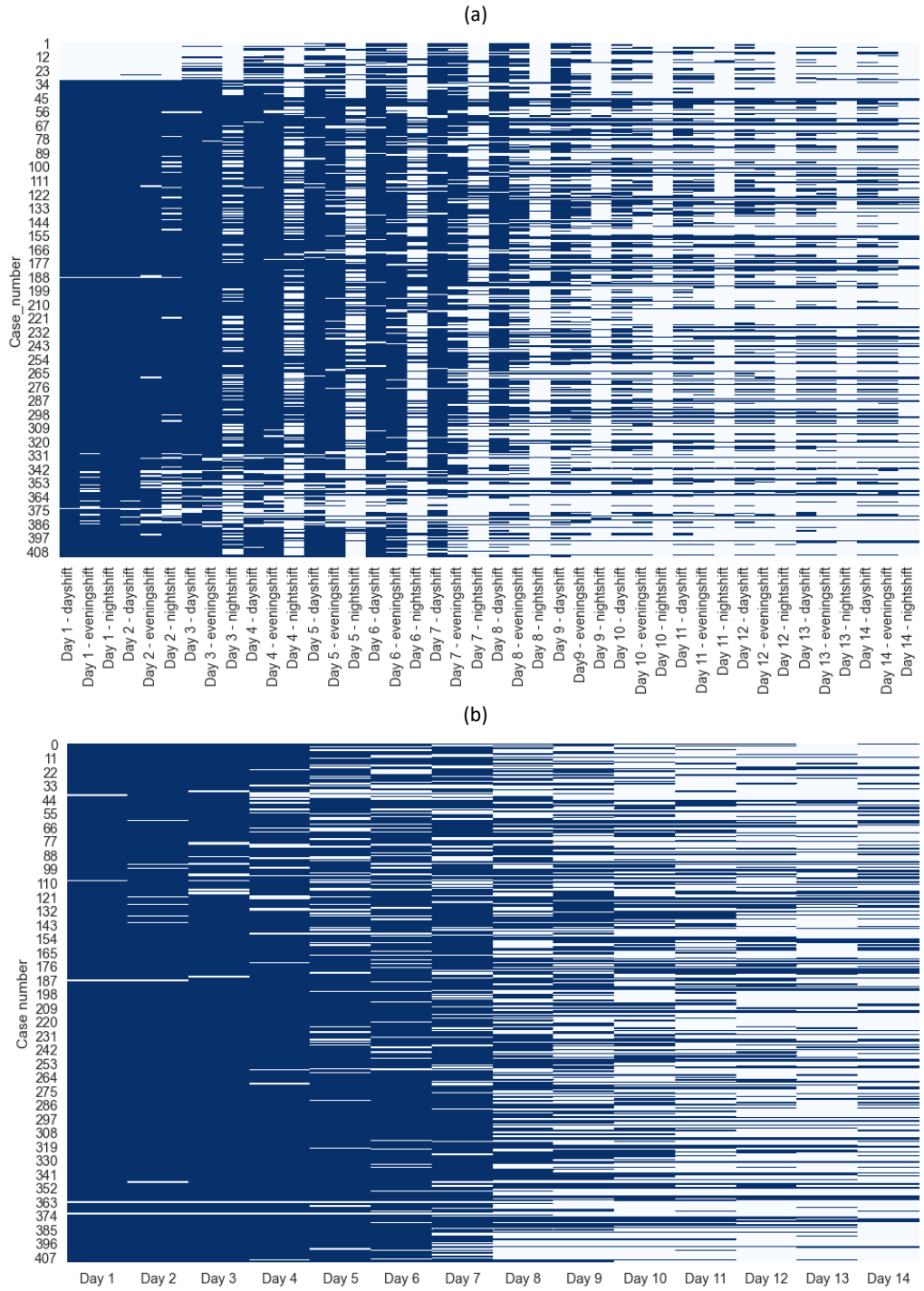


Figure 2 *– Mapping of missing values of (a) heart rate and (b) leukocytes count of each patient in the dataset. Every blue line represents a unit that is present in our dataset, while white lines represent missing values.*

parameter. In figure 2a, an overview of the incidence of missing values in the HR data is depicted. This plot shows that the number of missing values increases over time resulting in a more fragmented appearance of the data. This is due to the continuous monitoring at the ICU in the first few days after surgery. Interestingly, this plot shows a gap in the data in the first few days in the first cases of our study population. This is most probably the result of the use of different ICU monitors in that year, which did not export the data to the EHR of these patients. The systolic BP and temperature resemble the same pattern of missing values, which is plausible as vital signs are mostly measured all together. These plots are shown in appendix 1. However, the number of missing values is significantly higher in the respiratory rate data and shows fragmentation earlier in the postoperative window. In addition,

the number of missing values in the leukocyte data are shown in figure 2b and illustrates that this parameter is measured on a daily basis. The CRP data looks very similar to figure 2b, while the amylase data shows a higher number of missing values and is measured less regularly, see also appendix 3.
Missing data need to be handled because many ML models do not support missing data, so imputation of missing values is a necessity. Another reason for imputation is to obtain an improved estimate of the underlying distribution of each variable. In other words, with imputation techniques we aim at population level and not at the level of a single unit. In this study, we used two imputation techniques, namely *k*-nearest neighbor (*k*-NN) imputation and multiple imputation. *k*-NN imputation was applied to CRP levels and leukocytes count. In *k*-NN missing values are estimated by values extracted from other similar cases. This could be either a value from one other case or the (weighted) average of *k* other cases. The advantages of NN are that imputed values are actual values that occurred in the dataset. This makes this method very suitable for CRP levels and leukocytes count, as the course in these variables is important to maintain. We used four nearest neighbors as samples to use for imputation, which were weighted equally. For HR, RR, systolic BP, temperature and amylase we used multiple imputation. Multiple imputation is considered to be a very powerful technique. This technique provides an unbiased and valid distribution of the data based on the available data and is often used to deal with missing data. The unstructured data contained missing values as well, even some patients had no written report at all. However, there are no imputations techniques available to impute this particular type of data. We solved this by using a bag of words model which uses zeros when text data is absent, this is further explained in the next chapter.

Chapter 4 – Materials and methods

In this chapter, we focus on the materials and methods we used to conduct our research. In section 4.1 we describe the design of this study. Subsequently, in section 4.2 we explain the selection of features to train the ML models, followed by section 4.3 which explains the machine learning algorithms used in this study and lastly, in section 4.3, the software that we applied is specified.

4.1 Study design

This study is a single-center retrospective cohort study regarding patients who underwent esophagectomy for esophageal cancer in ZGT Almelo, the Netherlands. As described in the previous chapter, we extracted postoperative data from these patients to construct a ML model to predict pneumonia or anastomotic leakage after esophagectomy. Input features used for training include vital signs (HR, systolic BP, RR and temperature), laboratory measures (amylase, CRP and leukocytes) and nursing assessment reports (activity, general information, health perception and observations). Furthermore, we calculated the increase in CRP and leukocytes between each day as additional features, by subtracting the value measured on each day by the value measured on the previous day.

4.2 Text features

Before features could be derived from the nursing assessments, pre-processing of the raw text was necessary. Text pre-processing consisted of lowering text and removal of punctuation and digits. Thereafter the text was tokenized, which means that the text was separated into (word)-tokens, which was followed by stopwords removal. The list of stopwords that we used is specified in appendix 2. Examples of commonly used stopwords are '*dat*', '*het*' or '*toen*'. In addition, a stemming algorithm, namely the Kraaij-Pohlmann stemmer, was used to reduce words to their stem. However, this stemmer made drastic reductions of words, making them difficult to read and was therefore turned off.

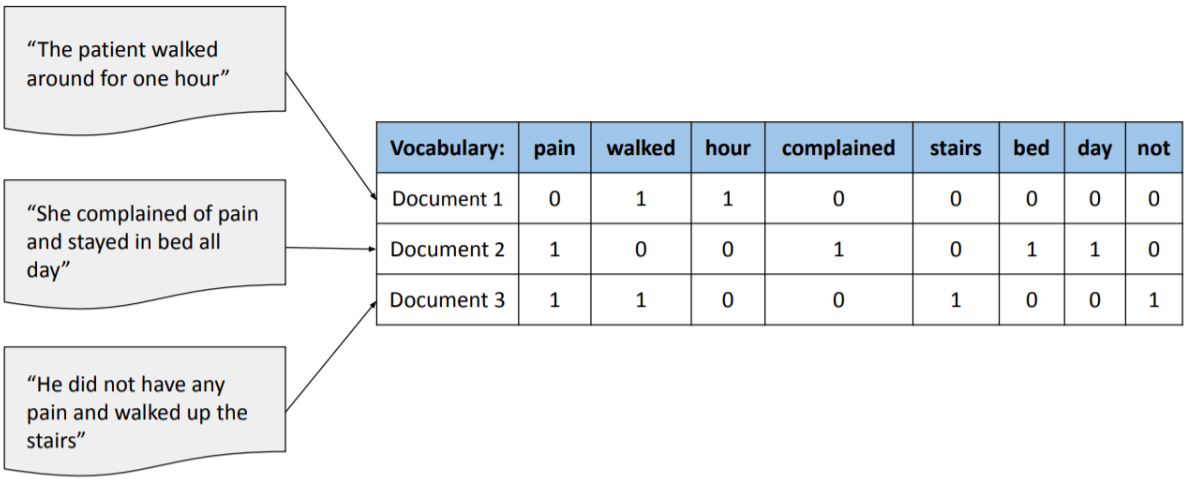
After pre-processing, the text data was further inspected by creating word cloud plots. Such plots visualize the most frequent words that occur in a certain text and could assist in familiarizing with the text content and finding the most discriminating tokens. To create these plots, the text data of each nursing assessment form were divided into a reference group and a complication group. Thereafter, the reports of each patient within each group were combined and converted into a word cloud. For every group and form, a word cloud of unigrams and bigrams was created.

4.2.1 Bag of words model

To extract features from the unstructured data a so-called bag-of-words (BoW) model was used. This model describes the occurrences of words (unigrams) or short phrases (n-grams) from a certain vocabulary within a document. By calculating the word frequencies, a frequency vector is obtained of each document within a collection of documents, named the corpus. A visual example of a BoW model is depicted in figure 2. The advantage of this method is that it is an easy, flexible and fast way to analyze and apply unstructured data. The disadvantage is that this method ignores the context in which words occur. To overcome this disadvantage, bigrams were included as well.

To use a BoW model, a vocabulary of unique n-grams is defined first. The size of the vocabulary is very important. When the size is too small, the vocabulary will lack discrimination. When the vocabulary is too elaborate, not only the computation load will increase dramatically, but also the number of features will grow significantly, which will affect model performance negatively. There are no guidelines to select the right size of the vocabulary. Therefore, we experimented with different sizes of vocabularies: namely 40, 80 and 200. The performance of these different vocabularies was evaluated during training and testing of the machine learning models. To select uni- and bigrams that indicate anastomotic leakage or pneumonia, we chose the top 40, 80 and 200 most frequent unigrams and bigrams of each form that were reported at the day the complication was diagnosed (i.e. we

created a dictionary of size 40, 80 and 200 for activity, health perception, general information and observations based on the day the complication was diagnosed).



Vocabulary:	pain	walked	hour	complained	stairs	bed	day	not
Document 1	0	1	1	0	0	0	0	0
Document 2	1	0	0	1	0	1	1	0
Document 3	1	1	0	0	1	0	0	1

Figure 3 *– An example of a bag of words model to create text features. First, the sentences are split into so-called tokens, single words. Then the frequency of each term of the vocabulary is determined per document. This results in a vector of term frequencies of each document based on the same vocabulary, that can be used as a text feature.*

Another disadvantage of using BoW is that it only represents the term frequency, which does not reflect the importance of a word. Therefore we calculated the term frequency - inverse document frequency (TF-IDF), which identifies the importance of a term within a document. Words that occur often in every document have less value, e.g. and, not, it, etc. TF-IDF will prevent that words with a high term frequency will dominate the vocabulary. After scoring the frequencies of uni- and bigrams, we calculated the TF-IDF of each term within each document according to the following equation:

$$w_{i,j} = tf_{i,j} \cdot log\left(\frac{N}{df_i}\right) \quad (1)$$

$$w_{i,j} = \text{TF} - \text{IDF score for term } i \text{ in document } j$$
$$tf_{i,j} = \text{frequency of term } i \text{ in document } j$$
$$df_i = \text{number of documents containing term } i$$
$$N = \text{total number of documents}$$

4.3 Feature selection

In this study, we compared the recovery from esophagectomy between the patients in group 0 and group 1 to predict postoperative complications. In the complication group, the postoperative day of the diagnosis of pneumonia or anastomotic leakage varied among patients, which is visualized in figure 3. Both pneumonia and anastomotic leakage occurred most frequently on day two and three. To create an accurate comparison it is important to acknowledge that patients, whether they are in group 0 or 1, are in a different condition a week after surgery compared to directly after esophagectomy. As a result, it is not fair to compare the data of e.g. a patient with a complication on day six to the data of a patient from group 0 on day two. Moreover, the feature vectors of a ML model need to be of the same size, therefore we have to set a limit to the amount of postoperative data we use of each patient. To do this, we focused on the initial four days after esophagectomy and selected the data until one day before a complication was diagnosed. We decided to use a feature size of two consecutive days, which means that we could include patients who were diagnosed with pneumonia or anastomotic leakages on day three, four and five after surgery. In other words, when a patient was diagnosed on

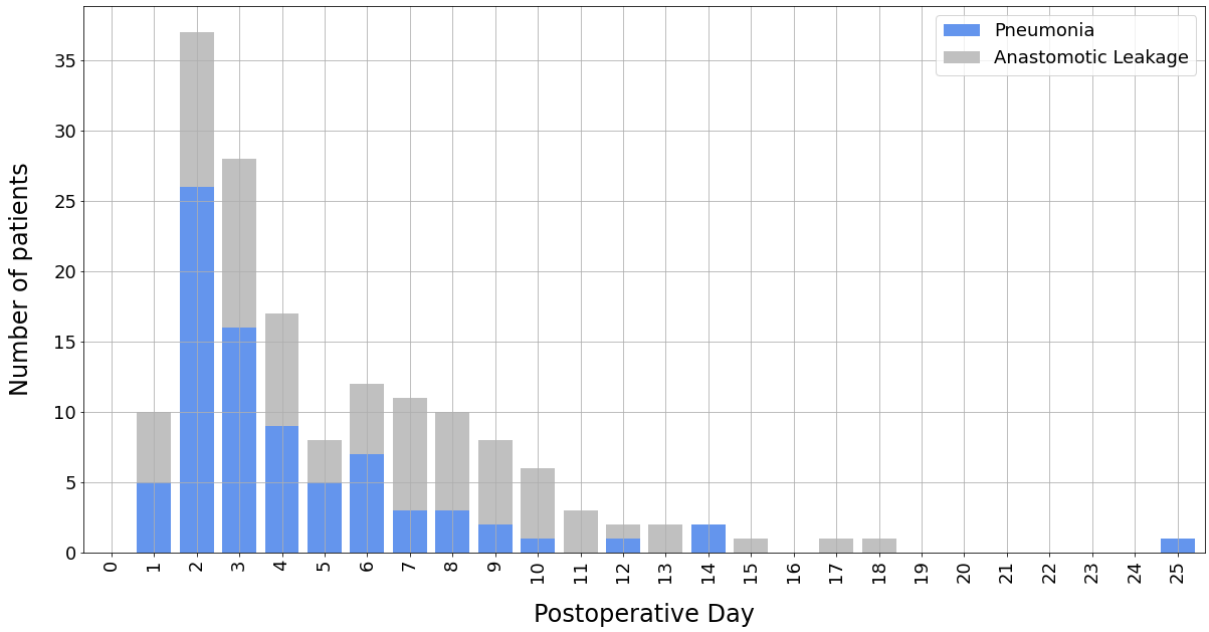


Figure 4 *– Number of patients diagnosed with pneumonia or anastomotic leakage per day after esophagectomy. Day 0 is the day of surgery.*

day four, the postoperative data of day two and three were selected as input. This is also visualized in figure 5. For the reference group 0, we randomly selected data of two consecutive days, either day one and two, day two and three or day three and four.

To train our ML models different combinations of features were used to discover the most important variables in predicting postoperative complications. At first, we used a dataset consisting of only the structured data (vital signs and laboratory values). Thereafter we added the BoW vectors and/or the TF-IDF vectors of a dictionary of different sizes (40, 80 and 200).

4.4 Machine learning algorithms

The input features were divided into a trainingset (70%) and a testset (30%). We used a stratified split to ensure a consistent distribution of patients from group 0 and 1 in both test- and trainingset. In this research, we utilized random forest, decision trees, logistic regression and support vector machine models to predict complications after esophagectomy. The settings of these models are further specified in appendix 3. The features of logistic regression and SVM were scaled, which was not necessary for decision trees and random forests. To establish a balanced dataset between class 0 and 1 an under-sampling method named NearMiss [45] was evaluated as well. This could be beneficial to the sensitivity of the model.

To evaluate the relevance of each feature in these models, we estimated the feature importance of each feature, structured and unstructured, after the models were trained and

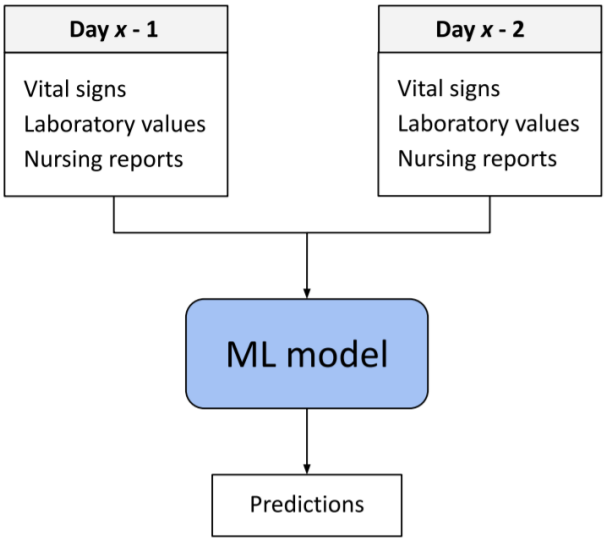


Figure 5 *– Diagram of the input and output of each ML model in this study. The postoperative data of two consecutive days of each patient from the trainingset is the input of the ML model to predict the outcome. Day x is defined as follows:*

$$x = \begin{cases} \{3,4,5\} & for\ group\ 0 \\ day\ of\ complication & for\ group\ 1 \end{cases}$$

tested. For logistic regression and SVM, we used the coefficients of these models to assess the importance of each feature. Coefficients of a regression model are the numbers by which the feature values are multiplied in the model equation and describe the magnitude and orientation of the relation between the independent variable (feature) and the dependent variable (group 0 or group 1). For the decision tree and random forest models, we used the impurity decrease to determine the importance of each feature.

The ML models were validated by using the test set. The performance of the models was evaluated using the receiver operating characteristic (ROC) curve and its area under de curve (AUC). Furthermore, the accuracy, precision, sensitivity (recall) and specificity were calculated by using the confusion matrix of the test set. For more information, appendix 4 shows the confusion matrix and equations of these metrics. We also calculated these metrics for the training set, to evaluate whether our models are overfitting. Overfitting occurs when a model is too complex for the amount of noisiness in the training data, which results in overgeneralization. This can be recognized by a much better performance of the trainingset compared to the testset.

4.5 Software and hardware

The structured data was extracted directly from our EHR (HiX, Chipsoft) using Microsoft SQL Server Management Studio, version 18.5.1, combined with Microsoft Visual Studio, version 16.9.4. After these extractions, the quality of the data was manually evaluated in the EHR application by cross-checking random samples in the data. The unstructured data was extracted from the EHR by using CTcue version 3.5.7 (CTcue B.V. Amsterdam, the Netherlands). This is a software tool developed to enable easy and custom data extraction from the EHR. Pre-processing of the data, feature generation and model construction were computed in Python version 3.9.0 (Python Software Foundation, Wilmington Del).

Chapter 5 – Results

In this chapter, we present the results we obtained in this research. We first discuss the population characteristics in section 5.1 and the results of the text analysis in section 5.2. Thereafter we present the performance of the ML models and the most important features to predict postoperative complications in section 5.3.

5.1 Population characteristics

From the dataset, described in chapter 3, 164 patients were selected to train and test our models. The majority of this population were men (79.1%) and the median age was 66. From this population, 112 patients were divided into the reference group, group 0. Group 1 contain the other 52 patients who suffered from anastomotic leakage (23) or pneumonia (29) in the first three to five days after esophagectomy.

5.2 Text analysis

Word clouds were generated from each nursing assessment form, for both the reference and complication group. In figure 6 word clouds of unigrams and bigrams are shown of the health perception report of the complication group. For these plots, we only used the reports written on the day the patient was diagnosed with either pneumonia or anastomotic leakage. This displays the vocabulary that is used by nurses to describe the patient's health perception while suffering from a major inflammatory complication. These word clouds reveal that pain and how the patient is feeling, e.g. short of breath, tired or good, are the most important themes that are discussed in the health perception reports. Furthermore, pain medication and vital signs such as temperature and oxygen requirement are discussed in these reports. Frequently used unigrams, such as 'goed' and 'pijn', can be used in different ways to describe the patient's circumstances. Therefore, the bigrams are important to understand the context in which these words are recorded. For example, the word pain could be used to describe the absence of pain or to report an increase in pain complaints.

Table 3 *– Overview of performances outcomes of several ML models based on two different feature sets. Set 1 contains the vital signs and laboratory data. Set 2 is an extension of set 1 with the TF-IDF vectors of the nursing assessment reports based on a vocabulary of 40 uni- and bigrams. The values in brackets are the performance outcomes for the trainingset to assess whether the models are overfitting.*

		Accuracy	***Precision***	***Sensitivity***	***Specificity***	***AUC***
Logistic regression						
	Set 1	0.80 (0.82)	0.80 (0.74)	0.50 (0.64)	0.94 (0.90)	0.78 (0.87)
	Set 2	0.80 (0.96)	0.67 (0.94)	0.75 (0.94)	0.82 (0.97)	0.84 (0.99)
SVM						
	Set 1	0.74 (0.81)	0.67 (0.75)	0.38 (0.58)	0.91 (0.91)	0.81 (0.86)
	Set 2	0.68 (0.89)	0.50 (0.90)	0.44 (0.72)	0.79 (0.96)	0.81 (0.94)
Decision Tree						
	Set 1	0.74 (0.82)	0.80 (1.00)	0.25 (0.42)	0.97 (1.00)	0.80 (0.84)
	Set 2	0.82 (0.85)	0.73 (0.79)	0.69 (0.72)	0.88 (0.91)	0.79 (0.84)
Random forest						
	Set 1	0.78 (0.92)	0.78 (1.00)	0.44 (0.75)	0.94 (1.00)	0.89 (0.97)
	Set 2	0.70 (0.82)	0.67 (1.00)	0.12 (0.44)	0.97 (1.00)	0.90 (0.97)

5.3 Model performance

Various features were evaluated to train machine learning models. In table 3 an overview is shown of the performance of four ML models when using two different feature sets. The first set is without the text data and the second set contains the TF-IDF vectors of a vocabulary of 40 words. Smaller and

larger vocabularies were also used in our analysis, but the model performance remained nearly the same. We also applied a NearMiss under-sampling algorithm, but this dramatically diminished the performance and was therefore turned off. In the brackets, the performance metrics of the training set are described. This shows that especially random forest and decision trees are prone to overfitting when using both our feature sets. Logistic regression and SVM show little overfitting when using only the structured data features, but do overfit when using the second set of features. Interestingly, most models have a rather low sensitivity against a high specificity. For example, the SVM model trained and tested with the first feature set shows a sensitivity of 38% and a specificity of 91%. When adding the text features in the second feature set, the sensitivity increase in every model, except for random forest. However, this is at the expense of the specificity and precision, that decrease when adding text features in these models.

(a)

(b)

Figure 6 – Word clouds of words and phrases associated with pneumonia or anastomotic leakage based on the form health perception. The size of the words or phrases indicates their frequency. These reports are written by nurses on the day the complication was diagnosed and show the most frequently used words to describe the patient's health perception. Figure (a) depicts the most occurring unigrams in these reports and (b) shows the most frequent bigrams used by nurses.

5.3.1 Feature importance

To gain more understanding in the features that are important in the prediction of the ML models, we evaluated feature importance. In figure 7, the coefficients of each feature from the structured data of the logistic regression model are shown. Positive coefficients indicate that, based on this feature, the output class is likely to be positive, so in group 1. Negative coefficients make it less likely for the case to be in group 1. Coefficients close to zero are less relevant in the classification. Note that logistic regression was trained and tested using scaled features. Based on figure 7, the following features are most important: temperature in the evening shift 48 and 24 hours before the diagnosis, the difference between CRP levels and amylase 24 hours before the diagnosis.

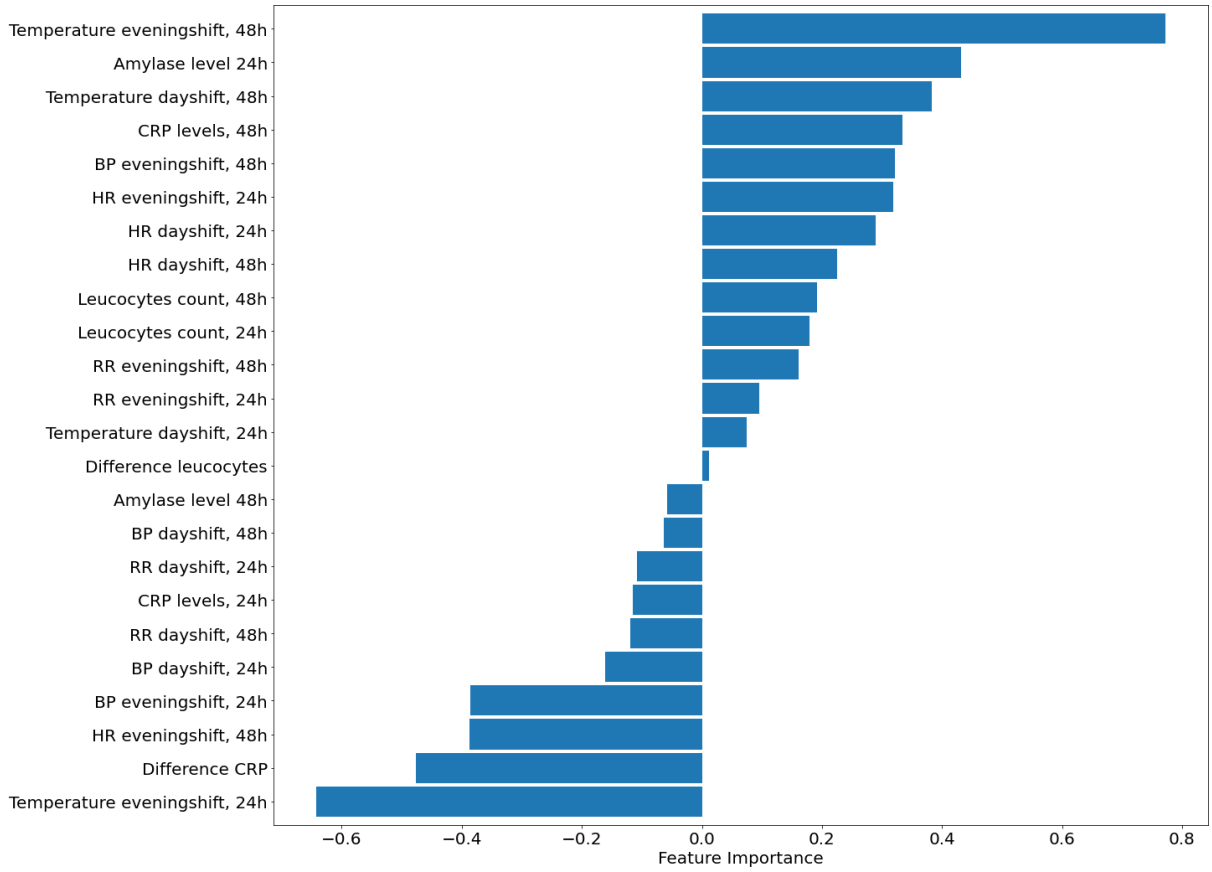


Figure 7 *– Coefficients of the vital signs and the laboratory measurements of our logistic regression model. Coefficients close to zero are less relevant in the classification of the two classes. Positive coefficients indicate that this particular feature will predict a sample as group 1 when all the other features are zero. Conversely, when the coefficient is negative, this feature will predict a sample into group 0.*

Moreover, we have also obtained the feature importance of the structured data features in a random forest model. In figure 8, the importance of each structured data feature is shown. The five most important features in order of importance are: CRP levels measured 48 hours in advance, the difference between CRP levels, amylase measured 48 hours in advance, the temperature measured in the evening shift 48 hours in advance and amylase measured 24 hours in advance.

Based on the most relevant features of logistic regression and random forest, we experimented with another feature set, containing only the temperature and CRP features. In figure 9 we have shown the performance of logistic regression and random forest based on this feature set compared to the performance of the entire structured dataset. Remarkably, the performance of logistic regression increases when selecting only CRP and temperature features, while random forest maintained a steady performance on both datasets. The true positive rate, sensitivity, on the other hand, decreases when using this selection of features.

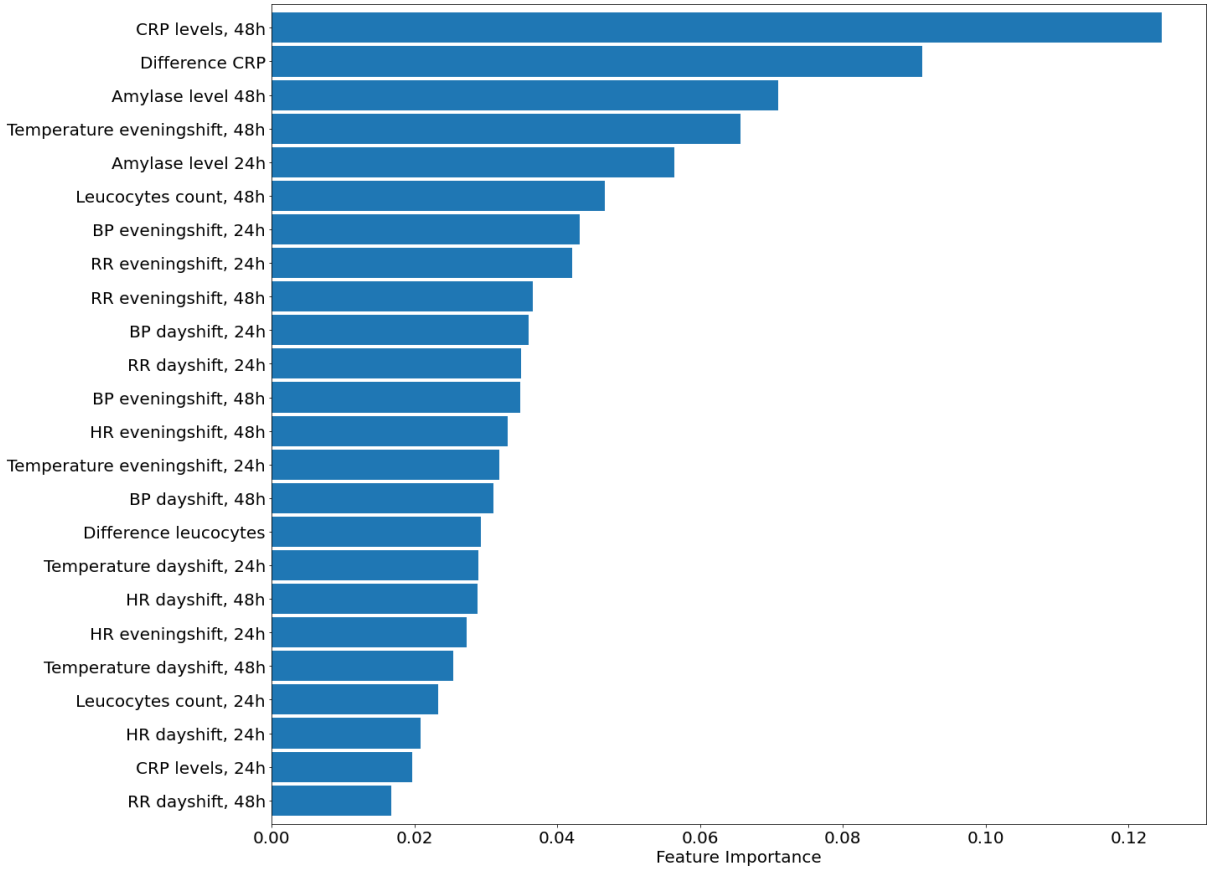


Figure 8 – Feature importance of vital signs and laboratory values in a random forest model based on impurity decrease.

Besides the structured data features, we also estimated the most prominent text features. In table 4, the top ten most important features of each model are given. Note that the same terms can occur in different forms and are probably used in a different context. Furthermore, the time window in which the feature is important is given, which could either be 24 or 48 hours before a potential postoperative complication. The weight of these terms is evaluated using the coefficient of logistic regression and SVM and the impurity decrease in random forest. These results data need to be interpreted with caution as the role of text features was of limited value in our models.

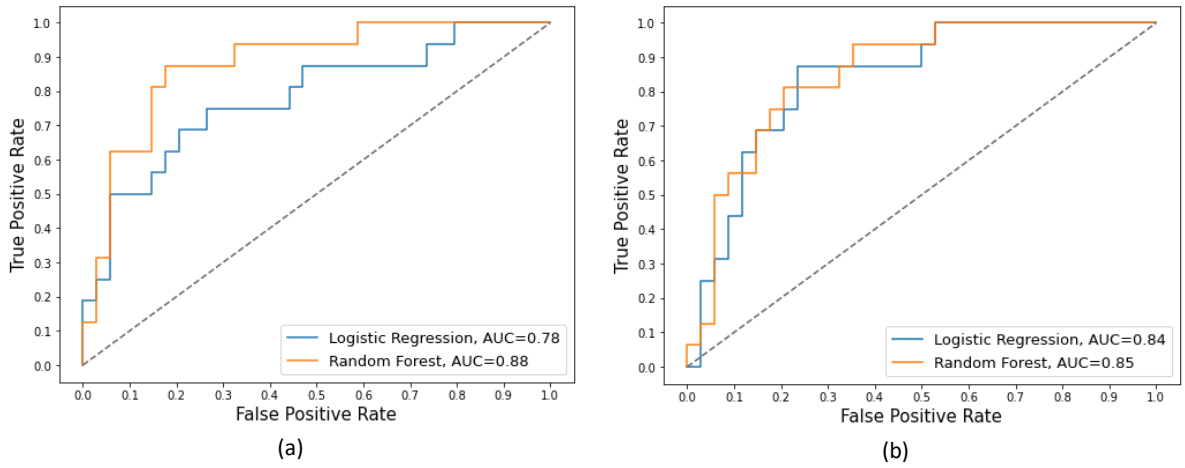


Figure 9 – ROC curve of the performance of logistic regression compared to random forest when using all the structured data (a) and using only CRP and temperature features (b).

What stands out is that the form health perception occurs most often as the source in the top ten most important features. The other forms occur far less often.

***Table 4** – Overview of the top ten most important uni- and bigrams for three different ML models. Per term, the form in which it occurred and the time window in which it was mentioned is given. GI = general information, HP = health perception, Ob = observations and Ac = activity. For logistic regression and SVM the coefficients are used and for random forest feature importance is evaluated using impurity decrease.*

Logistic Regression			Support Vector Machines			Random Forest		
Term	Form	Weight	Term	Form	Weight	Term	Form	Weight
Temp	GI 48h	0.810	*Voelt goed*	HP 48h	0.745	*Kortademig*	HP 24h	0.021
Kortademig	HP 24h	0.650	*Geeft goed*	HP 24h	0.630	*Geen klachten*	Ob 48h	0.013
Niet goed	HP 24h	0.558	*Temp*	GI 24h	0.556	*Zuurstof*	HP 24h	0.010
Hoest	HP 24h	0.545	*Verhoogd*	Ob 48h	0.529	*Benauwd*	HP 24h	0.010
Verhoogd	Ob 48h	0.519	*Verhoogd*	Ob 24h	0.514	*Niet benauwd*	HP 24h	0.010
Geen klachten	HP 48h	0.505	*Kortademig*	HP 24h	0.509	*Zuurstof*	Ob 24h	0.009
Goed voelen	HP 48h	0.437	*Hoest*	HP 24h	0.475	*Voelt goed*	HP 24h	0.006
Gezakt	Ob 24h	0.426	*Zuurstof*	Ob 48h	0.368	*Niet benauwd*	HP 48h	0.005
Beter voelen	HP 24h	0.424	*Erg vermoeid*	Ac 48h	0.344	*Ging goed*	Ac 24h	0.004
Voelt goed	HP 48h	0.421	*Niet goed*	HP 24h	0.340	*Hoest*	GI 24h	0.004

5.3.2 Classification using feature subsets

In our previous results, the complication group consisted of patients who were diagnosed with pneumonia or anastomotic leakage on day three, four or five after surgery. To assess whether this is a disadvantage, we split the structured features of group 0 and 1 into three groups according to the day of complication diagnosis (3, 4 or 5). Note that these subgroups are rather small, see figure 4, causing a substantial risk of overfitting. Especially on day five the amount of patients is too small and is therefore inadequate to train a ML model. In figure 10 the prediction by a logistic regression model of the patients diagnosed on day three (a) and day four (b) are shown. The prediction of a complication on day three is almost random, while the prediction on day 4 is much better and closer to the performance when using the entire patient population. Attention must be paid to the fact that these models both overfitted the data due to the small number of patients.

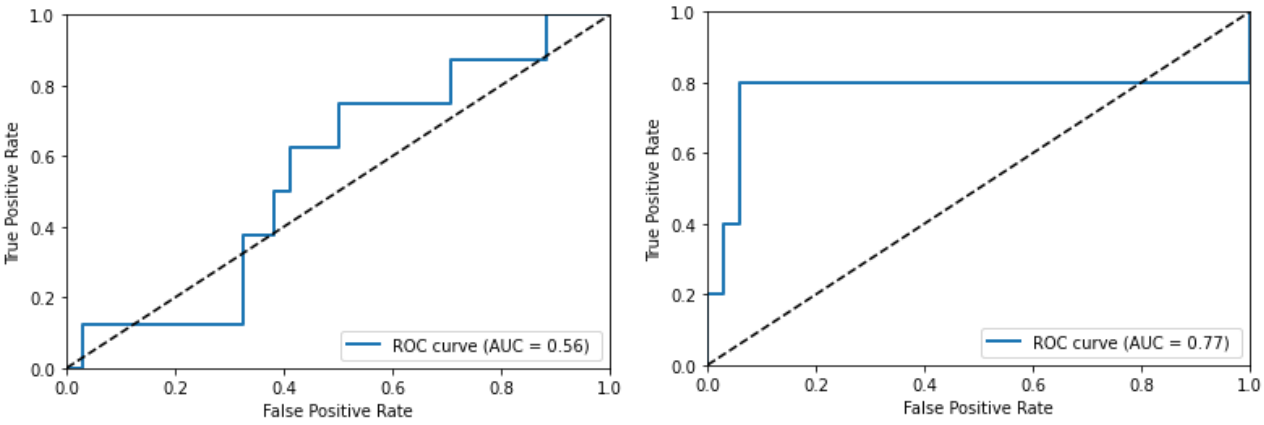


Figure 10 *– ROC curve of logistic regression model to predict pneumonia or anastomotic leakage on (a) day three and (b) day four after esophagectomy using structured data features.*

Chapter 6 – Discussion

This research aimed to explore the ability of machine learning algorithms to predict major complications in patients who underwent esophagectomy by using structured and unstructured EHR data. In this chapter, we further discuss the findings of this research in section 6.1 and evaluate the strengths and limitations of this study in section 6.2. Thereafter, in section 6.3, we give our view on future research to predict postoperative complications after esophagectomy.

6.1 Interpretation of results

In this study, we found that postoperative data is useful to predict postoperative complications in patients who underwent esophagectomy. Based on table 3, our random forest model revealed the highest AUC, namely 0.90, when using both structured and unstructured data. Nevertheless, the sensitivity was rather low, 12%, against an excellent specificity of 97%. Remarkably, all of our models revealed a high specificity, against a poor sensitivity. Although the sensitivity increased when adding text features from nursing assessments, except for random forest, the overall performance did not improve when adding text features. Furthermore, our results show that our ML models added a minimal improvement to the classification of postoperative complications after esophagectomy over more traditionally used logistic regression models.

In our view, machine learning is a powerful technique to discover important features to predict postoperative deterioration. This study confirmed the significance of CRP levels and temperature to predict anastomotic leakage or pneumonia. It also revealed that amylase measured 24 hours before a complication is an important indicator of postoperative complications. The coefficients of the logistic regression model in figure 7 demonstrated that amylase levels are important to confirm a complication, while the difference in CRP levels could help rule out postoperative complications. We also discovered that the performance of our random forest model remained similar when utilizing solely CRP and temperature features compared to the results based on the entire structured dataset, see figure 9. This suggests that a random forest model is more efficient in selecting the most discriminating features from our structured dataset compared to our logistic regression model. It also further substantiates that CRP and temperature together are the most important parameters to predict postoperative complications. These results are in agreement with what we expected at the beginning of our research. In our current clinical practice, we consider the level of CRP the most indicative biomarker of an inflammatory complication. Furthermore, high levels of amylase are considered alarming and indicative of anastomotic leakage. Many other studies have already indicated the predictive value of CRP [19]–[22] and drain amylase levels [26]–[28] to predict anastomotic leakage in an early stage. Aiolfi et al [19] showed in a meta-analysis that CRP levels measured on day three, four and five after esophagectomy are useful as a negative test to rule out anastomotic leakage. This also explains why it is easier to rule out anastomotic leakage than to detect it, which is why we found a far better true negative rate, than a true positive rate. However, when using the data of day three, we were unable to predict postoperative complications, see figure 10. It is possible that CRP levels on day three are not as discriminating compared to day four.

The sensitivity of our ML models was overall very low, while the specificity was rather high. This demonstrates that based on our dataset it is easier to classify the patients without complications than the patients with complications and that our classifiers generate more false negatives than false positives. This could be due to the size of group 0 (n=112), which was probably sufficient for the models to learn to classify these cases. The complication group (n=52) seemed to be too small to find patterns and indicators to classify them. This may also explain why the under-sampling algorithm, named NearMiss, deteriorated the classification and suggests that when the complication group grows, the sensitivity could increase as well. Another reason for the low sensitivity of our classifiers could be the heterogeneity among patients during their postoperative recovery after esophagectomy. Our models demonstrated in figure 10 that predicting a postoperative event on day three is much more difficult

than on day four. Even though the population of these two subgroups is far too small to draw any firm conclusions, it illustrates that on day three the overlap between group 0 and 1 is more substantial than on day four. In other words, it is plausible that normal recovery from a surgery of this extent overlaps with a complicated recovery, especially in the first few days after surgery. Kohl et al. [46] have investigated the normal inflammatory or stress response to surgery and trauma. They stated that all inflammatory markers peak about day two after surgery and return to normal levels around day six. This could explain the overlap between the two patient groups on day three postoperatively. Furthermore, the article explains that the ability to adequately respond to surgical trauma requires an integrated interaction between several organs and systems. As a result, an unrecognized disease, e.g. coronary artery disease, could interfere the normal recovery and potentially result in complications. It is therefore important for clinicians to be able to recognize deviations from the normal inflammatory response after surgery. Such deviations could indicate pre-existing medical diseases or postoperative complications. However, there is still little insight into the normal inflammatory reaction after esophagectomy to be prepared for expected responses and recognize abnormal time courses. On top of that, Urschel et al. [13] has shown in their research that the severity of the morbidity due to anastomotic leakage is variable and mostly dependent on the gastric viability, the site of the leak (neck or thorax), the timing of the leak and the containment of the leak by surrounding tissue. In other words, there are different responses to anastomotic leakage, which also makes it difficult to predict postoperative complications. When taking a closer look at the patients who were falsely classified as negative, we see that most of these cases are classified wrongly in all models. It is unclear why all models struggle to classify the same specific patients as most of them have abnormal inflammatory and temperature measurements. Patients that are truly predicted positive most often have very high levels of amylase or CRP, which are more obvious signs of postoperative inflammatory complications. We hypothesize that there is a substantial part of the complication group that do not show obvious signs of complications and are therefore wrongly predicted as normal. These particular patients are especially difficult to recognize in clinical practice. More research is therefore needed into these specific patients to find patterns of deviation from normal postoperative recovery after esophagectomy.

This study also showed that the role of RR, HR and systolic BP, is inferior to the importance of CRP and temperature in our models. We were expecting that these vital signs would be of more added value to detect serious illness in our patient population. Nevertheless, based on our current clinical experience, vital signs show abnormalities when a patient is already severely ill, which probably makes these parameters less useful in the early prediction of postoperative deterioration. Moreover, both Evans et al. [48] and Lockwood et al. [49] have systematically reviewed the importance of vital signs and have shown that vital signs measurements are only of limited value to detect serious illness in adults and that normal vital signs do not necessarily indicate stable physiological function. Since these studies, measuring techniques have improved to increase the reliability and usefulness of vital signs monitoring, but recording the RR remains challenging, especially on the ward [50]. This explains the increase in missing values in the RR data compared to the other vital signs, see figure 11 in appendix 1. The potential predictive value of RR could therefore be underestimated due to missing values and the challenges to correctly estimating the RR. Furthermore, in our study design, we chose to use a maximum of three measurements of each vital sign per day, while especially HR, RR and systolic BP show much more fluctuation during the day compared to laboratory values. As a result, our three measurements might be too small to resemble the physiological function of these patients. It is likely that the use of continuous monitoring of vital signs would therefore be more effective in detecting postoperative complications. However, this would dramatically increase the amount of data and would complexify our study design. Moreover, continuous data is currently only available during admission at the ICU, while most patients spend the majority of their hospital stay at the ward.

In regards to the unstructured data, we have shown in table 4 the ten most important uni- and bigrams of each ML model. Since the text features hardly improved the performance of our models, we cannot

draw strong conclusions from these results, other than that these results highlight that the nursing assessments of the patient's health perception play the most relevant role in our models. We were expecting a more meaningful addition of the text features to the classification. In our opinion, the reports of the nursing assessments capture the patients' clinical well-being, which could potentially give more context to the structured data. In our clinical practice, biomarkers are always evaluated together with a patients' clinical condition when diagnosing postoperative complications. One explanation of the limited value of text data could be the size of the vocabulary. There are no guidelines on selecting the appropriate size of the vocabulary of a BoW model. In this research, we used four different written forms, namely health perception, general information, activity and observations. We added text features with a vocabulary of size 40, this means that we had 40 features of each form of two days, resulting in a total of 320 features. Our population consisted of only 164 patients, so the number of features compared to the number of patients is way out of proportion. This has probably contributed to the overfitting of our models, which in case of logistic regression and SVM was quite substantial after adding the text features, resulting in poorer classification. To solve this we used only one of the forms, namely the patient's health perception, but this did not improve the performance of the models. Therefore it is also possible that the nursing assessments contain too little discriminative information to predict postoperative complications. We are aware that every nurse has a different approach towards recording information about patients. Although they often use the same words to describe certain conditions, some nurses will report more than others and have different priorities in recording patient information. When using a more uniform vocabulary and method of reporting the patient's well-being, the text features might become more useful to predict postoperative complications. However, in clinical practice this is probably not a feasible and efficient way to report for nurses, as they prefer to write in free text. It is also possible that our approach with a BoW model is not suitable for this type of data. Another approach to use clinical text data is shown by Goh et al [43]. They utilized Latent Dirichlet Allocation (LDA) which is a topic model that generates topics, based on patterns of word frequency from a set of documents. LDA is also considered a technique for dimensionality reduction, as it summarizes the discussion about a particular topic of each document. This technique could potentially solve our challenge with the vocabulary size. Furthermore, it is then easier to add other clinical unstructured data, without growing too many features and without the potential overlap found among clinical notes. However, this technique is much more complex compared to a BoW model and is therefore not suited for an explorative study of our size.

Similar studies that developed predictive models of the risk of postoperative complications used patient data from large national clinical registries, which often contain demographic characteristics, comorbidities, information about the surgical procedure and other pre-operative data [33]–[35], [37], [47]. Such databases often lack granular information, such as HR or leukocytes count, which we focused on in our study. However, we did not include any pre-operative factors, nor did we utilized demographic characteristics of our population. Based on the results of these studies, adding preoperative data, such as comorbidities, could contribute to the early prediction of complications, especially within the first few days after surgery, when little postoperative data is available.

6.2 Strength and limitations

To the best of our knowledge, this is the first study that utilizes structured and unstructured postoperative data to predict postoperative complications in patients who underwent esophagectomy. This study has demonstrated that postoperative data, especially CRP levels, amylase and temperature, together with ML algorithms are able to make a fair prediction of anastomotic leakage and pneumonia. Furthermore, we have discovered that the physiological diversity of the response to surgical trauma and postoperative complications forms a big challenge in the early detection of postoperative complications. ML could be a great technique to gain more insight into the inflammatory response of not only esophagectomy but on surgery in general.

We are aware that our research contains some limitations. To begin with, we were able to include all patients who underwent esophagectomy from 2010 until the end of 2020 in our research. However, the number of patients is still relatively small and our models are therefore prone to overfitting. Moreover, this is a single-center study, which limits the generalizability of our findings. Secondly, our reference group, group 0, did not necessarily contain merely patients with an uncomplicated recovery. Due to lack of time, we decided to use patients with an admission according to our fast track protocol. Although we are certain that group 0 does not contain patients with pneumonia or anastomotic leakage, patient with for instance a wound infection could still appear in this group. This perhaps blurred the discrimination between group 0 and group 1. Lastly, an existing limitation of working with clinical data is the number of missing values. In our missing analysis, we have seen that especially respiratory rate showed the most substantial number of missing values. This is because nurses on the ward have to count the RR by themselves, which requires focus and time. Since we focused on the first few days of the postoperative recovery, the number of missing values was less substantial compared to the end of our postoperative window. We, therefore, expect that together with our imputation techniques, missing data did not bias our results.

6.3 Recommendations

This retrospective study has shown that ML models can predict postoperative complications with high specificity based on postoperative data. But before we continue to apply ML models to postoperative data in these patients, we need a better understanding of the physiological variation among patients during uncomplicated recovery and severe inflammatory complications. To start with, a thorough trend analysis of CRP levels during postoperative recovery is beneficial to define the range of a normal inflammatory response after esophagectomy. It is important to not only focus on the absolute values but also zoom in on the slope of CRP levels during postoperative recovery because the difference in CRP levels between days was an important feature in our models. Furthermore, we recommend having a closer look at the patients with postoperative complications and to investigate the differences in inflammatory responses among them. ML techniques could then be a great tool to find new patterns that can discriminate this normal range from patients with a postoperative event. Ultimately, we believe that ML models could then be useful to support clinicians in recognizing patients that deviate from normal recovery after esophagectomy.

Lastly, this study does not rule out the potential of text data to predict postoperative complications. In further research, it is important to create a discriminative vocabulary when using BoW and to evaluate this with an experienced physician. Besides that, other sources of unstructured data are interesting to incorporate, for example, clinicians' notes or radiological reports could contain other aspects of the patients' (pre-operative) clinical condition.

Chapter 7 – Conclusion

In conclusion, this study revealed that ML models have an overall fair prediction of postoperative complications after surgery when using postoperative data. Within these models, CRP and temperature are important predictors of anastomotic leakage and pneumonia. Furthermore, text features could contribute to a better sensitivity of models to predict major postoperative complications, but its potential needs to be further researched. We also recommend to investigate the physiological differences in the inflammatory response to surgical trauma and postoperative complications to enable better recognition of deviation from normal recovery after esophagectomy.

Despite the limitations of this research and our future recommendations, our results indicate that ML, based on both structured and unstructured data, may improve the ability to predict postoperative complications after esophagectomy. Our approach is not only useful in patients who underwent esophagectomy, but can also be generalized to other clinical areas.

Chapter 8 – References

[1] M. José *et al.*, "Esophageal cancer: Risk factors, screening and endoscopic treatment in Western and Eastern countries," *World J Gastroenterol*, vol. 21, no. 26, pp. 7933–7943, 2015.

[2] Integraal-Kankercentrum-Nederland, "NKR Cijfers." [Online]. Available: https://iknl.nl/nkr-cijfers. [Accessed: 24-Jun-2021].

[3] F. D. en D. L. van der P. Daan M. Voeten, Chantal M. den Bakker, Ruben S.A. Goedegebuure, David J. Heineman, "Niet-gemetastaseerde slokdarmkanker | Nederlands Tijdschrift voor Geneeskunde," *24-09-2019*. .

[4] J. Straatman *et al.*, "Minimally Invasive Versus Open Esophageal Resection," *Ann. Surg.*, vol. 266, no. 2, pp. 232–236, Aug. 2017.

[5] J. H. Kauppila, S. Xie, A. Johar, S. R. Markar, and P. Lagergren, "Meta-analysis of health-related quality of life after minimally invasive versus open oesophagectomy for oesophageal cancer," *British Journal of Surgery*, vol. 104, no. 9. John Wiley and Sons Ltd, pp. 1131–1140, 01-Aug-2017.

[6] S. S. A. Y. Biere *et al.*, "Minimally invasive versus open oesophagectomy for patients with oesophageal cancer: A multicentre, open-label, randomised controlled trial," *Lancet*, vol. 379, no. 9829, pp. 1887–1892, May 2012.

[7] W. Zhang, D. Yu, J. Peng, J. Xu, and Y. Wei, "Gastric-Tube versus whole-stomach esophagectomy for esophageal cancer: A systematic review and meta-Analysis," *PLoS One*, vol. 12, no. 3, p. e0173416, Mar. 2017.

[8] D. Vetter and C. A. Gutschow, "Strategies to prevent anastomotic leakage after esophagectomy and gastric conduit reconstruction," *Langenbeck's Archives of Surgery*, vol. 405, no. 8. Springer Science and Business Media Deutschland GmbH, pp. 1069–1077, 01-Dec-2020.

[9] E. Booka *et al.*, "Meta-analysis of the impact of postoperative complications on survival after oesophagectomy for cancer," *BJS Open*, vol. 2, no. 5, pp. 276–284, Sep. 2018.

[10] M. K. Ferguson and A. E. Durkin, "Preoperative prediction of the risk of pulmonary complications after esophagectomy for cancer," *J. Thorac. Cardiovasc. Surg.*, vol. 123, no. 4, pp. 661–669, 2002.

[11] C. D. Wright, J. C. Kucharczuk, S. M. O'Brien, J. D. Grab, and M. S. Allen, "Predictors of major morbidity and mortality after esophagectomy for esophageal cancer: A Society of Thoracic Surgeons General Thoracic Surgery Database risk adjustment model," *J. Thorac. Cardiovasc. Surg.*, vol. 137, no. 3, pp. 587–596, 2009.

[12] C. T. Bakhos *et al.*, "Impact of the surgical technique on pulmonary morbidity after esophagectomy," *Ann. Thorac. Surg.*, vol. 93, no. 1, pp. 221–227, 2012.

[13] J. D. Urschel, "Esophagogastrostomy anastomotic leaks complicating esophagectomy: A review," *Am. J. Surg.*, vol. 169, no. 6, pp. 634–640, 1995.

[14] M. Derogar, N. Orsini, O. Sadr-Azodi, and P. Lagergren, "Influence of major postoperative complications on health-related quality of life among long-term survivors of esophageal cancer surgery," *J. Clin. Oncol.*, vol. 30, no. 14, pp. 1615–1619, 2012.

[15] R. Bhagat *et al.*, "Postoperative Complications Drive Unplanned Readmissions After Esophagectomy for Cancer," *Ann. Thorac. Surg.*, vol. 105, no. 5, pp. 1476–1482, 2018.

[16] F. G. Fernandez *et al.*, “Hospital Readmission Is Associated With Poor Survival Following Esophagectomy For Esophageal Cancer,” *Ann Thorac Surg*, vol. 99, no. 1, pp. 292–297, 2016.

[17] S. Y. Chen, D. Molena, M. Stem, B. Mungo, and A. O. Lidor, “Post-discharge complications after esophagectomy account for high readmission rates,” *World J. Gastroenterol.*, vol. 22, no. 22, pp. 5246–5253, 2016.

[18] S. R. Markar, A. Karthikesalingam, and D. E. Low, “Enhanced recovery pathways lead to an improvement in postoperative outcomes following esophagectomy: systematic review and pooled analysis,” *Dis. Esophagus*, vol. 28, no. 5, pp. 468–475, 2015.

[19] A. Aiolfi, E. Asti, E. Rausa, G. Bonavina, G. Bonitta, and L. Bonavina, “Use of c-reactive protein for the early prediction of anastomotic leak after esophagectomy: Systematic review and bayesian meta-analysis,” *PLoS One*, vol. 13, no. 12, pp. 1–13, 2018.

[20] E. Asti *et al.*, “Utility of C-reactive protein as predictive biomarker of anastomotic leak after minimally invasive esophagectomy,” *Langenbeck’s Arch. Surg.*, vol. 403, no. 2, pp. 235–244, 2018.

[21] J. K. Park, J. J. Kim, and S. W. Moon, “C-reactive protein for the early prediction of anastomotic leak after esophagectomy in both neoadjuvant and non-neoadjuvant therapy case: A propensity score matching analysis,” *J. Thorac. Dis.*, vol. 9, no. 10, pp. 3693–3702, 2017.

[22] A. C. Gordon, A. J. Cross, E. W. Foo, and R. H. Roberts, “C-reactive protein is a useful negative predictor of anastomotic leak in oesophago-gastric resection,” vol. 88, pp. 223–227, 2018.

[23] S. H. Hoeboer, A. B. J. Groeneveld, N. Engels, M. van Genderen, B. P. L. Wijnhoven, and J. van Bommel, “Rising C-Reactive Protein and Procalcitonin Levels Precede Early Complications After Esophagectomy,” *J. Gastrointest. Surg.*, vol. 19, no. 4, pp. 613–624, 2015.

[24] P. Vulliamy, S. McCluney, S. Mukherjee, L. Ashby, and T. Amalesh, “Postoperative Elevation of the Neutrophil: Lymphocyte Ratio Predicts Complications Following Esophageal Resection,” *World J. Surg.*, vol. 40, no. 6, pp. 1397–1403, 2016.

[25] B. Shi, X. Wang, Y. U. E. Zhao, X. Fei, J. I. Zhu, and F. U. Yang, “Preoperative neutrophil to lymphocyte ratio predicts complications after esophageal resection that can be used as inclusion criteria for enhanced recovery after surgery A Multi-center Observational Study.”

[26] Y. Perry, C. W. Towe, J. Kwong, V. P. Ho, and P. A. Linden, “Serial Drain Amylase Can Accurately Detect Anastomotic Leak After Esophagectomy and May Facilitate Early Discharge,” *Ann. Thorac. Surg.*, vol. 100, no. 6, pp. 2041–2047, 2015.

[27] L. Giulini *et al.*, “Prognostic Value of Chest-Tube Amylase Versus C-Reactive Protein as Screening Tool for Detection of Early Anastomotic Leaks After Ivor Lewis Esophagectomy,” vol. 29, no. 2, pp. 192–197, 2019.

[28] W. Sik, J. Jung, H. Shin, Y. Roh, G. Eun, and D. Joon, “Amylase level in cervical drain fluid and anastomotic leakage after cervical oesophagogastrostomy †,” vol. 56, no. January, pp. 301–306, 2019.

[29] V. Arvind, J. S. Kim, E. K. Oermann, D. Kaji, and S. K. Cho, “Predicting surgical complications in adult patients undergoing anterior cervical discectomy and fusion using machine learning,” *Neurospine*, vol. 15, no. 4, pp. 329–337, 2018.

[30] K. Adnan, R. Akbar, S. W. Khor, and A. B. A. Ali, “Role and Challenges of Unstructured Big Data in Healthcare,” in *Advances in Intelligent Systems and Computing*, 2020, vol. 1042, pp. 301–

323.

[31] D. Talmor and B. Kelly, “How to better identify patients at high risk of postoperative complications?,” *Curr. Opin. Crit. Care*, vol. 23, no. 5, pp. 417–423, 2017.

[32] K. Merath *et al.*, “Use of Machine Learning for Prediction of Patient Risk of Postoperative Complications After Liver, Pancreatic, and Colorectal Surgery,” *J. Gastrointest. Surg.*, vol. 24, no. 8, pp. 1843–1851, 2020.

[33] A. Bihorac *et al.*, “MySurgeryRisk: Development and Validation of a Machine-learning Risk Algorithm for Major Complications and Death After Surgery,” *Ann. Surg.*, vol. 269, no. 4, pp. 652–662, Apr. 2019.

[34] B. A. Fritz *et al.*, “Using machine learning techniques to develop forecasting algorithms for postoperative complications : protocol for a retrospective study,” pp. 1–7, 2018.

[35] M. Bronsert, A. B. Singh, W. G. Henderson, K. Hammermeister, R. A. Meguid, and K. L. Colborn, “Identification of postoperative complications using electronic health record data and machine learning,” *Am. J. Surg.*, vol. 220, no. 1, pp. 114–119, Jul. 2020.

[36] A. Rajkomar, J. Dean, and I. Kohane, “Machine Learning in Medicine,” *N Engl J Med*, vol. 380, pp. 1347–58, 2019.

[37] S. Bolourani *et al.*, “Using machine learning to predict early readmission following esophagectomy,” *J. Thorac. Cardiovasc. Surg.*, 2020.

[38] N. Altorki and A. Sedrakyan, “Commentary: Can machine learning reduce readmissions after esophagectomy? A consummation devoutly to be wished,” *J. Thorac. Cardiovasc. Surg.*, pp. 1–3, 2020.

[39] P. C. Müller *et al.*, “Fit-for-Discharge Criteria after Esophagectomy: An International Expert Delphi Consensus,” *Dis. Esophagus*, Sep. 2020.

[40] Y. Nevo *et al.*, “Activity Tracking After Surgery: Does It Correlate With Postoperative Complications?,” *Am. Surg.*, p. 000313482098881, Jan. 2021.

[41] C. Boer, H. R. Touw, and S. A. Loer, “Postanesthesia care by remote monitoring of vital signs in surgical wards,” *Current opinion in anaesthesiology*, vol. 31, no. 6. NLM (Medline), pp. 716–722, 01-Dec-2018.

[42] K. Jensen *et al.*, “Analysis of free text in electronic health records for identification of cancer patient trajectories,” *Nat. Publ. Gr.*, pp. 1–12, 2017.

[43] K. H. Goh *et al.*, “Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare,” no. 2021, pp. 1–10.

[44] E. L. Barber, R. Garg, C. Persenaire, and M. Simon, “Natural language processing with machine learning to predict outcomes after ovarian cancer surgery,” *Gynecol. Oncol.*, vol. 160, no. 1, pp. 182–186, 2021.

[45] S. Yen and Y. Lee, “Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset,” vol. 344, pp. 731–740, 2006.

[46] B. A. Kohl and C. S. Deutschman, “The inflammatory response to surgery and trauma,” *Curr. Opin. Crit. Care*, vol. 12, no. 4, pp. 325–332, 2006.

[47] Y. Ohkura *et al.*, “Development of a model predicting the risk of eight major postoperative complications after esophagectomy based on 10 826 cases in the Japan National Clinical Database,” no. July 2019, pp. 313–321, 2020.

[48] D. Evans, B. Hodgkinson, and J. Berry, “Vital signs in hospital patients: A systematic review,” *Int. J. Nurs. Stud.*, vol. 38, no. 6, pp. 643–650, 2001.

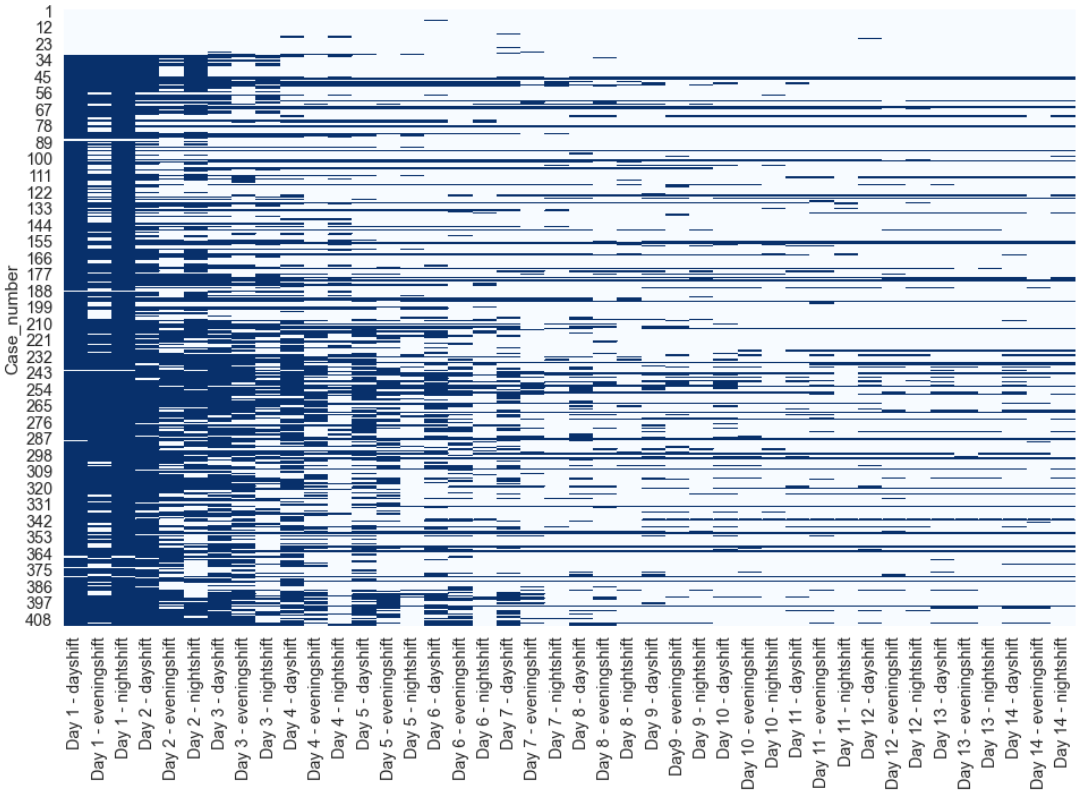
[49] C. Lockwood *et al.*, “Vital signs,” pp. 207–230, 2004.

[50] S. C. Gandevia and D. K. McKenzie, “Respiratory rate: The neglected vital sign,” *Med. J. Aust.*, vol. 189, no. 9, p. 532, 2008.

Appendices

Appendix 1 – Missing values plots

(a)



(b)

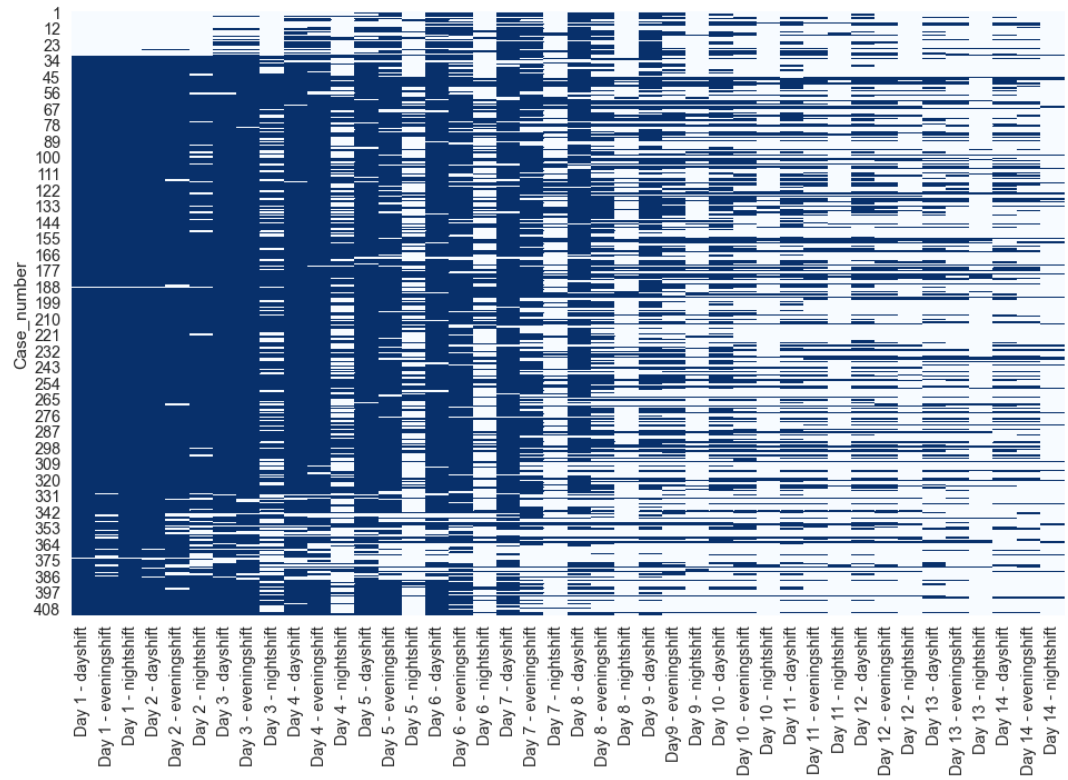
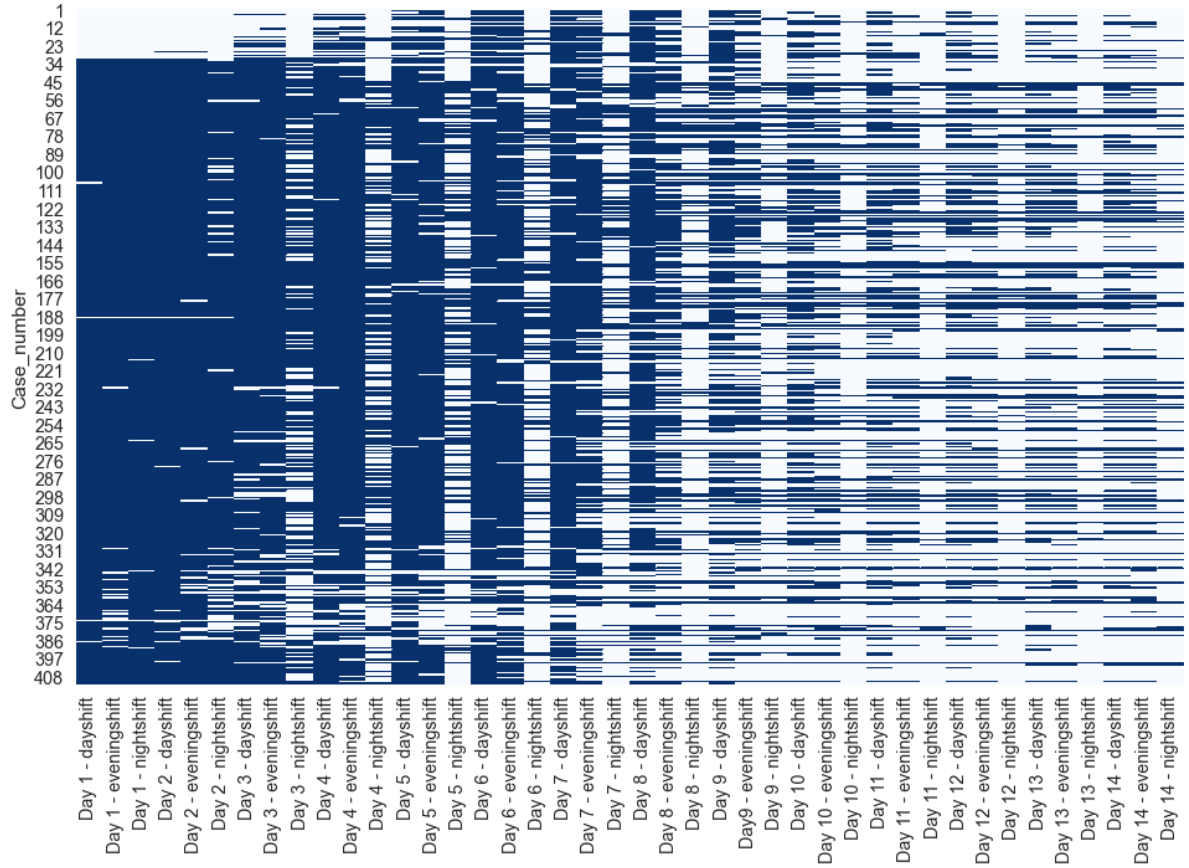


***Figure 11** – Missing values plot of (a) respiratory rate and (b) systolic blood pressure*

(a)



(b)

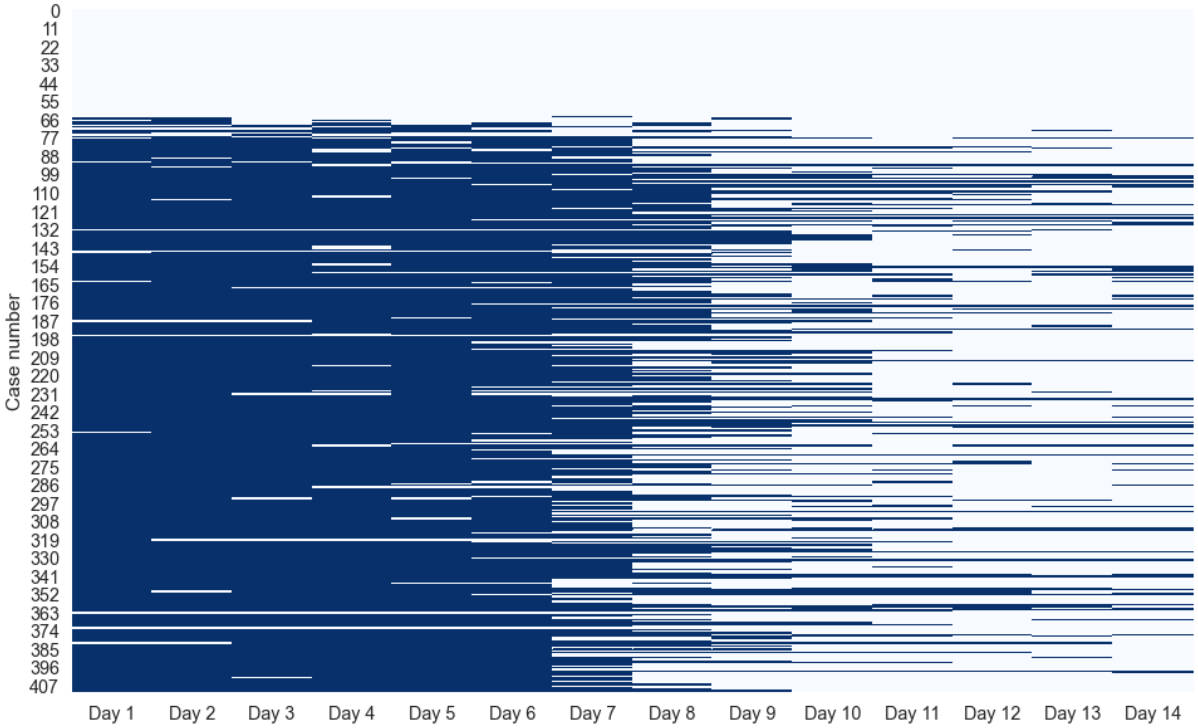


Figure 12 *– Missing values plot of (a) temperature and (b) amylase levels*

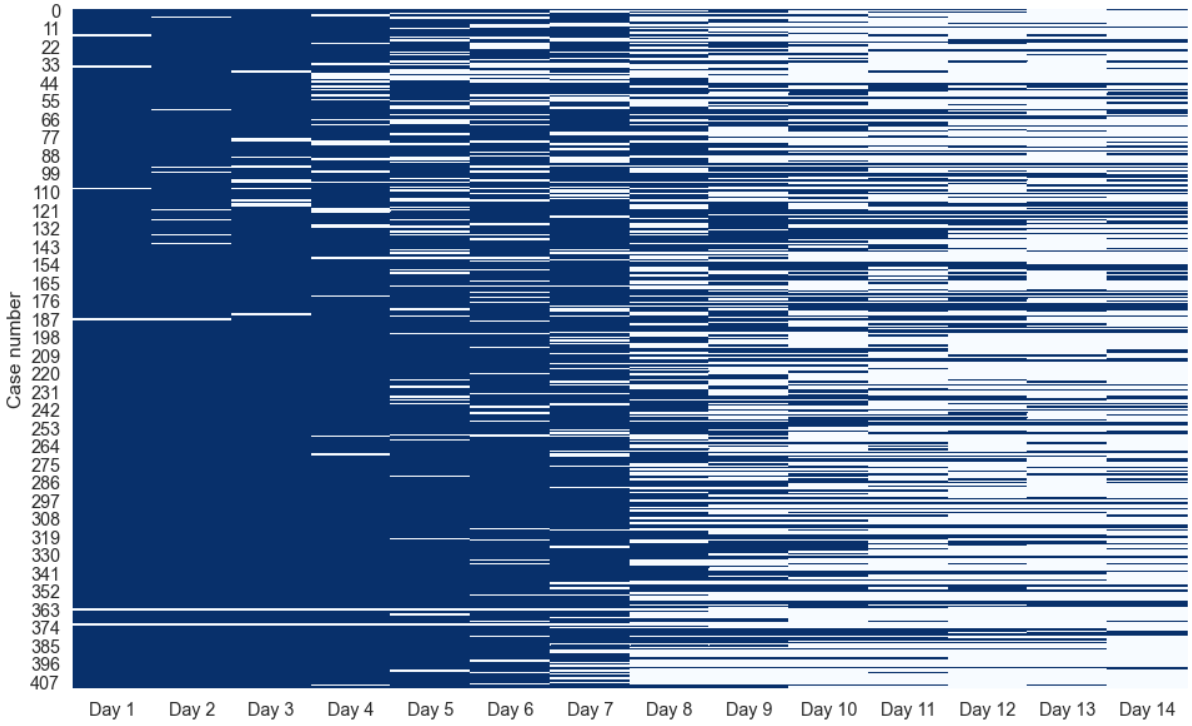


Figure 13 *– Missing values plot of CRP levels*

Appendix 2 – List of stopwords

'aan',
aangaande',
aangezien',
achter',
achterna',
af',
afgelopen',
akturk',
al',
aldaar',
aldus',
alhoewel',
alle',
allebei',
alleen',
alles',
als',
alsnog',
althans',
altijd',
ander',
andere',
arts',
arts-assistent',
artsass',
ass',
assistent',
averdijk',
behalve',
beide',
ben',
betreffende',
bij',
binnen',
boven',
bovendien',
bovenstaand',
cc',
cm',
daar',
daarheen',
daarin',
daarna',
daarnaast',
daarnet',
daarom',
daarop',
dan',
dat',
de',
der',
det',
deze',
dhr',
die',
dienst',
dikwijls',
dit',
doch',
doen',
dokter',
door',
doorgaans',
dr',
dus',
echter',
een',
eens',
eerdat',
eerder',
eerst',
elk',
elke',
en',
enige',
enigzinds',
enkel',
enkele',
enz',
enzovoorts',
er',
erdoor',
etc',
etcetera',
even',
eveneens',
evt',
gauw',
gedurende',
geen',
geweest',
haar',
had',
hare',
heb',
hebben',
heeft',
heer',
hem',
het',
hier',
hierboven',
hierin',
hij',
hijzelf',
hoe',
hoewel',
hr',
hun',
i',
i.o',
iemand',
iets',
ik',
in',
inmidddels',
io',
iom',
ipv',
irza',
is',
ivm',
ja',
je',
jezelf',
jij',
jijzelf',
jou',
jouw',
juist',
jullie',
kan',
klaar',
kon',
kouwenhoven',
kunnen',
l',
later',
leoniek',
liever',
liter',
ltr',
maar',
mag',
mbt',
mbv',
me',
meer',
men',
meneer',
met',
mevr',
mevrouw',
mezelf',
mg',
mgr',
mij',
mijn',
mijzelf',
miligram',
mililiter',
misschien',
ml',
mn',
moet',
morgen',
mv',
mvr',
mw',
na',
naar',
nabij',
nadat',
nadien',
net',
niets',
nog',
nogal',
nu',
o',
obv',
of',
om',
omdat',
omhoog',
omlaag',
omtrent',
onder',
ondertussen',
ongeveer',
ons',
onszelf',
ook',
op',
opnieuw',
opzij',
over',
overigens',
pas',
precies',
reeds',
rondom',
seldam',
sinds',
sindsdien',
slechts',
sommige',
spoedig',
steeds',
svp',
tamelijk',
te',
tegen',
ten',
tenzij',
ter',
terwijl',
tgv',
tijdens',
toch',
toe',
toen',
tot',
totdat',
tov',
tussen',
u',
uit',
uitgezonderd',
uw',
vaak',
van',
vanavond',
vandaan',
vanmiddag',
vanmorgen',
vanuit',
vanwege',
vb',
vd',
veel',
verder',
vervolgens',
via',
volgens',
voor',
vooraf',
vooral',
vooralsnog',
voorbij',
voordat',
voorheen',
vrij',
vroeg',
waar',
waardoor',
waarom',
waarschijnlijk',
wanneer',
want',
waren',
was',
wat',
we',
weer',
wegens',
wel',
welke',
werd',
wezen',
wie',
wij',
wijzelf',
wil',
worden',
wordt',
ws',
zal',
ze',
zelf',
zelfs',
zich',
zichzelf',
zij',
zijn',
zo',
zodra',
zonder',
zou',
zowat'

Appendix 3 – Specifications of ML models

Logistic Regression
Model: sklearn.linear_model.LogisticRegression
Settings:

- Random state = 0
- other = default

Support Vector Machines
Model: sklearn.svm.SVC
Settings:

- kernel type = polynomial
- degree = 1
- Regularization parameter = 3
- Other = default

Decision Tree
Model: sklearn.tree.DecisionTreeClassifier
Settings:

- Function to measure the quality of a split = gini
- maximum depth of the tree = 3
- Other = default

Random Forest
Model: sklearn.ensemble.RandomForestClassifier
Settings:

- number of trees = 500
- maximum depth of the tree = 3
- random state = 42
- Other = default

Appendix 4 – Performance metrics

	Predicted (PN): 0	**Predicted (PP):** 1
Actual (N): 0	True negative (TN)	False positive (FP)
Actual (P): 1	False negative (FN)	True positive (TP)

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

$$\text{Precision} = \frac{TP}{PP}$$

$$\text{Sensitivity} = \frac{TP}{P}$$

$$\text{Specificity} = \frac{TN}{N}$$