Cost-Effectiveness of Expensive Drugs

Analysing dispensing practices, costs
and effectiveness of expensive drugs

Author
Jeroen van Kasteren

Intership Report
Vrije Universiteit Amsterdam
Dutch Hospital Data

December 18, 2018

Cost-Effectiveness of Expensive Drugs

Analysing dispensing practices, expenses and effectiveness of expensive drugs

Submitted by:

B.Sc. Jeroen E.L. van Kasteren

Submitted to:

Vrije University Amsterdam
faculty of science
Business Analytics
De Boelelaan 1081a
1081HV Amsterdam

Host organization:

Dutch Hospital Data
team informatieproducten
Oudlaan 4
3515 GA Utrecht

graduation supervisor:
Dr. Rikkert Hindriks

external supervisor:
M.Sc. Emile Strijbos

second reader:
Dr. René Bekker

December 18, 2018

1 Preface

This thesis is written to complete my two-year master Business Analytics at the Vrije Universiteit Amsterdam and to obtain the degree of MSc. Business Analytics is a multidisciplinary program, aimed at improving business processes with mathematics, computer science and economics, while developing quantitative and communication skills. This research is conducted during a six-month internship at Dutch Hospital Data. Dutch Hospital Data gathers and manages data of Dutch hospitals and provides the hospitals with insights from the data. This supports the hospitals with the continuous improvement of the quality of health care.

This research focuses on the costs of expensive drug dispenses in Dutch hospitals and the effectiveness of those drug treatments. The relations between the cost and effectiveness of those drug treatments are investigated. More specifically, it is addressed how a cost-effectiveness analyses can be conducted on drug treatments that can be generalised over multiple disorders.

This research would not be possible without the collaboration with DHD. Therefore, I would like to thank DHD for providing me the opportunity and utilities to conduct this research[1]. In particular I would like to thank my external supervisor Emile Strijbos of Dutch Hospital Data for his continuous supervision and help throughout this project and my graduation supervisor Rikkert Hindriks of the Vrije Universiteit Amsterdam for his guidance and ideas.

[1] In this research the data were managed with Visual Studio and analysed with the open source R software. The thesis was written with the open-source language LaTeX in Overleaf. I would like to thank all the contributors to the open-source softwares.

2 Management Summary

In the Netherlands, general hospitals and university medical centres are providing health care on a national level. Currently, one of the major challenges for health care are the rising expenditures of expensive drugs. Of the government's health budget, 7% is spent solely on expensive drugs and the these expenditures are growing by 7.7% every year. Hospitals indicate that they are forced to reduce general institution costs due to these rising expenditures. The hospitals and the government want more insights in expensive drug dispense practices and the cost-effectiveness of drug treatments. This will support negotiations with pharmaceutical companies and health insurers to organise better drug procurement and will help reduce the number of ineffective treatments that are prescribed. Therefore, this research answered the question:

How can a cost-effectiveness analysis be conducted on the dispense practices of expensive drugs in Dutch hospitals with the current available data?

This concerns a cost-effectiveness analysis that concludes per patient if an expensive drug treatment was cost-effective. It was preferable that the set-up can be generalised over multiple diagnoses and can be conducted with data of Dutch Hospital Data describing drug dispenses and admissions from, among others, all general hospitals and all university medical centres in the Netherlands.

To conduct the cost-effectiveness analysis, the treatment costs per patient were obtained by calculating the total expenses of drug dispenses and the treatment effectiveness was obtained by combining admission quality indicators describing in-hospital mortality, admission duration, readmissions and the total number of admissions. To fairly compare treatments of patients, the effects in treatment costs and treatment effectiveness due to patient characteristics, as age and disorder, should be filtered out, this is known as risk adjusting or case mix correction. This research focused on risk adjusting the treatment costs with various regression models. The results showed that LASSO Linear Regression was most suitable to do the risk adjustment.

To answer the research question, the data were first pre-processed by imputing or excluding invalid and missing values and risk adjusting the treatment costs and effectiveness. Subsequently, the cost-effectiveness was determined by measuring linear dependencies between the treatment costs and the treatment effectiveness and by analysing regression models that modelled the treatment costs. This shows how it is possible to do generic cost-effectiveness analyses on dispense practices of expensive drugs.

The results showed that a lot of noise was present in the data, preventing the regression models from properly risk adjusting the treatment costs. This can be solved by taking more variables into account. It is discouraged to use variables that indirectly measure treatment costs and treatment effectiveness, but variables are needed that directly measure the stage of a disorder, the patient's fitness, how well the patient recovered and how the patient experienced the treatment.

Future work is recommended to focus on the valuable information that is already in the data and fully exploit it, instead of indirectly extracting features from the data. Information should be extracted about which treatments are prescribed in combination with which interventions, combined with the number of admissions patients underwent with which complications. When this information is available to medical personnel, they can make conclusions about which treatments are cost-effective and this will help with better drug procurement and reduce the number of ineffective treatments that are prescribed. Eventually, this will reduce the expenditures of expensive drugs and improve the quality of health care.

Contents

3 Introduction

In the Netherlands there are different types of hospitals. The main types are general hospitals, specialised clinical hospitals and university medical centres. The 57 general hospitals are responsible for providing regular care in the Netherlands [66]. The 26 specialised clinical hospitals can provide complex care in one or more care areas, besides providing regular care [84]. Moreover, the specialised clinical hospitals also perform scientific research in one or more care areas. The eight academic university medical centres perform scientific research, educate medical specialists and provide highly specialised care, besides providing regular care [65].

The hospitals are providing health care on a national level. As an example, the Dutch general hospitals treated 8.3 million patients in 2017 only [67]. To keep providing the best possible health care to every patient, it is necessary to keep health care affordable and continuously improve health care. For the continuous improvement of health care, the registration load has to be reduced, policy-making has to be improved and cost reduction has to be accomplished [2]. In the Netherlands, Dutch Hospital Data (DHD) is a institution founded by the NVZ[1] and the NFU[2] to gather, process and manage anonymised patient data of hospitals in the Netherlands. In this way, DHD provides hospitals insights in the data to support hospitals with continuous improvement by strengthening the information position of hospitals on a national level and providing anonymised data[3] for scientific research [22].

Currently, one of the major challenges for health care in the Netherlands are the rising prices of expensive drugs [60]. Of the government health budget, 7%[4] is spent solely on expensive drugs. Furthermore, the expenditures of expensive drugs are growing by 7.7% every year[5] [89, 26, 67]. Which drugs qualify as expensive drugs is determined by the NZa[6] and ZiN[7] on behalf of the Ministry of Health, Welfare and Sport[8]. They consider drugs and coagulation factors as expensive drugs when they cost a hospital on average more than €1000,- per year per patient[9]. General hospitals indicate that they are forced to reduce general institution costs due to the rising prices of expensive drugs [68]. Cabinet Rutte III wants to negotiate better drug procurement on a national and international level and appeal the pharmaceutical industry to be transparent in their expenditures for expensive drugs [60, 61].

Providing an overview of current drug dispense practices with benchmark information on other hospitals supports hospitals in negotiations with health insurers on drug prices. Therefore, on behalf of Dutch hospitals, insights on a national level are needed in the cost-effectiveness of dispense practices of expensive drugs. DHD already provides insights in costs of dispense practices of expensive drugs through the GMM[10]. Adding insights about the effectiveness of drug treatments is currently a main

[1] Nederlandse Vereniging van Ziekenhuizen, translates to the Dutch Association of Hospitals.

[2] Nederlandse Federatie van Universitair Medisch Centra, translates to the Netherlands Federation of University Medical Centres.

[3] The data are presented at a group level, to ensure it can not be traced to individual patients or hospitals, unless explicit permission has been granted by the hospital.

[4] 4.8 billion of 73.5 billion in 2017 of health budget (Budgettair Kader Zorg) [26].

[5] The growth in expenditures are a result of drugs being developed for very specific conditions, concerning small specific populations. The new drugs are more effective in comparison with traditional less specific drugs. The consequence is that the new drugs are less often dispensed, making it harder to cover developing costs, increasing the price of the drug. Moreover, the number of patients that use expensive drug is growing, increasing the total price.

[6] Nederlandse Zorgautoriteit (NZa), translates to the Dutch Health Authority.

[7] Zorginstituut Nederland (ZiN), translated to Dutch National Heath Care Institute.

[8] Ministerie van Volksgezondheid, Welzijn en Sport (VWS).

[9] A drug or coagulation factor that costs a hospital more than €1000,- per year per patient is also known as an add-on drug [68].

[10] Geneesmiddelenmonitor (GMM, Drug Monitor). Through the GMM hospitals can gain insights by comparing quantities and costs of expensive drug dispenses in their own hospital with other hospitals in the Netherlands.

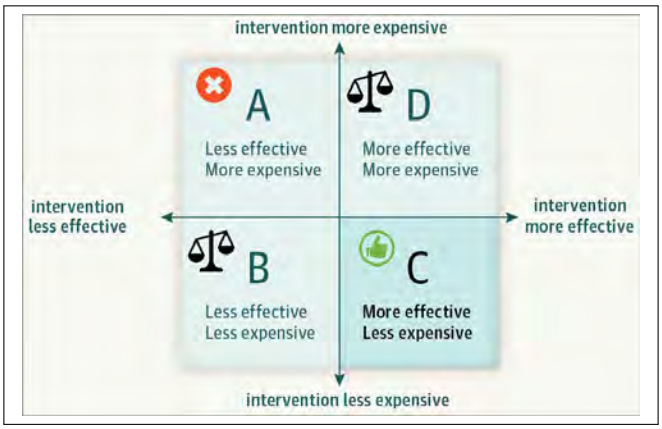

Figure 3.1: A visualisation of the outcome from a cost-effectiveness analysis where an intervention is compared with current practices [73].

point of focus. Zhang et al. (2010) [94] outlined that little was known about how the effectiveness of drug treatments varied among hospitals in the USA and whether any of the variation in effectiveness is associated with variation in drug expenses. A cost-effectiveness analysis could provide insights in drug expenses and the effectiveness of drug treatments. Rawlins (2016) [73] illustrated in a diagram the insights that can be obtained from a cost-effectiveness analyses, as shown in Figure 3.1.

The World Health Organisation points out that many cost-effectiveness analyses have already been conducted in the area of health care [4]. To date, however most cost-effectiveness analyses are qualitative and not quantitative and most analyses are focused on new specific interventions and specific diagnoses and cannot be generalised to the entire population of diagnoses [79]. In these studies, a new intervention for a specific diagnosis is evaluated on costs and health effects and compared with the costs and health effects of current practices. Examples of such analyses about drug treatments are McCune et al. (2013) [59] and Blommenstein et al. (2016) [8]. These studies are used to create national and international guidelines [39]. However, Hulscher et al. (2010) [38] and Shalit et al. (2008) [79] pointed out that still 30% to 40% of all patients do not receive drug treatments following guidelines, even though the guidelines are based on scientific evidence. This demonstrates the need to give hospitals an overview and better insights in the cost-effectiveness of current drug dispense practices.

Therefore, this research aims to examine the cost-effectiveness of drug dispense practices in Dutch hospitals. The main question of this research is:

How can a cost-effectiveness analysis be conducted on the dispense practices of expensive drugs in Dutch hospitals with the current available data?

This research will investigate if a cost-effectiveness analysis can be conducted that can be generalised over multiple diagnoses. The main focus of this research is to make sure patients are comparable by risk adjusting the treatment cost and treatment effectiveness derived from expensive drug dispense data and admission data.

This paper is organised as follows. First, Chapter 4 gives an overview of related work. Thereafter, the dataset is described in more detail with applied pre-processing steps in Chapter 5. In Chapter 6 an extensive explanation is given of all methods and models used for pre-processing the data, modelling treatment costs and measuring relations. The experimental set-up to evaluate the algorithms and answer the research question is described in Chapter 7 whereas the results are presented in Chapter 8. Chapter 9 presents all conclusions and answers the research question. This is followed by the discussion in Chapter 10 and all recommendations for further research in Chapter 11.

4 Related Work

This chapter will review a variety of studies that have been performed in the field of cost-effectiveness analyses in health care. Several studies will be addressed in Section 4.1, *Data of Cost-Effectiveness Analyses*, that outline the data characteristics of medical data used in cost-effectiveness analyses. Thereafter, studies will be discussed that investigated quality indicators which can be used to measure the effectiveness of treatments in Section 4.2, *Quality Indicators*. At last, the ecological fallacy is discussed in Section 4.3, *Ecological Fallacy*.

4.1 Data of Cost-Effectiveness Analyses

Various cost-effectiveness analyses are studied to learn about how a cost-effectiveness analysis can be conducted. Examples of drug dispense cost-effectiveness analysis in the literature are the study of McCune et al. (2013) [59], about the cost-effectiveness of a specific intervention for acute myeloid leukemia and Blommenstein et al. (2016) [8], about the cost-effectiveness of interventions for multiple myeloma. Such cost-effectiveness analyses only concern a specific diagnosis and specific new intervention, which is most common as indicated by the World Health Organization [4]. However, the approaches and results of these studies are hard to generalise to multiple diagnoses. Moreover, due to the sensitive nature of medical data, most studies rely on small datasets. To counter this problem, multiple studies generated data with Monte Carlo simulation [55, 53].

Health care expenditures and treatment effectiveness indicators are typically characterised by non-negative observations with many zero outcomes [55, 53]. As illustration, only 2.0% of the patients treated by Dutch hospitals received expensive drugs in 2017[1]. Therefore, 98.0% of the observations will be zero when analysing the costs of expensive drug dispenses per patients. The non-zero observations can be analysed by only taking patients into account that actually received an expensive drug. Usually, the non-zero observations have a positively skewed distribution with a heavy tail[2]. An example is given in the left graph of Figure 4.1, where the right graph of Figure 4.1 clarifies what it means to have a positively skewed distribution. Malehi et al. (2015) [53] state that the skewed nature of health care cost is one of the main issues when health care costs are modelled. They discuss various models that can cope with the skewed nature of the costs. Two of the models that performed well on health care data are also used in the experimental set-up of this research. Namely, Gamma Regression and Linear Regression with a logarithmic transformation of the response variable.

4.2 Quality Indicators

To conduct a cost-effectiveness analysis that can be generalised to multiple diagnoses, quality indicators are needed that can be generalised to multiple diagnoses and describe the effectiveness of drug treatments. The literature is reviewed to learn about which quality indicators should be used. As example, several studies use the number of occurrences of prescribing errors as quality indicator for the effectiveness of drug treatments. Dean et al. (2000) defined a prescribing error as, "A prescribing error occurs when, as a result of a prescribing decision or prescription writing process, there is an unintentional significant reduction in the probability of treatment being timely and effective or increase in the risk of harm." Lewis et al. (2009) [48] reported that prescribing errors mainly happen through prescribing drugs for which a patient is allergic, prescribing the wrong doses or prescribing with the wrong frequency, prescribing without proper instructions, or prescribing while a patient takes conflicting drugs. Their study pointed out that prescribing the wrong dose is most common. Moreover, they remarked that off-label use is not necessarily wrong and it is not a prescribing error when a patient wishes to use the drug.

[1] This percentage is obtained by analysing the data of DHD available for this research.

[2] A heavy tail means that many extreme values are present at that side of the distribution.

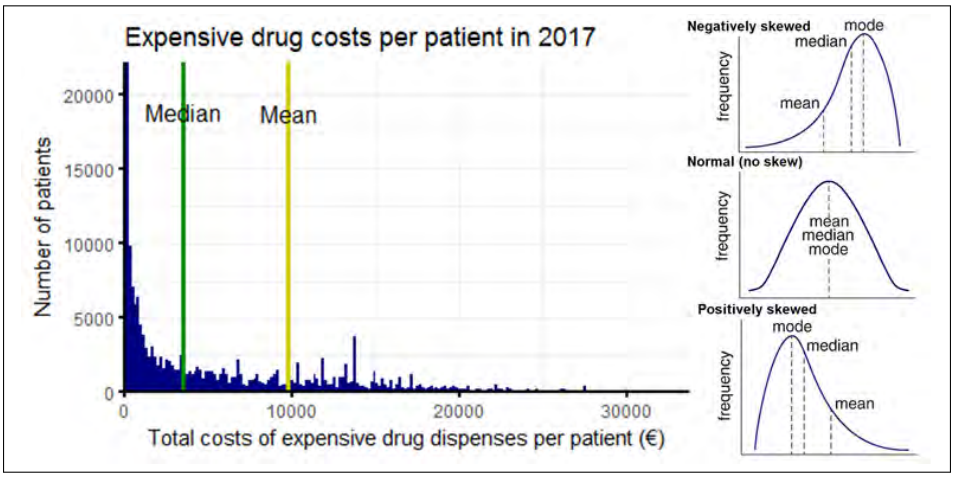

Figure 4.1: The left graph shows the total costs of expensive drug dispenses in 2017 per patient that received expensive drugs in 2017. The figure clearly shows a positively skewed distribution. The vertical green line shows the median and the yellow line shows the mean. Source: expensive drug data of DHD.
The right graphs present illustrations of a negatively skewed distribution, a distribution without a skew and a positively skewed distribution. It is also shown how the mean, median and mode are influenced by the skewness of the distributions. The skewness of a distribution can also be measured by the kurtosis.

The quality indicators used to describe effectiveness should contain enough variation for drug treatments, making them useful to identify effective and ineffective treatments. Zhang et al. (2010) [94] studied variation in prescribing high-risk drugs in the USA, by investigating the number of prescribing errors[3]. They found big differences between hospitals in prescribing practices, even after risk adjusting for differences in hospitals their patient populations. Additionally, the study of Smith et al. (2016) [82] also concluded that hospitals differ in treatment preferences[4]. This shows that analysing the effectiveness still contains insights that are needed in health care, to prevent prescribing errors.

Remark that hospitals have different patient populations and different patients need different drug treatments, resulting in varying treatment costs and effectiveness. When conducting a cost-effectiveness analysis, patients and hospitals are compared. Zhang et al. and Smith et al. both indicated that treatment costs and treatment effectiveness need to be risk adjusted for patient characteristics to fairly compare patients and hospitals[5].

Furthermore, Zhang et al. also concluded that less drugs is cheaper and can avoid adverse events while prescribing. Therefore, using the amount of drugs can be used as quality indicator. Finally, Zhang et al remarked that differences were found between hospitals in coding of similar patients. This made the patient population risk adjustment less accurate and less reliable. It should be discussed with experts if differences in coding are present between hospitals when doing a cost-effectiveness analysis and what the influence is on the analysis.

Next to prescribing errors, other quality indicators were found in the literature that could be used in cost-effectiveness analyses. For instance the indicators used in the Transparantiekalender[6] of ZiN[7].

[3]In the research of Zhang et al. (2010) [94], prescribing errors were defined as prescribing high-risk drugs to elderly, while the drugs are harmful for elderly or prescribing high-risk drugs to patients with a comorbidity, while the drugs are harmful for patients with that comorbidity.

[4]Smith et al. even found that the same disorder was treated differently dependent on the ethnicity of a patient. Although ethnicity is a patient characteristic, it should not be part of the characteristics that are used for risk adjustments.

[5]Lane-Fall and Neuman (2014) [46] explain: "Risk adjustment (also known as severity adjustment) is the process of statistically accounting for differences in patient case mix that influence health care outcomes."

[6]Translates to Transparency calendar.

[7]Zorginstituut Nederland, Translates to Dutch National Health Care Institute.

However, these indicators mostly focus on what care a hospital can provide and which facilities are present and not on the effectiveness of the provided care. Furthermore, the indicators of the Basisset Medisch Specialistische Zorg Kwaliteitsindicatoren[8] of the IGJ[9] is focusing on what percentage of the patients had which treatments [40]. Nevertheless, the indicators of both sets are for specific disorders and cannot be generalised. This shows that not every quality indicator can be used for this research. Forster et al. (2012) [25] even criticised the use of indicators to measure patient well-being and treatment effectiveness. They argue that most indicators are based on what can be measured and not on what should be measured. Moreover, it can create a perverse incentive to push for better indicator outcomes[10] instead of pushing for real effectiveness of care. This should be taken into account when define the quality of treatments.

When the treatment effectiveness is measured by quality indicators, the quality indicators are compared with the treatment costs to infer conclusions about the cost-effectiveness of treatments. Therefore, the quality indicators should be relevant to compare the indicator with the treatment costs. For instance, Dedier et al. (2001) [20] and Schouten et al. (2005) measured effectiveness of drug treatments by investigating if several interventions were performed within a limited time since the start of the admission. However, only modest correlations were found, indicating that the quality indicators were not relevant. They highlighted the difficulty of linking general outcome measures to the actual effectiveness of care. Which should be taken into account when doing a cost-effectiveness analysis.

More research was reviewed about other general quality indicators describing the quality of admissions. Which was relevant for this research, because those quality indicators are present in the available dataset. Cihangir et al. (2013) [16] demonstrated that the Unexpected Long Length of Stay (UL-LOS) admission quality indicator can be used to detect adverse events. The UL-LOS is a risk adjusted outcome and thus can be used for comparing hospitals with different patient populations. Remark that risk adjusting is statistically accounting for differences in patient characteristics that influence the outcomes, to make sure different patients can be compared fairly, because the differences in outcome due to patient characteristics as age and disorder are removed from the outcome[46]. The risk adjustment for the UL-LOS is done with the Grouped Mean[11] and is based on a patient's diagnoses and age. Fischer et al. (2014), Shams et al. (2015) [80] and Hekkert et al. (2017) [32] address that the potentially preventable readmission ratio can be used as an admission quality indicator to identify the quality of an admission. Van der Laan et al. (2015) [44] discuss the risk adjustment of the Readmission Ratio, conducted with Logistic Regression. Furthermore, Van der Laan et al. (2015) [45] devoted attention to the Hospital Standardized Mortality Ratio (HSMR) as risk adjusted quality indicator for admissions. The HSMR is also risk adjusted with Logistic Regression.

Lingsma et al. (2018) [50] discuss the complex correlations between the HSMR, UL-LOS and readmission ratio. A composite quality indicator is proposed based on patient surveys and expert reviews. The composite quality indicator can be used to show the effectiveness of treatments in a compact and clear way. The composite quality indicator is as follows, ordered from best to worst outcome of a patient's treatment:

(A) Survived treatment without UL-LOS and without readmission.
(B) Survived treatment with UL-LOS, but without readmission.
(C) Survived treatment without UL-LOS, but with readmission.
(D) Survived treatment with UL-LOS and with readmission.
(E) Did not survive treatment.

[8] Translates to Basic Set Specialised Medical Care Quality Indicators.
[9] Inspectie Gezondheidszorg en Jeugd, Translates to the National Healthcare Institute.
[10] Hospitals can influence indicator outcomes by using different admission policies or different policies in coding.
[11] The Grouped Mean is also used in this research and is explained in Subsection 6.2.2, *Grouped Mean*.

4.3 Ecological Fallacy

When performing an analysis were multiple levels exits, the ecological fallacy should be taken into account. This research, dealing with health care data, can analyse relations on a hospital level and on a patient level. Hofstede et al. (2017) [36] discussed that relations between indicators deduced at a grouped level, should not be used to conclude about relations at an individual level, this is known as the ecological fallacy. Hofstede et al. showed that significant correlations on a hospital group level can even be inversely correlated on an individual patient level.

Lastly, the study of Marang-van de Mheen and Shojania (2014) [56] demonstrated how Simpson's paradox applies to health care quality indicators, even with perfect risk adjustment. This should be taken into account when measuring quality of diverse groups. A theoretical example from Manktelow et al. [54] is given in Table 4.1. Two hospitals are presented that both treat a certain number of patients. The patients are from two risk groups and per risk group the hospitals have an identical HSMR. However, Hospital B has on average a higher HSMR, because Hospital B treats more high risk patients. Therefore, hospitals with more high risk patients are penalised by a higher HSMR. This is not overcome by risk adjustment, as shown in the example. Marang-van de Mheen and Shojania argue that the HSMR should not be used to compare hospitals. It should only be used for a specific hospital to quantify their own quality of care.

Table 4.1: Simpson's paradox for the HSMR.

	Hospital A	**Hospital B**
Low-risk patients	100	50
Deaths	10 (10%)	5 (10%)
Number expected	10 (10%)	5 (10%)
HSMR	100 ($\frac{10}{10} \cdot 100$)	100 ($\frac{5}{5} \cdot 100$)
High-risk patients	50	100
Deaths	15 (30%)	30 (30%)
Number expected	10 (20%)	20 (20%)
HSMR	150 ($\frac{15}{10} \cdot 100$)	150 ($\frac{30}{20} \cdot 100$)
All patients	150	150
Deaths	25	35
Number expected	20	25
HSMR	125 ($\frac{25}{20} \cdot 100$)	140 ($\frac{35}{25} \cdot 100$)

5 Data Description and Preparation

In this chapter the available dataset will be described in detail in Section 5.1, *Data Description*, first. Second, all applied pre-processing steps will be discussed in Section 5.2, *Pre-processing*.

5.1 Data Description

In this section the dataset, known as the LBZ[1] will be described first in Subsection 5.1.1, *LBZ*. Second, the quality indicators that can be derived from the LBZ will be addressed in Subsection 5.1.2, *Quality Indicators*. The LBZ is the registration of medical, administrative and financial data of general hospitals, specialised clinical hospitals and university medical centres. Figure 5.1 illustrates the LBZ database and its components. For this research, LBZ data of the years 2015, 2016 and 2017 are used.

5.1.1 LBZ

Currently, the LBZ contains data of 88 hospitals[2]. In the LBZ, a hospital is identified with an unique AGB code[3]. Next to the AGB code, the hospital name and hospital type are documented. Furthermore, an indication of the size of every hospital was needed for this research. This could be derived from the LBZ by calculating the number of patients treated per hospital per year. A patient was defined as treated by a hospital when the patient was admitted to the hospital, received ambulant care or got an intervention[4]. Table A.1 in Appendix A.1.1 summarises the above described general data available per hospital.

Hospitals provide admission data for the LBZ. The admission data consider clinical admissions, extended observations, day nursing, outpatient treatments and ambulant care. From 2015 till 2018 there are 8.2 million recorded admissions of 5.5 million patients in the dataset. Per admission the hospital is registered that admitted the patient with start and end date of the admission. The patient that is admitted is registered with an patient ID[5] and a pseudonymized BSN[6]. Moreover, the determined main diagnosis for which the patient was admitted is stored, with corresponding severity[7]. The data also contain the place of origin of the patient before the admission and if the admission is urgent[8]. For every admission it is known what the expected mortality rate is and if the patient survived the admission. Furthermore, it is stored what the expected Length of Stay (LOS) of the admission is in

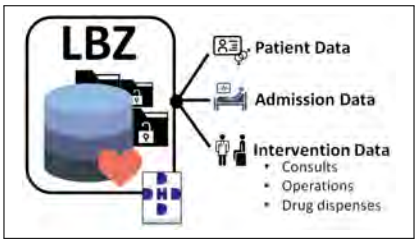

Figure 5.1: This figure shows the LBZ database containing medical data. The LBZ contains data about patients, admissions and interventions.

[1] Landelijke Basisregistratie Ziekenhuiszorg, translates to National Basic Registration of Hospital Care.
[2] Remark that almost all hospitals supply data to the LBZ.
[3] Algemeen GegevensBeheer-code (AGB), translates to General Data Management code. Every health care institution has a unique AGB code commissioned by Vektis.
[4] Note that outpatient care is not included.
[5] The Patient ID is only unique within a hospital.
[6] Burger Service Nummer, translates to Citizen Service Number
[7] The historic mortality rate for this diagnosis, annually calculated by the CBS.
[8] Urgent indicates that care was absolutely needed within 24 hours.

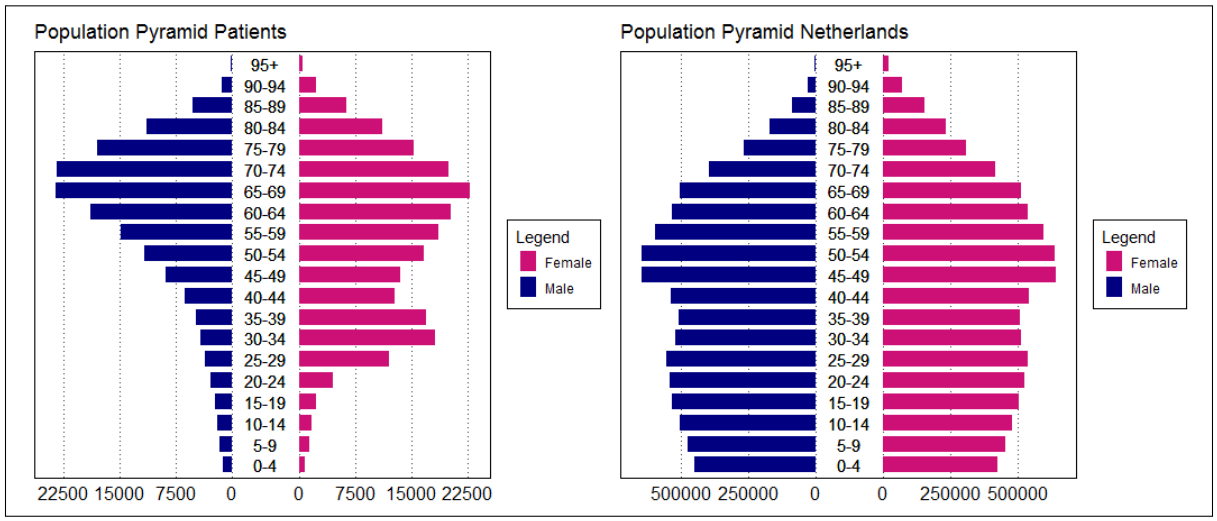

Figure 5.2: On the left side, this figure shows the population pyramid of all patients that received expensive drugs in 2015, 2016 and 2017. As reference, the population pyramid of all Dutch citizen in 2017 is presented on the right.

days and what the realised LOS is. Finally, the chance the patient is readmitted within 30 days after discharge is present with an indicator recording if the patient was actually readmitted. Table A.2 in Appendix A.1.2 summarises the above described general data available per hospital.

Next to admission data, hospitals also supply data about interventions to the LBZ. Interventions are for example medical operations and expensive drug dispenses. The dataset covers 4 million drug dispenses prescribed to 4 hundred thousand patients in 86 different hospitals from 2015 till 2018[9]. As an illustration, Figure 5.2 shows the pyramid plot of all patients that received expensive drugs from 2015 till 2018. Per drug dispense, data are recorded about which drug is dispensed on which date in which quantity for what costs. Furthermore, the patient that receives the drug is registered by a patient ID[5] along with the patient's age, sex, zip code and nationality. The hospital that dispenses the drug is registered with the name of the hospital and its AGB code. The diagnosis for which the drug is prescribed is stored along with the specialist who prescribed the drug. Extra information is kept indicating if the drug was prescribed for the condition it was meant for[10] and if the costs are covered by the insurance. Table A.3 in Appendix A.1.3 summarises the above described general data available per hospital.

5.1.2 Quality Indicators

There are no indicators available that measure the effectiveness of expensive drug dispenses. However, some general indicators are available that measure the quality of admissions. This considers the Hospital Standardized Mortality Ratio (HSMR), the percentage of Unexpected Long Length of Stays (UL-LOS) and Readmission Ratio, which are also addressed in Chapter 4, *Related Work*. Research has shown that these indicators correlate with adverse events of treatments [16, 32]. Assuming that adverse events indicate a poor quality of care and an ineffective treatment, these indicators can be used to derive the effectiveness of the drug treatment.

The HSMR is the ratio of observed in-hospital mortality divided by expected in-hospital mortality [45]. In-hospital mortality is registered by hospitals and provided to the LBZ and the expected in-hospital mortality is calculated by the CBS. Similar, the Readmission Ratio is the ratio of observed readmissions

[9] Remark that not all hospitals that supply LBZ data, also supply data about expensive drug dispenses.

[10] A drug has one or more conditions it is meant to be used for. If the drug is used for one of those conditions, the product is on-label. Otherwise, it is off-label.

divided by the expected readmissions [44]. Which admissions qualify as readmissions is determined by the CBS and the CBS calculates the number of expected readmissions. Lastly, the UL-LOS is the percentage of UL-LOS admissions to the number of admissions [28]. Which admissions are UL-LOS is determined by DHD[11].

5.2 Pre-processing

To make the data analysable, various tables were combined and the data were cleaned to create the dataset suitable for this research. All pre-processing steps are described in this section[12]. First will be discussed how invalid values are handled in Subsection 5.2.1, *Invalid Values*. Afterwards will be addressed how outliers are detected and managed in Subsection 5.2.2, *Outlier Detection*.

5.2.1 Invalid Values

In the available data, not every value is correctly registered for every observation. To cope with missing, inconsistent and wrong values, the following checks and clean ups are incorporated in the pre-processing steps.

First, the recorded zip codes were not always in the same format and sometimes the zip codes were missing. All valid zip codes are transformed to four digit zip codes, because the first four digits of a valid Dutch zip code[13] are needed to link patients to their Social Economic Status (SES). The SES is dependent on time, and calculated every four years. For all years without SES, the last known SES of that zip code is used. When the zip code was invalid and no SES could be linked, the SES was imputed with the average SES, in line with earlier research [44, 45].

Second, it is desirable to have one specialist per drug dispense record. Nevertheless, the specialist is documented in three different ways. The first way is through the specialist who prescribed the drug, which is missing in 5% of the dispenses. The second way is through the DBC code[14] of the diagnosis which is accompanied with a DBC specialist code, which is missing for 12% of the dispenses. The third way is through the specialist who dispensed the drug, which is missing for less than 1% of the dispenses. In line with existing procedures of DHD, the prescribing specialist was chosen to be the primary specialist[15]. When the primary specialist was missing, it was first filled with the specialist who dispensed the drug. When the specialist who dispensed the drug was missing too, the DBC specialist was used. Afterwards, it was checked if the new specialist code with the DBC diagnose code still was a valid DBC diagnose specialism combination, following the DBC methodology of the NZa.

Third, the dispensed drug was registered based on its active substance with the Anatomical Therapeutic Chemical (ATC) Classification System [93]. With the drug database G-Standaard of Z-Index[16] was checked if a drug dispense was registered with a valid ATC code. Analyses showed that 0.1% of the ATC codes were invalid. This only considered missing ATC codes, when an ATC code was registered, it always was a valid ATC code.

[11] An admission is UL-LOS when the length of stay is 50% longer than expected. The expected length of stay of an admission is the average length of stay of similar historic admissions in the LBZ. An admission is similar when the admitted patients have the same characteristics. The patient characteristics are diagnosis and age.

[12] The data could be obtained and manipulated with the Structured Query Language (SQL). SQL is a standardized language for defining and manipulating data tables in a relational database [14].

[13] A Dutch postcode is only valid when it is registered in the Register of Addresses and Buildings maintained by The Netherlands Cadastre, Land Registry and Mapping Agency, in short the Kadaster.

[14] Diagnose Behandeling Combinatie code (DBC, Diagnosis Treatment Combination code) determined by the NZa.

[15] Noteworthy is that the specialist who prescribed the drug was often the pharmacist. Knowing that the pharmacist dispensed the drug would not add valuable information. Therefore, the specialist who dispensed the drug was chosen as primary specialist.

[16] The G-Standaard is a drug database of all products dispensed or used in the pharmacy. The G-Standaard is maintained by Z-Index on behalf of the NZa.

Fourth, the costs for a drug dispense could be derived in two ways. The first way is by using the maximum tariff of the NZa[17]. By multiplying the NZa maximum tariff with the amount dispensed, the costs of the drug dispense could be obtained. The second way is by using the costs registered by the hospitals. Nevertheless, the costs registered by hospitals are invalid in 31% of the records, while the costs calculated with the NZa maximum tariffs are only invalid in 7% of the records. Moreover, the amount registered by hospitals is suspiciously low[18] for one third of the registered records. Therefore, for this research the costs obtained through the NZa maximum tariff is used as the costs for a drug dispense. It is necessary to remark that the drug dispense costs and the amount dispensed are only valid when they are not missing and are greater than zero.

Fifth, the pseudonymized BSN of a patient is registered when the patient is admitted. For 8% of the admitted patients the BSN is not registered. Nevertheless, when receiving an expensive drug, the patient does not have to be admitted. Consequently, 33% of the patients that receive expensive drug have a missing BSN code.

In several cases it was possible to impute invalid values. Those cases are summarised below:

- When the patient's nationality was missing, it was imputed with the Dutch nationality, because 99.8% of the patients are Dutch residents.
- When the patient's zip code was invalid, the SES of the patient could not be derived. In those cases the SES was set to *Average*, this is in line with research of the CBS [44, 45].
- When the On-/Off-label is missing, it is replaced by on-label. For on-label is the most common label, covering 67% of the dispenses.
- When the insurance status is missing, it is set to insured. For insured is the most common label, covering 98% of the dispenses.
- When the BSN of a patient is missing, a unique patient identifier is created based on the hospital code and the patient ID. In a hospital exceeding analyses, this results in all patients without BSN being viewed as patients that are treated in only one hospital.

Finally, observations are excluded from the dataset when certain irreplaceable variables are not validly registered. Imputing these variables could assign dispenses to wrong groups with incorrect costs, which could influence results too much. The irreplaceable variables are summarised in Table 5.1. In addition, per irreplaceable variable is presented what percentage of the observations do not have this value validly registered. In total 8.2% of the observations are excluded due to invalid irreplaceable variables.

Table 5.1: Essential variables

Value name	**% not valid**
Name of the hospital	0
Unique identifier for the patient	0
Age of the patient	0.1
ATC code of the dispensed drug	0.1
Dispensed quantity	0.1
Cost of the drug dispense	7.1
DBC diagnose specialism combination	2.4

[17] Based on the purchase price of the pharmacy, a maximum is set on the amount charged by the hospital for dispensing the drug by the NZa. The NZa maximum tariff is usually used as cost for dispensing the drug.

[18] Less than 50% of the NZa maximum tariff.

5.2.2 Outlier Detection

An outlier can be defined as follows [31],

"An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism."

Following this reasoning, outliers should be removed from the dataset, to ensure that analyses are not influenced by factors outside the scope of the analyses. In this section will be addressed how outliers are detected in continuous and categorical variables.

The drug dispense data and the admission data contain various continuous variables. No extreme values were found for most variables[19]. Nevertheless, several dispenses had extreme values in dispensed drug quantity and drug costs, that would classify as outlier. It was not straightforward to remove these outliers from the data, because it was not clear which observations were extreme due to real rare cases and due to registration errors. Consequently, the set of outliers are divided into potential outliers and problematic outliers, following the terminology of High (2000) [34]. Potential outliers are outliers that are likely to be real rare cases, while problematic outliers are the most extreme cases that are almost certainly generated by registration errors.

Various methods can be used to detect outliers, a clear summary is given by Seo (2002) [78]. To be applicable to medical data, the outlier detection methods need to cope with several characteristics of the data:

- The data distribution is depended on the dispensed drug and the diagnosis for which the drug is dispensed.
- The data have a skewed distribution for most diagnoses and is not normally distributed.
- Duplicate values often occur due to dispenses in predetermined quantities.

Currently, the Z-score method[20] is used by DHD to detect outliers in the drug dispense data. Nevertheless, the assumption of the Z-score method about normality of the data is violated. Analyses showed that this results in a less reliable behaviour, labelling too many or too few observations as problematic outliers. Therefore, more robust methods are studied. A robust detection method is the LOF method[21]. In theory, the LOF method can distinguish a small group of rare observations from lone problematic outliers, which is essential for not labelling specific treatments of small groups as problematic outliers. In practice, it turned out that the method requested a lot of computational power, making it unsuitable for the large drug dispense dataset. Eventually, the POD method[22] was developed for this research to detect problematic outliers and cope with the characteristics of the drug dispense data. This method is based on the Adjusted Boxplot method[23].

Figure 5.3 shows the amount dispensed of an expensive drug. By visual analysing the distribution, the outliers are detected. The potential outliers are blue encircled and the problematic outliers are red encircled. First, the normal boxplot, known as Tukey's Method, is used to detect outliers. All observations outside the wiskers of the boxplot were labelled as problematic outliers, which contained many potential outliers. The Adjusted Boxplot Method clearly has wider whiskers than Tukey's Method, but still labels potential outliers as problematic outliers. Finally, in the bottom right, the POD method has the largest box and the method labels only actual problematic outliers as problematic outliers. This is not a coincidence, because the POD method was tuned to fit the drug dispense data.

[19] This is due to checks of DHD to ensure data quality.

[20] The Z-score method is explained in the Appendix A.3.1, *Z-score Method*.

[21] Local Outlier Factor method of Breunig et al. (2000) [10], explained in depth in Appendix A.3.2, *Local Outlier Factor Method*.

[22] Problematic Outlier Detection method, explained in Subsection 6.1.3, *Problematic Outlier Detection Method*.

[23] The Adjusted Boxplot is explained in Subsection 6.1.2, *Adjusted Boxplot Method*. This method is based on Tukey's method, explained in Subsection 6.1.1 *Tukey's Method*.

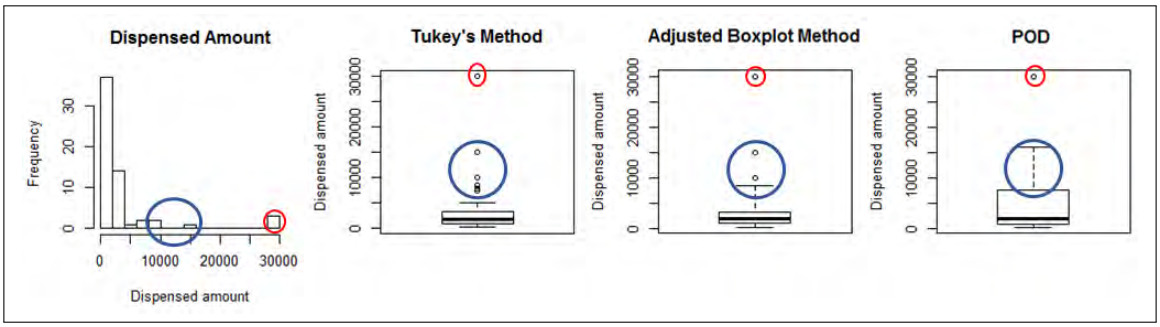

Figure 5.3: All figures show the amounts dispensed of the coagulation factor Octocog Alfa for patients with chronic kidney failure in 2016. The potential outliers are blue encircled and the problematic outliers are red encircled. The first figure from the left shows a histogram of the amounts dispensed. The second figure presents how Tukey's method detected outliers and the third figure shows how the Adjusted Boxplot Method detected outliers. Finally, the last figure illustrates how the POD method detected outliers.

Stratification was used to divide all drug dispenses into multiple homogeneous sub-populations, to improve the accuracy of the outlier detection. In line with the approach of DHD and the CBS [44, 45], stratification was done by creating a group per year that a drug was dispensed, per ATC code of the drug, and per CCS group[24] the drug was prescribed for. Per group was checked if the group contained enough different observations. If the group contained at least 25 observations and the most common value did cover less than 50% of the group, problematic outliers were detected with the POD method. The outlier detection was not conducted on groups with less than 25 observations, because in those groups too few observations were present to make a clear distinction between potential outliers and problematic outliers. Moreover, when more than 50% of the observations had the same value, almost all other observations were labelled as outlier, making the outlier detection unreliable. Afterwards, 22% of all drug dispenses could not be checked on problematic outliers. This was due to 86% of the groups[25] not fulfilling the restrictions[26]. Most groups contained too few patients to do a reliable outlier detection on. All dispenses that are could not be checked for outliers are labelled as valid observations. By performing outlier detection on all other groups, 0.6% of all drug dispenses were labelled as problematic outliers due to an outlier in drug dispense cost or an outlier in the quantity dispensed. In this way the observations with a problematic outlier could be removed.

Most categorical variables did not need outlier detection, for the values are predefined. As example, the SES can only have values 'lowest', 'below average', 'average', 'above average' and 'highest'. Nevertheless, hospitals treat different populations. Consequently, certain populations are hardly represented in some hospitals. When analysing specific populations, hospitals that treat less than 1% of the population do not add value to overall analyses. Similar, if a drug is prescribed to less than 1% of the population, it adds more noise to the analyses, more than it adds value.
For this reason, the categorical variables DBC specialism, DBC diagnosis, ATC drug code and hospital are revised with two specific restrictions. Namely:

- The value needs to cover more than 1% of all selected dispenses.
- When ordering the categorical values in descending order of most occurrences, the value needs to be part of the biggest 90%[27].

Given a selection of the variables DBC specialists, DBC diagnoses, ATC drug codes and hospitals, it is tested per variable if their values meet the requirements. When a value fails both requirements, the value is not taken into account in further analysis.

[24] Clinical Classifications Software group, a diagnosis and procedure categorization scheme for health care. All DBC and ICD-10 codes are divided in 259 clinically meaningful groups.
[25] 14.199 of 16.583 groups.
[26] The restrictions that a group contained at least 25 observations and the most common value would cover less than 50% of the group.
[27] For deciding which hospitals are taken into account, a limit of 95% is used, preventing excluding too many hospitals.

6 Methodology

To conducting a cost-effectiveness analysis on drug dispense practices, regression models will be used to risk adjust the treatment costs for patient characteristics. To increase the accuracy of the regression models, problematic outliers should be removed from the dataset. As addressed in Subsection 5.2.2, removing problematic outliers ensures that the models are not influenced by factors outside the scope of the analyses.

First, the methods to detect problematic outliers will be described in Section 6.1, *Outlier Detection*. Second, the regression models to model the treatment costs will be explained in Section 6.2, *Prediction Costs Models*. Third, K-Fold Cross Validation and various error measures to measure the predicting performance of the regression models will be addressed in Section 6.3, *Prediction Measures*. Finally, Section 6.4, *Correlation Measures*, will review methods to measure correlations. The correlations will be used to measure the cost-effectiveness of drug dispenses.

6.1 Outlier Detection

The Problematic Outlier Detection (POD) method was developed in this research to detect outliers in the drug dispense data. The POD method is based on the Adjusted Boxplot method and the Adjusted Boxplot method is based on Tukey's method. Tukey's method is addressed first in Subsection 6.1.1. Second, The Adjusted Boxplot method is addressed in Subsection 6.1.2. Last, Subsection 6.1.3 will explain the Problematic Outlier Detection method.

All Methods are discussed given the observed variable $X = \{x_1, \ldots, x_N\}$. Define $x_i \in \mathbb{R}$ as observation i, for $i \in \{1, \ldots, N\}$, with N the total number of observations. Define a problematic outlier as $\tilde{x}_i$.

6.1.1 Tukey's Method

Tukey (1977) [86] proposed a way to display continuous univariate data, now known as the boxplot or the box and whisker plot. An theoretical boxplot is presented in Figure 6.1. The boxplot is based on quantiles. Given $\alpha \in [0, 1]$, the α-quantile is the value for which $\alpha\%$ of the data are smaller or equal to that value. In formula form, define the cumulative distribution function $F(x_i)$ and α-quantile function $Q(\alpha)$[1] as follows,

$$\begin{aligned} F(x_i) &= \tfrac{1}{N} \textstyle\sum_{j=1}^{N} \mathbb{1}_{\{x_j \leq x_i\}}, \\ Q(\alpha) &= \inf_i \{x_i, F(x_i) \geq \alpha\}. \end{aligned} \tag{6.1}$$

As shown in Figure 6.1, a box is presented in a boxplot with as lower limit B_{lower} and as upper limit B_{upper}. Furthermore, a line is drawn at the median, $Q(0.50)$. The whiskers of the boxplot are calculated with the Inter Quartile Range (IQR). The lower limit of the whiskers is W_{lower} and the upper limit is W_{upper}. The whisker limits can be calculated using the infimum and supremum[2]. To detect problematic outliers, the follow constants are proposed [86],

$$\begin{aligned} B_{lower} &= Q(0.25); \\ B_{upper} &= Q(0.75); \\ \text{IQR} &= Q(0.75) - Q(0.25); \\ W_{lower} &= \inf_i \{x_i, x_i \geq Q(0.25) - 3 \cdot \text{IQR}\}; \\ W_{upper} &= \sup_i \{x_i, x_i \leq Q(0.75) + 3 \cdot \text{IQR}\}. \end{aligned} \tag{6.2}$$

[1] Remark, the infimum is the smallest value present in a subset. Therefore, $Q(\alpha)$ is the smallest value of all x_i for which $F(x_i) \geq \alpha$.

[2] The supremum is the largest value present in a subset.

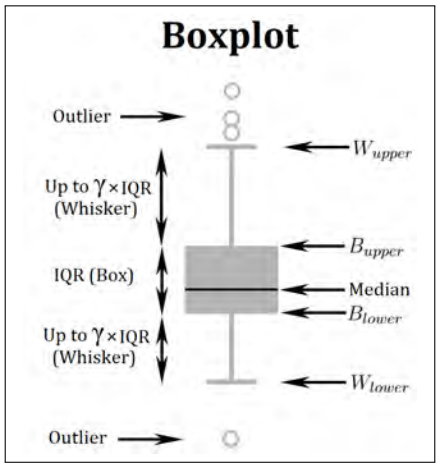

Figure 6.1: This figure shows the basic elements of Tukey's method. $\gamma = 1.5$ is used for regular boxplots, while $\gamma = 3$ is used to detect problematic outliers, also known as far outs.

When an observation x_i lies outside the interval $[W_{lower}, W_{upper}]$, it is presented as dot, and labelled as problematic outlier. When x_i lies between the whisker, but outside the box, it is labelled as potential outlier.

An advantage of Tukey's method is that it is based on quantiles. Therefore, the method is not sensitive to outliers and does not assume the data to have a specific distribution. A disadvantage is that it tends to label observations in the tail of skewed data as problematic outlier. This problem is addressed by the Adjusted Boxplot method.

6.1.2 Adjusted Boxplot Method

The Adjusted Boxplot method adjusts Tukey's method for better outlier detection in case of skewed data [37]. The Adjusted Boxplot changes the limits of the whiskers. For the change, the medcouple (MC) of Brys et al. (2004) is used [12],

$$\begin{aligned} h(x_i, x_j) &= \frac{(x_j - Q(0.50)) - (Q(0.50) - x_i)}{x_j - x_i}, \\ \text{MC} &= \underset{x_i \le Q(0.50) \le x_j}{\text{median}} h(x_i, x_j). \end{aligned} \tag{6.3}$$

Remark that $h(x_i, x_j)$ lies between $[-1, 1]$. Consequently, the MC also lies between $[-1, 1]$, where right skewed distributions will have a positive MC, left skewed distributions a negative MC, and symmetric distributions a MC near zero.

Based on the recommendations of Hubert and Vandervieren (2004) [37], the new boxplot parameters are,

$$\begin{aligned} B_{lower} &= Q(0.25); \\ B_{upper} &= Q(0.75); \\ \text{IQR} &= Q(0.75) - Q(0.25); \\ W_{lower} &= \inf_i \{x_i,\ x_i \ge Q(0.25) - 1.5e^{-4 \cdot MC} \cdot \text{IQR}\}; \\ W_{upper} &= \sup_i \{x_i,\ x_i \le Q(0.75) + 1.5e^{4 \cdot MC} \cdot \text{IQR}\}. \end{aligned} \tag{6.4}$$

When an observation x_i lies outside the interval $[W_{lower}, W_{upper}]$, it is labelled as a problematic outlier. As remarked by Hubert and Vandervieren, the Adjusted Boxplot does not account for tail heaviness. As a consequence, the method can in practice label potential outliers as problematic outliers, when the data have a heavy tail. Furthermore, the method could not cope with the many duplicate values in the data. As a result, this method labelled too many potential outliers as problematic outliers.

6.1.3 Problematic Outlier Detection Method

To create a detection method that could cope with the characteristics of the drug dispense data, the Problematic Outlier Detection (POD) Method was developed for this research. The POD method is based on the idea of the Adjusted Boxplot method to adjust the parameters of Tukey's boxplot. By choosing the right parameters of the boxplot, the accuracy increased of detecting problematic outliers, without detecting potential outliers. The parameters are chosen as follow,

$$\begin{aligned} B_{lower} &= Q(0.10); \\ B_{upper} &= Q(0.90); \\ \text{IQR} &= Q(0.90) - Q(0.10); \\ W_{lower} &= \inf_i\{x_i, x_i \geq Q(0.10) - 2 \cdot \text{IQR}\}; \\ W_{upper} &= \sup_i\{x_i, x_i \leq Q(0.90) + 2 \cdot \text{IQR}\}. \end{aligned} \tag{6.5}$$

When an observation x_i lies outside the interval $[W_{lower}, W_{upper}]$, it is labelled as a problematic outlier. By choosing a wider IQR than Tukey's method, the POD method could deal with data containing a large number of identical values and data with a skewed distribution.

The parameters of the POD method were tuned to fit the drug prescription data. First, the problematic outliers were labelled by visually analysing the distribution of drug dispenses of homogeneous groups[3]. Second, the best parameters of the POD method were found by a local search strategy with trial and error. The POD method was tuned such that not more than 1% of the data[4] are labelled as problematic outliers and potential outliers were not being labelled as problematic outliers.

6.2 Predicting Costs Models

Patients with different characteristics need different treatments, which results in treatment costs varying from case to case. To compare patients fairly, the treatment costs need to be adjusted for patient characteristics. Various regression models are used to model the treatment costs. The outcomes are used to risk adjust the treatment costs for patient characteristics and the outcomes are used to examine which factors have an influence on the treatment costs. The regression models are explained in depth in this section. Remark that for the regression models, the response variable is the treatment costs.

First, two benchmark models are discussed in Subsection 6.2.1, *Grand Mean*, and Subsection 6.2.2, *Grouped Mean*. These models are solely based on averaging historic data. Second, Ordinary Linear Regression is addressed in Subsection 6.2.3, *Linear Regression*. Ordinary Linear Regression is used to investigate if linear dependencies are present between the response variable and explanatory variables. Nevertheless, Figure 4.1 in Chapter 4, *Related Work*, showed that the treatment costs on itself are non-negative and positively skewed, which is line with the literature. Therefore, Ordinary Linear Regression may not be appropriate to model the treatment costs. As indicated in the literature, when dealing with non-negative skewed data, it is more appropriate to log-transform the response variable before conducting the Linear Regression. This variant is also discussed in Subsection 6.2.3, *Linear Regression*. Another method proposed in the literature, to deal with non-negative skewed data, is by using Gamma Regression.

Furthermore, the goal is to construct a cost-effectiveness analyses that can be generalised over multiple diagnoses. Therefore, it is desirable to create a pipeline in which multicollinearity between explanatory variables does not have to be investigated manually. This is solved by using the variable selection methods LASSO Regression and Stepwise Regression. These methods are discussed in Subsection 6.2.3, *Linear Regression*, because they are used for Linear Regression. Additional, in Subsections 6.2.5,

[3] A group was identified as patients with the same diagnosis and the same drug.

[4] When more than 1% of the data are labelled as problematic outliers, it would indicate that too many potential outliers are labelled as problematic outliers.

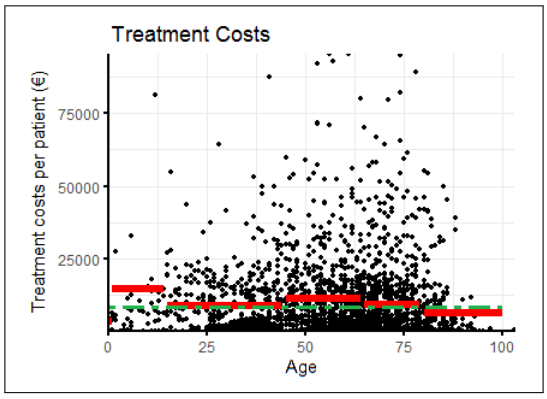

Figure 6.2: This figures shows the drug dispense treatment costs of drug dispenses per patient. The green dotted line shows the Grand Mean and the Red lines the Grouped Mean per age category, [0, 1-14, 15-44, 45-64, 65-79, 80+].

Principal Component Analysis Regression, and 6.2.6, *Partial Least Squares Regression*, two other regression models are addressed that are developed to handle multicollinearity between explanatory variables.

Last, non-linear relations can exist between the response variable and explanatory variables. Although various Regression models will already be used that can handle non-linear relations, the machine learning model Random Forest Regression can learn a large diversity of non-linear patterns. Therefore, Random Forest Regression is also tested in this research and explained in Subsection 6.2.7, *Random Forest Regression*.

For all models the dependent response variable is annual treatment costs per patient, presented as $\mathbf{Y}$. Given a treatment and a patient, various features can be constructed. These are the independent explanatory variables, presented as $\mathbf{X}$. Let there be N observations and V explanatory variables. Then the data can be presented as in equation 6.6. An example of observation i would be $(y_i, x_{i1}, x_{i2}, \ldots, x_{iV})$. Every y_i is one observation of the annual costs of a patient. And x_{ij} is variable j of observation i.

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1V} \\ x_{21} & \cdots & x_{2V} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{NV} \end{pmatrix}. \tag{6.6}$$

When a model predicts the treatment costs for patient $p \in \{1, \ldots, N\}$, the predicted value is presented as $\hat{y}_p$. Define the error the model for this observation as ε_p.

6.2.1 Grand Mean

The Grand Mean, also known as the Null Model, is the most naif model. It calculates β_0, the mean over all historic observations. Then β_0 is the prediction for the treatment costs. An example of the Grand Mean is given in Figure 6.2, where the Grand Mean is shown for the treatment costs per patient.

The Grand Mean can be presented as in formula 6.7,

$$\begin{aligned} \beta_0 &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} y_i, \\ \hat{y}_p &= \beta_0 + \varepsilon_p. \end{aligned} \tag{6.7}$$

The Grand Mean model assumes that the data are homogeneous and that there are no dependency between the response variable and the explanatory variables.

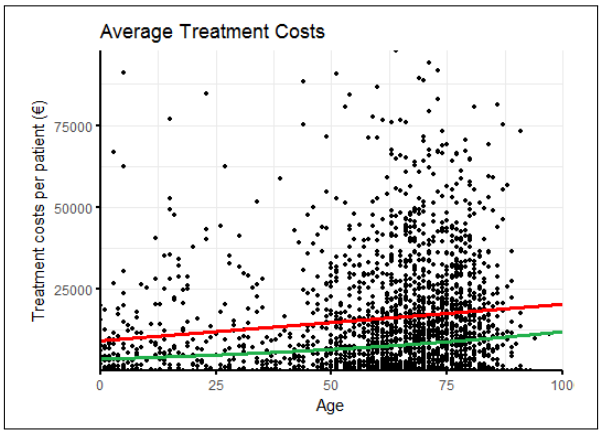

Figure 6.3: This figures shows the drug dispense treatment costs of drug dispenses for leukaemia patients. The red line shows the prediction of Linear Regression and the green line shows the prediction of Linear Regression with a log-transformation of the response variable. Remark that the green line shows a slight curve, showing the non-linear relation that can be modelled using a transformation of the data.

6.2.2 Grouped Mean

The Grouped Mean is an enhanced Grand Mean model. Instead of taking the mean over all historic observations, the observations are first divided into sub-populations on the basis of several explanatory variables. Per sub-population the Grand Mean is used. Therefore, the predicted treatment costs is the mean over all observations in the corresponding sub-population. When there were no historic data available in a sub-population, the Grand Mean prediction β_0 is used. An example of the Grouped Mean is given in Figure 6.2, where the Grouped Mean is shown for the treatment costs per patient.

Let there be L different sub-populations, $1, \ldots, L$. Define C_l as the collection of all patients part of sub-population l, and $|C_l|$ be the number of patients in sub-population l. The Grouped Mean can be presented as in formula 6.8.

$$\begin{aligned} \beta_l &= \begin{cases} \frac{1}{|C_l|} \sum_{p \in C_l} y_p, & \text{if } |C_l| > 0, \\ \frac{1}{N} \sum_{i=1}^{N} y_i, & \text{if } |C_l| = 0, \end{cases} \\ \hat{y}_p &= \beta_l + \varepsilon_p. \end{aligned} \tag{6.8}$$

The Grouped Mean model assumes that the data are homogeneous per sub-population and that there is no dependency between the response variable and the explanatory variables within a sub-population. Moreover, the model assumes that the response variable is only influenced by the variables that define a sub-population.

6.2.3 Linear Regression

Linear Regression[5] is a model where per explanatory variable the best fitting linear relation is calculated with the response variable. The best linear relation is defined as the relation where the total distance to the linear line over all observations is minimised. Given a patient and the patient's features, the prediction for the treatment costs can be calculated by taking the sum over all linear relations. As example, Figure 6.3 shows.

In formula form, Linear Regression can be modelled as,

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_V x_{iV} + \varepsilon_i, \quad i \in 1, \ldots, N \\ \varepsilon_i &\sim N(0, \sigma^2), \quad i \in 1, \ldots, N \end{aligned} \tag{6.9}$$

[5] In this research, multiple linear regression is used, because multiple explanatory variables are considered at once. For more information about Linear Regression: Bijma et al. (2013) [6].

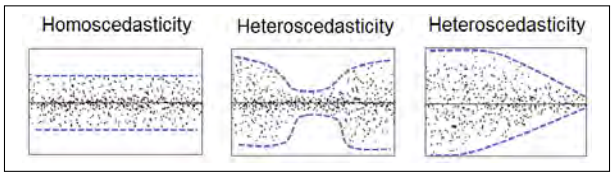

Figure 6.4: Here the difference between homoscedasticity and heteroscedasticity is shown. A variable is homoscedastic when its variability is equal, independent of other variables. When the variability is dependent on another variable, the variable is heteroscedastic.

Note that all coefficients $\hat{\beta}_j$ have an accent circonflexe, to indicate that they are estimates of the best linear relation, based on historic data. The following assumptions are needed for linear regression:

1. Explanatory variables or a combination of explanatory variables are not correlated with other explanatory variables[6]. Otherwise, the coefficients $\hat{\beta}_i$ will be unreliable.
2. There is no omitted variable bias. Omitted variable bias is present when variables are not taken into account that influence the response variable[7]. Not taking these variables into account results in an overestimation of the estimated coefficients $\hat{\beta}_i$.
3. The residuals ε_i are independent homoscedastic distributed, following a normal distribution, $\varepsilon_i \sim N(0, \sigma^2)$. This also indicates that outliers are rare and the residuals are uncorrelated. Moreover, when the residuals are normally distributed, the response variable y_i will be normally distributed too. An example of homoscedasticity and heteroscedasticity is presented in Figure 6.4.

When the explanatory variables in the model are correlated with each other, the estimated coefficients $\hat{\beta}_i$ will be underestimated. When the residuals are not independent normally distributed, the estimated coefficients $\hat{\beta}_i$ are unreliable. The coefficients are estimated[8] by minimising the sum of squared errors, $\sum_{i=1}^{N} {\varepsilon_i}^2$.

Additionally, Linear Regression is also modelled with a logarithmic transformation[9] of the response variable. This extension is chosen to counter the skewed nature of the treatment costs. The logarithmic transformation can be presented in formula form as follows,

$$\begin{aligned} \ln(\hat{y}_i) &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_V x_{iV} + \varepsilon_i, \quad i \in 1, \ldots, N \\ \varepsilon_i &\sim N(0, \sigma^2), \quad i \in 1, \ldots, N \end{aligned} \tag{6.10}$$

Furthermore, insignificant explanatory variables or explanatory variables whose effect are better modelled by other variables, can lower model performance and should not be taken into the regression. To automatically select which explanatory variables are used, the following methods are considered,

- Stepwise Regression[10] [88]. This method starts without any explanatory variable in the regression model and iterative adds explanatory variables one by one. The added variable is the one which results in the model with the lowest AIC value[11] [1]. When one of the added variables becomes insignificant, due to multicollinearity, it is left out of the regression model.

[6] Note that the intercept is one of the explanatory variables. When a combination of variables form a perfect multicollinearity with the intercept it is known as the Dummy Variable Trap. This can be solved by leaving one of the explanatory variables out of the regression.

[7] Remark that in practice there are always variables that cannot be measured. A solution can be to add as much explanatory variables as possible. On the other hand, every added explanatory variable has a certain variability which causes a random error.

[8] Iteratively Reweighted Least Squared (IWLS) is used to estimate the coefficients, based on the Generalised Linear Model (GLM) implementation in R [21, 30, 72].

[9] The natural logarithm with base e is used in this research.

[10] Bidirectional Elimination is used.

[11] The Akaike Information Criterion value decreases by a better fit, but increases when adding more variables.

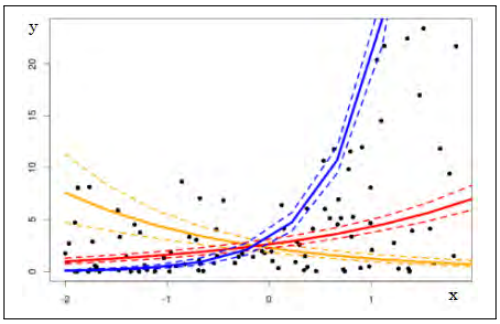

Figure 6.5: This figure of Hertzog (2016) [33] shows different non-linear relations that can be modelled by Gamma Regression.

- LASSO Regression[12] [85, 81] changes the way the coefficients are calculated, by penalising large coefficients and forcing the coefficients of unimportant explanatory variables to zero. After estimating the coefficients with LASSO, these coefficients are used in an ordinary Linear Regression for prediction. Remark that although all explanatory variables are used in this regression, some will not influence the prediction, because their coefficients are set to zero by the penalty of LASSO. To determine the strength of the penalty, 10-Fold Cross Validation[13] is used with 100 different strength values[14]. Furthermore, it is remarked by Lockhart et al. (2014) [51]: "The usual constructs like p-values, confidence intervals, etc., do not exist for lasso estimates". This is because the LASSO estimator is not differentiable.

6.2.4 Gamma Regression

Gamma Regression is a part of the group of Generalised Linear Models (GLM) [58], where the relation between the explanatory variables and the response variable is non-linear. For Gamma Regression the response variable is assumed to follow a Gamma distribution. Moreover, a linear relation between the natural logarithm of the response variable[15] and the explanatory variables is assumed. A theoretical example of a prediction conducted with Gamma Regression is given in Figure 6.5.

In formula form, Gamma Regression can be modelled as,

$$\begin{aligned} \ln(\hat{y}_i + \varepsilon_i) &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_V x_{iV}, \quad i \in 1, \ldots, N \\ \hat{y}_i &\sim \Gamma(a, b), \quad i \in 1, \ldots, N \end{aligned} \tag{6.11}$$

The main assumption of Gamma Regression is that the response variable follows a Gamma distribution. Additionally, the explanatory variables should not be correlated with each other and there should be no omitted variable bias, as in Linear Regression. The coefficients are again estimated by Iteratively Reweighted Least Squared (IWLS) and the variable selection is done with Stepwise Regression.

6.2.5 Principal Component Analysis Regression

Principal Component Regression (PCR) [7, 41] first performs Principal Component Analysis (PCA) on the explanatory variables[16] to obtain the principal components[17]. The most important principal components[18] are used as explanatory variables to predict the response variable with Linear Regression.

[12] Least Absolute Shrinkage and Selection Operator (LASSO) penalises large weights by adding the sum of absolute coefficients to the Least Squared Error.

[13] K-Fold Cross Validation is explained in Subsection 6.3.1, *K-Fold Cross Validation*.

[14] The 100 strength values are determined by the algorithm of Simon et al. [81].

[15] The natural logarithm is the chosen link function.

[16] For PCR it is advised to first normalise the explanatory and response variables [76].

[17] To obtain the principal components, Singular Value Decomposition (SVD) is used. A clear explanation of SVD is given by Lay (2002) [47].

[18] The components with highest variance.

Let there be M principal components[19] of which the first J are used in the regression. Every principal component, z_m, is a linear combination of the original explanatory variables, x_j. Define γ_{mk} as the coefficient of the m^{th} principal component and k^{th} explanatory variable. Moreover, define β_m as coefficient of the m^{th} principal component. Note that γ_{mk} is estimated by PCA and that β_m is estimated by Linear Regression. PCR can be presented as follows,

$$\begin{aligned} z_{mi} &= \gamma_{m1}x_{i1} + \ldots + \gamma_{mV}x_{iV}, \quad m \in 1,\ldots,M, \quad i \in 1,\ldots,N \\ \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \ldots + \hat{\beta}_J z_{iJ} + \varepsilon_i, \quad i \in 1,\ldots,N \\ \varepsilon_i &\sim N(0,\sigma^2), \quad i \in 1,\ldots,N \end{aligned} \tag{6.12}$$

The principal components are by construction uncorrelated, also known as orthogonal. and sorted on greatest variance. When the original K explanatory variables are correlated or have little variance, it is sufficient to only use the first most important principal components in the regression. The first J principal components will contain the most important patterns of the original explanatory variables. The other principal components can be left out of the regression with minimal loss of information. This reduces the number of components the Linear Regression has to take into account.

To determine J, the number of principal components used in the Regression, 10-Fold Cross Validation[20] is used. First, PCR is executed with $J = 1,\ldots,M$. Second, the reference model is chosen as the model with the lowest Mean Squared Error[21] over all 10 folds. Using the randomisation test approach [90], the smallest model[22] not significant worse than the reference model is actually chosen. Finally, J is set to the number of principal components in the chosen model.

In summary, the advantage of PCR is that it can deal with correlated explanatory variables. The drawback is that PCR does not do feature selection on the explanatory variables[23] and PCR is not used to detect which explanatory variables directly influence the response variable to which extend[24]. Lastly, for PCR it is advised to mean-center the response variable and explanatory variables before applying PCR [76].

6.2.6 Partial Least Squares Regression

Partial Least Squares Regression (PLSR) [92] first performs Partial Least Squares (PLS) on the explanatory variables and the response variable[25] to obtain the latent variables[26]. The most important latent variables[27] are used as explanatory variables to predict the response variable with Linear Regression.

Latent variables in PLSR are similar to principal components in PCR, for both are linear combinations of the original explanatory variables, both are by construction uncorrelated, and both are used as input for a Linear Regression. The difference is that principal components are solely based on the explanatory variables to maximise variance, while latent variables are based on the explanatory variables and the response variable to maximise covariance[28].

[19] $M = \min(N-1, K)$

[20] K-Fold Cross Validation is explained in Subsection 6.3.1, *K-Fold Cross Validation*.

[21] This is the Root Mean Squared Error, explained in Subsection 6.3.5, *Root Mean Squared Error*, without a square root.

[22] The model with the least number of principal components, smallest J.

[23] PCR extracts the most important patterns as principal components. However, these principal components are a linear combination of all original explanatory variables. Leaving out some principal components does not reduce the number of original explanatory variables used.

[24] The Linear Regression indicates which principal components are significant to explain the variability in the response variable. Nevertheless, the principal components are not linked one-to-one to the original explanatory variables.

[25] For PLSR it is advised to first normalise the explanatory and response variables [76].

[26] To obtain the latent variables, the kernel algorithm [49] is used.

[27] The latent variables with highest covariance with the response variable.

[28] This is the covariance between the explanatory and response variables.

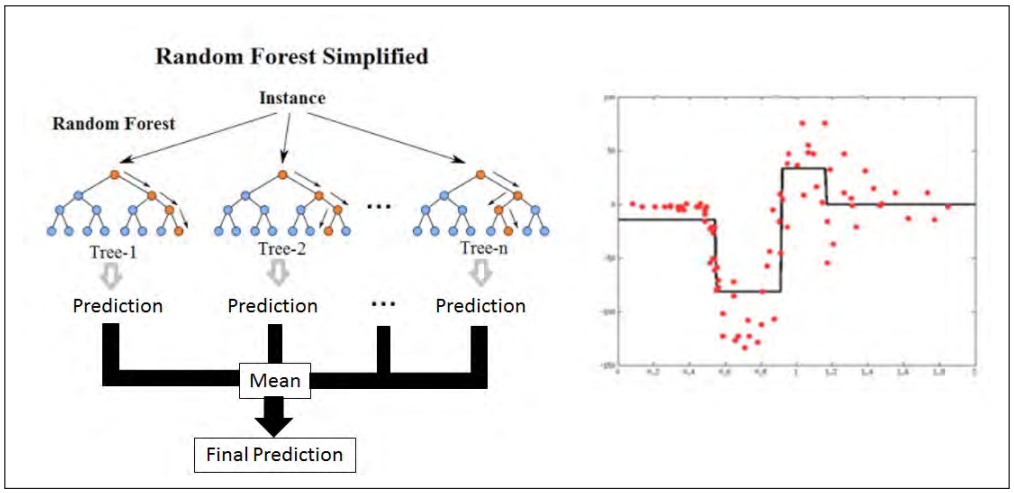

Figure 6.6: In the left illustration is shown how multiple decision trees are combined. Every decision tree gives a prediction and all predictions are averaged to give one final prediction. In the right graph is shown how Random Forest Regression can learn very complex non-linear relations.

Due to the similarities with PCR, PLSR can also be presented as in equation 6.12. For this, redefine z_m as a latent variable and let there be M latent variables of which the first J are used in the regression. Moreover, let γ_{mk} be the coefficient of the m^{th} latent variable and k^{th} explanatory variable. Lastly, redefine β_m as the estimated coefficient of the m^{th} latent variable.

Similar to PCR, it is sufficient in PLSR to only use the first J most important latent variables in the regression. The number of latent variables is chosen in the same way as the number of principal components are chosen in PCR. Moreover, the advantage and drawbacks for PCR described in Subsection 6.2.5, *Principal Component Analysis Regression*, also apply to PLSR.

6.2.7 Random Forest Regression

Random Forest Regression (RFR) is a machine learning algorithm that ensembles decision trees [11, 77]. The prediction of the Random Forest Regression is the average over the outcomes of all decision trees in the forest. A theoretical example of Random Forest Regression is shown in Figure 6.6.

Every decision tree is constructed based on a subset of the data. Per tree, a population of N observations are sampled with replacement. Per node, one third of the explanatory variables[29] are sampled without replacement. Consequently, the split per node is based on sampled observations and a subset of explanatory variables. This ensures that many different decision trees are present in the Random Forest. Furthermore, per node the split is based on the explanatory variable which after the split on that variable results in the lowest RMSE of the out-of-bag sample[30].

The decision trees are grown to full depth. Moreover, 100 decision trees are used in the Random Forest. Using more decision trees did not increase performance, while less decision trees would decrease performance.

[29] With V the number of explanatory variables, $\lfloor \frac{V}{3} \rfloor$ explanatory variables are sampled per node.

[30] By sampling the observations with replacement, a part of the observations will not be used for building the decision tree. As approximation, this regards $\lim_{n\to\infty}(1 - \frac{1}{N})^N \approx 36.8\%$ of the data. This part of the data is used to test performance.

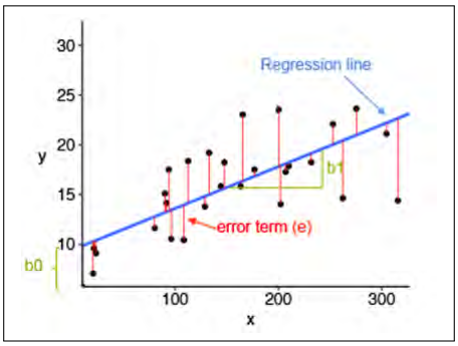

Figure 6.7: The error is the distance between the actual observation, black dot, and the regression line, blue line.

6.3 Prediction Measures

With 5-Fold Cross Validation the predictive performance of every model is tested. In Subsection 6.3.1, *K-Fold Cross Validation*, 5-Fold Cross Validation is explained and is explained why 5 folds were chosen. Within 5-Fold Cross Validation, prediction measures are needed to measure the predictive performance on the test sets. Taking into account that the response variable is continuous, various prediction measures were available [6]. In this section the actual response variable y_i is compared with the predicted value $\hat{y}_i$. Define the error as the difference, $\varepsilon_i = \hat{y}_i - y_i$. As example, Figure 6.7 shows the errors made by a prediction with a linear regression line.

First, the Mean Error (ME) is addressed in Subsection 6.3.2, *Mean Error*. Remark that the ME can provide insights in the average prediction, but should not be used to actually measure predictive performance. Therefore, the ME is only used to conclude if a model predicts costs too high or too low on average. Second, to actually know the magnitude of the error per prediction, the Mean Absolute Deviation (MAD) is used. The MAD is described in Subsection 6.3.3, *Mean Absolute Deviation*. Third, it is preferable to also measure the magnitude of the error relative to the actual treatment costs. The Mean Absolute Percentage Error (MAPE) can be used for that purpose and is addressed in Subsection 6.3.4, *Mean Absolute Percentage Error*. Fourth, Subsection 6.3.5, *Root Mean Squared Error*, explains the Root Mean Squared Error (RMSE), which penalises larger errors. It is noteworthy that the RMSE is sensitive to outliers and that the RMSE can be used to measure how well models predict outliers. Finally, the Coefficient of Determination (R^2) is described in Subsection 6.3.6, *Coefficient of Determination*. The R^2 measures how well the variation of the data is explained by the regression model. By combining all insights from the prediction measures, a complete overview can be obtained of model performance.

6.3.1 K-Fold Cross Validation

In K-Fold Cross Validation (CV) [62] the observations are partitioned into K equal sized subsets as shown in Figure 6.8. Sequential, one of the K subsets is chosen as test set. The other $K-1$ subsets are used as training sets. The models are trained on the train set. The trained models are then used to predict the treatment costs for the test set. Remark that the test set consists of observations the model has not seen before. The procedure of training and testing is repeated till all subsets have been chosen once as test set.

The models will be retrained and tested K times. With prediction measures, the results on the test sets can be analysed. This will provide an indicator of how well the models perform. The hypothesis is that when a model performs well on the test set, it learned relations from the train set that actually exist and are no coincidence[31]. When a model learned relations from the train set that not actually existed, the model underfitted or overfitted on the training data. This phenomenon is shown in Figure 6.9.

[31] In this way it is tested if models perform properly, even when model assumptions are violated.

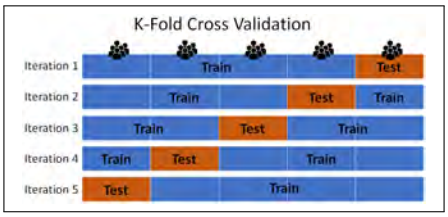

Figure 6.8: Illustration of K-Fold Cross Validation, where $K = 5$. Per iteration the data are divided in a train set and a test set.

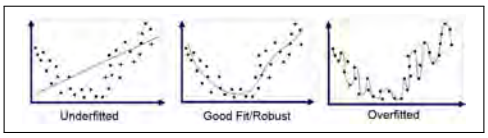

Figure 6.9: Illustration of overfitting. Overfitting and underfitting are examples of dependencies that are learned from the training data that are not actual present. Commonly, models that overfit or underfit show poor results on the test set. This figure is from Bhande, A. (2018) [5].

In this research, $K = 5$ is used for Cross Validation, considering the sample size of the data and the data characteristics. This divided the data in 80% training data and 20% test data, which is common practice when Cross Validation is used to measure prediction performance [43].

6.3.2 Mean Error

The Mean Error (ME) is a measure that that indicates if a model predicts too high or too low on average. As remarked before, the ME should not be used to actually measure predictive performance. The ME is calculated with equation 6.13.

$$\begin{aligned} \text{ME} &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \varepsilon_i, \\ \text{ME} &\in \mathbb{R}. \end{aligned} \tag{6.13}$$

Note that the unit of the ME is the unit of the response variable. A negative ME shows that the model predicts on average too low, while a positive ME shows that the model predicts on average too high. The ME should be close to zero[32].

6.3.3 Mean Absolute Deviation

The Mean Absolute Deviation (MAD) is a prediction measure that represents the average error. This shows the magnitude of the error per prediction. The MAD is calculated with equation 6.14,

$$\begin{aligned} \text{MAD} &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} |\varepsilon_i|, \\ \text{MAD} &= \in \mathbb{R}^{+}. \end{aligned} \tag{6.14}$$

Note that the unit of the MAD is the unit of the response variable. The MAD should be as small as possible.

6.3.4 Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) is a prediction measure that measures the error relative to the actual costs. The MAPE is calculated with equation 6.15,

$$\begin{aligned} \text{MAPE} &= \tfrac{100}{N}\textstyle\sum_{i=1}^{N} \left|\frac{\varepsilon_i}{y_i}\right|, \\ \text{MAPE} &\in \mathbb{R}^{+}. \end{aligned} \tag{6.15}$$

[32] Note that the ME can be misleading. When a model predicts poorly, the ME can still be zero when the model predicts as much too high as too low.

The MAPE is in percentages relative to the actual costs and should be as small as possible. In the literature has been addressed that MAPE favours lower predictions[33].

6.3.5 Root Mean Squared Error

The Root Mean Squared Error (RMSE) is a prediction measure that represents the absolute fit. The RMSE is calculated with equation 6.16,

$$\begin{aligned} \text{RMSE} &= \sqrt{\tfrac{1}{N}\sum_{i=1}^{N}(\varepsilon_i)^2}, \\ \text{RMSE} &\in \mathbb{R}^+. \end{aligned} \tag{6.16}$$

The RMSE puts an extra penalty on large errors and should be as small as possible.

6.3.6 Coefficient of Determination

The Coefficient of Determination (R^2), also known as the R-squared, is a prediction measure that represents to what extend the model performs better than the baseline model (the Grand Mean). First, define the mean response variable as $\bar{y} = \sum_{i=1}^{N} \frac{1}{N} y_i$. Next, the R^2 is calculated with equation 6.17.

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{N}(\varepsilon_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \\ R^2 &\in (-\infty, 1]. \end{aligned} \tag{6.17}$$

If the model predicts every observations perfectly, the $R^2 = 1$. When the model performs as well as the baseline model, the $R^2 = 0$. In the unusual situation where the model performs worse than the baseline model, the $R^2 < 0$. However, a negative R^2 indicates severely poor performances. Cohen (1988) [17] suggests that for social science a R^2 of 0.26 indicates substantial performances, of 0.13 indicates moderate performances, and of 0.02 indicates weak performances. It differs over disciplines what values for R^2 indicate acceptable performances. Cohen's reference measure is used, because treatment costs are inferred from prescriptions rather than directly observed and patients are highly variable[34].

6.4 Correlation Measures

Correlation Measures are used to test whether relationships exists between features. One can measure if there are linear dependencies between continuous variables[35]. Furthermore, it can be tested if a numeric variable is dependent on a categorical variable or two categorical variables are dependent on each other. By calculating the correlations between the response variable and every explanatory variables it will be measures if dependencies exist between the variables; if the correlations are positive or negative; and how strong the correlations are. This allows to validate the conclusions drawn from analysing the Regression models. Additionally, relations can be measured between the treatment cost and quality indicators to gain insights about the cost-effectiveness. Remark that calculating the correlations is similar to performing univariate regressions.

Which correlation test to use is dependent on the nature of the variables that are compared, as shown in Table 6.1.

[33] Despite the MAPE often being used, Makridakis (1993) [52] remarked that $y_i = 100$ with $\hat{y}_i = 150$ results in MAPE = 50%, while the same error where $y_i = 150$ with $\hat{y}_i = 100$ results in MAPE = 33%. This shows that an error on a relatively small actual value results in a large MAPE. If $y_i = 0$, the MAPE can even become infinite.

[34] Two patients with exactly the same characteristics can have totally different treatment costs. This is also a consequence of the fact that not all variables that influence the cost are measured.

[35] Remark that correlations found do not have to be causal, but can also be a coincidence.

[36] For completeness, the Chi-squared test is presented in the table. However, this test was not used in this research, because there were no cases where categorical variables needed to be compared with categorical variables.

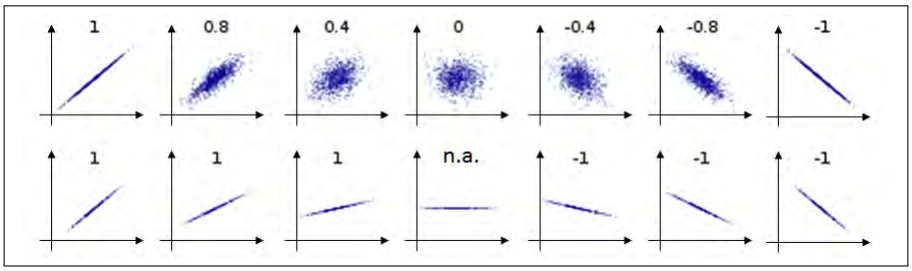

Figure 6.10: In this figure by D. Boigelot, the Pearson's Correlation Coefficient is shown for different theoretical examples. Remark how the correlation coefficient is influenced by the direction of the relation and the noise in the data. However, the bottom row shows that the slope of the relationship has no influence.

Table 6.1: Correlation measures

Variable nature	**Test**
Continuous - Continuous	*Pearson, Spearman or Kendall*
Continuous - Categorical	*Kruskal-Wallis or ANOVA*
Categorical - Categorical	*Chi-squared test*[36]

When both variables are numeric, Pearson's Correlation Coefficient and Kendall's Rank Correlation Coefficient are used[37]. Pearson's Correlation is most reliable in the case that both variables are normally distributed. On the other hand, when the data are not normally distributed, Kendall's Rank Correlation is most reliable. When one of the variables is numeric and the other is categorical, the ANOVA test is used[38].

It is necessary to view the conclusion of every significance test in isolation of others. As example, let there be m significance tests, all testing with a 95% confidence level. Combining the m conclusions, does also combine the confidence level of all confidence levels and reduces the confidence level to $(95\%)^m$. Combining the conclusions of for instance 5 significance test with a 95% confidence level, would drop the confidence level to $(95\%)^5 = 77\%$. This problem was addressed by Bonferroni (1936) [9], and can be easily solved by testing with $1 - \frac{5\%}{m}$ confidence per significance test. This solution is known to be very conservative and less conservative methods exists. Nevertheless, the Bonferroni correction does not assume anything about the significance test that is conducted and is widely used due to its simplicity.

In this section the following definitions are used. Let $X = (x_1, \ldots, x_N)$ be a sorted variable, $x_1 \leq x_2 \leq \ldots \leq x_N$, and let $Y = (y_1, \ldots, y_N)$ be the corresponding values, sorted by X[39]. The data consist of observations (x_i, y_i), for $i = 1, \ldots, N$.

6.4.1 Pearson's Correlation Coefficient

Pearson's Correlation Coefficient ρ [70] measures to what degree two numeric variables X and Y have a linear relationship. An example of correlation strengths is given in Figure 6.10.
The Pearson's Correlation Coefficient between two variables is the covariance between two variables divided by the standard deviation of both variables multiplied, as shown in equation 6.18.

$$\rho = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})}} \tag{6.18}$$

[37] Kendall's and Spearman's Rank Correlation do coincide in most cases, calculating both would not add insightful information. Kendall's Rank Correlation is chosen for it is more reliable in small groups and can handle duplicates better than Spearman's Rank Correlation.

[38] Based on Glass et al. (1972) [29], in large groups the ANOVA is still farily reliable when the normality assumption does not hold. Therefore, the ANOVA test was favoured over the Kruskal-Wallis test.

[39] Remark that with $i > j$ and $i, j \in \{1, \ldots, N\}$, y_i can be smaller than y_j, because Y is sorted on X.

Note that $\bar{x}$ is the mean of x, $\bar{x} = \sum_{i=1}^{N} \frac{1}{N} x_i$. Moreover, the coefficient falls in the interval $\rho \in [-1, 1]$. A negative coefficients implies a negative linear relation between the two variables, and conversely a positive coefficient implies a positive linear relation. Hinkle et al. (2003) [35] provide a rule of thump for interpreting the strength of the correlation. Their standards are shown in Table 6.2.

Table 6.2: Correlation Coefficient Interpretation

Correlation Coefficient	**Interpretation**
(0.9, 1.0], (-0.9, -1.0]	Very high correlation
(0.7, 0.9], (-0.7, -0.9]	High correlation
(0.5, 0.7], (-0.5, -0.7]	Moderate correlation
(0.3, 0.5], (-0.3, -0.5]	Low correlation
[0.0, 0.3], [0.0, -0.3]	Negligible correlation

Pearson's Correlation Coefficient assumes a normal distribution of both variables and homoscedasticity of the linear relation between the variables.

6.4.2 Kendall's Rank Correlation Coefficient

Kendall's Rank Correlation Coefficient τ [42] measures to what degree two numeric or ordinal variables X and Y are related. To handle ties, a specific version of Kendall's Rank Correlation is used, known as Kendall τ_b.
Kendall τ_b is based on pairs $\{(x_i, y_i), (x_j, y_j)\}$, with $i \neq j$ and $i, j \in \{1, \ldots, N\}$. There are $\binom{N}{2} = \frac{N(N-1)}{2}$ pairs. To calculate Kendall τ_b, the following quantities are needed:

- Number of concordant pairs C. A pair is concordant when,
  - $x_i > x_j$ and $y_i > y_j$ or
  - $x_i < x_j$ and $y_i < y_j$.
- Number of discordant pairs D. A pair is discordant when,
  - $x_i > x_j$ and $y_i < y_j$ or
  - $x_i < x_j$ and $y_i > y_j$.
- Number of pairs not tied in X, N_x. A pair is not tied in X when $x_i \neq x_j$.
- Number of pairs not tied in Y, N_y. A pair is not tied in Y when $y_i \neq y_j$.

With these quantities Kendall τ_b can be calculated with equation 6.19,

$$\tau_b = \frac{C-D}{\sqrt{N_x - N_y}} \tag{6.19}$$

The coefficient falls in the interval $\tau_b \in [-1, 1]$. A negative coefficients implies a negative relation between the two variables, conversely, a positive coefficient implies a positive relation. Again, Table 6.2 can be used for the interpretation of correlation strength. Last, besides continuous variables, Kendall's Rank Correlation Coefficient can deal with any kind of ordinal variables.

6.4.3 Confidence Interval Correlation Coefficients

In the case that both values are numeric, the confidence interval of a test statistic was obtained by bootstrapping [64, 23]. Bootstrapping is sampling with replacement from a sample, to obtain multiple representative samples. Bootstrapping assumes that all observations in the original sample are independent and identically distributed.

Let there be N observations in the sample, then N observation are drawn with replacement from the sample. By repeating this process B times, B representative samples of size N can be obtained. These representative samples are bootstrapped samples of the original sample. A test statistic can be

calculated for all B samples. This can be any test statistic. A bootstrap confidence interval of the test statistic is then obtained by analysing the B test statistic outcomes. With the bootstrap confidence interval, the sensitivity of the test statistic is measured. First, the B test statistic outcomes are arranged in ascending order. Next, calculate the following quantiles of the B ordered test statistic outcomes,

$$\begin{array}{rl} Q(\frac{\alpha}{2}) & = \frac{\alpha}{2}\%\text{-quantile}, \\ Q(1-\frac{\alpha}{2}) & = (1-\frac{\alpha}{2})\%\text{-quantile}. \end{array} \tag{6.20}$$

Then, the bootstrap confidence interval is

$$[Q(\frac{\alpha}{2}), \quad Q(1-\frac{\alpha}{2})]. \tag{6.21}$$

This interval represents the sensitivity of the test statistic on a $(1-\alpha\%)$-confidence level. When zero is not part of the bootstrap confidence interval, it can not be stated that the test statistic is not zero with $\alpha\%$-certainty.

6.4.4 Analysis of Variance

The Analysis of Variance (ANOVA) consists of statistical models that can test if the mean of a response variable significantly differs between groups. This subsection addresses the one-way ANOVA, which is used in this research.

Let there be K groups, within every group n_k observations. Define every observation as y_{ki}, with $k = 1, \ldots, K$ and $i = 1, \ldots, n_k$. The null hypothesis H_0 is that the mean of the response variable is the same for every group. Consequently, the alternative hypothesis H_1 is that the mean is not the same for every group. This can be summarised as follows,

$$\begin{array}{rll} H_0: & \mu_1 = \ldots = \mu_K, & \\ H_1: & \exists\, \mu_i \neq \mu_j, & i,j = 1,\ldots,K, i \neq j \\ \mu_k = & \frac{1}{n_k}\sum_{i=1}^{n_k} y_{ki}, & k = 1,\ldots K, \\ \mu = & \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k} y_{ki}, & \end{array} \tag{6.22}$$

Remark that the alternative hypothesis is true when $\mu_i \neq \mu_j$ holds for at least one pair of (i,j). To test the hypothesis, the variability of the observations is needed and is measured with the Total Sum of Squares, SS_T. Furthermore, the variability of the observations within groups is needed and measured with the Sum of Squares Residual, SS_R. The null hypothesis is rejected when SS_R is much smaller than SS_T, because that indicates that there are significant differences between groups. The SS_T and SS_R can be calculated as follows,

$$\begin{array}{rl} SS_T & = \sum_{k=1} K \sum_{j=1} n_k (y_{ki} - \mu)^2 \\ SS_R & = \sum_{k=1} K \sum_{j=1} n_k (y_{ki} - \mu_k)^2 \end{array} \tag{6.23}$$

Next, the Mean Squared Error between groups, MSE_R, and within groups, MSE_{T-R}, are needed to test significance. The Mean Squared Error (MSE) is calculated by dividing the sum of squares by the corresponding degrees of freedom[40]. The formulas of the MSE_R and MSE_{T-R} are,

$$\begin{array}{rl} MSE_{T-R} & = \frac{SS_T - SS_R}{(k-1)} \\ MSE_R & = \frac{SS_R}{(N-k)} \end{array} \tag{6.24}$$

A summary of the one-way ANOVA is given in Table 6.3. The test assumes that the observations are independent normally distributed per group, with common variance over all groups. Under this

[40] For ANOVA, all errors are, by assumption, independent normally distributed. Therefore, the Sum of Squares (SS) is the sum of normal random variables, which is a chi-squared distribution (χ^2). However, when estimating the Mean Squared Error, one observation is 'fixed' and is not considered to be a random variable that is allowed to vary. The number of observations that are free to vary are known as the degrees of freedom when estimating statistical parameters. For a full explanation about the degrees of freedom, I refer to Walker (1940) [91].

assumption, the ratio $f = \frac{MSE_{T-R}}{MSE_R}$ has a F-distribution with $K-1$ and $N-K$ degrees of freedom, $f \sim F_{I-1,n-I}$. The null hypothesis is rejected on an $(1-\alpha)$%-confidence level if and only if $f > F_{K-1,N-k;\alpha}$. It is noteworthy that Glass et al. (1972) [29] showed that ANOVA is still fairly reliable, even when the normality assumption does not hold. Yet, this is not the case for too small groups[41] or when the data distribution deviates to much from a normal distribution.

Table 6.3: One-way ANOVA table.

	Degrees of Freedom	**Sum of Squares**	**Mean Squared Error**	**F value**
Between groups	$K-1$	$SS_T - SS_R$	MSE_{T-R}	$\frac{MSE_{T-R}}{MSE_R}$
Within groups	$N-K$	SS_R	MSE_R	
Total	$N-1$	SS_T		

[41] A group is small when approximately less than 30 observations are in the group.

7 Experimental Set-up

In this chapter will be explained how the cost-effectiveness analyses of expensive drug treatments are conducted. In Section 7.1, *Modelling Treatment Costs*, shall be described how a target group of patients is obtained. This section also addresses how the treatment costs were extracted from the data and which features were used to model the treatment costs. Second, Section 7.2, *Regression Models*, addresses how the regression models are used model the treatment costs, to make sure the treatment costs could be risk adjusted for patient characteristics. Remark that risk adjusting is statistically accounting for differences in patient characteristics that influence the outcomes, to make sure different patients can be compared fairly, because the differences in outcome due to patient characteristics as age and disorder are removed from the outcome [46]. Afterwards shall be described in Section 7.3, *Correlations*, how relations between the treatment costs and quality indicators were measured to analyse the effectiveness of expensive drug treatments. Figure 7.1 presents a schematic view of the experimental set-up.

7.1 Modelling Treatment Costs

This section reviews how the treatment costs will be modelled by regression models. First, Subsection 7.1.1, *Target Group*, describes how a target group is obtained. Next, given a target group, the treatment costs can be calculated, as addressed in Subsection 7.1.2, *Response Variable*. The response variable is modelled with explanatory variables, which are described in Subsection 7.1.3, *Explanatory Variables*.

7.1.1 Target Group

Which drugs are prescribed to a patient, is depended on the patient's disorder and characteristics[1]. Cost-effectiveness analyses would be insightful, if all patients in the target group have the same disorder and when different drug treatments are possible. Then, differences in treatment costs and quality indicators could be analysed to compare the effectiveness of treatments. In this research, a subset of patients was selected based on one or more DBC diagnoses[2], this subset is the target group. Noteworthy is that the data consider admissions and drug dispenses from the LBZ dataset during the years 2015, 2016 and 2017. For consistency reasons, the results shown in this research will mainly focus on the target group of leukaemia patients, because Leukaemia patients are treated with expensive drugs and are often admitted to the hospital. This results in a lot of data about leukaemia patient available for this research.

7.1.2 Response Variable

Let there be a target group based on one or more DBC diagnoses. To select relevant drug dispenses, drug dispenses are only considered if they are prescribed for the DBC diagnoses of the target group. With a selection of relevant drug dispenses, patient's treatment costs can be derived by calculating the total expenses of all relevant drug dispenses from a patient. In this research the years 2015, 2016 and 2017 from the LBZ dataset are used. This spans multiple years. To keep treatments comparable, the yearly treatment costs are used. To obtain yearly treatment costs, the treatment costs were scaled to 12 months when the treatment lasted longer than 12 months. On the other hand, when a treatment did not last 12 months, the treatment cost were not scaled, because in those cases the treatment costs already presented how much such a treatment would costs in a year.

[1] Note that for a disorder, certain variants of the disorder also influence the treatment. Additional, which drug treatment is used is depend on patient characteristics like age, sex, comorbidities, allergies and the patient's current fitness.

[2] Additionally, the selection could also be based on one or more specific expensive drugs. However, an expensive drug could be used for various diagnoses. Therefore, making a selection based on solely expensive drugs would result in a diverse group of patients, making cost-effectiveness analyses less insightful.

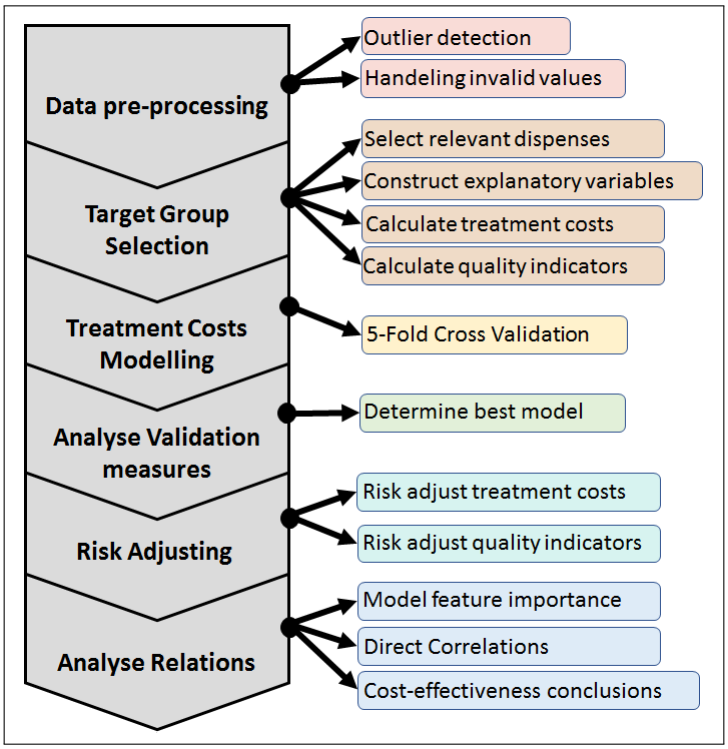

Figure 7.1: This figure shows the whole set-up of this research. First, the data are pre-processed by detecting outliers and handling invalid values, as addressed in Chapter 5.2, Pre-processing. Second, given a target group, their data are obtained and features are constructed. Third, various risk adjustment models are tested to predict treatment costs. Fourth, the model performances are measured and analysed to determine the best model. Fifth, the best model is used to actually risk adjust the treatment costs for patient characteristics. Sixth, the relations found by the models and the relations calculated with correlation statistics are analysed. From these analyses, the cost-effectiveness of treatments is derived.

Naturally, patients with different characteristics need different treatments and different treatments result in different treatment costs. Therefore, patient characteristics, such as kind of disorder, age and fitness, affect the treatment costs. To compare patients fairly, the treatment costs need to be risk adjusted for patient characteristics. This is similar to how quality indicators are risk adjusted for patient characteristics by DHD and the CBS [28, 44, 45]. By modelling the treatment costs, the treatment costs can also be predicted. The predicted treatment costs could be used as the expected treatment costs. The treatment costs risk adjusted for patient characteristics are obtained by subtracting the expected treatment costs from the actual treatment costs. The risk adjusted treatment costs allow for fair comparisons between hospitals. In the literature, the binary quality indicators mortality, UL-LOS and readmission are risk adjusted. Nevertheless, the treatment costs are a continuous response variable. Various regression models are known that can model continuous depend variables. Moreover, it is essential that the regression models are robust enough to cope with noisy observations. As an example, similar patients with the same DBC diagnose can be treated quite differently, due to unmeasured differences as allergies and how much a disorder is developed.

7.1.3 Explanatory Variables

The treatment costs can be risk adjusted for patient characteristics by only taking patient characteristics as explanatory variables in the regression models. The patient characteristics describe the patient's disorder and fitness. The patient's disorder is inferred from the diagnosis for which the drugs were prescribed, whereas the patient's fitness was not registered and could only be derived from other features. Therefore, the fitness of a patient was indirectly measured by the patient's age, comorbidities

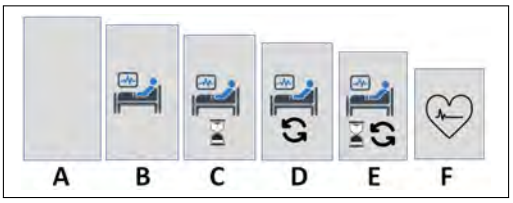

Figure 7.2: An illustration of the composite quality indicator ordered from best (A) to worst (F).

and history[3]. A complete overview of all features to correct the cost for patient characteristics is given in Appendix A.2.1, *Patient Features for Cost Correction.*

Besides correcting the costs for patient characteristics, other features were constructed to explain the treatment costs. This includes information about hospital characteristics, drug choices and admission information. To select only relevant admissions that are influenced by the treatment, all admissions during the treatment till 30 days after the last prescription are taken into account. In line with research of quality indicators, admissions till 30 days after the treatment are likely to be influenced by the treatment [44]. Admissions after 30 days since the last prescriptions are plausible to not be affected by the treatment anymore. Including features these explanatory features as explanatory variables, could show which factors influence the treatment costs, next to patient characteristics. All explanation features used in this research features are described in Appendix A.2.2, *Features for Cost Explanation.*

At last, indicators about the quality of admissions were extracted. The quality indicators describe in-hospital mortality, admission duration, readmissions and the total number of admissions. These quality indicators were already risk adjusted for patient characteristics by the CBS and DHD [44, 45, 28]. By adding the quality indicators as explanatory variables, the relations between the quality indicators and the treatment costs could be measured. In this way, the cost-effectiveness could be analysed. An overview of all quality indicators is presented in Appendix A.2.3, *Features for Cost-Effectiveness.* Remark that quality indicators describing clinical admissions and extended observations contain insightful information, because those admissions take at least several hours to multiple days. When a patient did not have a clinical admission or extended observation, the patient is assumed to have survived the treatment without readmissions or UL-LOS. Lastly, based on research of Lingsma et al. (2018) [50], the composite quality indicator is also determined per patient. The composite quality indicator[4] is presented in Figure 7.2 and consists of the following classes, ordered from best to worst outcome of a patient's treatment:

(A) Survived treatment without clinical admission and without extended observation.
(B) Survived treatment without UL-LOS and without readmission.
(C) Survived treatment with UL-LOS, but without readmission.
(D) Survived treatment without UL-LOS, but with readmission.
(E) Survived treatment with UL-LOS and with readmission.
(F) Did not survive treatment.

This composite quality indicator has several quality classes, where every patient will be part of one of the classes based on the patient's treatment outcomes. From the indicator, the quality of the treatment's outcome for the patient is measured.

7.2 Regression Models

Various regression models have been trained and tested to model the treatment cost. The Grand Mean and Grouped Mean model are used as benchmark. Whereas, among others Linear Regression and

[3]It is assumed that more complex patients did already visit the hospital for other disorders. Moreover, the year a patient is treated is also taken as a feature, for it describes when the patient had the disorder.

[4]Note that the proposed composite quality indicator differs from the composite quality indicator of Lingsma et al. (2018) [50]. The difference is an extra class added to the indicator.

Random Forest Regression are implemented as reliable cost modelling models. The models have been validated with 5-Fold Cross Validation[5] and several predictions measures including the Mean Absolute Deviation, the Root Mean Squared Error and the Coefficient of Determination. An overview of all prediction measures with a short description is given in Appendix A.5, *Prediction Measures*. Moreover, an overview of all implemented models is presented in Appendix A.4, *Regression Models*, with a short description of the models.

Three feature sets about explanatory variables have been described in the last section, Section 7.1, *Modelling Treatment Costs*: features for cost correction, features for cost explanation and features for cost-effectiveness. First, the models are tested with only cost correction features, to select the model that is the best in modelling how the costs are influenced by patient characteristics. Second, the models are tested with all three sets of features, to find the model that is the best in explaining which features are correlated with the treatment costs.

7.3 Correlations

Based on the prediction measures the overall best performing models are chosen as most reliable and worthy to analyse. By analysing these models, one could find out which features correlate with the treatment costs; find out if the influence of the feature is positive or negative; and find out the size of the influence. After the prediction, the residuals are analysed to investigate if and which assumptions are violated. Which can be used to validate the model assumptions. Additionally, the correlations are calculated between every feature and the treatment costs. This allows to validate the conclusions drawn from analysing the model, by comparing correlations with the results of the model.

Afterwards, the dataset is randomly divided in a train set containing 80% patients and a test set containing the other 20%. Then the treatment costs are risk adjusted by the model that is the best in modelling the treatment cost with patient characteristics. This is realised by training this model on the train set and actually risk adjusting the cost on the test set. The risk adjusting is conducted to make sure different patients can be compared fairly, because the differences in outcome due to patient characteristics as age and disorder are removed from the outcome. Now, the correlations can be calculated between the quality indicators and the risk adjusted treatment costs. This allows to derive what has an impact on the cost-effectiveness by analysing the correlations between the quality indicators and risk adjusted treatment cost. Moreover, the correlations can be analysed per hospital, to investigate if there are differences between hospitals[6].

When both variables are continuous, Pearson's and Kendall's Rank Correlation Coefficient[7] are used to calculate correlations. A 95% confidence bootstrap interval is obtained with the bootstrapping method explained in Subsection 6.4.3, *Confidence Interval Correlation Coefficients*. Correlations were only significant when both correlations coefficients showed a significant correlation and both were on average at least 0.2[8], based on Hinkle et al. (2003) [35]. When one of the variables is continuous and the other is categorical, the ANOVA is used, as explained in Subsection 6.4.4, *Analysis of Variance*. For the ANOVA, a 95% confidence level was used to test the significant relations. It was taken into account that conclusions drawn from the significance test should be viewed in isolation as addressed in Section 6.4, *Correlation Measures*. However, when comparing hospitals, the Bonferonni correction was used [9] to decrease the chance of incorrectly judging a hospital's treatments.

[5] K-Fold Cross Validation is explained in Subsection 6.3.1, *K-Fold Cross Validation*.

[6] Note that the analyses concerning the risk adjusted treatment costs only concern patients that were in the test set.

[7] Pearson's Correlation Coefficient is described in Subsection 6.4.1, *Pearson's Correlation Coefficient*, and Kendall's Rank Correlation Coefficient is explained in Subsection 6.4.2, *Kendall's Rank Correlation Coefficient*.

[8] Although Hinkle et al. (2003) [35] used 0.3 as cut-off point, 0.2 was used, taking into account that even very weak relations are already insightful. This also considers the amount of noise in medical data.

8 Results

In this section the results of the cost-effectiveness analysis are addressed. First, the performances of the models are presented and the influence of correcting the treatment costs for patient characteristics are addressed in Section 8.1, *Model Performances*. Second, relations between the treatment cost and constructed features were obtained via calculating direct correlations coefficients and by modelling the treatment costs. The relations are shown in Section 8.2, *Feature Correlations*. Last, the correlations between quality indicators and the corrected treatment costs are addressed in Section 8.3, *Results Effectiveness*.

For consistency, all results are focused on the target group of 4,833 leukaemia patients[1]. An example of the treatment timeline of all events in 2017 of a leukaemia patient is given in Figure 8.1. Noteworthy is that the cost-effectiveness analyses have been conducted for multiple different target groups. When important insights are obtained from the other analyses, they will be added to the results as a remark.

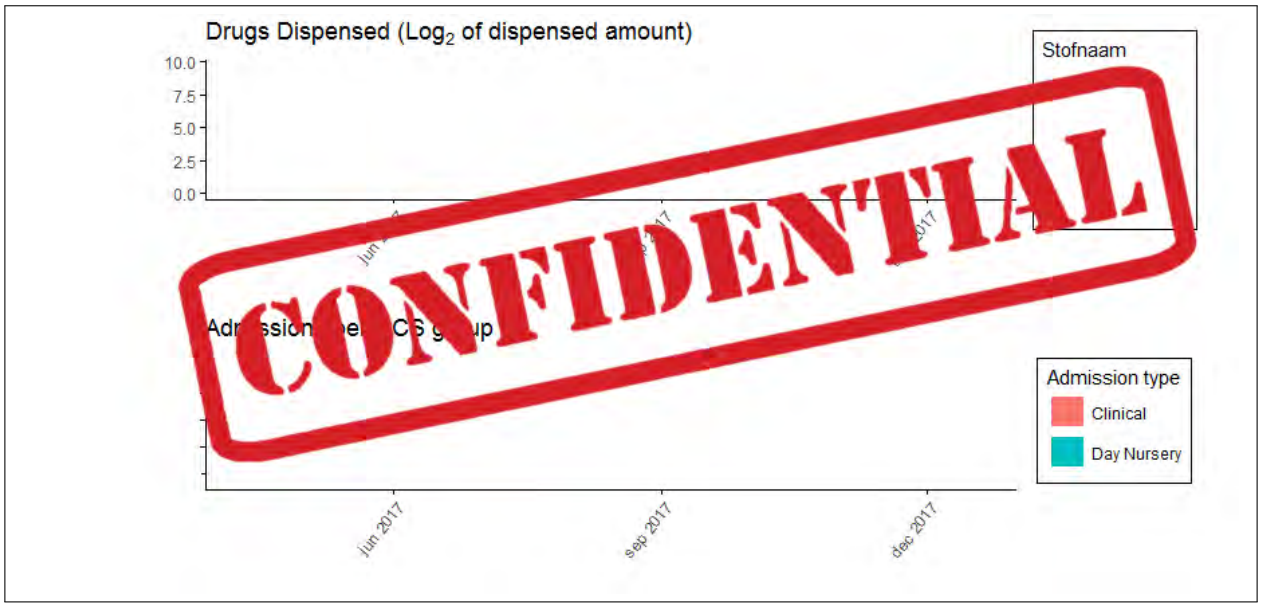

Figure 8.1: Example of the treatment timeline of a leukaemia patient. All drug dispenses and admissions since 2017 are shown for one patient. In the top timeline, all drug dispenses are shown per expensive drug. In the bottom timeline, all admissions are shown and per admission is shown if it was a day nursery or a clinical admission.

8.1 Model Performances

First, the models will be addressed that only use cost correction features describing patient characteristic. The performances of the models are based on test set results and are shown in Table 8.1. A complete table with bootstrap confidence intervals per measure per model is shown in Appendix A.6, *Result Tables*, Table A.9, with the train set results in Table A.10. All prediction measures are explained in Section 6.3, *Prediction Measures*.

Consider the test set results shown in Table 8.1. The Mean Error (ME)[2] is around zero for most mod-

[1] This regards all leukaemia diagnoses within the CCS group leukaemia. Leukaemia patients are chosen, because they are often admitted for extended observations or clinical admissions. This is essential for insightful effectiveness analyses, because the quality indicators are based on admissions.

[2] Remark that, as stated in Section 6.3, *Prediction Measures*, the ME provides insights in the average prediction, but should not be used to actually measure predictive performance.

Table 8.1: Test set results for models using only cost correction features. The models are predicting the treatment costs of leukaemia patients. Per test statistic the outcome is shown bold for average best performance. The Mean Error (ME) provides insights in how the model predicts, but does not show model performance. The Mean Average Percentage Error (MAPE) is a prediction measure, but favours lower predictions. The Mean Absolute Deviation (MAD) shows the prediction error per patient in euros. The Root Mean Squared Error (RMSE) is sensitive to outliers. The Coefficient of Determination (R^2) shows performance in comparison with the Grand Mean. Used abbreviations: Linear Regression (LR), Gamma Regression (GR), Stepwise Regression (Step) and Logarithmic-transormation (Scaled).

	ME	MAPE	MAD	RMSE	R^2
Grand Mean	**0**	22318	13320	19942	0.00
Grouped Mean	18	18883	12979	20267	-0.04
LR	-10	14721	12272	19303	**0.06**
Step LR	6	15069	12267	**19280**	**0.06**
LASSO LR	-5	15763	12276	19283	**0.06**
Scaled LR	-8532	**4019**	11560	21244	-0.14
Scaled Step LR	-8565	4218	**11556**	21247	-0.14
Scaled LASSO LR	-8779	4138	11573	21313	-0.15
GR	97	15910	12332	19356	0.05
Step GR	91	16738	12305	19304	**0.06**
RFR	104	22880	12383	19573	0.03
PCR	-2	16582	12361	19386	0.05
PLSR	**0**	15863	12342	19379	0.05

els, indicating that on average the models do not predict too low or too high. The exception is Scaled Linear Regression (Scaled LR), with a large negative ME. This indicates that Scaled Linear Regression predicts too low on average, which is in line with the high Root Mean Squared Error (RMSE), which penalises large errors. Scaled Linear Regression has the best prediction scores for the Mean Average Percentage Error (MAPE) and Mean Absolute Deviation (MAD). Furthermore, Linear Regression, Gamma Regression, Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) perform the best when the RMSE and R^2 are reviewed. When comparing the models, taking all measures into account, the models do differ on average, but most bootstrap confidence intervals do overlap. Indicating that the differences are not significant.

Overall, the models with the most robust performance[3] are Linear Regression with variable selection, PCR and PLSR. These models did on average outperform the benchmark models. The good performance of PCR and PLSR can indicate that there was multicollinearity between explanatory variables. However, this can be argued, because LR outperformed PCR and PLSR. Furthermore, which variable selection method was used for Linear Regression and Gamma Regression did not had much influence. Nevertheless, using a variable selection method did improve performance on average. Considering computation time, LASSO Regression was substantial faster than Stepwise Regression. Moreover, LASSO Linear Regression was preferred over PCAR and PLSR, because PCAR and PLSR are not created to analyse which features are important and LASSO Linear Regression does have that possibility. Therefore, LASSO Linear Regression was eventually used to do a risk adjustment on the treatment costs.

The risk adjusted treatment costs are shown in Figure 8.2. The treatment costs risk adjusted for patient characteristics are obtained by subtracting the expected treatment costs from the actual treatment costs. The expected treatment costs are the treatment costs predicted by the regression model. Note that the risk adjusted treatment costs are the residuals of the regression, and should be normally distributed, following the assumption of Linear Regression. The fact that the residuals are not normally distributed, indicates that the results of the LASSO Regression are less reliable. However, the results were validated using 5-Fold Cross Validation, which showed that that LASSO Regression was relatively

[3]Performance is robust when the confidence interval is small and multiple prediction measures show good performance.

the best model in comparison with the other models.

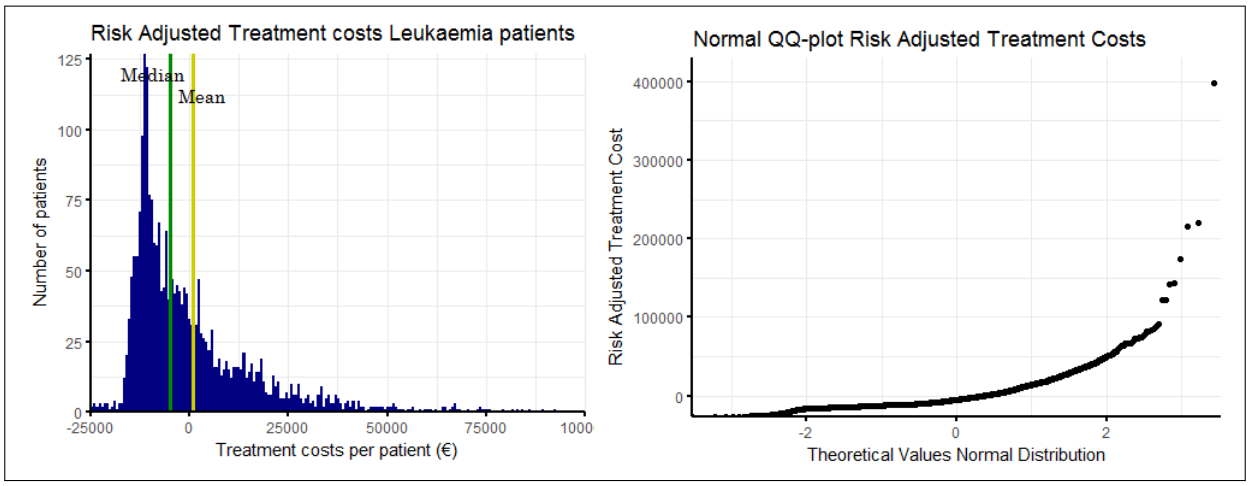

Figure 8.2: In the left graph, the risk adjusted treatment costs are shown for leukaemia patients. Visually can be concluded that the risk adjusted treatment costs are not normally distributed. This can also be derived from the gap between the mean and the median. Moreover, the normal QQ-plot on the right clearly shows that the adjusted treatment costs is not normal [6]. For the points are not approximately arranged around a straight line.

In Figure 8.3 the influence is shown of risk adjusting the treatment costs per hospital. In the top graph the treatment costs of leukaemia patients is shown and in the middle graph the risk adjusted treatment costs are presented. Remark that the risk adjustment does not influence the spread in treatment cost much, when comparing the difference per hospital between the treatment costs and risk adjusted treatment costs. This shows that a big part of the treatment costs can not be explained with the available explanatory variables describing patient characteristics. The main influence of the risk adjustment is the fact that the risk adjusted treatment costs are mean centred. These results are in line with the small coefficient of determination, R^2, shown in Table 8.1. The model centres the treatment costs better than the Grand Mean, but cannot explain the variation in treatment costs much. Furthermore, when comparing hospitals, Figure 8.3 shows that there is significant difference between hospitals in the distribution of treatment costs. This difference is tested by comparing a hospital with all other hospitals. A 95% confidence level was used and was corrected with the Bonferroni correction, which is discussed in Section 6.4, *Correlation Measures*.

Next, the models are also used to model the treatment cost using all three feature sets: cost correction features, cost explanatory features and cost-effectiveness features. The performances of the models are shown in Table 8.2. Significant improvement was found in every prediction measure in comparison with the results from Table 8.1, where the models only used cost correction features. This indicates that cost explanatory features and cost-effectiveness features contain valuable information that is needed to model the treatment costs[4]. Moreover, the regression models did significantly outperform the benchmark models. Of the regression models, RFR showed on average the best results on the test set.

Noteworthy is that RFR can learn the most complex patterns of all regression models. This could indicate that very complex dependencies exist between the treatment costs and the explanatory variables. Furthermore, good performance of PCR and PLSR can again indicate that there was multicollinearity between explanatory variables. But again, this can be argued, because LR outperformed PCR and PLSR. From the good performances of LR and RFR can be deduced that modelling simple linear dependencies is already sufficient to find relations between the treatment costs and the explanatory fea-

[4] Adding more variables only worsen the Grouped Mean model. Therefore, the Grouped Mean model was not provided with any explanatory features or cost-effectiveness features, to keep the benchmark performance as good as possible.

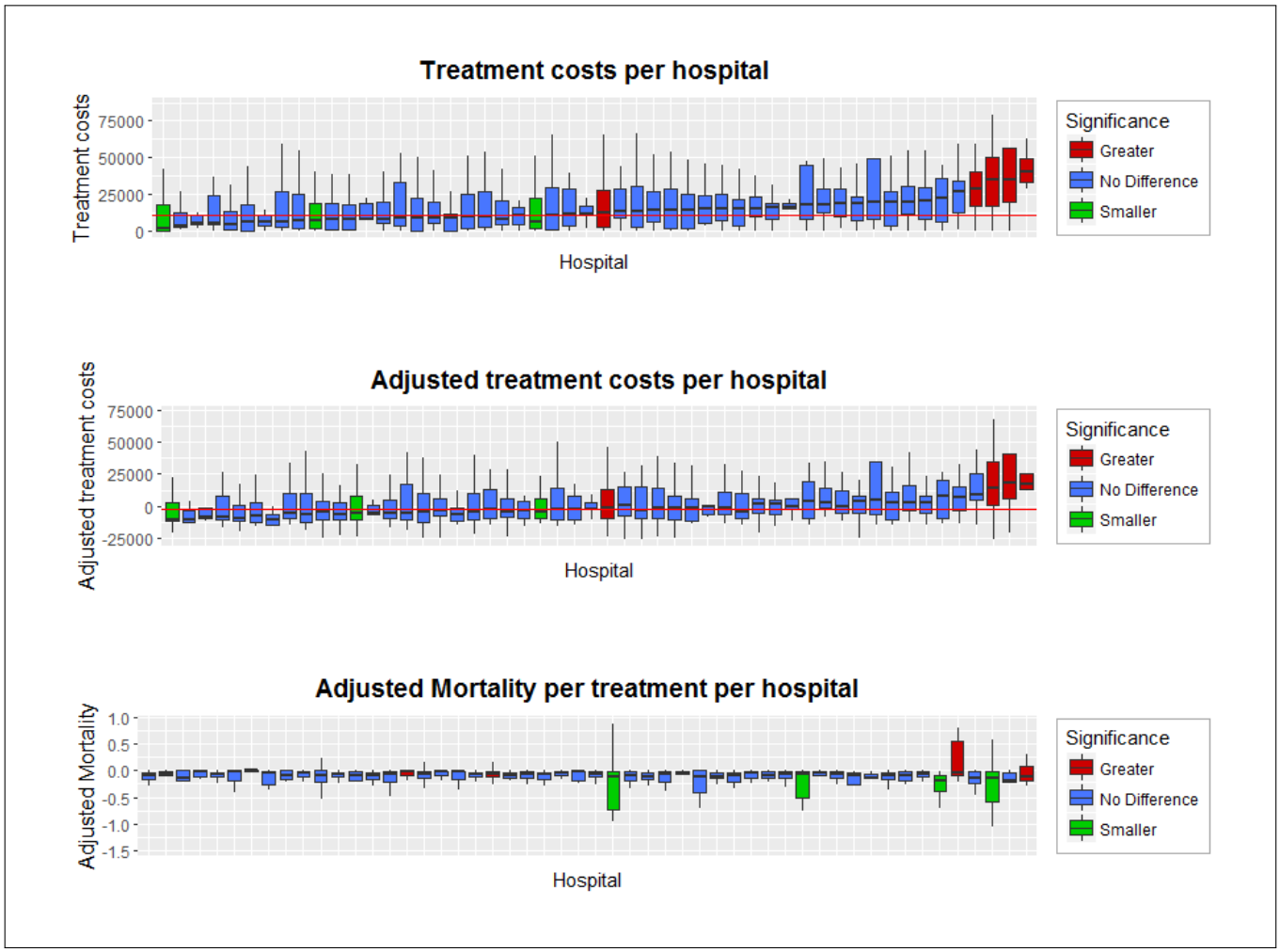

Figure 8.3: In this figure all hospitals that treat leukaemia patients are shown. In all three plots the hospitals are sorted on lowest to highest median treatment cost, thus using the order of the top plot. A red line displays the median of the treatment costs of all hospitals. In the top plot, boxplots show how the treatment costs distributions differ per hospital. In the middle plot, boxplots show how the adjusted treatment costs distributions differ per hospital. In the bottom plot, boxplots show how the adjusted mortality distributions differ per hospital For clarity reasons, the observations that fall outside the boxplot whiskers are not shown. For all three figures, the ANOVA test was used per hospital with a Bonferroni correction on the 95% confidence level to test if there was a significance between a hospital and all other hospitals.

tures. The good results of LR also show how robust LR is. Nevertheless, improvement in performance is found when a model as RFR is used that can learn more complex relations. Additionally, RFR is also a robust model, because it uses the Out-Of-Bag error to train the model. Lastly, a complete table of the performances of all models on the test set with confidence intervals per prediction measure can be found in Appendix A.6, *Result Tables*, Table A.11.

Table 8.2: Test set results for models using all three feature sets, cost correction features, explanatory features and cost-effectiveness features. The abbreviations of model names and prediction measures are addressed in Table 8.1.

Model	ME	MAPE	MAD	RMSE	R^2
Grand Mean	**-1**	21758	12399	19771	-0.00
Grouped Mean	-46	20358	12350	20221	-0.06
LR	14	9427	8931	16063	0.34
Step LR	18	7942	8857	16029	0.35
LASSO LR	22	7927	8785	16040	0.34
Scaled LR	11561	**474**	23693	201556	-117.03
Scaled Step LR	10796	500	22908	181824	-94.36
Scaled LASSO LR	9624	520	21918	174215	-85.48
GR	16856	1309	24645	181968	-97.89
Step GR	15903	1332	23651	169070	-84.64
RFR	338	1260	**7283**	**15519**	**0.39**
PCAR	2	7714	8864	16041	0.35
PLSR	9	8771	8852	16010	0.35

8.2 Feature Correlations

Next, the relations between the constructed features and the treatment costs were analysed. First, the correlations were calculated between the features and the treatment costs. The results are shown in Table 8.3, per feature is indicated if there was a significant correlation between the treatment costs and the feature. Furthermore, the influence of the features on the treatment costs was also analysed by investigating the models. For this, all three feature sets were provided to the models.

After modelling the treatment costs with all feature sets, the best performing models, LASSO LR and RFR, were analysed. For both models the feature importance[5] was analysed, to conclude if a relation existed between a feature and the treatment costs. When a feature is important for both models, the feature is shown in italic font in Table 8.3. Which features had a relation with the treatment cost according to the models was derived from the feature importance of the models. A complete list of the feature importance of all variables for RFR and LASSO LR is shown in Table A.12 in Appendix A.6.2, *Variable Importance*.

The results in Table 8.3 show that dependencies are found between the treatment costs and patient characteristics by both calculating direct correlations and analysing the regression models[6]. The important patient characteristics are diagnose[7], age and Charlson Comorbidity. Moreover, which drugs were used in the treatment also influenced the treatment costs. Lastly, the treatment duration had a positive relation, showing that longer treatments lead to higher costs.

[5] As remarked in Subsection 6.2.3, *Linear Regression*, p-values cannot be calculated for LASSO Linear Regression [51]. The feature importance is derived from the estimated coefficients multiplied by the mean of the corresponding variable. RFR bases the feature importance on how many times features are used in which tree nodes [11].

[6] Remark that the regression models also adjust the importance for explanatory dependencies found in other explanatory variables.

[7] The severity and DBC diagnose both describe the diagnose of a patient.

Table 8.3: This table shows if a correlation between a feature and the treatment costs is significant. In Section 7.3, Correlations, is addressed when a correlation is significant. The conclusions should be viewed in isolation of each other, because the correlations have not been tested with the Bonferroni correction [9]. When the features were also important for both Random Forest Regression and LASSO Linear Regression, the feature's name is italic and shown with an astrisk.

Significant	**Direction**	**Cost Correction Features**
Yes	Positive	Expensive Drug History
Yes	Negative	-
Yes	Multilevel	*Age Category**
		SES
		*DBC Diagnose**
		*Charlson Comorbidity**
No	None	Gender
		Dutch Citizen
		Admission history
		Year
Significant	**Direction**	**Cost Explanation Features**
Yes	Positive	*Treatment Duration**
		Urgency
Yes	Negative	-
Yes	Multilevel	*Severity**
		Peer Group
		*Drug prescribed**
No	None	Place of Origin
		Hospital Experience
		Hospital Size
		Label
		Insurance
		Number of Drugs
		Admitted
		Mortality
Significant	**Direction**	**Cost-Effectiveness Features**
Yes	Positive	Readmissions
		Number of Admissions
Yes	Negative	Adjusted Mortality
		Adjusted UL-LOS
Yes	Multilevel	-
No	None	Mortality
		Adjusted Readmissions
		UL-LOS
		Admission Duration
		Adjusted Admission Duration

8.3 Results Effectiveness

Eventually, the cost-effectiveness of treatments was analysed for multiple hospitals. An cost-effectiveness analyses is interesting when differences exists in risk adjusted cost and risk adjusted effectiveness between hospitals. The results in Figure 8.3 show that there are significant differences in treatment costs, also after risk adjusting for patient characteristics. The bottom plot in Figure 8.3 shows that there are also significant differences between hospitals in the quality indicator adjusted mortality. On the other hand, the cost-effectiveness can also be analysed for all patients, independent of the hospital where a patient is treated. In Table 8.3 is shown if significant correlations were present between quality indicators and the treatment costs. Furthermore, Figure 8.4 presents the relation between the adjusted mortality and the adjusted treatment costs for leukaemia patients.

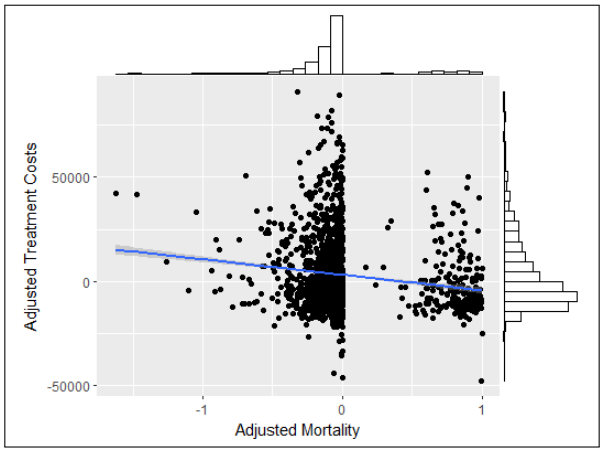

Figure 8.4: The adjusted mortality is often zero or smaller, which indicates that most patients survive all admissions while there was an expectation they would not survive. When the adjusted mortality is positive, it shows patients that did not survive their treatment. For both the negative and positive adjusted mortality applies that the higher the expectation that the patient would not survive the treatment, the lower the adjusted mortality.

It is noteworthy that it is was very dependent on the diagnose of the target group which correlations between the adjusted treatment costs and adjusted quality indicators were significant. Moreover, it also differed per hospital if the correlation was positive, negative or if there even was a significant correlation. In Table 8.4 is shown per quality indicator what percentage of the hospitals have a significant correlation between the quality indicators and the adjusted treatment costs. Noteworthy is that most quality indicators are indirectly influenced by the treatment duration. During longer treatments, patients get more often admitted and get a higher chance of being often readmitted or having a UL-LOS. Additional, every time the patient is admitted, there is a certain chance the patient will not survive the admission. Therefore, the adjusted mortality also reflects the times a patient is admitted, because the adjusted mortality lowers every time a patient is admitted and survived the admission. This is also due to the rarity of in-hospital mortality. Lastly, a complete overview about how many times correlations between quality indicators and the treatment costs are positive or negative, is shown in Table A.15 in Appendix A.7, *Leukaemia Hospital Facts*.

Finally, the quality of treatments was analysed with the composite quality indicator of Lingsma et al. (2018) [50]. In Figure 8.5 the distribution of the treatment costs and adjusted treatment costs are presented per class of the composite quality indicator. Remark that the composite quality indicator can show a quick overview of characteristics of the different classes, to analyse differences between the groups and keep track of treatment outcomes. Moreover, although the risk adjusted treatment costs are centred in comparison with the treatment costs, the distributions per class do relatively not change much by the risk adjustment.

Table 8.4: Per quality indicator the percentage of the hospitals is shown that have a significant correlation with the adjusted treatment costs. In Section 7.3, Correlations, is addressed when a correlation is significant.

Indicator	Percentage of Hospitals
Adjusted UL-LOS	0.33
Number of Admissions	0.29
Adjusted Admission Duration	0.29
Adjusted Mortality	0.25
Readmissions	0.25
Adjusted Readmissions	0.21
UL-LOS	0.21
Admission Duration	0.17
Quality Classes	0.12
Mortality	0.08

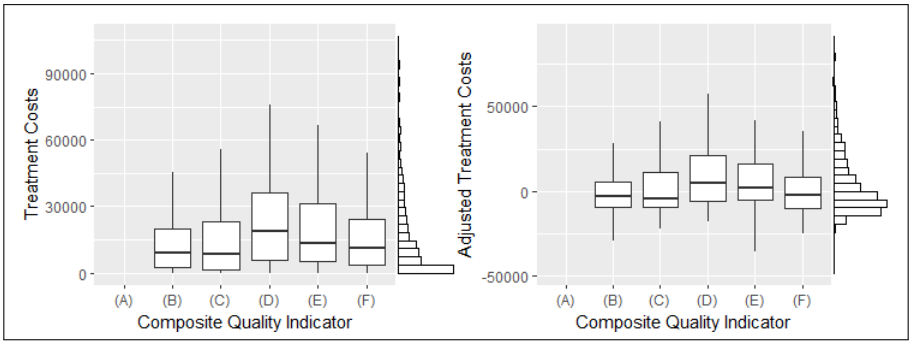

Figure 8.5: The treatment cost and risk adjusted treatment class shown per class of the composite quality indicator. Remark that all leukaemia patients underwent at least one clinical admission, resulting in an empty class (A).

9 Conclusion

This research focussed on the question: *How can a cost-effectiveness analysis be conducted on the dispense practices of expensive drugs in Dutch hospitals with the current available data?* The goal was to create a cost-effectiveness analysis that could be generalised over multiple diagnoses. In this chapter, the research question will be answered by summarising how a cost-effectiveness analysis can be set-up.

To conduct a cost-effectiveness analysis, the data need to be pre-processed first. This involves detecting problematic outliers and handling invalid values. Problematic outliers can be detected with the Problematic Outlier Detection method developed in his research and invalid values can be detected by using reference tables containing the allowed values. When an observation contains an invalid value or problematic outlier, it can be imputed by the most common valid value. However, when the invalid value or problematic outlier is irreplaceable, the whole observation needs to be excluded from the analysis.

Second, the treatment costs and treatment effectiveness need to be derived per patient. The treatment costs can be defined as the total expenses of expensive drug dispenses and the treatment effectiveness can be derived from admission quality indicators. To allow fair comparisons between patients and hospitals, a risk adjustment for patient characteristics[1] needs to be conducted on the treatment costs and treatment effectiveness. The quality indicators can be risk adjusted with Logistic Regression and the treatment costs can be risk adjusted with LASSO Linear Regression[2].

Third, relations between the treatment costs and treatment effectiveness can be measured with ANOVA, Pearson's Correlation Coefficient and Kendall's Correlation Coefficient, to conclude about the treatment cost-effectiveness. From the results can be concluded that the expensive drug choice and the duration of the treatment have a significant impact on the treatment costs. During longer treatments, patients have a higher treatment cost, because they receive more expensive drugs over the course of the treatment. Furthermore, during longer treatments, patients are admitted more often and those patients had more readmissions and more UL-LOS. On the other hand, mortality was often negatively correlated with the treatment costs. This indicates that treatments of patients that pass away are cheaper, because those treatments are aborted prematurely and would be more expensive when those patients survived.

[1] Lane-Fall and Neuman (2014) [46] explain: "Risk adjustment (also known as severity adjustment) is the process of statistically accounting for differences in patient case mix that influence health care outcomes."

[2] LASSO Linear Regression was compared with other models: the Grand Mean, Grouped Mean, Gamma Regression, Partial Least Squares Regression, Principal Component Regression and Random Forest Regression. Based on the predictive performance of LASSO Linear Regression, it was chosen as the most suitable model to do the risk adjustment.

10 Discussion

In this research several limitations of the data and of the used models were encountered. These limitations are discussed in this chapter and unrealistic assumptions are outlined. First, the quality of the available dataset is addressed in Section 10.1, *Available Data*. Second, the problem of premature withdrawals is discussed in Section 10.2, *Premature Withdrawals*. Third, Section 10.3, *Treatment Effectiveness*, reviews the limitations of how the treatment effectiveness was derived. Fourth, the downsides of the validation method for the regression models, that are used for risk adjusting the treatment costs, are discussed in Section 10.4, *Model Validation*. Fifth, the causality of significant correlations is addressed in Section 10.5, *Correlation Measures*.

10.1 Available Data

The LBZ registration is not flawless. Despite of it being a treasure of information, it is also a registration done by people, which includes a subjective bias. When hospitals do not gain added value from registering certain data, the corresponding fields in the data are often poorly registered. Furthermore, the hospitals manually supply data to DHD on a monthly basis. Consequently, months are sometimes missing due to, for instance, system failures. This creates inconsistency between hospitals registration. Experts remarked that differences in coding of hospitals and differences in how hospitals provide data result in inconsistency[1]. The consequence is that noise is present in the data, which makes the risk adjustments of the models less accurate and the calculated correlations less representative for the true causal dependencies. The same obstacle was remarked by health care studies in the literature and can be tackled by only using data of hospitals that correctly supplied all months of data and by only using field that are sufficiently registered. However, this limits the amount of data that can be used and makes the pre-processing of the data more time consuming, decreasing the possibilities to extract information from the data and limiting the time that can be used to actually conduct analyses.

Furthermore, the LBZ is limited in what it contains, and is not primarily meant for cost-effectiveness analyses. When modelling treatment costs, it is certain that not all features that influence the type of treatment are included in the dataset. For instance, the stage of how far a disorder is developed is not incorporated in the data, the same is true for a patient's allergies to certain drugs. The registered DBC diagnose can also still cover multiple variants of the same disorder, where every variant needs a different treatment. Modelling the treatment costs while correlated features are not included is known as Omitted Variable Bias [15]. Omitted Variable Bias results in overestimating the influence of the included features. This made the risk adjustment less accurate and conclusions based on feature importance less reliable. This problem can be solved by including the correlated features in the analyses or by including features that are correlated with the omitted feature.

10.2 Premature Withdrawals

Another point of discussion in this research is that it has not been taken into account that patients can be treated in multiple hospitals. For example, the most complex patients are often transfered to specialised hospitals, resulting in premature withdrawals in the hospitals that referred the patient. In a cost-effectiveness analyses, this can lead to penalising specialised hospitals for providing sub-optimal health care. The sub-optimal health care is indicated by poor quality indicator outcomes and the poor outcomes are due to patient characteristics that are not considered in the risk adjustment and the poor outcomes are not due to treatment choices of the hospital.

[1] As example, sometimes a month of data is not supplied by a hospital and some hospitals are not consistent in coding the comorbidities of patients.

Furthermore, Oostenbrink and Maiwenn [69] note that premature withdrawals lead to incomplete costs and can cause problems in analyses in health care, making the results of a cost-effectiveness less reliable. A premature withdrawal for instance happens when a patient switches from treatment, is referred to another hospital or passes away. A solution would be to track patients, independent of the hospitals where they are treated. Which would be possible if the BSN is registered for every drug dispense. For premature withdrawals not due to transfers to other hospitals, the premature withdrawals can be handled by leaving those treatments out of the analyses, or by adjusting the treatment cost and effectiveness for the fact that the treatment was prematurely withdrawn or by analysing premature withdrawals separately from treatments that are not prematurely withdrawn. This will make treatments more comparable and cost-effectiveness relations more reliable. Therefore, it is recommended to register when treatments start and end with information indicating if the treatment end was a premature withdrawal or not.

10.3 Treatment Effectiveness

A downside of how the treatment effectiveness was measured in this research, is that the effectiveness is indirectly derived from admission quality indicators and these indicators do not actually describe the quality of drug treatments. Furthermore, these quality indicators are not flawless. As example, only in-hospital mortality is registered. Thus, a patient who passes away outside the hospital due to the treatment, is not registered as passed away patient. This results in incomplete data and this means that differences in discharge policies between hospitals can influence the HSMR[2]. Moreover, it was encountered that most expensive drug dispenses do not happen during clinical admissions or extended observations. Using admission quality indicators to derive treatment effectiveness resulted in too many treatments without a quality indicator, resulting in weak correlations from which no conclusions could be drawn about the cost-effectiveness of treatments. By using quality indicators directly registered with the drug dispenses would lead to less noise and more reliable conclusions that can be drawn from the cost-effectiveness analyses.

Remark that the used quality indicators were not chosen because they measure the quality that should be measured, but they measure the quality that could be measured [44]. It would be in line with Value Based Health Care to directly measure a patient's well-being during and after the treatment[3] and to measure how well the patient physically recovered after the treatment[4] [71]. Nevertheless, no indicators are available that directly indicate the well-being of a patient and how well the patient recovered. Moreover, nowadays most quality indicators are concentrated on how the patient is treated and not on how the patient should be treated to maximise the patient's well-being. As an example, the quality indicators of the Transparantiekalender[5] of ZiN[6] focus on what care hospital can provide and which facilities are present and the quality indicators of the Basisset Medisch Specialistische Zorg Kwaliteitsindicatoren[7] of the IGJ[8] only focus on what percentage of the patients had which treatments.

As long as treatment effectiveness is not directly registered, it is expected to only find negligible or modest correlation between the treatment costs and effectiveness, as shown in this research and in the studies of Dedier et al. (2001) [20] and Schouten et al. (2005). All studies highlighted the difficulty of linking quality indicators to the actual effectiveness of care. Therefore, it is discouraged to try and find general methods that can indirectly derive effectiveness of treatments and try to measure effectiveness directly by adding the needed information to the dataset.

[2] A (30-day) post-discharge mortality would be more robust, but would also increase the registration load.
[3] Measuring a patient's well-being is also referred to as a Patient-reported Experience Measure (PREM).
[4] Measuring how well a patient recovered is also referred to as a Patient-reported Outcome Measure (PROM).
[5] Translates to Transparency calendar.
[6] Zorginstituut Nederland, Translates to Dutch National Health Care Institute.
[7] Translates to Basic Set Specialised Medical Care Quality Indicators.
[8] Inspectie Gezondheidszorg en Jeugd, Translates to the National Healthcare Institute.

10.4 Model Validation

In this research, the regression models used for risk adjusting the treatment costs were tested with 5-Fold Cross Validation, where the regression models were validated based on their prediction performance on test sets. The decision to use Cross Validation was made, because the distributions of the treatment costs could differ per target group and Cross Validation could test the robustness of models, even when model assumptions were violated. For instance, various models were used that assumed a normal distribution of the residuals. However, the residuals, the risk adjusted treatment costs, turned out not to be normally distributed. Additionally, Cross Validation penalises models that are overfitting on the training data.

A downside of using Cross Validation is that it solely tests predicting performance and does not take into account in what extend the assumptions of the regression models are violated. Therefore, the results of the models about feature importance are less reliable. This can be solved by more extensively analysing the distributions of the treatment costs and of the explanatory variables and only selecting suitable regression models, where no assumptions are violated. The violation of assumptions is, for instance, tested by analysing the residuals. Noteworthy is that most models assumed that no multicollinearity existed between the explanatory variables. This assumption was not tested, but instead was tackled by incorporating models that did not have this assumption[9], assuming that those models would outperform the others when multicollinearity existed between the explanatory variables. Remark that when multicollinearity exists between explanatory variables and a models assumes that no multicollinearity exists, it reduces the reliability of conclusions based on the feature importance.

After measuring the prediction performances of the regression models, the treatment costs were risk adjusted for patient characteristics. The results showed that risk adjusting the treatment costs did not have a lot of influence on the variation in treatment costs. However, it was expected that the treatment costs were very dependent on the patient's disorder and fitness and risk adjusting for patient characteristics would significantly decrease the variation in treatment costs. Three plausible conclusions are deduced, either the treatment costs are too noisy to be predictable with solely patient characteristics or the available data are too limited to describe patient characteristics or patient characteristics do not influence treatment costs that much. To investigate which is true, cost-effectiveness analyses were studied in the literature review. It was clear that those studies focused on a specific patient group with a specific disorder. This was done by for instance only selecting patients that received a certain treatment, knowing this treatment is only used when a disorder is in a certain stage. To perform cost-effectiveness analyses of similar quality, more specific information is needed about the disorder and the stage the disorder is in, to allow better treatment costs risk adjustment. Furthermore, the results of this research showed that the treatment cost is significantly influenced by patient characteristics as age, DBC diagnose and Charlson Comorbidity. This shows that patients characteristics do influence the treatment costs.

Next to modelling the treatment costs with only patient characteristic features, the treatment costs were also modelled with explanatory features describing treatment choices and hospital characteristics. From the results can be concluded that the expensive drug choice and the duration of the treatment had a significant impact on the treatment costs. However, it is not known if this is due to different treatment preferences of doctors or this is due to the fact that the chosen treatment indicates the stage of the patient's disorder better than the available DBC diagnose in combination with other patient characteristics.

[9] The problem of multicollinearity was, for instance, solved by some models by leaving out explanatory variables that are correlated with other explanatory variables.

10.5 Correlation Measures

A technical issue is that a correlation between two variables, does not mean that there is a causality between the two variables. When doing a top down investigation, as in this research, one may find unexpected correlations. Many correlations will exist, but the same correlations can be inverse for other disorders. It is inadvisable to search for correlations without strong believes, this will often leads to finding substantial correlations that exists by chance and are not causal. This problem is known as the multiple comparison problem, and can be solved by using the Bonferroni correction. Additionally, it is always recommended to first establish an underpinned argument for a possible correlation, before measuring it, because assuming that there is a causal link between two variables, because of a substantial correlation can be deceiving. Finding correlations should be tested with hypotheses based on facts or strong arguments.

In this research, correlations between the treatment costs and the quality were studied. The results showed that the correlations between the quality indicators and the treatment costs were namely due to quantities. More specifically, during longer treatments, patients have a higher treatment cost, because they receive more expensive drugs over the course of the treatment. Furthermore, during longer treatments, patients are admitted more often and more often admitted patients had more readmissions and more UL-LOS. On the other hand, mortality was often negatively correlated with the treatment costs, indicating that patients that pass away are cheaper. A possible explanation would be that the treatments of patients that pass away are aborted prematurely and these treatments would be more expensive when those patients survived. Noteworthy is that most correlations are negligible or low, due to the noise present in medical data.

11 Recommendations

In this chapter recommendations will be addressed that describe how valuable information can be obtained from the available data and how the data can be enhanced. First, Section 11.1, *Cost-Effectiveness Analysis* addresses recommendations considering future cost-effectiveness analyses. Second, Section 11.2, *LBZ Data*, lists general recommendations considering insights that can be obtained from the LBZ dataset. Finally, Section 11.3, *Ecological Fallacy*, addresses the ecological fallacy, to make sure it is avoided in future work.

11.1 Cost-Effectiveness Analysis

To properly adjust for patient characteristics, data need to be gathered about patient's fitness, patient's allergies and the condition of the disorder[1]. Furthermore, multiple studies highlighted the difficulty of linking general quality indicators to treatment effectiveness. It is discouraged to indirectly measure treatment effectiveness, but quality indicators should be gathered that directly describe treatment effectiveness by measuring the state of the disorder over time with the patient's well-being and information patient recovery. Furthermore, it is recommended to use Disability-adjusted life years[2] (DALY's) in cost-effectiveness studies [4].

11.2 LBZ Data

Further research is recommended to focus on the information that is already in the data and fully exploit it, instead of indirectly extracting features from the data. For instance, it recommended to extract information about which combinations of expensive drugs are prescribed for which diagnose to which patient groups and how this corresponds with the guidelines. This information can be used for hospitals to benchmark their treatments with those of other treatments.

It would be relevant for medical personnel to construct an interactive dashboard where an overview of meta-data about treatments, patient populations and complications are shown in a comprehensive and compact way. The dashboard would be most insightful when it shows the information on a high level, and dependent on the interest of the user, the user could select specific patient groups. In line with current practices, the selection of patient groups could be based on the year of the treatment and patients characteristics as disorder, age and comorbidity. Next, meta-data about treatments of this group of patients should be shown based on historic data. This meta-data considers how many times patients underwent which intervention, how many of what drugs were prescribed to patients, how long and how many admissions patients underwent and which complications happened during those admissions. Lastly, visualising the data is always preferred over presenting large tables.

11.3 Ecological Fallacy

When patients or hospitals are compared in dashboards, studies or other reports, the researcher should be aware of what the ecological fallacy is and how it can be avoided. As discussed in Chapter 4, *Related Work*, relations derived from a group level should not be used to conclude about relations on a patient level. It is recommended to always provide the possibility to dive into the data and view relations on a patient level, to ensure that conclusions about patient are derived from a patient level. Aggregating data should always be done with caution, taking into account the diversity of the observations that are aggregated.

[1] This improves the identification of patient's groups and helps to identify why certain patient groups will get or cannot get certain drug treatments.

[2] DALYs express the overall burden in years lost due to ill-health, disability or early death. A patient's DALYs can be calculated by adding the expected number of years of life lost due to early death to expected number of years lost due to disability. Here the years lost due to disability are weighted for the severity of the disability.

A Appendix

A.1 Available Data

In this section the data from DHD available for this research are presented. The hospital data are presented first. Next, the admission data are shown and afterwards the drug dispense data are addressed.

A.1.1 General hospital data

Table A.1 presents the relevant general hospital data. Per variable its name is shown with a description of the variable and with the number of unique values the variable has. When the variable is continuous, the number of unique values are not counted.

Table A.1: Hospital data description

Name	Description	Unique
Hospital name	*Name of the hospital with AGB code.*	86
Hospital type	*Type of the hospital: academic, specialised clinical or general.*	3
Hospital size	*The size of the hospital measured in number of patients treated per year*[1].	Continuous

[1]It is defined that a patient is treated by a hospital when the patient is admitted, received ambulant care, or got an intervention. Note that outpatient care is not included.

A.1.2 Admission data

Table A.2 summarises the relevant data that are recorded per admission. Per variable its name is shown with a description of the variable and with the number of unique values the variable has. When the variable is continuous, the number of unique values are not counted.

Table A.2: Admission data description

Variable name	Variable description	Unique
Hospital name	*Name of the hospital with AGB code.*	88
Patient identifier	*Identifier of the patient unique per hospital.*	5,910,260
Patient BSN	*Pseudonymized BSN of the patient.*	4,630,308
Patient year of birth	*Birth year of the patient.*	Date
Patient Comorbidity	*Indicating which of the seventeen Charlson Comorbidities a patient has.*	17
Admission type	*Indicating the type of hospital visit*[2]*.*	3
ICD-10 Diagnosis	*The main diagnosis, described by the tenth edition of the International Statistical Classification of Diseases (ICD-10)*	10,022
CCS Diagnosis	*The main diagnosis for which the drug was dispensed. Documented with the international Clinical Classifications Software.*	259
Place of Origin	*Place of origin of the patient when admitted* [3]*.*	4
Urgency	*Indicating if the patient was admitted with urgency. This indicates if care was absolutely needed within 24 hours.*	2
Severity	*Category of severity of the ICD-10 diagnosis based on the historic mortality rate.*	9
Mortality chance	*Indicates the chance a patient has to survive the admission based on the ICD-10 diagnosis.*	Continuous
Mortality	*Indicates if the patient survived the admission.*	2
Readmission chance	*Indicates the chance a patient has to be readmitted within 30 days after discharge, based on the ICD-10 diagnosis.*	Continuous
Readmission	*Indicates if the patient is readmitted.*	2
LOS	*Length of stay (LOS) in days.*	Continuous
LOS ratio	*Length of stay (LOS) in days divided by the expected length of stay in days.*	Continuous
UL-LOS	*Unexpected long length of stay (UL-LOS) indicates if the length of stay was at least 1.5 times longer than expected.*	2
Date	*Start and end date of admission*	Date

[2]A patient is registered as admitted to a hospital if the patient has a clinical admission, day nursing or extended observation.

[3]Categories for place of origin are home, other hospital, other health care institution, born in hospital

A.1.3 Drug Dispense Data

Table A.3 presents the relevant data available per drug dispense.

Table A.3: Dispense drug data description

Variable name	Variable description	Unique
Hospital name	*Name of the hospital with AGB code.*	86[4]
Patient identifier	*Identifier of the patient unique per hospital.*	413,771
Patient age	*Age of the patient.*	Numeric
Patient gender	*Gender of the patient.*	2
Patient zip code	*Zip code of the patient.*	3954
Patient nationality	*Nationality of the patient.*	82
Patient SES	*SES of the patient.*	5
Drug	*Drug dispensed based on the active substance. Documented with the Anatomical Therapeutic Chemical (ATC) Classification System.*	226
Drug age	*Number of years since the active substance was first dispensed.*	Continuous
Quantity	*Number of drug entities that are dispensed in milligram*[5]*.*	Continuous
Cost	*Cost of the drug in Euro, based on the maximum tariff of the NZa.*	Continuous
Label	*The label of the drug indicating, if the drug is on- or off-label.*	2
Insurance	*If the cost of the drug is covered by the insurance or not.*	2
Specialist	*The specialist who prescribes the drug.*	31
DBC Diagnosis	*The main diagnosis for which the drug was dispensed. Documented with DBC coding.*	2036
CCS Diagnosis	*The main diagnosis for which the drug was dispensed. Documented with the international Clinical Classifications Software.*	223
Date	*Date of drug dispense.*	Date

[4] Although 83 hospitals are present in the database, some do not yet provide data about expensive drug dispensing.
[5] Although almost all quantities are recorded in milligram, some drug are recorded in international units (iU).

A.2 Features

In this section all features used in the models are presented. First, the features to adjust treatment cost for patient characteristics are shown. Second, the features to explain the differences in treatment costs are addressed. Last, the features describing the quality of admissions are described.

A.2.1 Patient Features for Cost Correction

Given are one or more diagnoses to analyse. Next, the treatments of the patients are defined by all drug dispenses prescribed for one of those diagnoses. The response variable is the sum of the cost of all drug dispenses per patient. Table A.4 presents the explanatory variables that are used to adjust the response variable for patient characteristics.

Table A.4: Features for cost correction.

Feature	**Description**
Gender	*Gender of the patient.*
Age Category	*The age of the patient divided in the categories [0, 0], [1, 15), [15, 45), [45, 65), [65, 80), [80,80+)*[6]*.*
Dutch Citizen	*Indicates if a patient lives in the Netherlands or not.*
SES	*Social Economic Status of the patient.*
Charlson Comorbidities	*Indicating which of the seventeen Charlson Comorbidities a patient has.*
DBC Diagnose	*The main diagnosis for which the drug was dispensed. Documented with DBC coding.*
Year	*Treatment year, the year in which most dispenses took place.*
Admission history	*If the patient was admitted before the treatment started*[7]*.*
Expensive drug history	*If expensive drugs were dispensed to the patient earlier*[8]*.*

A.2.2 Features for Cost Explanation

Besides features to adjust treatment cost for patient characteristics, more features could be given to the models to explain treatment cost based on hospital characteristics and drug choices. When admissions are taken into account, only the admissions with a discharge date after the start of the treatment and admission date before 30 days after the treatment started are considered. Moreover, day nurseries are excluded. The explanatory features are presented in Table A.5.

[6] These categories are based on age categories of Ghielen (2014) [28].

[7] For this research, an admission is defined as being before the treatment when the discharge date of the admission is at least one day before the first dispense of the treatment.

[8] By construction, this are drug expenses for other DBC diagnoses that are dispensed at least one day before the first drug dispense of the treatment.

Table A.5: Features for cost explanation.

Feature	Description
Admitted	*If the patient had a clinical admission or extended observation since the start of the treatment.*
Mortality	*If the patient died during an admission.*
Severity	*Category of severity of the most frequent ICD-10 diagnosis of the admissions. The severity is based on the historic mortality rates.*
Place of Origin	*Place of origin of the patient when admitted.*
Urgency	*Indicating if the patient was admitted with urgency.*
Expensive drug	*Which drugs are dispensed to the patient.*
Number of drugs	*The number of different expensive drugs dispensed to the patient.*
Drug prescribed	*Which drug is dispensed.*
Label	*The label of the drug.*
Insurance	*If the cost of the drug is covered by the insurance or not.*
Treatment duration	*Number of days between first and last drug dispense.*
Hospital Experience	*Number of patients with the same DBC Diagnose the hospital treated since 2015.*
Hospital type	*Type of the hospital: general, specialised clinical or acadamic.*
Hospital size	*The number of patients treated per year.*

A.2.3 Features for Cost-Effectiveness

By deriving quality indicators from the data, it was possible to analyse the cost-effectiveness by calculating the correlation between the treatment cost and the quality indicators. The features for cost-effectiveness are showed in Table A.6.

As remarked in Section 7.1, *Modelling Treatment Cost*, patients do not have quality indicators when they had no clinical admission or extended observation during the treatment. In such cases, the quality indicators are set to zero, meaning the patient survived the treatment without readmissions or UL-LOS. This also means the patient had zero chance to die, zero chance on being readmitted, zero admissions and zero admission days. Furthermore, not all admissions are used for deriving the quality indicators. First, the admission needs to have a discharge date after the start of the treatment. Second, the admission also needs to have an admission date before 30 days after the treatment ended. Third, the admission is a clinical admission or extended observation. These rules are included to ensure that the admissions are almost certainly part of the treatment and that the admissions have relevant quality indicators.

Table A.6: Features for Cost-Effectiveness.

Feature	Description
Mortality	*Indicates if the patient survived all admissions.*
Adjusted Mortality	*Over all admissions, taking the sum of the difference between mortality and the mortality chance,* $\sum(Mortality - Mortality_{chance})$.
Readmissions	*The number of times the patient is readmitted.*
Adjusted Readmissions	*Over all admissions, taking the sum of the difference between Readmission and the Readmission chance,* $\sum(Readmission - Readmission_{chance})$.
UL-LOS	*Number of times the patient has a 50% unexpected long length of stay (UL-LOS) than expected.*
Adjusted UL-LOS	*Taking the difference between the number of UL-LOS and the number of clinical admissions and observations,* $\#$ *UL-LOS* $-$ $\#$ *admissions.*
Admission Duration	*Total duration of all admissions in days.*
Adjusted Admission Duration	*Over all admissions, taking the sum of the difference between and the number of clinical admissions and observations,* $\sum(Admission\ Duration\ - Admission\ Duration\ _{expected})$.
Number of Admissions	*Number of clinical admissions and extended observations.*

A.3 Outlier Detection

In this section three outlier detection methods are described that are studied in this research. Although these methods were implemented, they were not used in the final implementation. Therefore they are not addressed in Section 6.1, *Outlier Detection*. First the Z-score method is discussed in Subsection A.3.1, then the Local Outlier Factor method is described in Subsection A.3.2 and lastly Cook's Distance is addressed. All Methods are discussed given the observed variable $X = \{x_1, \ldots, x_N\}$. Define $x_i \in \mathbb{R}$ as observation i, for $i \in \{1, \ldots, N\}$, with N the total number of observations. Define a problematic outlier as $\tilde{x}_i$.

A.3.1 Z-score Method

The Z-score method detects outliers by locating all observations that are unlikely far away from the mean of the data [6]. The method is visualised in Figure A.1. Given the sample mean μ and sample standard deviation σ, the Z-score for observation i, z_i, is calculated with formula A.1.

$$\begin{aligned} \mu &= \tfrac{1}{N}\textstyle\sum_{i=1}^{N} x_i \\ \sigma &= \sqrt{\tfrac{1}{N-1}\textstyle\sum_{i=1}^{N}(x_i - \mu)^2} \\ z_i &= \tfrac{x_i - \mu}{\sigma} \end{aligned} \tag{A.1}$$

Now z_i indicates the number of standard deviations a point deviates from the mean, assuming that the variable $X = \{x_1, \ldots, x_N\}$ is normally distributed. A rule of thumb is that 0.3% of the observations lie outside three times the standard deviation and are problematic outliers. Thus, $\tilde{x}_i$ is problematic when $|z_i| > 3$.

A downside of this method is that it assumes a normal distribution of the data and can have unwanted results when distributions are skewed or extreme outliers are present. However, the drug dispense data are not normally distributed and extreme outliers are present and the data.

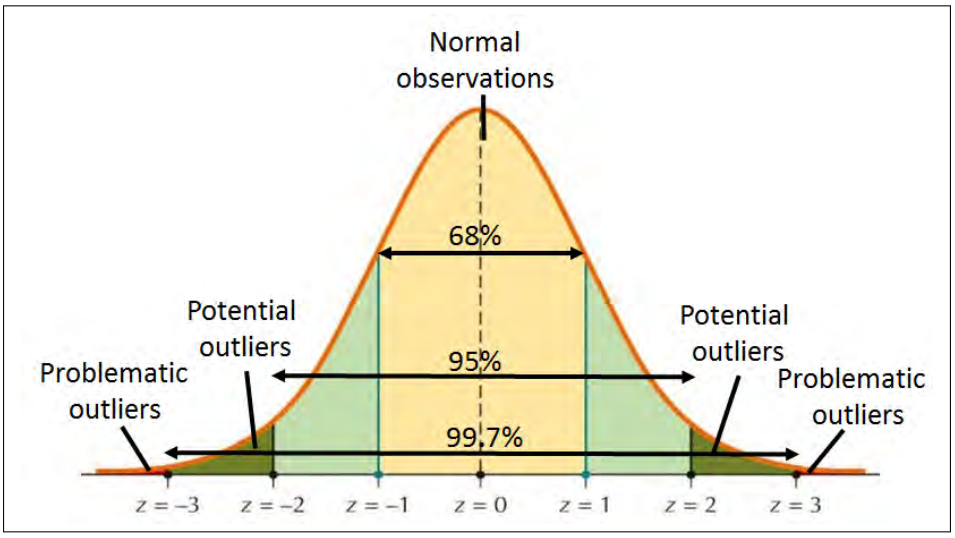

Figure A.1: The Z-score method is based on the normal distribution. This figure shows which observations would qualify as problematic outliers following the Z-score method.

A.3.2 Local Outlier Factor Method

A robust outlier detection method is the LOF method, Local Outlier Factor method [10]. Define d_{ij} as the distance between x_i and x_j. Furthermore, the *k-distance* of x_i is the distance to the k^{th} closest point. When the observations are sorted, such that $d_{i1} \leq d_{i2} \leq \ldots \leq d_{iN}$[9], then the *k-distance* of x_i is defined as,

$$kdist(i) = d_{ik}. \tag{A.2}$$

Using the definition with sorted observations, the *k-distance neighbourhood of* x_i are all x_j with $d_{ij} \leq d_{ik}$. Define N_{kdist} as the number of points with $d_{ij} \leq d_{ik}$[10].

Again using the definition with sorted observations, the Local Outlier Factor (LOF) of x_i is defined as,

$$LOF_k(i) = \frac{1}{N_{kdist}} \sum_{j=1}^{N_{kdist}} \frac{kdist(j)}{kdist(i)} \tag{A.3}$$

The $LOF_k(i) \in \mathbb{R}^+$, when $LOF_k(i) \leq 2$ it indicates that x_i belongs to a cluster. On the other hand, when $2 < LOF_k \leq 3$ it indicates a potential outlier. When $LOF_k(i) \geq 3$ it indicates that x_i is a problematic outlier.

The advantage of LOF is that it defines outliers by their local neighbourhood and does not assume a distribution. This allows to detect outliers between multiple clusters. Nevertheless, the downside is that for every point the *k-distance* is needed of all points in their *k-distance neighbourhood*. Which requires much computational power in large datasets. Moreover, k needs to be predefined. Different values of k can have significant impact. A value between 10 and 40 is recommended by Breunig et al (2000) [10]. Remark that a larger k increases the number of computation needed.

A.3.3 Cook's Distance

Outliers can have an unwanted influence on the outcome of a regression [18]. Cook's Distance compares the outcome of a regression with and without a given observation.

[9] The point closest to x_i is x_i itself, therefore i can be redefined as $i = 1$.
[10] Note that N_{kdist} can be greater than k when duplicate values are present.

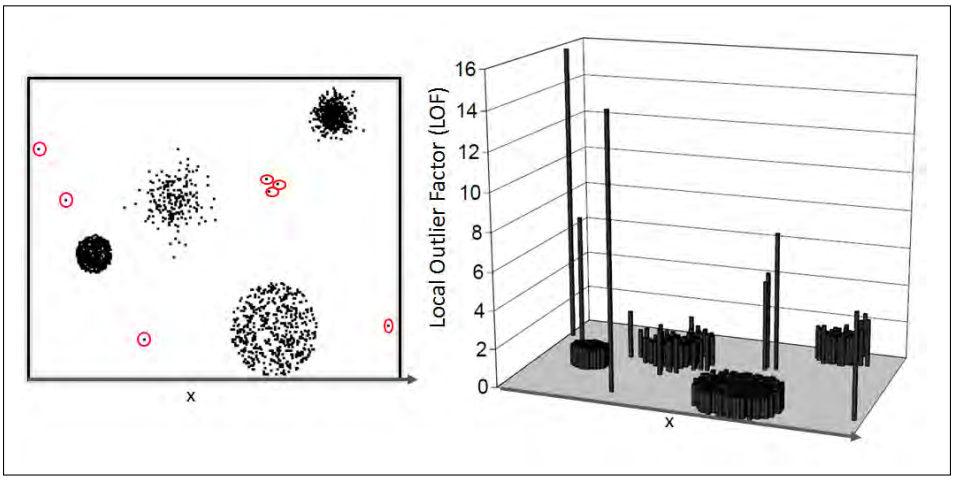

Figure A.2: The LOF method is able to detect which observations are part of clusters and which are not. Observations with a relative high LOF can be classified as problematic outlier. In the figure these points are red encircled.

Define the response variable $Y = \{y_1, \ldots, y_N\}$ and define V, the number of explanatory variables in the regression. Next, let $\hat{y}_j$ be the predicted value and let $\hat{y}_{j(i)}$ be the predicted value where observation i has been left out of the regression. Using the Mean Squared Error (MSE), the Cook's Distance for observation i, D_i, is calculated with,

$$\begin{aligned} D_i &= \frac{\sum_{i=1}^{N}(\hat{y}_j - \hat{y}_{j(i)})}{(V+1)\cdot MSE}, \\ \text{MSE} &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \end{aligned} \tag{A.4}$$

Several ways exist to label observations as problematic outliers using Cook's Distance. Two of them are,

- Observations with Cook's Distance larger than three times the mean Cook's Distance, $D_i > 3\frac{1}{N}\sum_{j=1}^{N} D_j$, can be labelled as outliers.
- Observations with Cook's Distance of $D_i > \frac{4}{N}$ are labelled as outliers.

When performing a regression, an advantage of using Cook's Distance to detect outliers is that it takes the response variable and all explanatory variables into account when detecting outliers. Yet, a disadvantage is that there is no general way to determine for which values of D_i observation i is an outlier.

A.4 Regression models

In this section all regression models are presented in Table A.7 with short descriptions of the models and the sections in which the models are explained.

Table A.7: Regression model overview.

Model	Description	Section
Grand Mean	*Uses the mean of all observations as prediction.*	6.2.1
Grouped Mean	*Uses the mean per group, based on diagnose, age category and treatment year, as prediction.*	6.2.2
LR	*Linear Regression (LR) with all explanatory variables.*	6.2.3
Step LR	*LR where a subset of explanatory variables is used, selected with Bidirectional Elimination (Stepwise Regression).*	6.2.3
LASSO LR	*LR where a subset of explanatory variables is used, selected with LASSO regression.*	6.2.3
Scaled LR	*LR with all explanatory variables and the response variable scaled with the natural logarithm.*	6.2.3
Scaled Step LR	*LR where a subset of explanatory variables is used, selected with Bidirectional Elimination. The response variable is scaled with the natural logarithm.*	6.2.3
Scaled LASSO LR	*LR where a subset of explanatory variables is used, selected with LASSO regression. The response variable is scaled with the natural logarithm.*	6.2.3
GR	*Gamma Regression (GR) with all explanatory variables.*	6.2.4
Step GR	*GR where a subset of explanatory variables is used, selected by the Bidirectional Elimination.*	6.2.4
PCAR	*Regression with the most important principal components as explanatory variables.*	6.2.5
PLSR	*Regression with the most important latent variables as explanatory variables.*	6.2.6
RFR	*Random Forest Regression (RFR) ensembles decision trees.*	6.2.7

A.5 Prediction Measures

In this section all prediction measures are presented in Table A.8 with short descriptions of the measures and the sections in which the measures are explained.

Table A.8: Prediction measures overview.

Measure	Description	Section
ME	Mean Error.	6.3.2
MAD	Mean Absolute Error.	6.3.3
R^2	Shows performance in comparison with Grand Mean.	6.3.6
MAPE	Absolute error relative to actual value.	6.3.4
RMSE	Absolute error with an extra penalty on large errors.	6.3.5

A.6 Result Tables

In this section results are presented. The prediction measure results are shown for modelling the treatment cost.

A.6.1 Prediction Measure Results

In Table A.9 the results of the prediction measures of the test set predictions are shown with confidence interval[11] for the models using only patient characteristic features. The value shown is the mean over the five results, which are obtained with 5-fold Cross Validation. Next, the corresponding train set results[12] are presented in Table A.10. Moreover, Figure A.3 gives a visualisation of a prediction on a test set, compared with the actual cost and Grouped Mean benchmark. Last, the prediction measure test set results are shown for the models using all features in Tables A.11. Figure A.3 presents a visualisation of a prediction on a test set, compared with the actual cost and Grouped Mean benchmark.

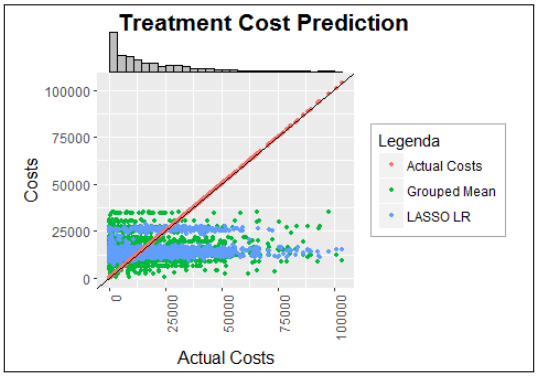

Figure A.3: Visualisation of a prediction of the treatment cost of the best model using only explanatory features compared with the Grouped Mean benchmark model. The actual costs are presented on the horizontal axis. Per actual cost, the predicted value is presented on the vertical axis. By predicting the actual cost with the actual cost itself, the most ideal prediction would be to be as close as possible to the identity line.

[11] With 5-Fold Cross Validation, Five results are obtained per measure per model, the confidence interval shows the worst and best result per measure per model.

[12] To keep the table concise, the confidence interval is not displayed.

Table A.9: Test set results for models using only cost correction features describing patient characteristic.

	ME	**MAPE**	**MAD**	**RMSE**	**R^2**
Grand Mean	0 [-2825, 1114]	22318 [9569, 35195]	13320 [12863, 14315]	19942 [17646, 25635]	0 [-0.01, 0]
Grouped Mean	18 [-2787, 1458]	18883 [10195, 28188]	12979 [11862, 14757]	20267 [16899, 26269]	-0.04 [-0.16, 0.09]
LR	-11 [-2673, 1018]	14721 [7901, 22353]	12272 [11670, 13280]	19303 [16754, 24717]	0.06 [0.02, 0.1]
Step LR	7 [-2647, 1054]	15069 [7887, 22347]	12267 [11660, 13281]	19280 [16733, 24699]	0.06 [0.02, 0.11]
LASSO LR	-5 [-2659, 1032]	15763 [7886, 23600]	12276 [11710, 13282]	19283 [16752, 24785]	0.06 [0.03, 0.1]
Scaled LR	-8533 [-10873, -7465]	4019 [1939, 5225]	11560 [10791, 13517]	21244 [18859, 27223]	-0.14 [-0.15, -0.12]
Scaled Step LR	-8566 [-10868, -7489]	4218 [1902, 5614]	11556 [10768, 13498]	21247 [18855, 27234]	-0.14 [-0.15, -0.12]
Scaled LASSO LR	-8780 [-11226, -7806]	4138 [1912, 5471]	11573 [10764, 13558]	21313 [18855, 27416]	-0.15 [-0.16, -0.13]
GR	97 [-2463, 1061]	15910 [7773, 22990]	12332 [11730, 13363]	19356 [16855, 24538]	0.05 [0.01, 0.09]
Step GR	92 [-2461, 1085]	16738 [7780, 25591]	12305 [11694, 13345]	19304 [16776, 24547]	0.06 [0.01, 0.1]
RFR	105 [-2498, 1128]	22880 [6778, 48938]	12383 [11726, 13446]	19573 [17163, 25160]	0.03 [-0.05, 0.09]
PCAR	-2 [-2666, 1105]	16582 [7523, 24449]	12361 [11790, 13354]	19386 [16872, 25035]	0.05 [0.03, 0.09]
PLSR	0 [-2689, 1092]	15863 [7616, 23997]	12342 [11815, 13258]	19379 [16934, 24998]	0.05 [0.03, 0.08]

Table A.10: Train set results for models using only cost correction features describing patient characteristic.

	ME	**MAPE**	**MAD**	**RMSE**	**R^2**
Grand Mean	0	22289	13302	20124	0
Grouped Mean	0	17206	11981	18895.6	0.12
LR	0	14672	12158	19241.4	0.09
Step LR	0	15592	12174	19258	0.08
LASSO LR	0	15457	12174	19258	0.08
Scaled LR	-8550	796	11500	21395	-0.13
Scaled Step LR	-8583	3924	11504	21406	-0.13
Scaled LASSO LR	-8784	3962	11534	21477	-0.14
GR	98	15731	12214	19320	0.08
Step GR	79	16189	12217	19317	0.08
RFR	94	11833	8943	14388	0.49
PCAR	0	17151	12308	19513	0.06
PLSR	0	16285	12253	19449	0.07

Table A.11: Test set results for models using all features.

	ME	MAPE	MAD
Grand Mean	0 [-2825, 1114]	22318.6 [9569, 35195]	13320.2 [12863, 14315]
Grouped Mean	18 [-2787, 1458]	18883.4 [10195, 28188]	12979.8 [11862, 14757]
LR	14 [-2121, 1021]	8896 [2463, 21503]	9030.8 [8574, 9596]
Step LR	41 [-2127, 1048]	7771.2 [2437, 15914]	8991.4 [8505, 9595]
LASSO LR	18 [-2116, 991]	7374.8 [2115, 17902]	8907.2 [8486, 9509]
Scaled LR	8755 [5530, 12126]	441 [315, 581]	21549.4 [16801, 28821]
Scaled Step LR	8547 [6095, 12203]	459.2 [321, 611]	21361.8 [17474, 28958]
Scaled LASSO LR	7609 [4304, 11253]	455.2 [315, 595]	20515.2 [15728, 27984]
GR	13007 [8854., 17469]	1241.8 [794, 1682]	21055.6 [16575, 28663]
Step GR	12438 [8775, 16544]	1261.4 [805, 1727]	20476.6 [16461, 27702]
RFR	277 [-1797.62, 1210]	1153 [688, 1691]	7444.6 [6960, 8243]
PCAR	18 [-2086, 1029]	6755.6 [1939, 14155]	8972.2 [8528, 9587]
PLSR	6 [-2080, 1047]	7913.6 [2328, 19181]	8994.8 [8552, 9668]

	RMSE	R2
Grand Mean	19942 [17646, 25635]	0 [-0.01, 0]
Grouped Mean	20267 [16899, 26269]	-0.04 [-0.16, 0.09]
LR	15787.6 [13256, 21601]	0.38 [0.28, 0.43]
Step LR	15764 [13205, 21568]	0.38 [0.28, 0.44]
LASSO LR	15745 [13216, 21654]	0.38 [0.28, 0.44]
Scaled LR	141714 [81313, 207727]	-53.75 [-114.12, -20.31]
Scaled Step LR	137396 [95204, 214902]	-47.4 [-72.49, -28.22]
Scaled LASSO LR	128136 [71773, 199680]	-42.86 [-79.76, -15.6]
GR	107477 [65750, 194492]	-28.77 [-57.27, -10]
Step GR	102501 [65761, 182837]	-25.92 [-50.5, -10]
RFR	15129 [12689, 21232]	0.43 [0.31, 0.48]
PCAR	15825 [13226, 21854]	0.37 [0.26, 0.44]
PLSR	15792 [13195, 21845]	0.38 [0.26, 0.44]

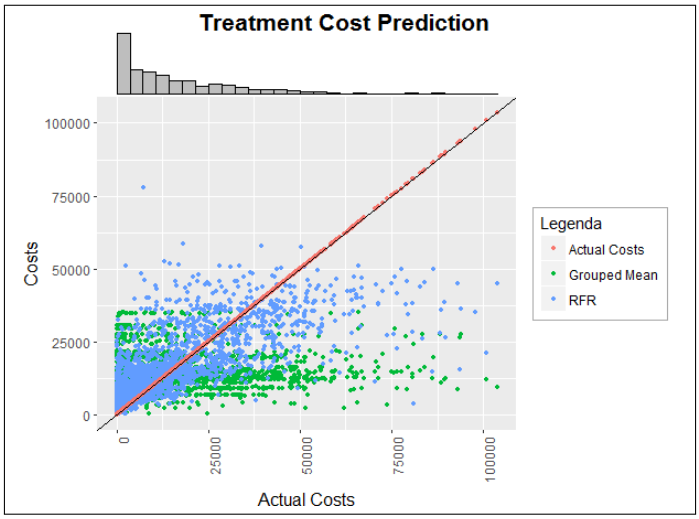

Figure A.4: Visualisation of a prediction of the treatment cost of the best model using all features compared with the Grouped Mean benchmark model. The actual costs are presented on the horizontal axis. Per actual cost, the predicted value is presented on the vertical axis. By predicting the actual cost with the actual cost itself, the most ideal prediction would be to be as close as possible to the identity line.

A.6.2 Variable Importance

The models use the features to predict the treatment cost, by analysing the influence every explanatory variable has on the response variable, feature importance was derived. For Random Forest Regression (RFR) the influence is derived from the number of times a feature is chosen to split on what depth of the tree. For LASSO Linear Regression (LASSO LR) the absolute coefficient values of the explanatory variables are used multiplied by the mean value of a variable. For categorical variables, the absolute importance of all separate categories are summed. The results are shown in Table A.12. When both models agree on a feature being important, the feature name is shown in bold face.

Table A.12: Influence of explanatory variables on the response variable. The direct importance is shown, as well as the percentage relative to the other variables. The Random Forest Regression Importance (RFR Imp.) is in 100,000,000. Note that the absolute coefficient values are used for LASSO LR Importance.

Feature	RFR Imp.	RFR %	LASSO LR Imp.	LASSO LR %
Drug	102849	68.6%	1810	23%
DBC Diagnose	14518	9.7%	490	6.2%
Age	8753	5.8%	206	2.6%
Treatment Duration	7421	5%	1916	24.3%
Comorbidity	6369	4.2%	228	2.9%
Place of Origin	4623	3.1%	67	0.9%
Severity	2799	1.9%	333	4.2%
Year	1318	0.9%	106	1.3%
Label	323	0.2%	130	1.7%
Hospital Experience	322	0.2%	222	2.8%
Expensive Drug History	158	0.1%	26	0.3%
Mortality	152	0.1%	29	0.4%
Adjusted UL-LOS	144	0.1%	158	2%
SES	57	0%	257	3.3%
Adjusted Admission Duration	47	0%	403	5.1%
Adjusted Readmission	7	0%	235	3%
Adjusted Mortality	0	0%	479	6.1%
Admission History	0	0%	35	0.4%
Dutch Citizen	0	0%	8	0.1%
Gender	0	0%	53	0.7%
Hospital Size	0	0%	347	4.4%
Insurance	0	0%	202	2.6%
Peer Group	0	0%	84	1.1%
Urgency	0	0%	41	0.5%
Number of Drugs	0	0%	292	5.1%

A.7 Leukaemia Hospital Facts

In this section facts are described about the selection of the target group. First, Table A.13 shows the number of patients per year for leukaemia patients. Second, Table A.14 presents the number of expensive drug dispenses per DBC diagnoses within the CCS group of Leukaemia. Last, Table A.15 shows for how many hospitals significant correlations were found between treatment costs and quality indicators for treatment effectiveness.

Table A.13: Overview showing the number of patients that received expensive drugs for leukaemia per year. Note that the total is the total number of patients for all years, which is not the sum of all years, because patients can be treated for multiple years.

Year	Total	% Male	% Female
2017	5834	0.60	0.40
2015	4202	0.58	0.42
2016	4892	0.60	0.40
Total	9562	0.60	0.40

Table A.14: Overview about the DBC diagnoses for which expensive drugs are dispensed.

DBC Diagnose	Dispenses	Percentage	Cumulative Percentage
0313.761 Acute myelode leukemie/RAEB-t	51069	29%	29%
0313.757 CLL, Waldenstrm, Hairy cell leukemie	47831	27%	57%
0313.771 Chronische myelode leukemie (CML)	26930	15%	72%
0316.6111 Leukemie	20222	12%	84%
0313.763 Myelodyplasie overige nno	10768	6%	90%
0313.756 Acute lymfatisch leukemie	7317	4%	94
0313.762 RAEB	6311	4%	98%
0313.773 CMMoL	3278	2%	100%

Table A.15: Overview about in how many hospitals correlations between quality indicators and the treatment costs are significant positive, negative or negligible for leukaemia patients.

Indicator	Positive	Not Significant	Negative
Mortality	2	45	5
Adjusted Mortality	1	37	14
Readmissions	12	38	2
Adjusted Readmissions	10	36	6
UL-LOS	9	38	5
Adjusted UL-LOS	0	34	18
Admission Duration	10	38	4
Adjusted Admission Duration	4	37	11
Number of Admissions	19	33	0
Composite Quality Indicator	8	39	5

Alphabetical Index

Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Petrov, B.N.; Csáki, F., 2nd International Symposium on Information Theory. Tsahkadsor, Armenia, USSR. Budapest: Akadémiai Kiadó, pp. 267281.

[2] Andrabi, I. (2012). A Culture of Continuous Improvement is Necessary for Success Under Value-Based Care. Retrieved September 12, 2018, from www.beckershospitalreview.com.

[3] Bakx, P., O'Donnell, O., & Van Doorslaer, E. (2016). Spending on health care in the Netherlands: not going so Dutch. Fiscal Studies, 37(3-4), 593-625.

[4] Baltussen, R. M., Adam, T., Tan-Torres Edejer, T., Hutubessy, R. C., Acharya, A., Evans, D. B. & Murray C. J. L. (2003). Making choices in health: WHO guide to cost-effectiveness analysis. World Health Organization.

[5] Bhande, A. (2018). What is underfitting and overfitting in machine learning and how to deal with it. Retrieved October 15, 2018, from www.medium.com.

[6] Bijma, F., Jonker, M. A., & van der Vaart, A. W. (2013). Inleiding in de statistiek. Epsilon Uitgaven.

[7] Bjørn-Helge Mevik, Ron Wehrens, & Kristian Hovde Liland (2016). pls: Partial Least Squares and Principal Component Regression. R package version 2.6-0.

[8] Blommestein, H. M., Verelst, S. G., de Groot, S., Huijgens, P. C., Sonneveld, P., & Uylde Groot, C. A. (2016). A costeffectiveness analysis of realworld treatment for elderly patients with multiple myeloma using a full disease model. European journal of haematology, 96(2), 198-208.

[9] Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilit, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

[10] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. ACM sigmod record (Vol. 29, No. 2, pp. 93-104). ACM.

[11] Breiman, L. (2001). Random Forest. Machine Learning, 45, 5-32.

[12] Brys, G., Hubert, M., & Struyf, A. (2004). A Robust Measure of Skewness. Journal of Computational and Graphical Statistics, Vol 13, 996 - 1017.

[13] Het Centraal Bureau voor de Statistiek. (2018) Mission statement. Retrieved October 19, 2018, from www.cbs.nl.

[14] Chamberlin, D. D., Gilbert, A. M., & Yost, R. A. (1981). A history of System R and SQL/data system. In Proceedings of the seventh international conference on Very Large Data Bases-Volume 7 (pp. 456-464). VLDB Endowment.

[15] Clarke, K. A. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. Conflict Management and Peace Science. 22: 341352.

[16] Cihangir, S., Borghans, I., Hekkert, K., Muller, H., Westert, G., & Kool, R. B. (2013). A pilot study on record reviewing with a priori patient selection. BMJ open, 3(7), e003034.

[17] Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Lawrence Erlbaum associates, publishers. 2nd.

[18] Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics, 19(1), 15-18.

[19] Dean, B., Barber, N., & Schachter, M. (2000). What is a prescribing error?. BMJ Quality & Safety, 9(4), 232-237.

[20] Dedier, J., Singer, D. E., Chang, Y., Moore, M., & Atlas, S. J. (2001). Processes of care, illness severity, and outcomes in the management of community-acquired pneumonia at academic hospitals. Archives of internal medicine, 161(17), 2099-2104.

[21] DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled clinical trials, 7(3), 177-188.

[22] Dutch Hospital Data (2018). Mission statement and experience. Retrieved from www.dhd.nl Retrieved April 3, 2018, from www.DHD.nl.

[23] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Volume 38. Siam.

[24] Fischer, C., Lingsma, H. F., Marang-van de Mheen, P. J., Kringos, D. S., Klazinga, N. S., & Steyerberg, E. W. (2014). Is the readmission rate a valid quality indicator? A review of the evidence. PloS one, 9(11), e112282.

[25] Forster, A. J., & van Walraven, C. (2012). The use of quality indicators to promote accountability in health care: the good, the bad, and the ugly. Open Medicine, 6(2), e75.

[26] Geneesmiddel Debat (2018). Kosten geneesmiddelen in Nederland. Retrieved September 12, 2018, from wwwgeneesmiddeldebat.nl.

[27] Geneesmiddel Debat (2018). Inkoop zorgverzekeraars. Retrieved September 12, 2018, from www.geneesmiddelendebat.nl.

[28] Ghielen, J. (2014). Toelichting bepaling verwahte verpleegduur. DT Healthcare Solutions and Dutch Hospital Data.

[29] Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of educational research, 42(3), 237-288.

[30] Hastie, T. J. & Pregibon, D. (1992) Generalized linear models. Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

[31] Hawkins, D. (1980). Identification of Outliers. Chapman and Hall, London.

[32] Hekkert, K., Van der Brug, F., Borghans, I., Cihangir, S., Zimmerman, C., Westert, G., & Kool, R. B. (2017). How to identify potentially preventable readmissions by classifying them using a national administrative database. International Journal for Quality in Health Care, 29(6), 826-832.

[33] Hertzog, L. (2016). Bayesian regression with STAN Part 2: Beyond normality. Retrieved October 15, 2018, from www.r-bloggers.com.

[34] High, R. (2000). Dealing with outliers: How to maintain your datas integrity. University of Oregon, available at www.uoregon.edu.

[35] Hinkle, D. E., Wiersma, W., & Jurs, S.G. (2003). Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin.

[36] Hofstede, S. N., van Bodegom-Vos, L., Kringos, D. S., Steyerberg, E., & Marang-van de Mheen, P. J. (2017). Mortality, readmission and length of stay have different relationships using hospital-level versus patient-level data: an example of the ecological fallacy affecting hospital performance indicators. BMJ Qual Saf, bmjqs-2017.

[37] Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. Computational statistics & data analysis, 52(12), 5186-5201.

[38] Hulscher, M. E., Grol, R. P., & van der Meer, J. W. (2010). Antibiotic prescribing in hospitals: a social and behavioural scientific approach. The Lancet infectious diseases, 10(3), 167-175.

[39] Institute for Medical Technology Assessment. Questionnaires for the measurement of costs in economic evaluations. Retrieved October 1, 2018, from www.imta.nl.

[40] Inspectie voor de Gezondheiszorg en Jeugd (2018). Basisset Medische Specialistische Zorg Kwaliteitsindicatoren. Retrieved Augustus 20, 2018, from www.igj.nl.

[41] Jolliffe, I. T. (2002). Graphical representation of data using principal components. Principal component analysis, 78-110.

[42] Kendall, M. (1938). A New Measure of Rank Correlation. Biometrika. 30 (12): 8189.

[43] Kop, R., Hoogendoorn, M., Ten Teije, A., Bchner, F. L., Slottje, P., Moons, L. M., & Numans, M. E. (2016). Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Computers in biology and medicine, 76, 30-38.

[44] van der Laan, J., Penning, C., de Bruin, A. (2017). Hospital Readmission Ratio Methodological report of 2015 model. (1-31). Statistics Netherlands.

[45] van der Laan, J., Penning, C., de Bruin, A., van den Akker, J., (2016). HSMR 2016: Methodological report. (1-44). Statistics Netherlands.

[46] Lane-Fall, M. B., & Neuman, M. D. (2013). Outcomes measures and risk adjustment. International anesthesiology clinics, 51(4).

[47] Lay, D.C., (2002). Linear Algebra and Its Applications. Pearson Education.

[48] Lewis, P. J., Dornan, T., Taylor, D., Tully, M. P., Wass, V., & Ashcroft, D. M. (2009). Prevalence, incidence and nature of prescribing errors in hospital inpatients. Drug safety, 32(5), 379-389.

[49] Lindgren, F., Geladi, P., & Wold, S. (1993). The kernel algorithm for PLS. Journal of Chemometrics, 7(1), 45-59.

[50] Lingsma, H. F., Bottle, A., Middleton, S., Kievit, J., Steyerberg, E. W., & Marang-van de Mheen, P. J. (2018). Evaluation of hospital outcomes: the relation between length-of-stay, readmission, and mortality in a large international administrative database. BMC health services research, 18(1), 116.

[51] Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. Annals of statistics, 42(2), 413.

[52] Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. International Journal of Forecasting, 9(4), 527-529.

[53] Malehi, A. S., Pourmotahari, F., & Angali, K. A. (2015). Statistical models for the analysis of skewed healthcare cost data: a simulation study. Health economics review, 5(1), 11.

[54] Manktelow, B. N., Evans, T. A., Draper, E. S., (2014). Differences in case-mix can influence the comparison of standardised mortality ratios even with optimal risk adjustment: an analysis of data from paediatric intensive care. BMJ Qual Saf 2014;23:7828.

[55] Manning, W. G.,& Mullahy, J. (2001). Estimating log models: to transform or not to transform? Journal of health economics, 20(4), 461-494.

[56] Marang-van de Mheen, P. J., & Shojania, K. G. (2014). Simpson's paradox: how performance measurement can fail even with perfect risk adjustment.

[57] McCue, T., Carruthers, E., Dawe, J., Liu, S., Robar, A., & Johnson, K. (2008). Evaluation of generalized linear model assumptions using randomization. Unpublished manuscript. Retrieved September 6, 2018, from www.mun.ca/biology.

[58] McCullagh, P., & Nelder, J.A. (1989). Generalized Linear Models, 2nd Edition. Chapman & Hall, London.

[59] McCune, J. S., Baker, K. S., Blough, D. K., Gamis, A., Bemer, M. J., KeltonRehkopf, M. C., Winter, L., & Barrett, J. S. (2013). Variation in prescribing patterns and therapeutic drug monitoring of intravenous busulfan in pediatric hematopoietic cell transplant recipients. The Journal of Clinical Pharmacology, 53(3), 264-275.

[60] Ministerie van Volksgezondheid, Welzijn en Sport (2018). Betaalbaar houden van medicijnen. Retrieved September 12, 2018, from www.rijksoverheid.nl.

[61] Ministerie van Volksgezondheid, Welzijn en Sport (2018). Schippers wil meer helderheid over kosten medicijnen. Retrieved September 12, 2018, from www.rijksoverheid.nl.

[62] Mosteller F., & Tukey J.W. (1968). Data analysis, including statistics. In Handbook of Social Psychology. Addison-Wesley, Reading, MA.

[63] Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal, 24(3), 69-71.

[64] Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. Philosophical Transactions of the Royal Society of London. Series A. 236(767), 333-380.

[65] Nederlandse Federatie van Universitair Medisch Centra (2018). Mission statement. Retrieved August 8, 2018, from www.nfu.nl.

[66] Nederlandse Vereniging van Ziekenhuizen (2018). Mission statement. Retrieved August 8, 2018, from www.nvz-ziekenhuizen.nl.

[67] Nederlandse Vereniging van Ziekenhuizen (2018). Brancherapport algemene ziekenhuizen. Editie 2018. Retrieved August 8, 2018, from www.nvz-ziekenhuizen.nl.

[68] Nederlandse Zorgautoriteiten (2018). Mission statement. Retrieved August 8, 2018, from www.nza.nl.

[69] Oostenbrink, J. B., & Al, M. J. (2005). The analysis of incomplete cost data due to dropout. Health economics, 14(8), 763-776.

[70] Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character, 187, 253-318.

[71] Porter, M. E. (2009). A strategy for health care reformtoward a value-based system. New England Journal of Medicine, 361(2), 109-112.

[72] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[73] Rawlins, M. D. (2016). Cost, Effectiveness, and Value: How to Judge?. Jama, 316(14), 1447-1448.

[74] Rosipal, R., & Trejo, L. J. (2001). Kernel partial least squares regression in reproducing kernel hilbert space. Journal of machine learning research, 2(Dec), 97-123.

[75] Schouten, J. A., Hulscher, M. E., Kullberg, B. J., Cox, A., Gyssens, I. C., van der Meer, J. W., & Grol, R. P. (2005). Understanding variation in quality of antibiotic use for community-acquired pneumonia: effect of patient, professional and hospital factors. Journal of Antimicrobial Chemotherapy, 56(3), 575-582.

[76] Seasholtz, M. B., & Kowalski, B. R. (1992). The effect of mean centering on prediction in multivariate calibration. Journal of Chemometrics, 6(2), 103-111.

[77] Segal, M. R. (2003). Machine learning benchmarks and random forest regression. Division of Biostatistics, University of California, San Francisco, CA 94143-0560.

[78] Seo, S. (2006). A review and comparison of methods for detecting outliers in univariate data sets. Doctoral dissertation, University of Pittsburgh.

[79] Shalit, I., Low, M., Levy, E., Chowers, M., Zimhony, O., Riesenberg, K., Bishara, J., & Raz, R. (2008). Antibiotic use in 26 departments of internal medicine in 6 general hospitals in Israel: variability and contributing factors. Journal of antimicrobial chemotherapy, 62(1), 196-204.

[80] Shams, I., Ajorlou, S., & Yang, K. (2015). A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or COPD. Health care management science, 18(1), 19-34.

[81] Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, Journal of Statistical Software, Vol. 39(5) 1-13.

[82] Smith, M. C., Ben-Shlomo, Y., Dieppe, P., Beswick, A. D., Adebajo, A. O., Wilkinson, J. M., & Blom, A. W. (2017). Rates of hip and knee joint replacement amongst different ethnic groups in England: an analysis of National Joint Registry data. Osteoarthritis and cartilage, 25(4), 448-454.

[83] Spearman, C. (1904). The Proof and Measurement of Association between Two Things. American Journal of Psychology. 15, 72101.

[84] Samenwerkende Topklinische opleidingsZiekenhuizen (2018). Mission statement. Retrieved August 27, 2018, from www.stz.nl.

[85] Tibshirani, R. (1996). Regression Shrinkage and Selection via the lasso. Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 26788.

[86] Tukey, J. W. (1977). Exploratory data analysis (Vol. 2). Tukey, J. W. (1977).

[87] Vektis (2018). Algemeen GegevensBeheer-code. Retrieved August 21, 2018, from www.vektis.nl.

[88] Venables, W. N. & Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth edition. Springer.

[89] Vereniging Innovatieve Geneesmiddelen (2017). Medicijn monitor. Editie 2017. Retrieved September 12, 2018, from www.verenigginnovatievegeneesmiddelen.nl.

[90] van der Voet, H. (1994). Comparing the predictive accuracy of models using a simple randomization test. Chemometrics and intelligent laboratory systems, 25(2), 313-323.

[91] Walker, H. M. (1940). Degrees of freedom. Journal of Educational Psychology, 31(4), 253.

[92] Wold, S., Ruhe, A., Wold, H., & Dunn, III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing, 5(3), 735-743.

[93] World Health Organization. (1996). Guidelines for ATC classification and DDD assignment. In Guidelines for ATC classification and DDD assignment. World Health Organization.

[94] Zhang, Y., Baicker, K., & Newhouse, J. P. (2010). Geographic variation in the quality of prescribing. New England Journal of Medicine, 363(21), 1985-1988.

[95] Zorginstituut Nederland (2018). Mission statement. Retrieved August 20, 2018, from www.zorginstituutnederland.nl.

[96] Zorginstituut Nederland (2018). Farmacotherapeutisch Kompas. Retrieved August 20, 2018, from www.farmacotherapeutischkompas.nl.