UNIVERSITY OF TWENTE.

performation
improving healthcare

Using statistical learning to predict the probability of a parallel sub-trajectory is getting wrongfully declared *at Performation*

A thesis presented for the master degree of Industrial Engineering and Management

J.H.A. Grimm
S1952420
16th of May 2023

Supervisory Committee:
Ella Sijbema, Business Area Manager Zorgregistratie, Performation
Milou Rossenaar, Consultant Zorgregistratie, Performation
Dr. Julia Mikhal, Assistant Professor, University of Twente
Prof.Dr.Ir. Erik Koffijberg, Full Professor, University of Twente

Faculty of Behavioural, Management and Social sciences (BMS)
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

Healthcare costs in the Netherlands are increasing every year, where especially administration and ICT costs are raising the total. For hospitals to be financially stable, all performed and delivered care must get reimbursed by healthcare insurers. Rightful care registration makes this possible, where the registration is rightful when it follows the regulation set up by the NZa (NL: Nederlandse Zorgautoriteit, Dutch Healthcare Authority). The regulations concern multiple challenges, one of these is the registration regarding parallel sub-trajectories within one specialty. Parallelism is defined as the registration of multiple care sub-trajectories within one specialty, where there is overlap in duration. Parallel care sub-trajectories are multiple active (filled with at least one care activity) care trajectories that will be declared at the healthcare insurer by the hospital within one specialty. This thesis focuses on assessing and predicting whether a parallel sub-trajectory is rightfully opened based on historical judgments. The reason for this research is to help hospitals to detect risky parallel sub-trajectories and correct them before they are billed to the healthcare insurers.

The main goal of this research is to give hospitals insight into their check on parallel sub-trajectories and to make a model that predicts the probability of a parallel sub-trajectory being wrongfully opened. The main research question that will be used to support this goal is:

How can statistical learning be used to predict the probability of a parallel sub-trajectory getting wrongfully declared?

The required data used to answer the above question is collected per hospital that participated. In total 13 hospitals gave permission to use their data in this research. Two methods are used by the hospitals that determine how parallelism should be checked, Horizontal Supervision and Self-Examination. For Horizontal Supervision 3070 parallel sub-trajectories are collected, for Self-Examination 2080. This data was transferred from the local databases per hospital and anonymized. The variables that are selected are the care profile classes (ZPKs) (NL: Zorgprofielklasses), with a focus on the first 8, as they are the biggest. Also, the overlap time between the parallel sub-trajectories is selected, as the care type, and the number of activities of the observed and parallel sub-trajectory. The models that are suitable for this research are decision trees, logistic regression, and random forest. The decision tree is easy to interpret, which makes them especially useful for cases that require transparency and explainability. For the decision tree, the disadvantage that is important to take into account is that this algorithm is prone to overfitting. This disadvantage is where the random forest comes into play. As this algorithm overcomes this by averaging the predictions of multiple trees. However, this algorithm is harder to interpret. Lastly, logistic regression is modeled, for this algorithm the advantages are that it is computationally efficient and provides interpretable results. The data is split into all specialties, ophthalmology, surgery, internal medicine, and remaining specialties.

In consultation with Performation, the model should perform at a desired level for an accuracy of at least 0.75, achieving the highest possible specificity. This is to keep the number of false positives as low as possible, but not classify too many cases wrong, so both the number of false positives and the number of false negatives should be low. The most suitable statistical learning method depends on the data that is modeled. For all specialties logistic regression performs best, for ophthalmology this is the decision tree, and for surgery, internal medicine, and the remaining specialties the random forest is best. The advice for Performation is to implement the ophthalmology decision tree, and for the other algorithm import more data to get a higher specificity. The outcome of the model returns a probability of a parallel sub-trajectory is opened correctly. The lower the probability, the higher the chance that this parallel sub-trajectory is wrongfully opened, and will be rejected during checking of sample observations. This probability can be used in the decision making which parallel sub-trajectories should be checked before the declaration.

The first limitation of this thesis is the available data. For the model, only the data from Self-Examination could be used. Within the data, there is an imbalance between the distribution of positive and negative cases in the data. In our data the number of negative cases is much smaller than the number of positive cases, this can lead to the classifier being biased to the positive class. Therefore, it is important to balance the dataset in future research, so leaving out positive cases. The second limitation is the low AUC of the ROC, which indicated that the model's predictive power is not much better than random chance. The model can still be useful, as it is used to indicate parallel sub-trajectory that have a higher risk of being wrongfully opened, so rather as a decision-making tool which sub-trajectories to check, than to predict which sub-trajectories are wrongfully opened. The third limitation of this thesis is that the diagnosis is not taken into account. In total more than 700 different diagnoses are present in the collected data. The suggestion is to group the diagnoses, examples of this are the difference between general and specific diagnoses, or to group diagnoses that are related to each other.

However, taken the limitations into account, this thesis is a start in how statistical learning can be used to predict the probability of parallel sub-trajectory getting wrongfully declared. This research already contributes in detect risks per hospital and the model give an first indication of the probability. The availability of new data is already taken into account. This research has a big potential to help lower the workload and finincial impact of wrongfully opened parallel sub-trajectories.

Table of Contents

1 Introduction

Healthcare costs in the Netherlands are increasing every year, however, half of the costs are not patient-related. Administration and ICT costs are raising the costs (Gelder, 2022). In the Annual Budget for 2023 (NL: Miljoenennota), the gross healthcare expenditure is set at 94.9 billion euros, which is an increase of 9.5% compared to 2022 (Koenraadt, 2022). A part of these costs come from incorrect care registrations. Wrongful registrations can occur because there is more care registered than allowed, or not all delivered care is registered. Hospitals have to check on whether they have registered following the rules set up by the NZa (NL: Nederlandse Zorgautoriteit, Dutch Healthcare Authority) and the insurance companies. On the other hand, hospitals can also register not all the care that is performed, no payment will follow as it is the hospitals' responsibility that all care is declared. These wrongful registrations can be unintentional, after all the registration of care is human work, and humans can make mistakes since regulations are not always clear for everyone. Since 2005 the care administration in the Netherlands changed from activity-based to diagnosis/treatment based. The main reason is to make the financing not depend on volume to reduce overtreatment. Since then the DBC is introduced, which is the abbreviation in Dutch for *Diagnose-Behandelcombinatie*, which is a combination of the diagnosis and the treatment associated with this diagnosis. With this new system, care is no longer billed per activity, but as a package. How much the package will cost is an agreement between healthcare organizations and healthcare insurance companies. This market is focused on the purchase of healthcare, where regulating and controlling organizations keep track of this market. The healthcare insurance companies are private, and therefore competition arises for the prizes and the number of insured people. The patients (or just insured people) can choose which insurer they want to have. The biggest three health insurance companies in the Netherlands are Achmea, VGZ, and CZ, which together occupy more than 70% of the market (Zorgwijzer, n.d.). Between the patient and the healthcare organization stand the care consumption market, here the patient can choose where he/she receives the actual care. This all makes that within healthcare in the Netherlands, three markets work together to provide, purchase, and finance care.

To be financially stable as a healthcare organization, rightful administration is important. The term rightful administration is rather vague and should be explained. Rightful administration holds that all the delivered care is registered following the regulations. Where a hospital's role is to help sick patients, the administration is what keeps the hospital financially stable, this means that hospitals need rightful administration to receive reimbursement from the health insurers. Care administration is important because it helps manage the daily operations and processes of hospitals. This includes capturing, processing, and analyzing data and information about patients, medical treatments, and financial transactions for example. Where care administration focuses on the financial and administrative aspects, care registration focuses on logging information about the care provided. Care registration helps patients to receive appropriate treatment, as in the information systems the medical history of the patients can be found. With the medical history, the healthcare provider can adjust the treatments to what fits best for the patient. Care administration also helps to improve the quality of the care, this is because with adequate registration it is easier to identify where improvement is needed. Care registrations are also necessary to receive money from health insurance companies. On one hand, it provides insight into the financial information of a hospital. On the other hand, it uses that information to make forecasts and employee planning. Performation helps healthcare organizations to improve their care registrations with their tools and consultancy services. Hospitals have to show whether the declarations made are permitted and legitimate to the healthcare insurance companies. Right now the shift is going from self-examination (NL: Zelfonderzoek) to horizontal supervision (NL: Horizontaal Toezicht). This new program was introduced in 2014, with as goal to improve the quality of care and lower the administrative burden (Nagtegaal, 2018). The ambition is that all hospitals have switched to horizontal supervision in 2025. Both methods have the goal to ensure rightful care administration. The checking method is the same, samples are drawn from all observations. However, the frequency differs. Where self-examination is to detect the declared mistakes and pay back the insurers, horizontal supervision is more focused on detecting potential mistakes. After the first quartile of 2023, 58 out of the 75 hospitals have already switched, which is equal to 77% (Horizontaal Toezicht Zorg, 2023). With self-examination, hospitals have to check once a year (which must be done in the first quartile) based on samples of their care registrations and prove to the health insurers that they have rightfully registered care and no fraud was committed. In comparison with horizontal supervision, self-examination is focused on the hospitals themselves taking responsibility for examining and checking their expense claims and administration, this process takes up a lot of time and resources. For both self-examinations and horizontal supervision, a list of sample tests is set by the Nza. One example of a sample test is the check on rightful parallelism. Parallel care sub-trajectories are multiple active (filled with at least one care activity) care trajectories that will be declared at the healthcare insurer by the hospital within one specialty where there is an overlap in duration. The last part is concerned with wrongfully declared. Hospitals have to declare care in terms of DBCs (explained in Section 4.1), when multiple care trajectories are declared within one specialty there are regulations to check whether this is correct or not. Both sub-trajectories need to have their own diagnosis (and establishment of that diagnosis) and their own treatment. This thesis will focus on assessing and predicting whether a parallel sub-trajectory is rightfully opened. Within one specialty, there are two different types of parallelism. When a patient comes to a specialty for the first time, an initial sub-trajectory is opened with care type 11. If the sub-trajectory is closed because of maximum duration, a

follow-up sub-trajectory with care type 21 is opened subsequently. The two types of parallelism depend on the care type, as mentioned above, either type 11 or type 21. More on parallelism can be found in Section 3.1. Out of all parallel cases a sample test has to be done. For each sample observation, it has to be checked, documented, and justified whether the parallel sub-trajectory was opened correctly. From the rejected parallel sub-trajectories an error margin follows. This error margin is extrapolated over all parallel sub-trajectories. Hospitals pay back this error margin to the healthcare insurers based on the proportion of parallel sub-trajectories of patients insured there.
The given assignment is to investigate if it is possible to predict the probability of a parallel sub-trajectory getting wrongfully declared. When a sub-trajectory has an increased probability of getting wrongfully declared, hospitals can check these sub-trajectories before the declaration is done. This will lead to a lower error margin, so lower total costs to be paid back.

1.1 Reason for this research

The care control department of the hospitals has to check whether the parallel sub-trajectories are rightful or not. Parallelism is hard to quantify because the regulations do not give a clear case description of when it is right or when it is wrong. Every parallel case is unique and has therefore been checked individually by hand. The more errors found in the sample, the higher the extrapolation costs will be. This research will help hospitals detect parallel sub-trajectories with an increased risk of being wrong. From the available data, it is known if a parallel observation is marked as approved or rejected by the control department. This data is used to make a model that predicts the probability of wrongly opened parallelism. The available data are both from horizontal supervision and self-examination. For the prediction models, only the self-examination data is used, as here also the data of the parallel sub-trajectory is available, whereas for horizontal supervision this is not the case.

1.2 Problem definition

The main goal of this research is to present a prediction model that can help to detect parallel care trajectories at risk of getting wrongfully declared, using dashboards and statistical interactions. To translate this to a scientific research question this formulated as followed:

How can statistical learning be used to predict the probability of a parallel sub-trajectory getting wrongfully declared?

Let's parse this sentence. Statistical learning is the main focus of this research. Statistical learning refers to a set of tools for making sense of complex data sets (James et al., 2021). The second part is to predict the probability, this is done based on historical data where parallel sub-trajectories have been judged as approved or rejected. The probability will depend on different variables, on which elaboration will follow later on. The next part is "a parallel sub-trajectory", as explained before, a parallel sub-trajectory is a sub-trajectory that is opened when there is already an active sub-trajectory for that specialty. The last part is about getting wrongfully declared, this holds that within the sample test, the parallel sub-trajectory is judged as unjustified.

To answer the research question, multiple sub-questions are formulated:

1. *What are the current challenges for registering parallel sub-trajectories?*
   The answer to this question provides a layout of the current challenges that are faced regarding registering parallel sub-trajectories. Besides that, we also look into how these registrations have been checked and justified by the hospitals. Therefore, a literature review on what regulations have to be taken into account.
2. *How can we visualize parallel sub-trajectories that have been judged in the sample tests?*
   This gives an insight into the current situation regarding where parallelism occurs. To get this insight is gained by making two dashboards. One will be made for horizontal supervision and the other one for self-examination data. These dashboards can be used by the hospitals for benchmarking to compare how their hospital is doing, in comparison with other hospitals that took part in this research. The dashboards are also the start of statistical testing to see where problems or irregularities occur.
3. *What literature is available on using statistical learning to predict a binary outcome?*
   Here literature will be collected for a theoretical framework for statistical learning. Within this thesis, we are interested in how statistical learning can be applied to predict a binary outcome, in this case, whether a parallel sub-trajectory is getting rejected or approved.
4. *What models are suitable to use in this research?*
   Here the translation from the theoretical framework to the application for this research is done. The different models will be compared on their performance, but also how they can be implemented within Performation.
5. *Which quantitative variables can be used for predicting the outcome?*
   A lot of data is available in the registration system, however, not everything can be used. For this, the variables will be investigated which can be used for both the observed sub-trajectory and for the sub-trajectory that is parallel to the observed one.

6. *What are the results of the models?*
   The answer to this question gives an overview of the different models, here the similarities as well as the differences will be discussed.
7. *In which way can the model help improve care registrations?*
   The question concerns what the implementation is from the outcome of this research. Both for how this contributes to Performation, but also how hospitals benefit from this.

1.3 Structure

In the next chapter, the problem background will be discussed. Here information regarding Dutch healthcare, care trajectories, checks on parallelism, problem description, Performation, and the stakeholders can be found. This chapter answers the first sub-question. In the third chapter, the dashboard that are made during this research to answer the second sub-questions is reviewed. In Chapter 4, the available literature about the regulations and statistical learning will be reviewed to answer the third research question. In Chapter 5 the methods used during this research will be discussed, here sub-question 4 and 5 will be answered. Chapter 6 will answer sub-question 6, and describe the results of the different models. In the conclusion, Chapter 7, an answer will be provided for the last sub-question. In this chapter, the limitations, the main conclusion, and future research will be discussed.

2 Problem background

In this chapter, the background of the problem will be described. We will first explain how the Dutch healthcare system works, with a focus on care administration. Here Performation comes into place, as they help hospitals to get into control of their care administration with its tool Notiz. Also, other stakeholders will be described.

2.1 Dutch healthcare system

The Dutch health insurance system is based on market forces. Within this system, Dutch citizens are obligated to take out basic insurance, while healthcare insurers compete with each other about the prices and volumes of different healthcare providers (Koen Kuijper, 2022). As said before, this system consists of three primary parties. The first is the patient, or the insured to say. The second is the care provider: the hospital, general practitioner, and multiple health institutes. The last one is the health insurers: privately-owned companies that handle the payment of the delivered care. Between these parties, different collaboration is in place. Between the patient and the care providers stands the healthcare market. This means all the care that is available to the patient when in need. However, this is an ideal view of reality. Nowadays, patients have to be alert that the care they receive at a certain care provider is also reimbursed by the health care insurer they are insured at. Between the care provider and the health insurer, the purchasing of healthcare is done. The link between the patient and the health insurer appears in the insurance market with its insurance policies.
Until 2005 every care action was separately registered and billed. This meant that the patient, or the health insurer on behalf of, had to pay per care activity. In 2005 the DBC system was introduced. The DBCs are a combination between the diagnosis and the corresponding treatment. With this new system the patient, or the health insurer on behalf of, pays an amount for all the received care. A DBC is a package containing information regarding the diagnosis and treatment the patient receives. Every DBC has its price tag (Performation, n.d.-a). The price tag of a DBC is established as follows. The care provider registers the diagnosis and the care actions in a sub-trajectory. To go from a sub-trajectory to a billable DBC the grouper is needed. The grouper is a software program that translates a sub-trajectory into a standardized diagnosis-treatment combination. The care activities in the sub-trajectory, together with the diagnosis determine the price tag of the DBC. After the sub-trajectory is transformed into a DBC, the healthcare provider can claim the expenses from the healthcare insurance companies of the patient. (Performation, n.d.-a).

2.2 Care trajectories

A patient comes to the hospital with a care question, then a new care path is opened. This must be done by the professional who performs the gate function and follows the registration rules of the NZa. It is not allowed to open a parallel care path or a care sub-path within the same specialty when no new or own care question is in place. Here the difference is made that a new care question also yields a new diagnosis and anamnesis, which leads to a different policy of treatment concerning the care questions. An example of a common right parallelism is within ophthalmology. If a patient gets a cataract in one eye, a new sub-trajectory is opened and the treatment is started. When later also in the other eye cataract gets diagnosed, it is allowed to open a second sub-trajectory, as the care question is new. However, the treatment must be based on the two eyes separately, and not combined. Here the parallelism is justified because the diagnosis is two-sided. To also illustrate the wrong parallelism, when a patient is already under treatment for diagnosis X, however, a new care question Y rises, and therefore a new care trajectory is opened to perform a test to set a diagnosis. It turns out that the diagnosis belonging to care question Y is a complication of the treatment for diagnosis X. In this case the parallelism is wrong, and no new care trajectory should have been opened, and the care activities of the complication have to be registered in the trajectory of X. The core problem within this system is that only qualitative checking can with certainty tell whether the parallelism is correct or not. There are multiple reasons why a parallel trajectory is opened. Of course, when the rules allow it to be a parallel trajectory (for example when the diagnosis is two-sided) a new diagnosis with treatment, or an action is on the list of exceptions. But it can also be opened because it is not yet sure if it is a new care question or not. If the diagnosis and treatment do differ from the original trajectory later in the process, the parallel trajectory was rightfully opened. But when the new care question led to the same diagnosis as the patient is treated for in the original care question, the parallel trajectory is wrongly opened, so the care activities have to be moved to the original trajectory The care profile classes will be noted as “ZPK1 to ZPK8”, as ZPK is “zorgprofielklasse”, which is a care profile class in Dutch. ZPK1 refers to outpatient and emergency room visits. ZPK2 concerns activities within a nursing day. The third ZPK are activities within the clinic. ZPK4 concerns diagnostic activities. ZPK5 are operative operations. Activities with ZPK6 are classified as other therapeutic activities. ZPK7 are activities belonging to imaging diagnostics. And the last, ZPK8, is clinical chemistry and hematology (Performation, n.d.-a). Another case that can happen, is that there are two different care questions with their separate treatments within separate specialties. However, in this thesis, we focus on parallelism within one specialty, because parallelism between different specialties happens in specific cases and is therefore not as big of a ‘black box’ as within one specialty.

2.3 Checks on parallelism

Per patient, multiple DBCs can be opened, following the rules handed out by the NZa more on the rules can be found in Section 3.1. The rules are overarching for all the parallel trajectories, this makes it not easy to check automatically whether a parallel care registration is rightful or not. The regulations should provide enough guidance to assess whether the parallelism is correct or incorrect. However, this can't be done automatically. The parallel cases have to be evaluated and judged manually.

The hospital has to prove to the health insurers that care was lawfully declared. A way to do this is self-examination. Here samples are taken annually from all the care registration. There are two sample tests for parallelism as said before. Within the self-examination, the sample size is 125 sub-trajectories for both, so 250 sub-trajectories every year are checked for rightful parallelism. The sample size is set by the NZa.

2.4 Problem description

The core problem is that the error margin of the parallel trajectories being incorrect is higher than desired. The biggest cause of this is that in most cases it is not possible to automatically check whether the parallelism is correct. As stated before, the regulations are overarching and not specific to individual cases. The error in wrongful parallelism can be found in multiple causes. The first is that the diagnosis belonging to the observed sub-trajectory is incorrect. This means that the diagnosis of the observed sub-trajectory arises from the parallel (previously opened) sub-trajectory. Another cause for the error is that there are two different diagnoses, however, the different diagnoses do not have different treatments. For example, the second diagnosis cannot be treated, therefore opening a new care trajectory is wrong. Another wrongful parallel trajectory lies in documentation. When in the reports, a clear diagnosis and treatment are lacking, then it is not right to open a second trajectory. The above-mentioned is why parallelism can be wrong, but not how the mistake was made. At first, when a patient says he has a new care question, it is easier to open a new trajectory and start reporting than to wait to see whether it is a different care question. The NVZ, ZN, NFU, and FMS (abbreviations can be found in the Appendix) have made a registration flow to help gatekeepers see if it is justified to open a new parallel sub-trajectory. Here a gatekeeper should ask him/herself multiple questions to find out whether it is allowed to open a new DBC. However, here still the biggest 'if' is whether there is a new care question. The core problem will not concentrate on decreasing the number of wrongfully opened parallel DBCs, but on detecting parallel care trajectories at risk of being wrongly declared. This will support hospitals in correcting potential mistakes, thereby reducing their error margin.

2.5 Performation

Performation supports hospitals in reaching a higher efficiency rate, together with tools, data, and advice. They have multiple focus areas, which will be shortly discussed here. The first area is financial, here the focus is on identifying improvement options for calculating the cost of care within the care organization. The second area is capacity management, this area helps to optimize their capacity throughout the care organization. The third area is data analytics which helps care organizations get valuable information from their data, and this can be used as steering information. The last area is care registration, where the research was performed. Here the focus is on the right registration of the delivered care. This is a time-consuming task, which Performation provides consultancy and tooling for. For this, the tool Notiz is used (Performation, n.d.-b). Every day a lot of care is registered within hospitals. This registration can go wrong and leads to possible wrongful billing. On the other hand, sometimes care is delivered but not (completely) registered, this is also known as under-registration. Within Notiz the Daily Audit tool detects this. Every night the tool checks, through various checks on the registration, what goes wrong and right in the form of work lists. Based on these work lists, users (health providers) can make changes in electronic patient records.

2.6 Stakeholders

Other stakeholders that are involved are both the participating hospitals, as well as the hospitals that do not participate but use Notiz. More specific information about the participating hospitals can be found in section 4.2. Their data from the self-examination is used (with permission and randomized). For these hospitals, a dashboard is made during this research, which gives insight into how they score based on the self-examination of parallel trajectory checking. For all the hospitals using Notiz, so also the hospitals that did not participate in this study, new control checks will be available based on the outcome of the research, which they can use to detect possible wrong parallel registrations. Medical staff and patients are not considered stakeholders, as their input is not taken into account, nor will the outcome change something directly for them.

3 Current situation

In this chapter, data analysis of the available samples test is elaborated on. After the data was collected, two dashboards were made to visualize the current situation. These dashboards were made in Power BI, and the goal is two-sided. On one hand, the dashboards are used as a starting point for statistical learning, to provide better insight into which variables might be relevant. On the other hand, these dashboards give hospitals an insight into how they are performing based on rightful parallel registration, and how they are performing in comparison with other participating hospitals. First, the dashboard for the participating hospitals that have implemented horizontal supervision be discussed. After that, the dashboard for self-examination will be reviewed.

3.1 Horizontal supervision observations

From the sample tests of horizontal supervision in total 3070 observations were collected from 6 different hospitals. Within these observations, surgery (CHI), internal medicine (INT), and ophthalmology (OOG) are most present. The number of observations per specialty is equal to 577, 553, and 551 observations respectively. However, as can be seen in figure 1, the difference between the hospitals concerning the dominance of these three specialties differ.

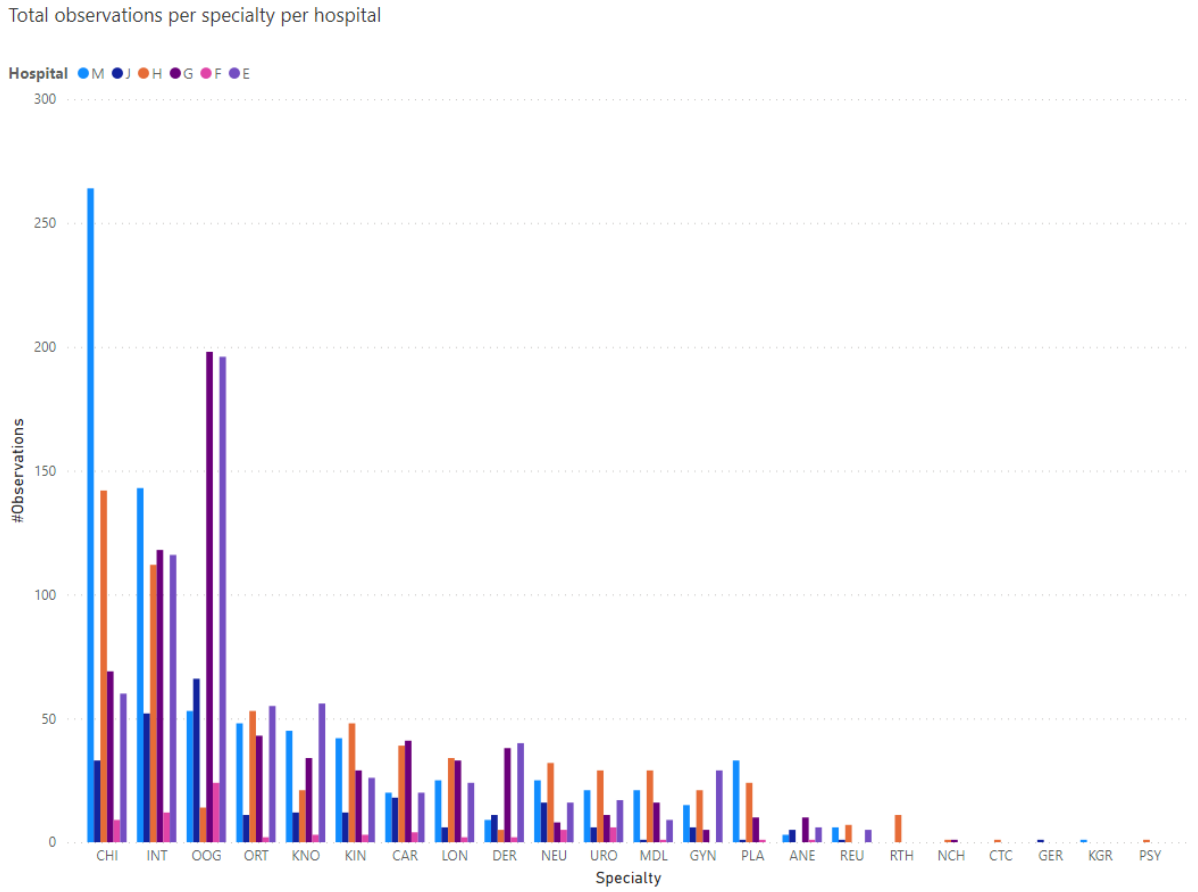

Figure 1: Overview of all observations per specialty per hospital for horizontal supervision (n = 3070), surgery (CHI) has the most observations, followed by internal medicine (INT) and ophthalmology (OOG)

To give an insight into how the proportion per hospital is, figure 2 can be best used. In this figure all the specialties are displayed, in proportion to the total number of observations per hospital. First, we look at surgery (CHI). Hospital E has the least observations, where this is 8.9%, which is also equal to 60 observations. In comparison with that, hospital M has the most observations for surgery with 34.1%, which is equal to 264 observations. Looking at internal medicine (INT) the proportions differ from 16.0% at Hospital F to 20.2% at Hospital J. For ophthalmology (OOG), the differences between the hospitals are large. This specialty is the least present in the sample tests of hospital H with 2.2% (14 observations), and the highest proportion can be found at hospital F with 32.0% (24 observations). This difference in percentage in comparison with the difference in the number of observations is not proportionally. This is due to that Hospital F has carried out fewer sample tests than the other hospitals. To put that into numbers, hospital F judged 75 observations, whereas Hospital H judged 624. The average number of judged observations per hospital is 512.

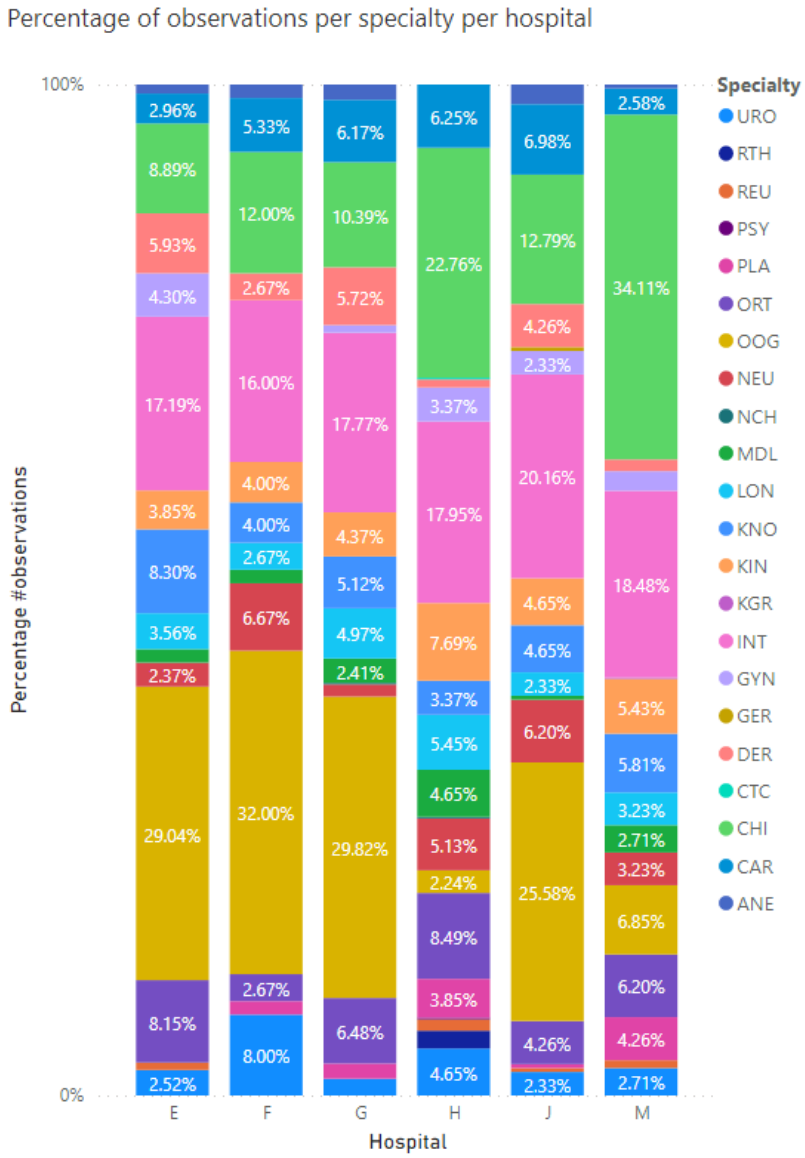

Figure 2: Percentage per specialty per hospital for horizontal supervision over all observations, Hospitals E, F, G, and J have more or less the same division, where Hospitals H and M both have little observations for ophthalmology (OOG) and in comparison more for surgery (CHI)

If we look at Figure 3, we see the number of approved and rejected observations per specialty. If we focus on the rejected number of observations, the same three specialties have the most observations.

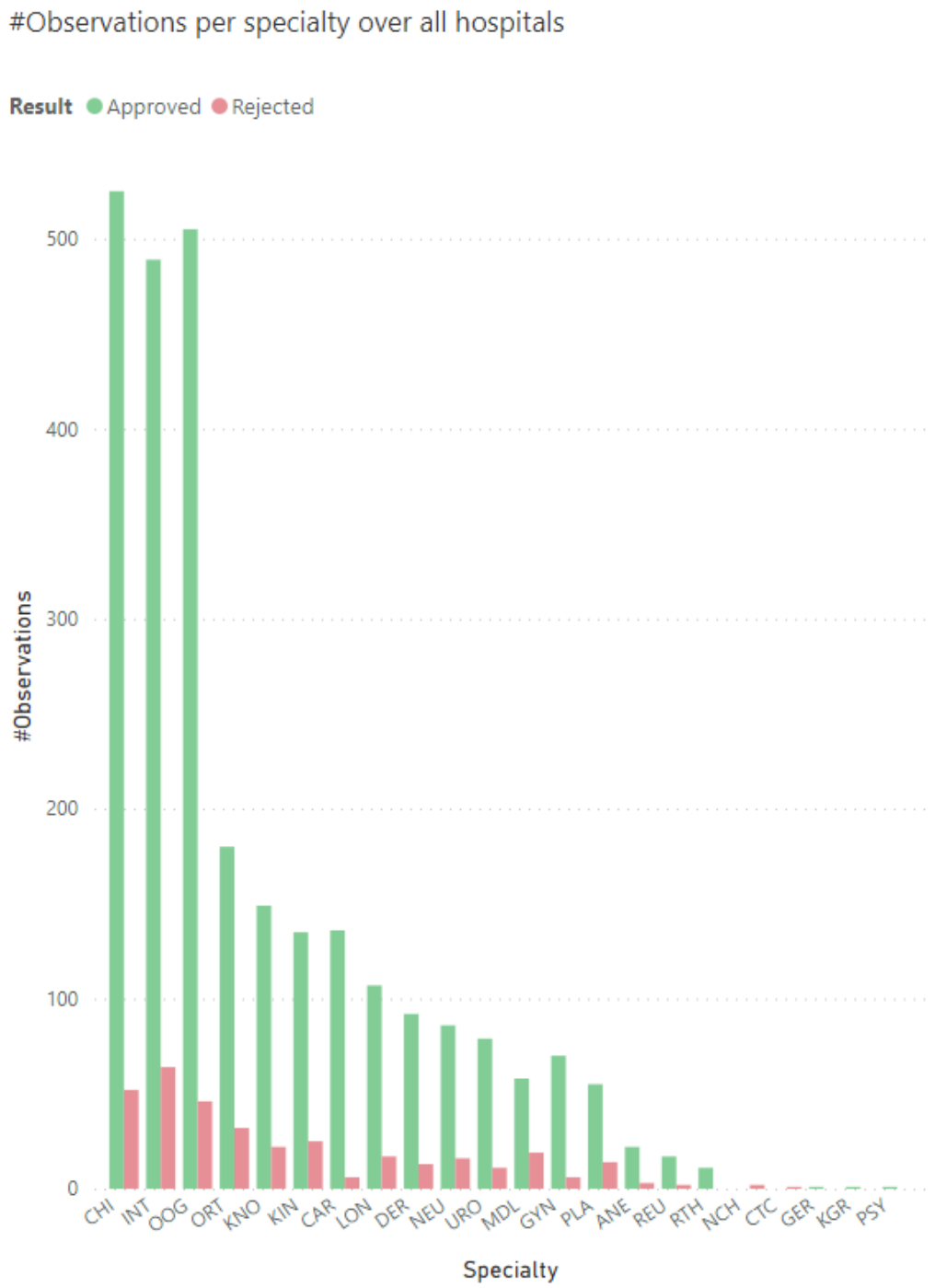

Figure 3: Number of observations both approved and rejected per specialty over all hospitals (n = 3070), where surgery (CHI) has the most total observations, the most rejected observations are for internal medicine (INT)

In Figure 4 the percentage of approved and rejected sub-trajectories is given. This figure is sorted with the most observations per specialty on the left and the least on the right. Out of all 3070 observations, 351 observations were judged as rejected, which is equal to 11.4%. The biggest error margin (number of rejected observations divided by the total number of observations) is for gastroenterology (MDL), which is 24.6% (19 rejected observations).

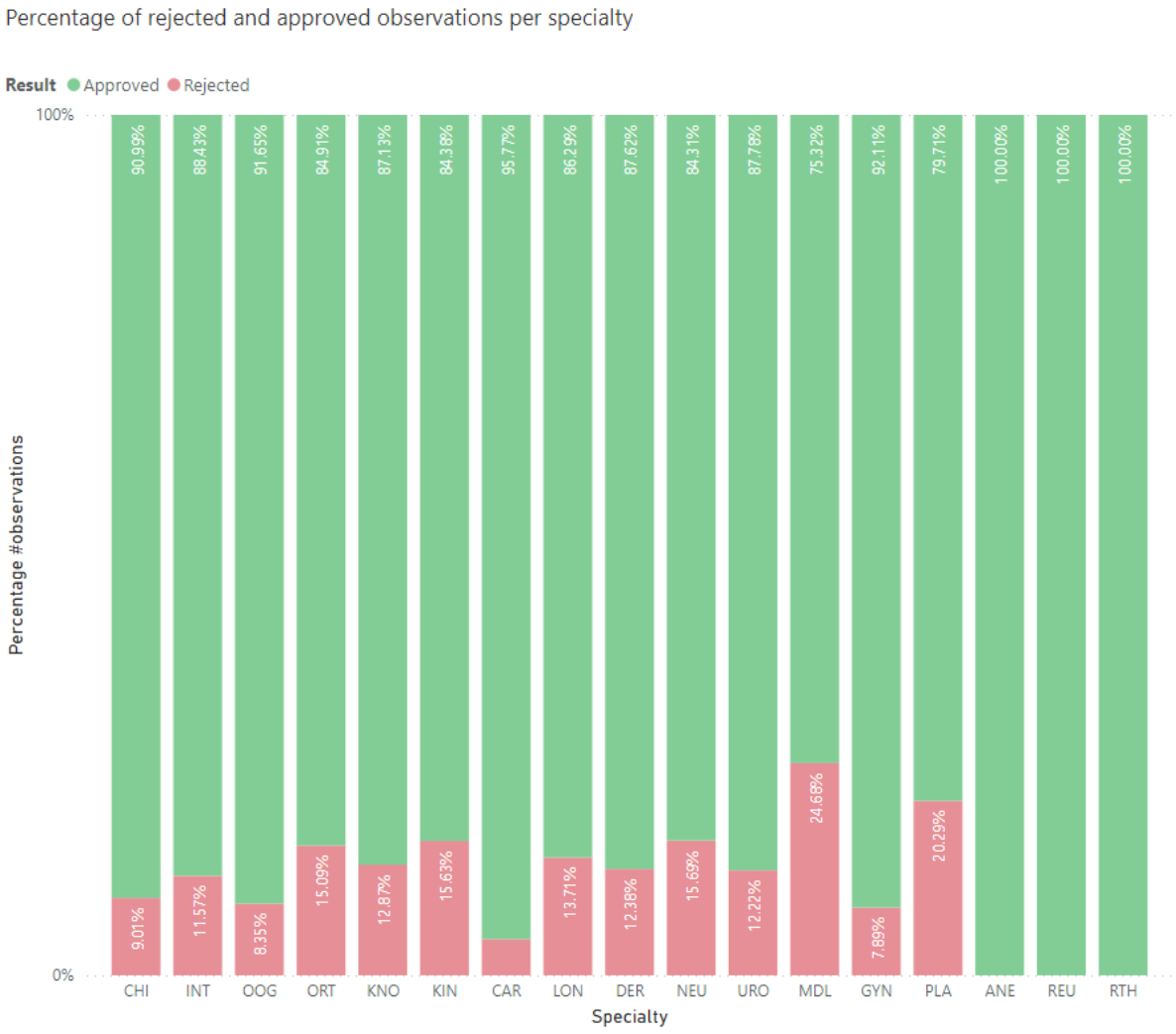

Figure 4: Percentage of approved and rejected observations per specialty over all hospitals (n = 3070), the highest error margin is encountered for gastroenterology (MDL)

When looking at the difference between the hospitals in Figure 5, the biggest error margin is for hospital F, which is 22.6% (17 rejected observations). Hospital F has just started with horizontal supervision, where to goal is to improve the quality of care registrations, so therefore this margin can be high as not yet much improvement could have been made. It can also be that the doctors at Hospital F register more parallel sub-trajectories, if this is not correct, then the hospital should invest in better training for the doctors. It can also be that the hospital just had bad luck with their sample. Hospital E has the lowest error margin with 3.7% (25 rejected observations). Why this error margin is so much lower than that of hospital F is unknown. Again luck, training, or the total number of parallel sub-trajectories could have played a role.

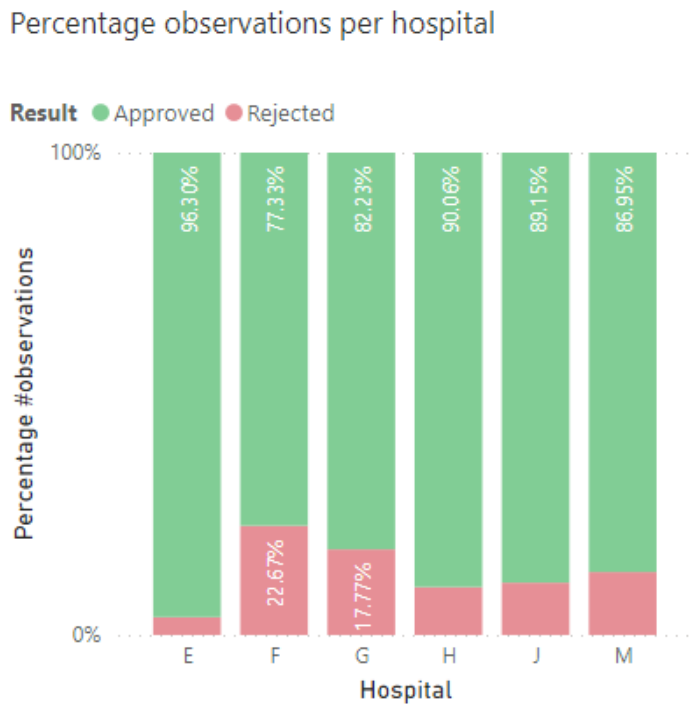

Figure 5: Percentage of approved and rejected observations over all specialties per hospital, hospital E has the lowest error margin, and hospital F has the highest

For horizontal supervision, we lastly look at the care type so between care types 11 and 21. Out of the 3070 observations, 1467 observations were care type 11, which is equal to 47.8%. Hospital F only judged observations with care type 21, as we can see in Figure 6. Hospitals are allowed to choose whether they judge the same amount of care types 11 and 21, or have different sample sizes based on risk.

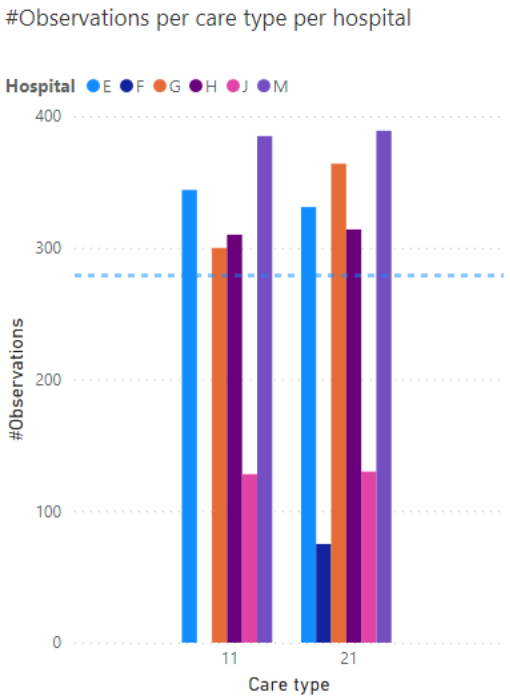

Figure 6: Number of observations per care type per hospital, most observations are of the care type 21

Figure 7 gives the comparison between approved and rejected observations per care type per hospital. For every hospital, the error margin differs per care type. On average, the error margin for sub-trajectories with care type 11 is 10.6%, and for care

type 21 this is 12.2%. For hospitals J and M the difference in error margin between care types 11 and 21 is 14.1% and 13.1% respectively. However, interesting to point out, for hospital J care type 11 contains more errors, whereas, for hospital M, this is care type 21.

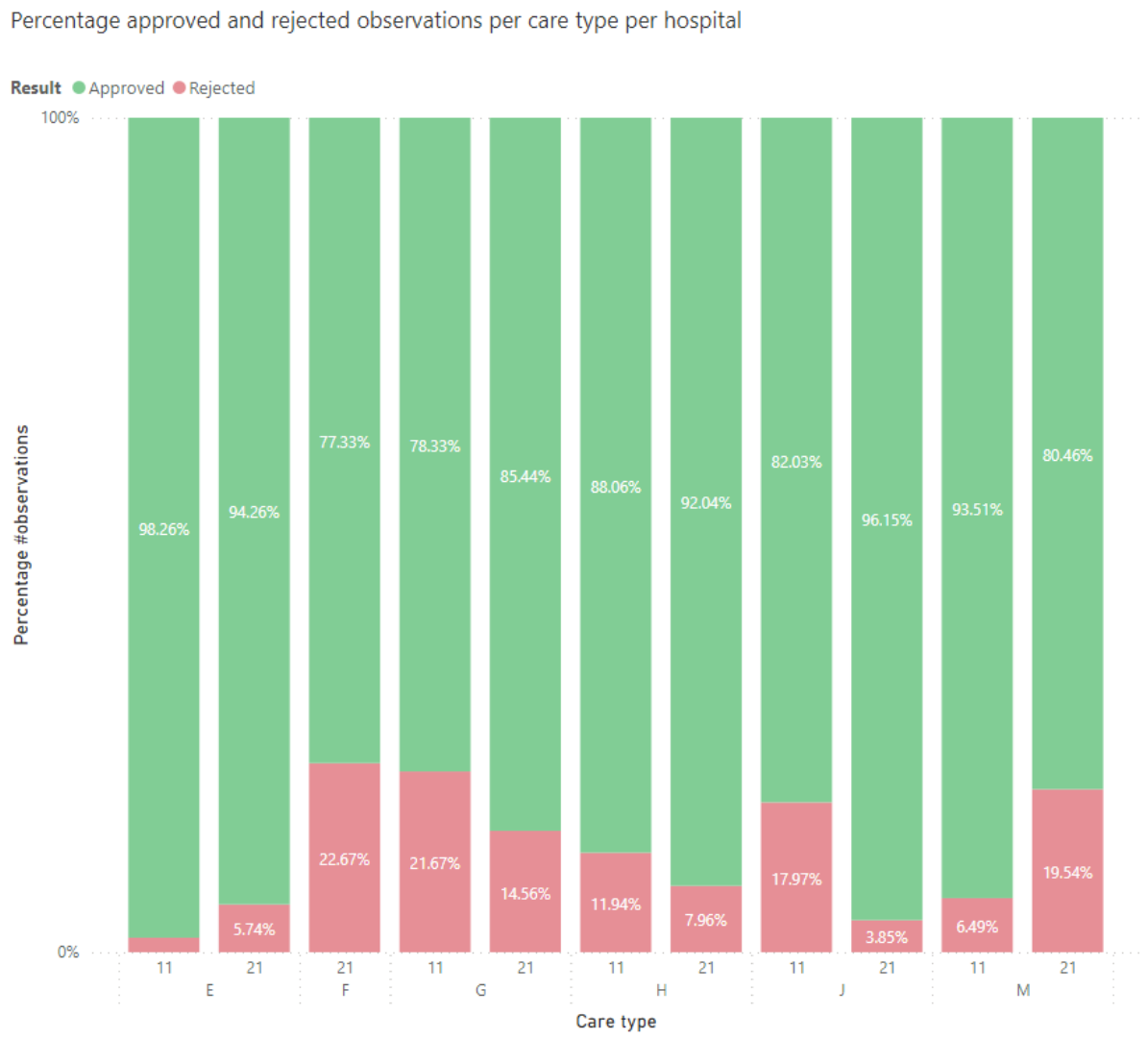

Figure 7: Percentage of approved and rejected observations per care type per hospital over all specialties for horizontal supervision, Hospitals G and J should focus on improving parallel registrations with care type 11, whereas Hospital M should improve the parallel registrations with type 21

3.2 Self-examination observations

For self-examination more information regarding the observed sub-trajectory as the parallel sub-trajectory was available. In total 2080 observations were judged at 8 different hospitals. The biggest specialty among the observations is ophthalmology, with in total of 596 observations, which is equal to 28.7%. This high number can be explained that Hospital B is specialized in ophthalmology (OOG), and therefore does not have any other specialties. In figure 8 the number of observations per specialty over all hospitals can be found. After ophthalmology (OOG), internal medicine (INT), and surgery (CHI) have the most observations in the samples.

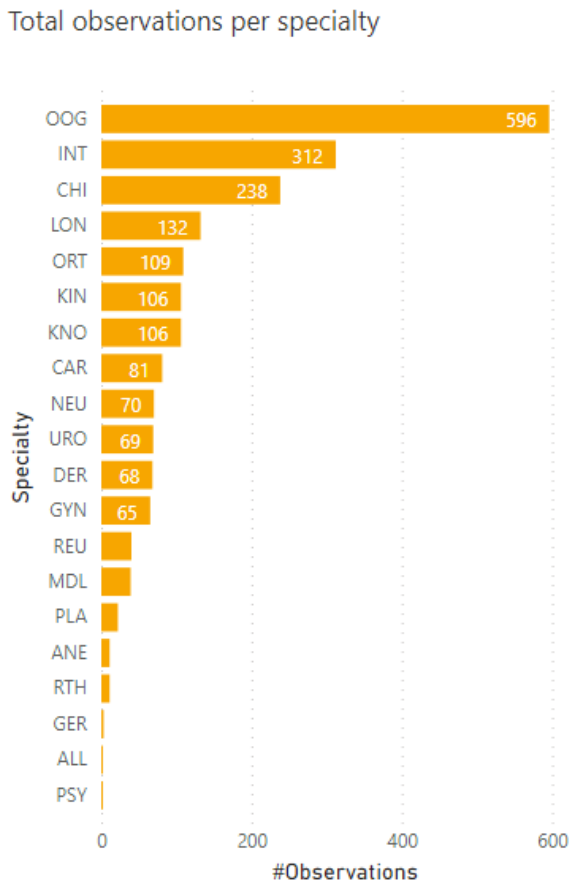

Figure 8: Number of observations per specialty over all hospitals for self-examination, the three biggest specialties are ophthalmology (OOG), internal medicine (INT), and surgery (CHI)

In figure 9, the proportion per hospital per specialty is given. When reviewing this graph, we leave hospital B out of consideration, as this hospital only has one specialty. Looking at the biggest specialty, ophthalmology (OOG), it is important to say that hospital A does not have this specialty in their hospital, and therefore no observations are found in their samples. Then comparing hospitals C, D, F, K, L, and N, the proportion differs from 14.2% (32 observations) at hospital F to 35.8% (81 observations). If we look at internal medicine, the proportion differs from 9.7% (20 observations) at Hospital K to 22.6% (115 observations) at Hospital C. Lastly looking at surgery, hospital N has the least percentage of observations for this specialty, namely 6.6% (15 observations), and hospital A has the most with 24.8% (59 observations).

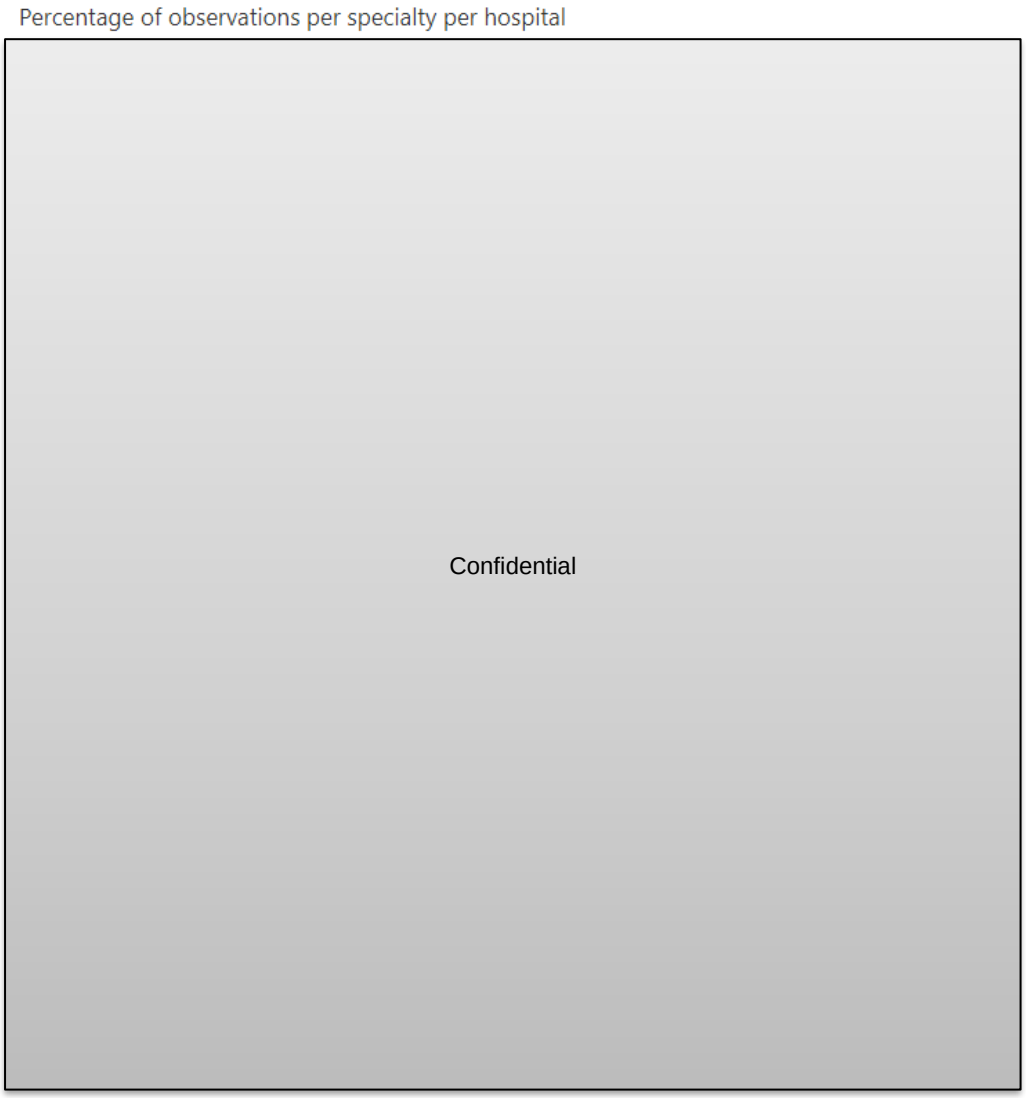

Figure 9: Percentage per specialty per hospital for self-examination overall observations

Figure 10 gives the number of approved and rejected observations per specialty. Here we also see that when we focus just on the rejected observations, the same three specialties are the biggest.

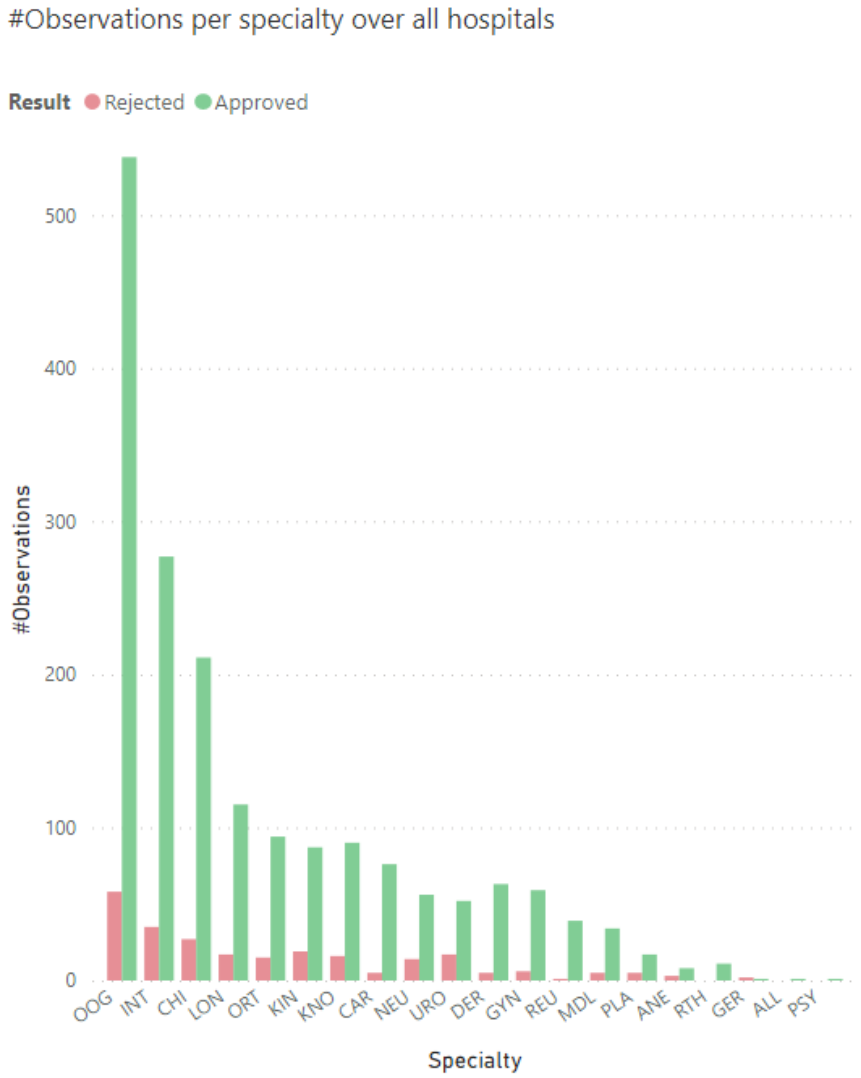

Figure 10: Number of observations both approved and rejected per specialty over all hospitals from self-examination (n = 2080), the biggest number of rejected observations are for ophthalmology

The average error margin over all hospitals and specialties is 12.0%. In figure 11 the error margins per specialty can be found. Again the x-axis is sorted with the most observations on the left and the least on the right. We first focus on the three specialties that have an error margin of 0, these specialties are radiotherapy (RTH), allergology (ALL), and psychiatry (PSY). These specialties have 11, 1, and 1 observations. On the contrary, the highest error margin is 66.7% for gastroenterology (MDL), but this is equal to 2 out of the total 3 observations that were rejected. If we look at the specialties that have at least 50 observations, 12 remain. Out of these 12, we see that the highest error margin is encountered for urology (URO), with 24.6% (17 rejected observations).

Figure 11: Percentage of approved and rejected observations per specialty over all hospitals (n = 2080)

In Figure 12 the error margin per hospital is found. The lowest error margin is for hospital K with 0.5%. This hospital only rejected 1 out of the 206 observations. Why this error margin is so low is unknown. The same possible explanations for the differences in horizontal supervision hold. Hospital N has the highest error margin with 22.1%, where 50 out of the 226 observations were judged as rejected. For this hospital, the attention points lie in the urology (URO) specialty, as here 9 out of the 16 observations were rejected.

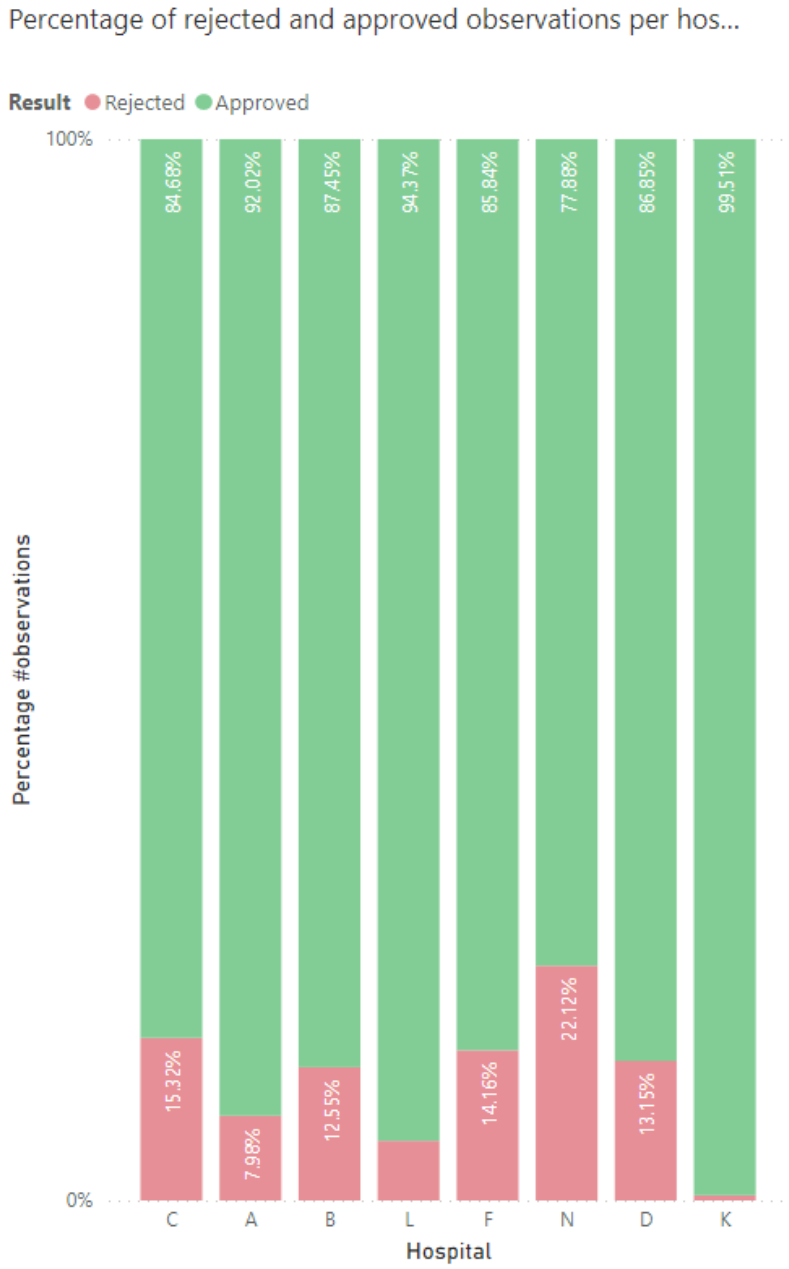

Figure 12: Error margin per hospital for self-examination that differ from 0.49% to 22.12%

In Figure 13 we compare the overlap time between the observed and parallel sub-trajectory. The overlap time is the time in days between the end date of the parallel sub-trajectory and the beginning date of the observed sub-trajectory. Over the hospitals there is no difference encountered between the approved and rejected sub-trajectories, however, when we look at the differences per hospital, we do see that the averages differ. The biggest difference is encountered at Hospital K. However, the average of the rejected sub-trajectory is only based on 1 observation.

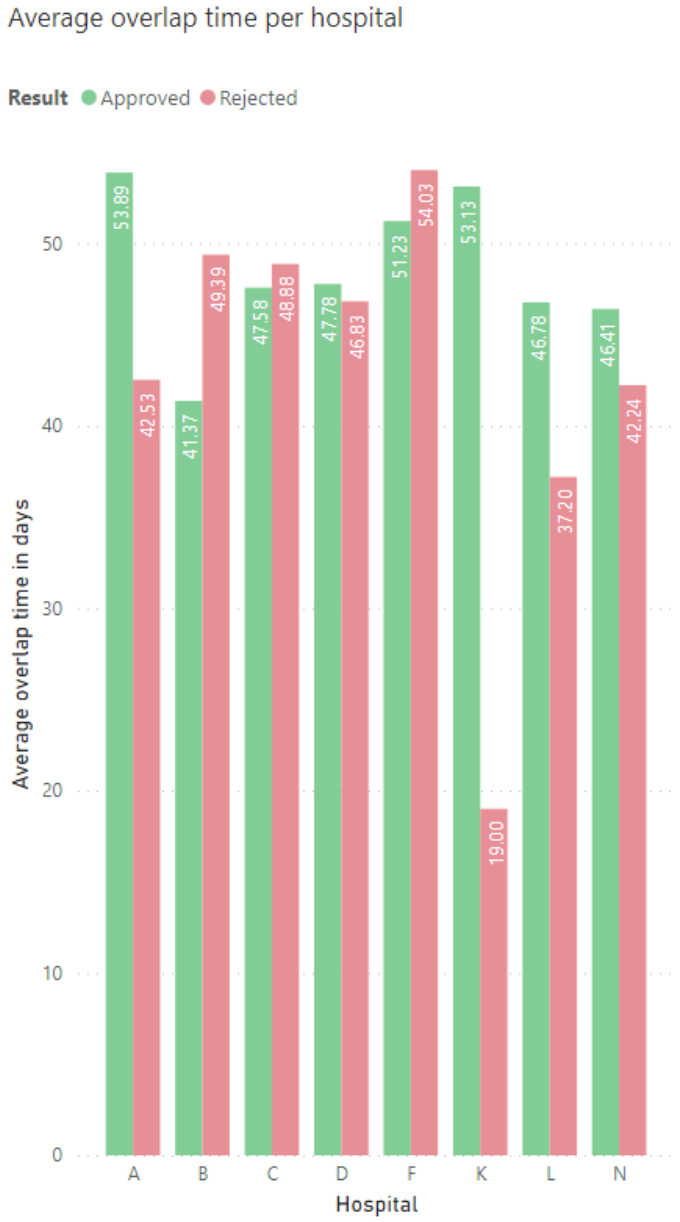

Figure 13: Average overlap time for approved and rejected observations per hospital for self-examination

The dashboard also gives insight into the difference in error margin on whether the diagnoses of the observed and parallel sub-trajectory are the same or not. When the diagnosis is the same, the error margin drops to 3.5%, based on 9 rejected observations out of 258. If the diagnosis is not the same, the error margin is 13.2%. That the diagnoses are the same mostly occur in ophthalmology, as here 238 out of the in total 258 observations with the same diagnoses are found.

3.3 Summary

The most important takeaways are that the three biggest specialties are ophthalmology (OOG), internal medicine (INT), and surgery (CHI), for both horizontal supervision and self-examinations. The average error margin for horizontal supervision is 11.4%, whereas for self-examination the average error margin is 12.0%.

4 Literature review

In this chapter, a literature review will be given on topics that are important in this thesis. First, the regulations regarding parallel care registration will be discussed. After that the core principles of statistical learning will be discussed, here different models and how to measure their performance will be mentioned.

4.1 Regulations

A care trajectory is opened with the first care activity of a new care question from the patient. This should be done by the professional practitioner who has the gate function and works following the registration rules of the NZa. It is not allowed to charge care at the expense of health insurers without a valid referral, the only exception is for acute care. This care trajectory contains all the treatments per care question of a patient. A care trajectory starts with an initial sub-trajectory (ZT11), which can take up to 90 days. After that, a follow-up sub-trajectory (ZT21) can be opened, which can take a maximum length of 120 days. After these 120 days again a follow-up sub-trajectory can be opened, and so on. The combination of the initial and (multiple) follow-up trajectories is called the DBC-care trajectory. For this duration a few exceptions hold. The first is a registration of the OR (operation room), here the duration after the date of operation is at most 42 days. When the patient is hospitalized, the duration of the DBC trajectory can be 42 days after discharge. When a medicine costs more than €1000 per patient per year, is it called 'expensive medicine' according to the Nza (Pfizer, n.d.), chemo, and dialysis are examples. For these DBC trajectories, the duration is at most until at most 1 day before the next expensive treatment (Performation, n.d.-a).

When a patient reports at the hospital, the first question is whether the patient is known within the given specialty. If the patient is new, so not known, a new care trajectory is opened with a sub-trajectory of type 11. In this sub-trajectory, the diagnosis and treatment are registered. When the patient is known within the given specialty, a second question arises. The question regards whether there is a new care question or not. When there is a new care question together with its own diagnosis and treatment a new trajectory can be opened. If the complaints are matching with an already known previous care question, then the registration has to be added to the ongoing trajectory. With the registration of care, the following patient details are noted; name, gender, BSN, birthday, address, phone number, multiple births, GP information, dentist, pharmacy, and health insurance. Opening parallel care and/or sub-trajectory within the same specialty is not allowed when there is no new or own care question, determining of diagnosis, and separately performed treatment from the other care trajectory.

Parallelism is defined as the registration of multiple care trajectories within one specialty, where there is overlap in duration. Within the "Nadere Regel", the rules regarding parallelism can be found (Nza, 2022). It is not allowed to open a parallel care path or a care sub-path within the same specialty when no new or own care question is in place. Here the difference is made that a new care question also yields a new diagnosis and anamnesis, which leads to a different policy of treatment concerning the care questions. When the diagnosis type is the same but the diagnosis can be two-sided (for example both left and right eye), it is allowed to open a sub-trajectory during the same duration. It is not allowed to open a parallel care trajectory:

1. When the combination of both diagnoses appears in the diagnosis-combination table, which can be found on the Nza site (Koen Kuijper, 2022).
2. When different care questions have the same diagnosis type with the duration of an existing trajectory.
3. Within the specialty of cardiology, unless it's a peer consult, cardiac rehabilitation, and guidance for cardiovascular disease and heart-lung transplant.
4. Within the specialty of clinical geriatrics, unless it's a peer consult or clinical co-treatment.
5. For neonatology within the specialty of pediatrics.
6. Within the specialty of geriatric rehabilitation care.
7. For the diagnosis of 'geriatric medicine' within the specialty of internal medicine, except for peer consult or co-treatment.
8. Within the specialty gynecology for the same phase* during one pregnancy, except for the phase before puerperium postnatal depression occurs after postnatal complications. *phases: pregnancy, childbirth, and puerperium.

   With some diagnoses, it is allowed to open a parallel care trajectory when:
9. Diagnoses are established during trauma care according to the "Advanced trauma life support" (ATLS). These diagnoses can be parallel registered, provided the conditions for parallelism are met when they are founded during the screening.
10. Diagnoses found during the population study "Screening colorectal carcinoma". These diagnoses can be parallel registered, provided the conditions for parallelism are met when they are founded during screening and a care trajectory should be started.

Every hospital has to check the correctness of parallel care trajectories. A sample test follows, with three different partial observations. The first partial observation checks all the sub-trajectories with ZT11 have parallelism with an earlier opened

care trajectory within the same specialty. The second partial observation contains sub-trajectories with ZT11 that are not rejected and are also not in partial observation 1. The last partial observation is from sub-trajectories with ZT21 that have parallelism with an earlier opened care trajectory within the same specialty. The first two partial observations have to be tested on rightful referral registration and whether a new care trajectory (ZT11) is opened correctly. The third partial observation is tested on whether the chosen trajectory is rightfully parallel (Nederlandse Zorgautoriteit, 2022). The checks on rightful parallelism are the first and the third. For self-examination, all three partial observations have a sample size of 125 observations each year. For horizontal supervision, the hospitals can choose their sample size per partial observation.

4.2 Statistical learning

Statistical learning is the overarching term for a set of tools that helps to understand data. These tools can be either supervised or unsupervised. Supervised statistical learning involves a prediction model of an output based on at least one input. The difference with unsupervised statistical learning lies in that there is no supervising output in unsupervised learning. Inputs are also known by the names of predictors, independent variables, features, or just variables. Where the outputs are often called the response or dependent variable. Variables are either quantitative or qualitative, the latter also called categorical. Quantitative variables are numerical values, examples of this are a person's age, income, or price. Qualitative variables are values of different classes or categories, examples of this are if a person has a typical diagnosis (yes or no) or what product brand is bought (brand A, B, or C). Problems with a quantitative response are often referred to as regression problems, whereas qualitative responses are referred to as classification problems (James et al., 2021). Below multiple models will be explained that were considered during this thesis based on knowledge and expertise.

4.2.1 Decision tree

Tree-based models stratify or segment the prediction space into regions. The decision tree can both be regression and classification based, both use recursive binary splitting for growing the decision tree. Looking at the example decision tree, the regions R_1, R_2, R_3, and R_4 are called terminal nodes or leaves of the tree. The points where the predictor spaces are split are called the internal nodes, an example is the node whether $X_2 \leq t_2$. The segments of the tree that are connected to the nodes are called branches. The predicted response of a regression tree is given by the mean response of the training dataset that belongs to the same terminal node. A regression tree is likely to be an oversimplification of the true relationships between the inputs. However, its advantage is that it is easy to interpret and gives a good graphical representation. A classification tree is similar to a regression tree, the difference is that the classification tree is used to predict a qualitative response, whereas a regression tree is focused on a quantitative response. Here the prediction is that each observation belongs to the most commonly occurring class of the training dataset in the region it belongs to. Trees are not very robust, this yields that a small change in the data can give a big change in the final estimation (James et al., 2021).

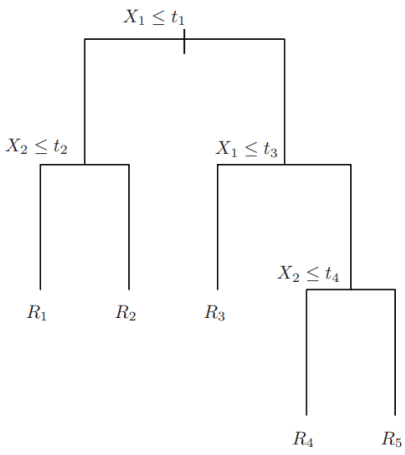

Figure 1: Example of decision tree

4.2.2 Linear and logistic regression

Linear regression is used to predict a quantitative response. First, we look at simple linear regression. Here a quantitative response, say Y, is predicted based on a single input, say X. Mathematically, the relationship of linear regression is written like this:

$$Y \approx \beta_0 + \beta_1 X$$

We look at an approximation, which is indicated by "≈". β_0 is the intercept, where β_1 is the slope of the linear model, both are called coefficients. However, more often more input variables estimate the output. A separate linear regression model could be fitted for each different predictor, however, you can also fit a multiple linear regression model. Then mathematically the relationship for Y with n variables is written like this:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

When the response is dichotomous, logistic regression is better fitted. With linear regression, estimates can be outside the [0,1] interval. A logistic regression model does not directly model the response of Y, it first calculates the probability of Y belonging to a category. Here the logistic function is written like this:

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

In order, to fit the above model, maximum likelihood is used. The intuition behind this is that for β_0 and the coefficients β_n estimates are found that correspond as closely as possible to the observed value. The above function only considers one predictor, however, more often multiple predictors are taken into account. When a response variable has more than two classes, for example with a ZPK1, here only the observed sub-trajectory can contain a ZPK1, only the parallel sub-trajectory

can contain a ZPK1, both or neither sub-trajectories can contain a ZPK1, this variable then has 4 different classes. The logistic regression function can then be extended to multinomial logistic regression (James et al., 2021).
Whether linear or logistic regression is best, depends on the type of response (James et al., 2021). Logistic regression is mostly used when you want to predict binary outcomes, such as whether a patient will get a certain disease or not. Logistic regression works by estimating the probability of a binary outcome rather than predicting the outcome. Therefore first the linear combination of the input variables and their associated coefficients are computed. This is then adapted in the logistic function to obtain the predicted probability of the binary outcome (Hastie, 2009).

4.2.3 Random forest

Decision trees suffer from high variance, as the tree is highly dependent on how the training data is split. Bootstrap is a way to lower this variance. With bootstrap repeated samples are taken from a training data set and all the predictions are averaged out. In a random forest, each decision tree is built on a randomly selected subset of the training data and a randomly selected subset of the input variables. This randomness is incorporated to prevent overfitting. The trees are built with the use of a recursive process where the algorithm splits the data into smaller and smaller subsets. To come from each different decision tree to a prediction, all the predictions of the tree are aggregated (Breiman, 2001).

4.2.4 Artificial neural networks

Another prediction model that can be used is artificial neural networks (ANN). This is a machine learning algorithm that is designed to stimulate the behavior of biological neural networks, like the brains of living organisms. ANN is known for recognizing patterns in data to make predictions. An ANN consists of layers of interconnected nodes, also known as neurons, which process and transmit information in the form of numerical values. Each neuron receives an input variable from other neurons, then performs a calculation on that input variable, and generates an output. The connections between neurons are represented by weights, which determine the strength and direction of the signal between neurons.

4.2.5 Comparison between different models

The above-mentioned algorithms have their strengths and weaknesses. The decision tree can handle both categorical and numerical data. Decision trees are easy to understand and interpret, which makes them useful for cases that require transparency and explainability. However, decision trees are prone to overfitting, meaning they can create overly complex trees that may not generalize well to unseen and/or new data. An expansion of this model is the random forest, where multiple decision trees are combined to create a more robust and accurate model. The overfitting issue is overcome by averaging the predictions of multiple trees. Another strength of the random forest is that it can handle large datasets and interactions between features. The weaknesses of random forests are that they can be computationally expensive and harder to interpret. The strengths of logistic regression are that it is computationally efficient and provides interpretable results in terms of feature importance. However, it is limited to binary classification and assumes a linear relationship between features, which may not always hold in real-world data. The ANN is the most complex algorithm, which is inspired by the human brain's neural networks. These networks are highly flexible and can model complex non-linear relationships, unlike logistic regression. The ANN requires large amounts of data and is computationally intensive. Due to the complexity, this algorithm is seen as a "black box", making it less interpretable. The choice of which algorithm depends on the specific problem, the characteristics of the data, the interpretability requirements, and the computational resources available.

4.2.6 Model performance

There are multiple ways to assess the accuracy of a classification model. The first is accuracy, which is equal to $\frac{number\ of\ true\ positive + number\ of\ true\ negatives}{positives + negatives}$. The accuracy gives the percentage of cases that were rightfully classified. Within this formula the true number of true positives and the number of true negatives are states, these are better known as true positives (TP) and true negatives (TN). With these numbers, the sensitivity can be calculated. The sensitivity refers to the proportion of actual positive cases that are correctly identified as positive by the model. The sensitivity is equal to $\frac{number\ of\ true\ positives}{number\ true\ positives + number\ of\ false\ positives}$. The specificity refers to the proportion of actual negative cases that are correctly identified as negative by the model. The specificity is equal to $\frac{number\ of\ true\ negatives}{number\ true\ negatives + number\ of\ false\ positives}$.
The accuracy, sensitivity, and specificity all range from 0 to 1, with higher values indicating better performance. However, the TP and TN are determined by the threshold, so the threshold must be set so that the performance value with the high priority has the highest value. Here it is good to point out that a model with high sensitivity but low specificity may identify too many false positives, while a model with high specificity but low sensitivity may miss too many true positives (Powers, 2011).

The ROC curve (receiver operating characteristic curve) displays how a model is performing in comparison with a random guess. If the area under the ROC curve (AUC) is bigger than 0.5, then the model is better than a random guess. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 – specificity). With changing the threshold of the classification model, the ROC curve changed, as both the sensitivity and specificity are displayed in the ROC curve. The ROC is strongly dependent on the chosen threshold. With the right threshold, the AUC can still be around 0.5, which suggests no predictive power. However, the model still can be used. If the consequences of a false positive or false negative are low, the model can be used as a decision-making tool. This type of model can also be used as a starting point, from where regularly monitoring the performance and updating the model as new data becomes available is important (Fawcett, 2006).

Figure 2: Example of a ROC curve

The higher the AUC, the better the performance of the model is as the predictions are better. The ROC is useful for comparing different thresholds. The ROC is strongly dependent on the chosen threshold. In consultation with Performation, specificity is most important.

4.3 Application of the literature

In this section, we will look into how the literature can be applied to the current situation. The goal of this research is to provide a model that can calculate the change of a parallel sub-trajectory wrongfully declared. For this, multiple models are implemented and compared. First, the decision tree will be tested. This model is chosen because of its simplicity, which makes it easier to implement into Notiz. Also, the logistic regression is implemented, as this is a more diverse method than the decision tree, as the probability can take more values, but is still doable to implement. Later during the research, both the random forest and artificial neural networks were considered. To expand the model and improve its performance, random forest was implemented. Artificial neural networks are the most comprehensive model, however, very hard to implement in Notiz, therefore not further considered suitable. For hospitals, it is considered most important that the number of false positives is as low as possible, without having to check all parallel sub-trajectory. This means that specificity is the most important. However, this goal is not to achieve the highest possible specificity, as this will involve a high number of cases to check, of which the majority will be correct. So both the specificity and the accuracy should be as high as possible to reach the best suitable threshold. The threshold is for checking the performance of the model, in the application of the model in the Notiz tool, the threshold can be chosen by the hospital.

5 Methods

In this chapter, we will look into the methods used in this thesis. First, the usage of Microsoft SQL Server Management Studio will be explained. Then the current situation for sample testing on parallel sub-trajectories is done, the results can already be found in Chapter 3. For this, two dashboards are made in Power BI. The statistical testing is done in R Lastly, the available data will be explained.

5.1 Tools used

To extract the data and prepare it, Microsoft SQL Server Management Studio is used. SQL Server Management Studio (SSMS) is an environment that integrates any SQL infrastructure. SSMS can be used to query, design, and manage databases and data warehouses (Microsoft, n.d.). We used this language to extract data from databases per hospital that are stored to use Notiz. First, we had to decide which data we wanted to extract and how this all is connected in the database. This had to be done for both horizontal supervision and self-examination. To extract the right data, two different queries were made for horizontal supervision and self-examination. For the horizontal supervision, the query was not too hard, as the data was only stored in one table. This query extracted whether the sample is marked as wrong or correct. Also, the comment and whether it was randomly reviewed, also with comment. When the approval status was filled, this is also extracted. About the sub-trajectory of the sample the specialty, the diagnosis, and the care type (either 11 or 21) are extracted. No patient-specific data (such as a patient number) is extracted, therefore there is no need for anonymization. For the self-examination, we extracted more data. As the data was stored in more tables, the patient number and the sub-trajectory number is needed as connectors. This export of the data contains the following: patient number, care trajectory number, sub-trajectory number, whether the sub-trajectory was added as parallel, specialty, diagnosis, care type, start and end date of the care trajectory, start and end date of the sub-trajectory, approved or rejected parallelism, and what had changed to the parallelism when rejected. For both the observed as the parallel sub-trajectory for each care activity the number, code, class, date, amount, executive specialty, and requested specialty are extracted. Also for each variable, the hospital (anonymized) is noted. Within SQL the number of care activities per sub-trajectory is determined. After all this, the specific patient data is anonymized, which means that the patient number, the care trajectory number, and the sub-trajectory number are replaced with anonymized random numbers, to prevent a traceback.

After that, all the data is combined into one database in Excel. Here the data is prepared for both the visualization as well as the statistical testing. In Excel, the anonymized number is combined with a letter belonging to the hospital. The overlap time is calculated in Excel, which is defined as the difference in days between the end date of the observed sub-trajectory and the starting date of the parallel sub-trajectory. Per sub-trajectory, it is extracted if they have care activities in certain care profile classes. We look at the first 8, as they are mostly used.
In Power BI two dashboards are built, one for horizontal supervision and one for self-examination. The student created these dashboards to visualize the current situation based on the available data and to compare the hospitals to each other. More about this can be found in Chapter 3, where the current situation is explained by these dashboards.

In R studio statistical learning is applied. First, we load the self-examination data. Then the total data set is divided into four different sets. The 3 biggest specialties within the data set, and the remaining specialties. For statistical learning, it is important to split the data into training and testing sets. This is important because the training set is used to train the data, and the testing set is used for validating the model. The test set should be of the division 30/70 or 20/80 (in the percentage of the total number of observations) for testing/training data (James et al., 2021). In this thesis, we use 30/70. All 5 data sets (total, eye, surgery, internal, and remaining specialties) are split into a training and testing set. After that for all the data sets three error margins are calculated, for the training, testing, and both together. The error margins are calculated by taking the number of wrongful registrations within the sample and dividing it by the total number of registrations in the sample.

Multiple models were investigated, however, not all were implemented. The outcome of the decision tree and logistic regression were both easy to implement in the Notiz tool. The artificial neural network was also implemented; however, the data did not seem to fit this algorithm, and the outcome of this model would be harder to implement for Performation. To make the model less dependent on the training data split, random forests were incorporated to compare how this might benefit the model's performance. The first is the decision tree, which classifies the training data. This is done by using the 'rpart' function together with the 'rpart.plot' function in R. The second model is the logistic regression model, this is chosen instead of linear because the desired outcome is between 0 and 1. We model the binomial fit with the 'glm' function in R. The same variables as for the decision trees are taken into account. Backward selection is used to select the significant variables. Backward selection starts with a model that includes all potential input variables and sequentially removes variables that do not contribute significantly to the model. In R, the function "step" can be used with the direction "backward" to obtain the variables that can be removed without significantly reducing the model fit. After that, we predict the outcome of the test data.

With the outcomes of the test data, it is necessary to classify the data as rejected and approved based on a certain threshold. This threshold is determined by varying this threshold to get the desired level of specificity and accuracy. The desired level is the combination where the combination of specificity and accuracy is as high as possible. The specificity and accuracy are sensitive to the threshold chosen. The best threshold is found in R with the "coords" package. Here the best threshold with a focus on specificity, with reasonable accuracy, is calculated by the ROC for multiple thresholds. The threshold chosen is the value with the best AUC. To give a value to the desired level is the highest possible specificity, with an accuracy of at least 0.75, this is chosen because still most of the incorrectly opened sub-trajectories are traced, but the number of checked sub-trajectories is not too high with too many false negatives in comparison with true negatives. In consultation with Performation the desired number of cases classified as "rejected", so the cases that should be checked, is 20% of the total number of parallel cases. Here also a confusion matrix is made to determine the different measures of accuracy. Lastly, for the testing set, the ROC curve is plotted. The last model is the random forest, here different trees are generated and the tree with the best accuracy is stored.

5.2 Available data

The data used in this research is historical. The horizontal supervision data collect all the samples that are finished within the check on parallelism. These are the tests from the moment the hospital started with horizontal supervision, this can therefore be of multiple years. The data from the self-examination is from the last year the hospitals completed all the checks, most of the time in 2021. In total, the sample of horizontal supervision contains 3070 observations of 6 different hospitals. 9 hospitals shared the data of their self-examination, which adds up to a total amount of 2080 observations. The difference between the number of observations lies in that self-examination only checks once per year, and horizontal supervision more often per year, which is why those observations are more.

In the Netherlands, there are 5 different types of hospitals, university medical centers, specialized hospitals, top clinical hospitals, general hospitals, and expertise centers. Out of the 13 hospitals participating in this research, 12 were general or top clinical, and 1 is an expertise center. Because of privacy reasons the hospitals' names are not mentioned and are replaced by letters. Below the number of observations per hospital can be found and whether data for horizontal supervision and/or self-examination was available.

Hospital	Horizontal supervision #observations	Self-examination #observations
A		238
B*		231
C**		509
D		213
E	675	
F***	75	226
G	664	
H	624	
J	258	
K		206
L		231
M	774	
N		226
Total	3070	2080

Table 1: Available data per hospital

* Specialized hospital, ** Two years of data available, *** Switched recently, therefore data for both horizontal supervision and self-examination

6 Results

In this chapter, we review the results of the model. We look at the different models and their performances. For this part, only the self-examination is taken into account. This is because the data from horizontal supervision is not static: this means that it is not known whether the registration is still the same as when the parallelism was judged. If for example the observed sub-trajectory was rejected, so the parallelism was not correct, this sub-trajectory is added to the parallel sub-trajectory and therefore no more parallelism is encountered. The decision to only focus on self-examination is that we can see for certain that the data is still the same as when it was judged.

6.1 Variable selection

Within the data set multiple variables were available, however not all were useful. First, we focus on the 8 biggest ZPKs, rather than the individual care activities. For the model, we do not take the specialty into account, other than the split in different data sets. The diagnosis is not considered, the reason for this is that this is too specific. We only focus on the overlap time between the observed and parallel sub-trajectory, we don't look at the overarching care trajectories. The variables we do take into account are the care type of the observed sub-trajectory, the overlap time between the observed and the parallel sub-trajectory, the number of care activities for both the observed as the parallel sub-trajectory, and whether the observed and parallel sub-trajectory had care activities for ZPK 1 to 8. For the last variable, there are four classes. The first is that only the observed sub-trajectory has that certain ZPK among the care activities, the second that only the parallel sub-trajectory has that certain ZPK among the care activities, the third is that both the observed and parallel sub-trajectory have that certain ZPK among their care activities, and the last one, that neither the observed nor parallel sub-trajectory have that certain ZPK among their care activities.

6.2 Error margins

The data sets are split into a training and testing set. This is done randomly, where 70% of the observations belong to the training set and the remaining 30% to the testing set. For all the data sets the error margin is calculated. This is done to compare the accuracy. Below a table can be found per data set of what the error margin is.

	Total	*Training set*	*Testing set*
All specialties	0.12 (n = 2193)	0.12 (n = 1535)	0.12 (n = 658)
Ophthalmology (OOG)	0.09 (n = 643)	0.10 (n = 450)	0.08 (n = 193)
Surgery (CHI)	0.11 (n = 261)	0.13 (n = 182)	0.09 (n = 79)
Internal medicine (INT)	0.12 (n = 328)	0.14 (n = 229)	0.08 (n = 99)
Remaining specialties	0.14 (n = 961)	0.14 (n = 672)	0.15 (n = 289)

Table 2: Error margin per data set, where the highest error margin is encountered for the remaining specialties

6.3 Classification tree

In R the training data is used to derive the classification tree. In this section, we will review these trees, each for every data set and its performance. To make the classification trees more readable, abbreviations are used, which can be found in the appendix. The threshold of the tree is based on the classification that belongs to approximately 15% with the highest risk of getting wrongfully declared. This 15% is chosen because it is equal to the highest error margin, as can be seen in Table 2. For all the trees the classification regions can be around 15% of the training data. If you decide to increase this to 20%, the number of false negatives and the number of checked cases might increase substantially.

6.3.1 All specialties

For all the specialties all variables as described in section 6.1 are taken into account. In figure 17 the classification tree for all specialties is shown. For this decision tree, only the presence of ZPK1 and the number of activities in the observed sub-trajectory is used to classify the data.

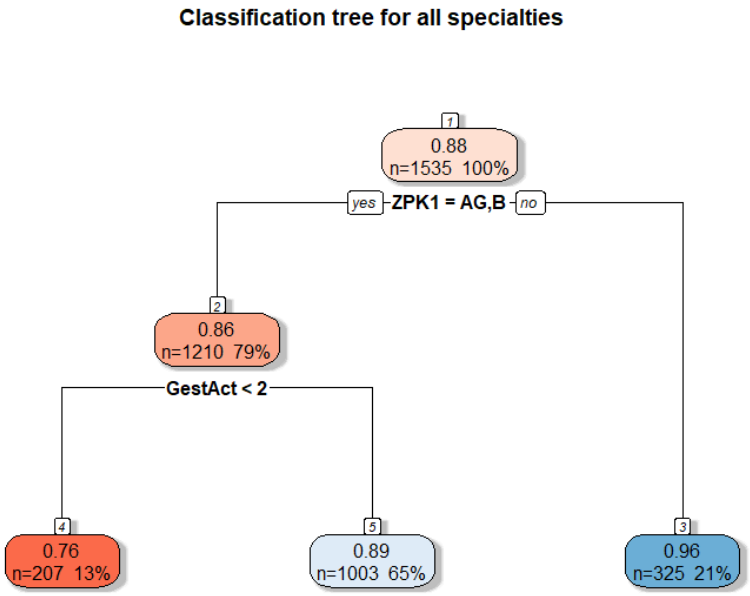

Figure 3: Classification tree for all specialties

To get the 15% of the observations with the biggest risk of getting wrongfully declared, we only classify all responses in region 4 as rejected. The accuracy of this is 0.7872, the sensitivity is 0.8726, and the specificity is 0.2118. If you want to increase the specificity, more regions can be classified as rejected. Here, also region 5 can be classified as rejected. The specificity then increases to 0.8941, however, the accuracy drops to only 0.3343. To get this specificity, in total 505 cases should have been checked. In practice, this is not doable, there the threshold for the classification tree is 0.76.

When we use this classification for the testing data, the below confusion matrix follows.

All specialties	Predicted rejected	Predicted approved
Actual rejected	18	67
Actual approved	73	500

6.3.2 Ophthalmology

For this classification tree also all variables as described in section 6.1 are taken into account. Figure 18 displays the classification tree for ophthalmology. In comparison with the classification tree for all specialties, these three classify the data based on more variables. The important variables according to the model are ZPK1, ZPK2, ZPK4, ZPK5, ZPK6, overlap time, number of activities in the observed sub-trajectory, and number of activities in the parallel sub-trajectory.

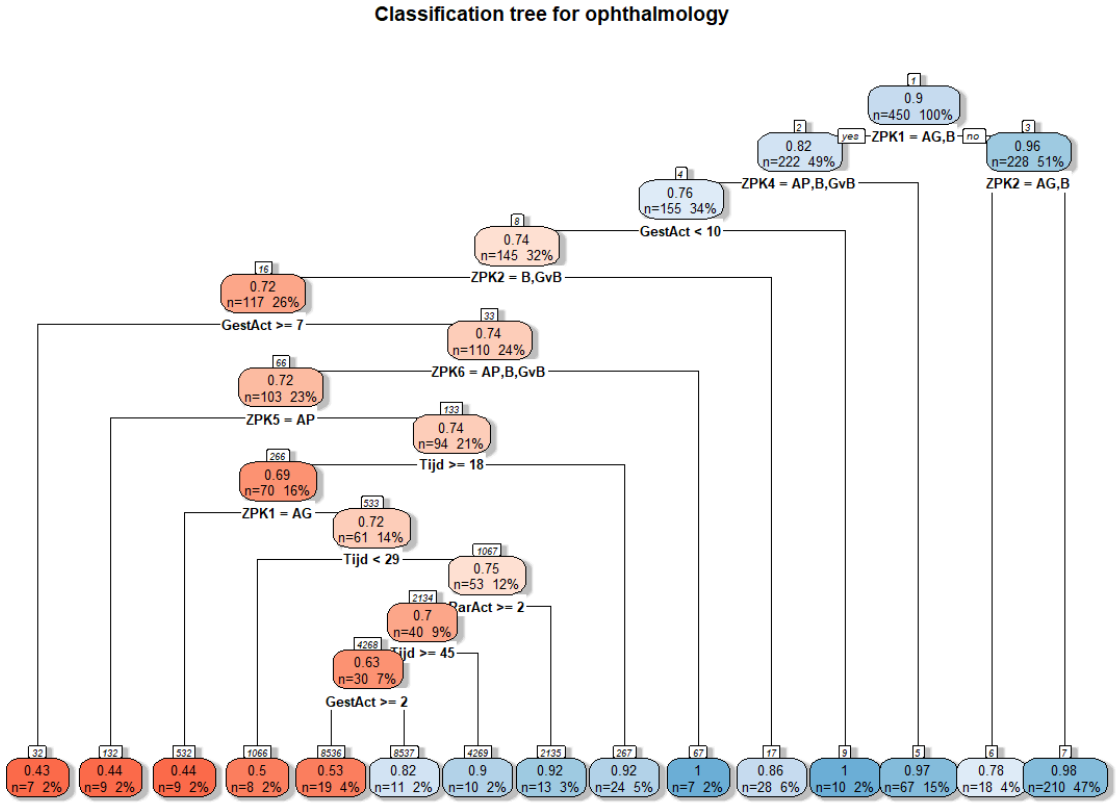

Figure 4: Classification tree for ophthalmology

When taking the 13% of classifications that have the highest risk, the regions 32, 132, 532, 1066, and 8536 will be predicted as rejected. The accuracy of this model is 0.8394, the sensitivity is 0.8827, and the specificity is 0.2857. To increase the specificity the threshold should be higher than 0.53. The next predicted value will be Region 6, which is 0.78. Then the specificity will stay the same, as no new true negatives will be classified. But with this threshold, the accuracy will decrease to 0.8083, as 6 new false negatives are classified. If also region 8537 is classified as rejected, the threshold is increased to 0.82. The belonging specificity when the threshold is set at 0.82 is 0.3571 and the accuracy is 0.7979. The cases that are predicted as rejected are 35. If we increase the threshold even further to 0.86 (so also include region 7), the accuracy is 0.7668, but the specificity increases to 0.5714. Increasing the threshold even further to 0.90 (including region 4269), the accuracy stays the same, however, the specificity will increase to 0.7143. If the threshold is increased further, the accuracy drops below the desired level of 0.75, so the advised threshold is 0.90.

When using the threshold of 0.90, the following confusion matrix will follow.

Ophthalmology	Predicted rejected	Predicted approved
Actual rejected	10	4
Actual approved	41	138

6.3.3 Surgery

For the surgery decision again all variables are taken into account. The important variables according to the model are ZPK4, ZPK6, ZPK7, overlap time, and the number of activities in the parallel sub-trajectory.

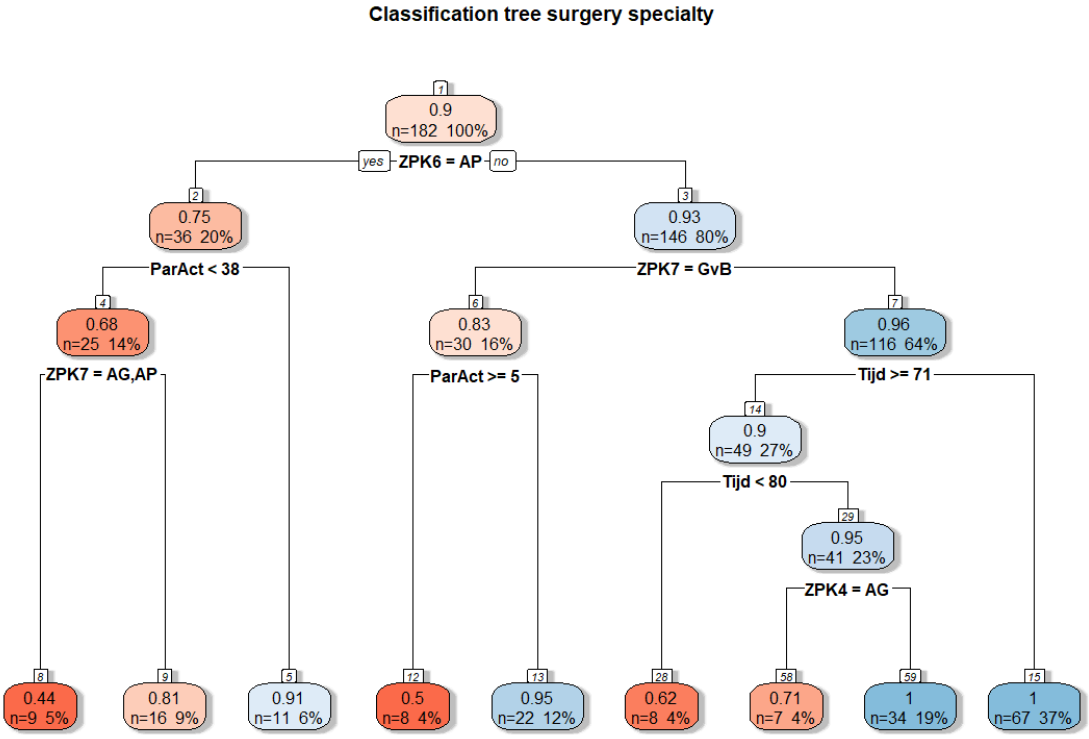

Figure 5: Classification tree for surgery

The responses in regions 8, 12, and 28 are classified as rejected because they are 13% of the observations with the highest risk. With a threshold of 0.63, the accuracy of this model is 0.7848, the sensitivity is 0.8971, and the specificity is 0.0909. When adding a region to the rejecting classification, the threshold increases to 0.71, where region 58 is added. The accuracy then decreases to 0.7468, but the specificity stays the same. If Region 9 is added as well, the threshold is increased to 0.81, However, the accuracy then drops below 0.75. So the advised threshold is 0.63.

Then the following confusion matrix can be constructed for the test data with a threshold of 0.63.

Surgery	Predicted rejected	Predicted approved
Actual rejected	1	10
Actual approved	7	61

6.3.4 Internal medicine

For internal medicine we don't look at the ZPK5, this is because internal medicine is a contemplative specialty, and there seldom operative. Following the model, 5 variables are important for the classification. These variables are ZPK6, ZPK7, ZPK8, the number of activities in the observed sub-trajectory, and the number of activities in the parallel sub-trajectory.

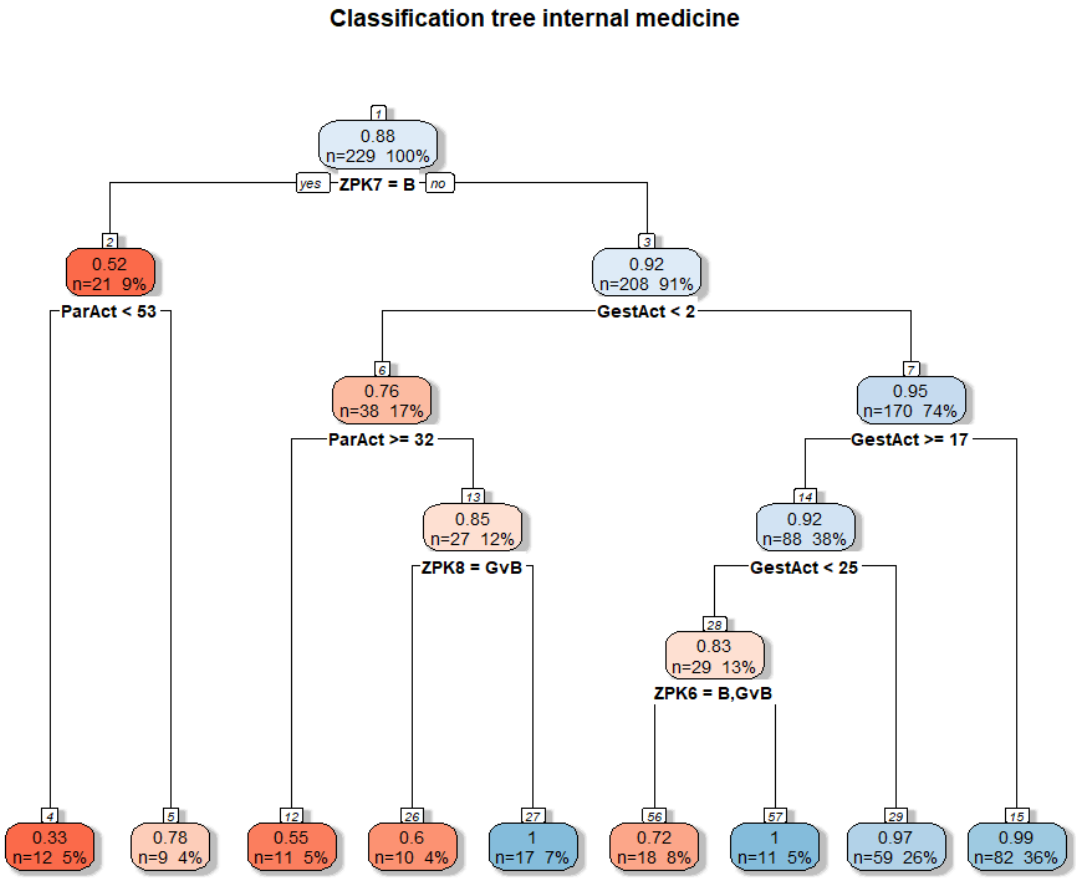

Figure 6: Classification tree for internal medicine

The regions that take up 14% with the highest risk are regions 4, 12, and 26, these get predicted as rejected. The accuracy of this model is 0.8182, the sensitivity is 0.8966, and the specificity is 0.2500 for the threshold of 0.61. Increasing the threshold to 0.73 yields an accuracy of 0.7879 and a specificity of 0.3333. Increasing the threshold further will yield an accuracy lower than 0.75, so the advised threshold is 0.73

Then the following confusion matrix of the testing data for internal medicine with a threshold of 0.73 is constructed.

Internal medicine	Predicted rejected	Predicted approved
Actual rejected	4	8
Actual approved	13	74

6.3.5 Remaining specialties

For the remaining specialties again all variables are taken into account. The important variable for this classification tree is ZPK3, ZPK5, overlap time, the number of care activities in the observed sub-trajectory, and the number of care activities in the parallel sub-trajectory.

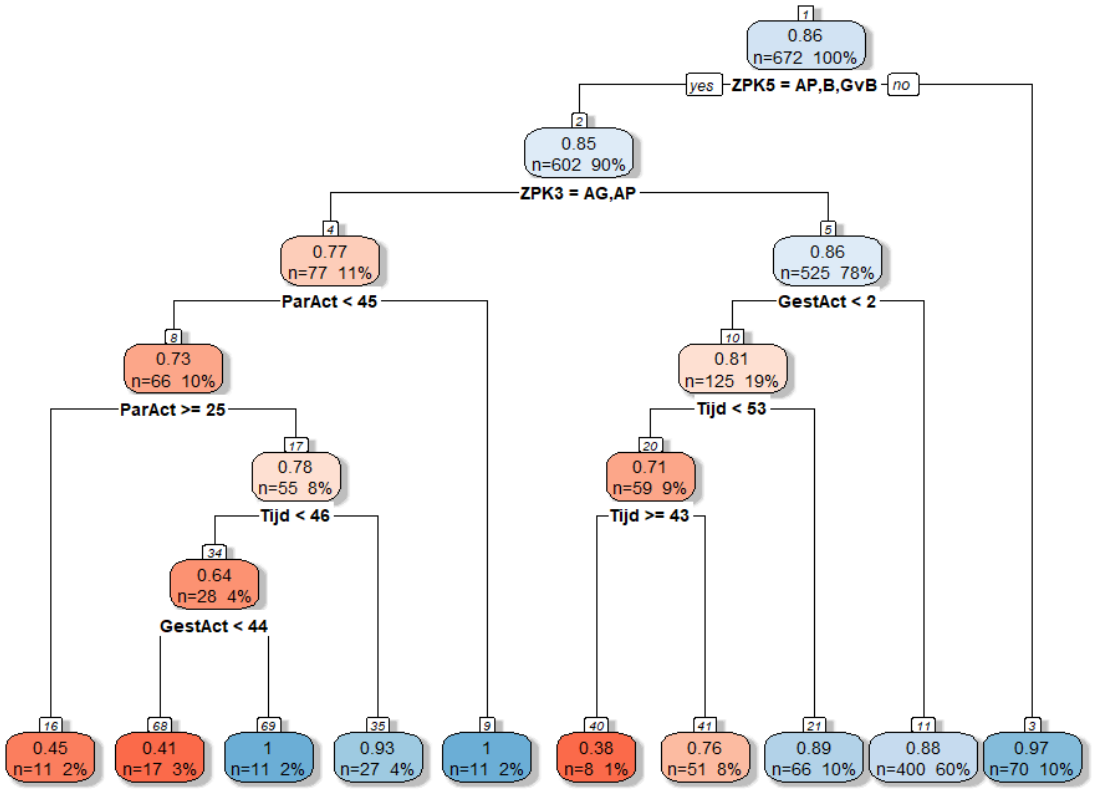

Figure 7: Classification tree of remaining specialties

The responses in regions 16, 40, 41, and 68 are classified as rejected, as they take up the bottom 14% of the predicted observations. With a threshold of 0.77, the accuracy of this model than in 0.7924, the sensitivity is 0.8866, and the specificity is 0.2381. Adding region 11, the accuracy goes to 0.3633, so the advised threshold is 0.77.

The following confusion matrix can be constructed with the testing data when using threshold 0.77.

Remaining specialties	Predicted rejected	Predicted approved
Actual rejected	10	32
Actual approved	28	219

6.4 Logistic regression

For logistic regression, the same variables are considered per data set as for the classification tree. For every data set, we will look into the coefficients of the model and its performance. In consultation with Performation, it is most important that the number of false positives is as low as possible, so the specificity is as high as possible. However, this might result in too many cases that should be checked, therefore, the accuracy should be at least 0.75.

6.4.1 All specialties

In R Studio the "glm" function is used to derive a logistic regression function. After backward selection, the following variables are determined to be significant; ZPK1, ZPK2, ZPK3, ZPK4, and the number of activities of the observed sub-trajectory. The estimated coefficients can be found in Table 3.

All specialties	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1933364	0.53213019	2.2425647	0.024924902
GestAct	0.011	0.005	2.268	0.023
ZPK1AP	1.214	0.364	3.330	0.001
ZPK1B	0.188	0.230	0.818	0.413
ZPK1GvB	3.221	1.029	3.132	0.002
ZPK2AP	-0.305	0.397	-0.768	0.443
ZPK2B	-0.644	0.645	-0.999	0.318
ZPK2GvB	0.365	0.293	1.246	0.213
ZPK3AP	0.045	0.411	0.108	0.914
ZPK3B	-0.255	0.630	-0.404	0.686
ZPK3GvB	0.806	0.330	2.443	0.015
ZPK4AP	-0.583	0.311	-1.875	0.061
ZPK4B	-0.596	0.346	-1.721	0.085
ZPK4GvB	-0.759	0.272	-2.794	0.005

Table 3: Coefficients for all specialties with logistic regression

When implementing the backward-selected variables, the threshold can be calculated. The optimal threshold only considering the specificity is 0.884. The accuracy belonging to this threshold is 0.4954, the sensitivity is 0.4468, and the specificity is 0.8235. In this case, the specificity is high, but the accuracy is low. By changing the threshold by steps of 0.001, the desired accuracy can be found. Setting the threshold at 0.858 gives an accuracy of 0.7568 and a specificity of 0.2824. However, the latter case returns a lower AUC of 0.5406.

When only specificity is most important, to minimize the number of false positives, the proposed threshold by the algorithm is 0.884. This returns that 387 out of the 658 cases should be checked, which is why more than the 20% that is set in consultation with Performation. When looking at the desired level of accuracy, the number of predicted rejected is 85 cases. This would still be within the capacity of hospitals to check this. To take one step further and look at what the threshold should be, corresponding with rejecting 20% of the lowest chances, this is 0.858. This threshold is set by changing the threshold by steps of 0.001. With this threshold, the accuracy is 0.7568 and the specificity is 0.2824.

The below confusion matrix follows then with the testing data using a threshold of 0.858:

All specialties	Predicted rejected	Predicted approved
Actual rejected	24	61
Actual approved	99	474

The ROC curve is displayed in figure 23. Here we see that the area under the curve (AUC) is equal to 0.541. The general rule of thumb is that the AUC-ROC should be above 0.7 to be considered fair. As this AUC is just above 0.5, it indicates some degree of discriminative ability, but it is not very useful for distinguishing positive and negative cases. As stated before, a low AUC still can be used, however, some limitations have to be taken into account. As these limitations hold for multiple models, this will therefore be considered in the discussion (Section 7.2).

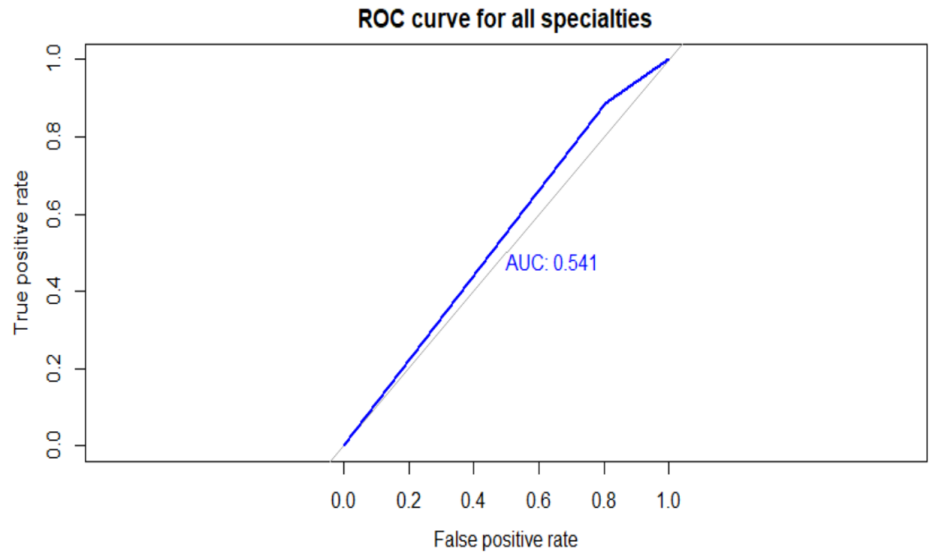

Figure 22: ROC curve for all specialties with logistic regression with threshold 0.858

6.4.2 Ophthalmology

With the backward selection, the variables that are selected for ophthalmology are ZPK2, ZPK3, ZPK4, ZPK5, and ZPK6. The estimated variable coefficients can be found in Table 4.

Ophthalmology	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-14.877109	1016.94855	-0.014629	0.988328
ZPK2AP	1.455	1.176	1.238	0.216
ZPK2B	-2.045	1.173	-1.744	0.081
ZPK2GvB	1.042	0.738	1.412	0.158
ZPK3GvB	18.554	1016.948	0.018	0.985
ZPK4AP	-1.097	0.697	-1.575	0.115
ZPK4B	-1.350	0.714	-1.891	0.059
ZPK4GvB	-1.416	0.663	-2.135	0.033
ZPK5AP	-1.890	0.740	-2.554	0.011
ZPK5B	1.044	0.869	1.202	0.230
ZPK5GvB	-1.337	0.557	-2.401	0.016
ZPK6AP	-1.131	0.714	-1.584	0.113
ZPK6B	2.059	0.917	2.245	0.025
ZPK6GvB	-1.021	0.588	-1.737	0.082

Table 4: Coefficients for ophthalmology with logistic regression

With the above-mentioned coefficients, the new model will predict the outcome values for the test data. The threshold is chosen to get the highest specificity. The accuracy belonging to this threshold is 0.7409, the sensitivity is 0.7430, and the specificity is 0.7143. The accuracy is almost at the desired level. To get the desired level, the threshold should decrease to 0.880, then the accuracy is 0.7617, with a specificity of 0.6429.

The threshold for this model is 0.880, which yields the following confusion matrix:

Ophthalmology	Predicted rejected	Predicted approved
Actual rejected	9	5
Actual approved	41	138

The ROC curve for threshold 0.880 is displayed in figure 23. The area underneath the curve is equal to 0.573, which is not high enough to call this model fair or good. The same as for the previous ROC, the limited use will be discussed in the discussion.

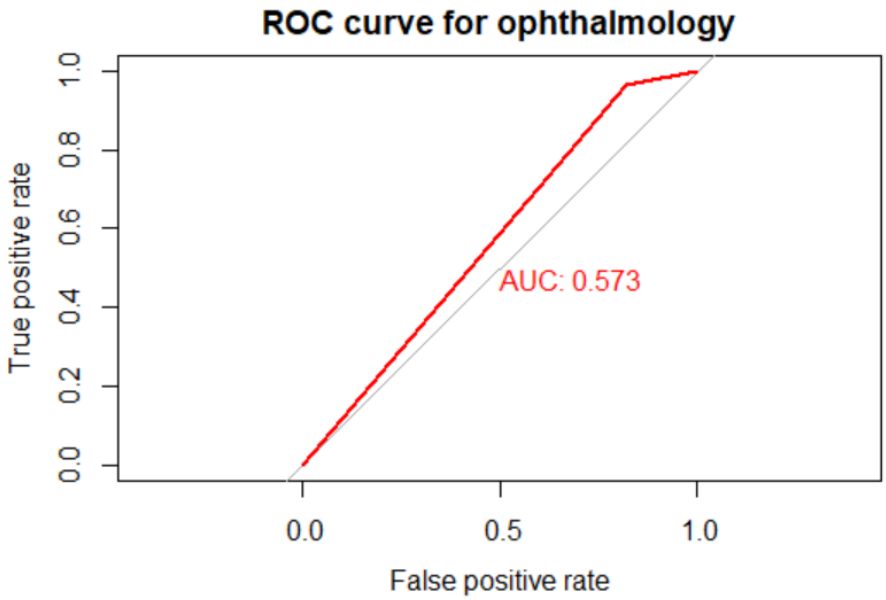

Figure 23: ROC curve for ophthalmology with logistic regression with threshold 0.880

6.4.3 Surgery

For the logistic model, the variables that have been selected with backward selections are care type and ZPK6. The coefficients for these variables can be found in table 5.

Surgery	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	18.769287	1221.40055	0.015367	0.987739
ZorgType21	-0.769	0.542	-1.419	0.156
ZPK6AP	-17.382	1221.401	-0.014	0.989
ZPK6B	-15.901	1221.401	-0.013	0.990
ZPK6GvB	-15.948	1221.401	-0.013	0.990

Table 5: Coefficients for surgery with logistic regression

For this model, only two input categorical variables are taken into account. This results in a small range of possible values. With this small range, the threshold is set rather high to achieve the highest sensitivity. The accuracy belonging to the above table is 0.2152, the sensitivity is 0.1324, and the specificity is 0.7273. This accuracy is not at the desired level. The threshold that gives an accuracy of the desired level is 0.800, this is 0.8101, where the specificity is 0.000.

The threshold is 0.800, which leads to the confusion matrix below.

Surgery	Predicted rejected	Predicted approved
Actual rejected	0	11
Actual approved	4	64

The ROC curve is displayed in figure 24. The area underneath the curve is equal to 0.427 which means that this model is worse than just random guessing. However, the positive and negative cases can be reversed, this means that the AUC is also inverted. The new AUC is 0.573, which is in the same order as the other AUCs. This can be because of a bad split between the training and testing data set, but mostly because only two variables are taken into account.

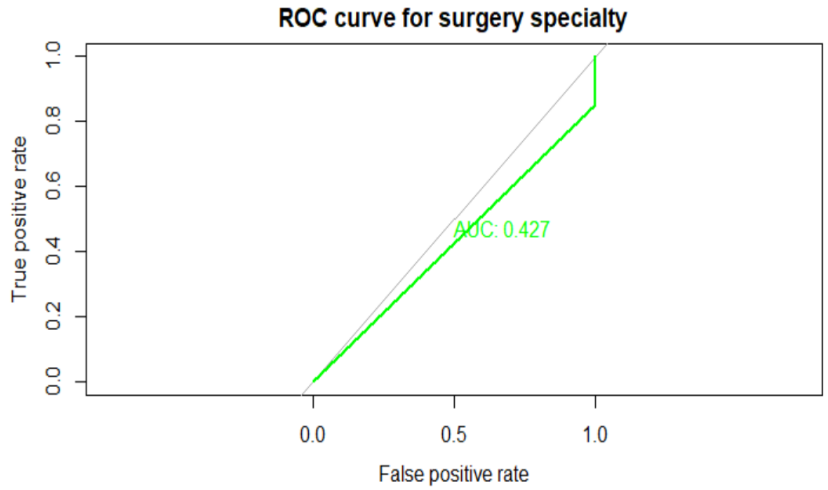

Figure 24: ROC curve for surgery with logistic regression with threshold 0.800

6.4.4 Internal medicine

For internal medicine, the variables that are selected with backward selection are overlap time, ZPK6, ZPK7, and ZPK8. Below the estimated coefficients of these input variables can be found.

Internal medicine	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.994921	1.25927683	3.172393	0.0015119
Tijd	0.0116	0.0084	1.3852	0.1660
ZPK6AP	-1.3850	0.8363	-1.6562	0.0977
ZPK6B	-0.7931	0.8861	-0.8951	0.3707
ZPK6GvB	-1.7706	0.7579	-2.3363	0.0195
ZPK7AP	-0.6476	1.2166	-0.5323	0.5945
ZPK7B	-3.6985	1.1640	-3.1776	0.0015
ZPK7GvB	-0.7834	1.1080	-0.7070	0.4795
ZPK8AP	0.0447	0.7399	0.0605	0.9518
ZPK8B	0.5392	0.7250	0.7438	0.4570
ZPK8GvB	-1.3097	0.7854	-1.6675	0.0954

Table 6: Coefficients for internal medicine with logistic regression

The threshold for internal medicine lies at 0.965 for retrieving the highest specificity. The accuracy belonging to this threshold is 0.3434, the sensitivity is 0.2529, and the specificity is 1.000. To get higher accuracy, the threshold should be lowered to at least 0.850, then the accuracy is 0.7576, and the specificity is 0.4167.

With this, the below confusion matrix of the testing data with a threshold of 0.850 is constructed.

Internal medicine	Predicted rejected	Predicted approved
Actual rejected	5	7
Actual approved	17	70

The ROC curve for threshold 0.850 is displayed in figure 25. The area underneath the curve is equal to 0.568 which means that this model is better than just random guessing, but still, as a classification model not very useful.

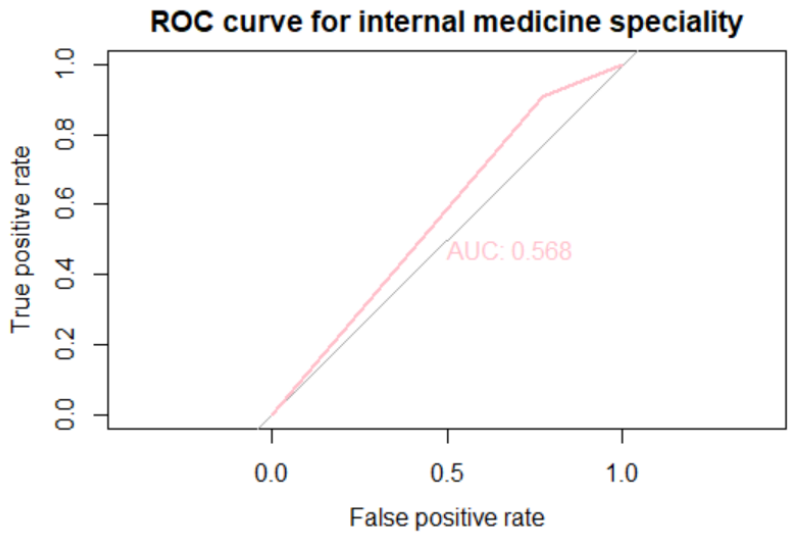

Figure 25: ROC curve for internal medicine with logistic regression with threshold 0.850

6.4.5 Remaining specialties

For the remaining specialties, the variables that are significant with backward selection are whether the diagnosis is the same between the observed and the parallel sub-trajectory, the number of activities in the observed sub-trajectory, the number of activities in the parallel sub-trajectory, ZPK1, ZPK3, and ZPK5. The coefficients can be found in table 7.

Remaing specialties	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	17.508854	932.359300	0.018779	0.985017
ZelfdeDiagNee	-14.9022	932.3589	-0.0160	0.9872
GestAct	0.0095	0.0062	1.5389	0.1238
ParAct	0.0154	0.0096	1.6136	0.1066
ZPK1AP	1.8998	1.1069	1.7163	0.0861
ZPK1B	-0.2235	0.3346	-0.6680	0.5041
ZPK1GvB	14.7584	1144.5430	0.0129	0.9897
ZPK3AP	-0.3900	0.6765	-0.5765	0.5643
ZPK3B	0.5711	1.1761	0.4856	0.6272
ZPK3GvB	1.0172	0.5260	1.9338	0.0531
ZPK5AP	-1.7839	0.9022	-1.9773	0.0480
ZPK5B	-3.1156	0.9069	-3.4353	0.0006
ZPK5GvB	-1.8011	0.7380	-2.4407	0.0147

Table 7: Coefficients for internal medicine with logistic regression

The threshold to get the highest sensitivity is set at 0.860. The accuracy belonging to this threshold is 0.4775, the sensitivity is 0.4899, and the specificity is 0.4048. To get the desired accuracy, the threshold should be lowered to 0.836, then the accuracy is 0.7543 and the specificity is 0.1190.

Taking this into account the below confusion matrix of the testing data with a threshold of 0.836 is generated.

Remaining specialties	Predicted rejected	Predicted approved
Actual rejected	5	37
Actual approved	34	213

The ROC curve of threshold 0.836 is displayed in figure 26. The area underneath the curve is equal to 0.490 which means that this model is better than just random guessing, but not much. This AUC can also be reversed, resulting in 0.510.

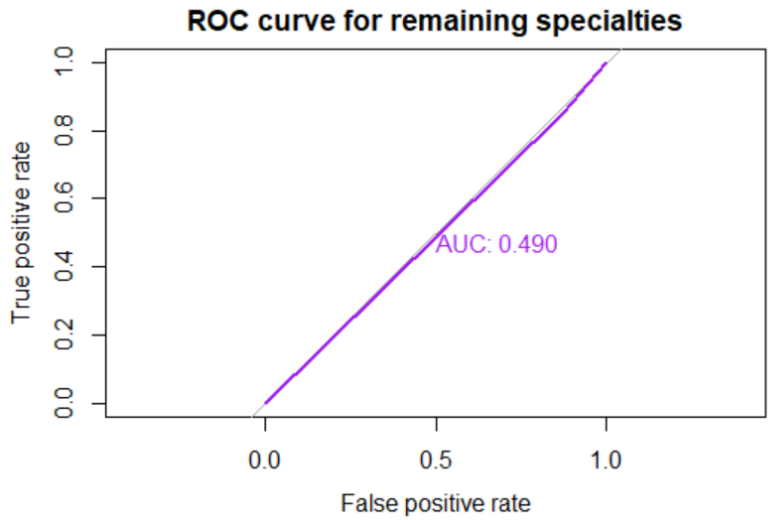

Figure 26: ROC curve for remaining specialties with logistic regression with threshold 0.836

6.5 Random forests

For this model, multiple random forests are performed where several values for the number of chosen predictors are selected and an Importance Plot is made. The results and performance will be discussed per data set.

6.5.1 All specialties

For all specialties, the Importance Plot can be seen in figure 27. The plot on the left gives the average decrease in accuracy if we leave out that variable. So the higher this value, the more important the variable. For all specialties, the number of activities in the observed sub-trajectory is the most important. The graph on the right gives the mean decrease Gini, this is a measure based on the gini impurity index used for calculating the splits in trees. Here also holds, the higher the variable, the more important the variable. When comparing these two, both give indicate the number of activities in the observed and parallel sub-trajectory, ZPK1, and ZPK4 in their top 6.

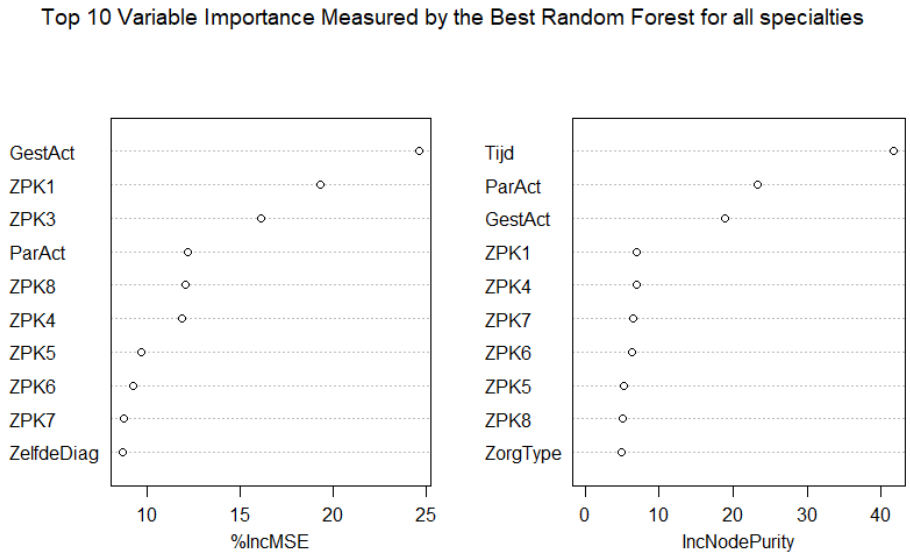

Figure 28: Importance plot for all specialties, where the top 3 returns different important variables

When predicting the outcome of the testing data following the random forests, the following threshold is calculated the same as in the logistic regression model. The best threshold found is 0.924 to reach the highest specificity. The corresponding accuracy is 0.4468, the sensitivity is 0.3944, and the specificity is 0.8000. When lowering the threshold to get better accuracy of at least 0.75, the threshold should be 0.780, then the specificity drops to 0.2588. With threshold 0.780 the below confusion matrix is constructed:

All specialties	Predicted rejected	Predicted approved
Actual rejected	22	63
Actual approved	98	475

The ROC curve is given in figure 29, the AUC is 0.533.

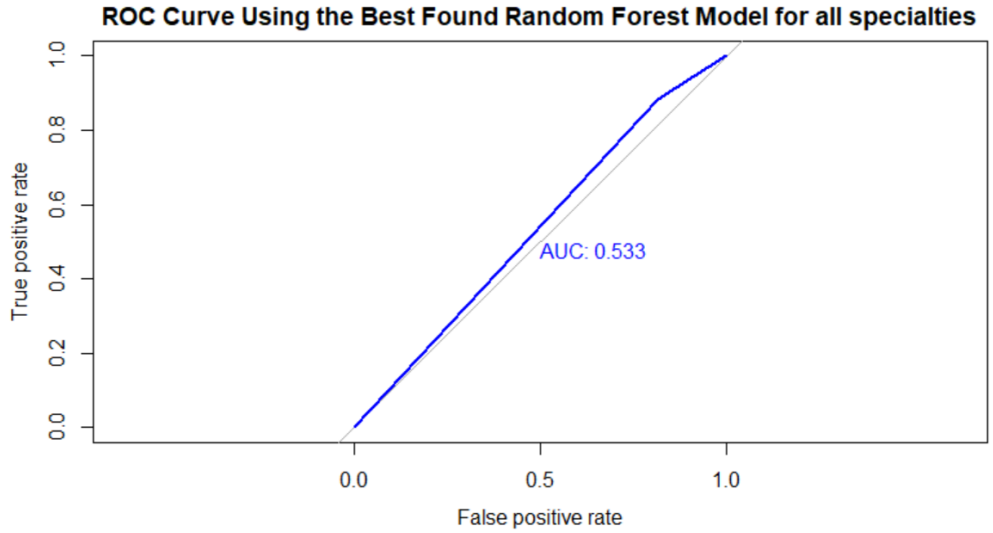

Figure 29: ROC curve for all specialties with random forest with a threshold of 0.780

6.5.2 Ophthalmology

For ophthalmology, the Importance Plot can be found in figure 30. When comparing the two measures of variable importance, both give indicate the number of activities in the observed and parallel sub-trajectory, ZPK1, and ZPK4 in their top 6.

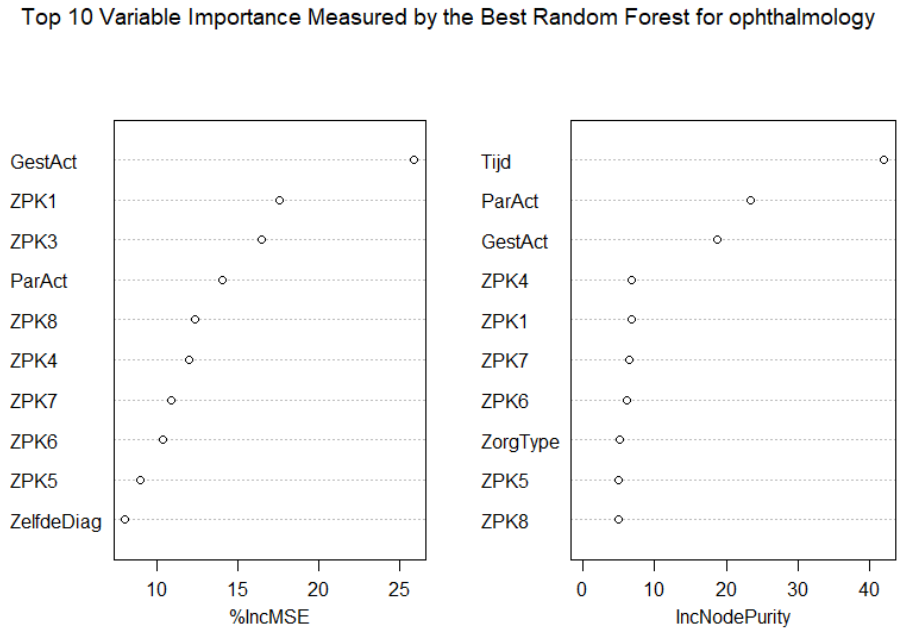

Figure 30: Variable Importance plot for ophthalmology, where the number of care activities in the observed sub-trajectory is important from both plots

The best threshold for the highest specificity found during the algorithm is 0.926, where the accuracy is 0.6114, the sensitivity is 0.5922, and the specificity is 0.8571. When lowering the threshold to 0.805, the desired level of accuracy is reached. With this threshold, the accuracy is 0.7668 and the specificity is 0.500.

Taking this threshold of 0.805 into account, the below confusion matrix is generated.

Ophthalmology	Predicted rejected	Predicted approved
Actual rejected	7	7
Actual approved	39	140

The ROC curve can be found in figure 31 for ophthalmology with a threshold of 0.805. Here again, the AUC is just above 0.5, with 0.552.

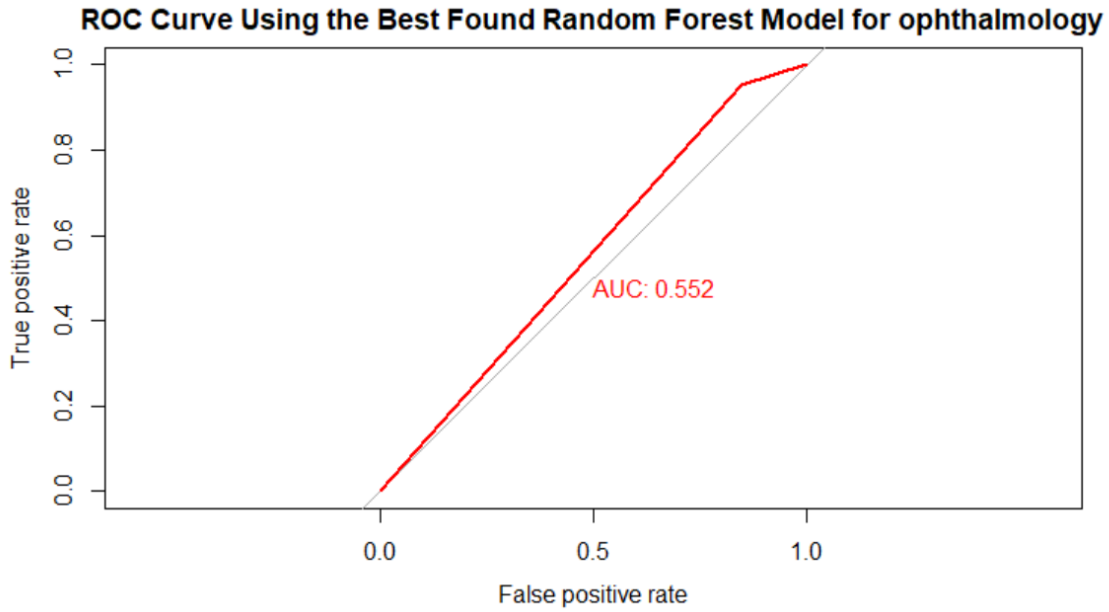

Figure 31: ROC curve for ophthalmology with random forest with threshold 0.805

6.5.3 Surgery

For surgery with random forest, 4 variables are most important comparing the two measures as can be seen in figure 32. These are the number of activities in the parallel sub-trajectory, ZPK4, ZPK6, and ZPK7.

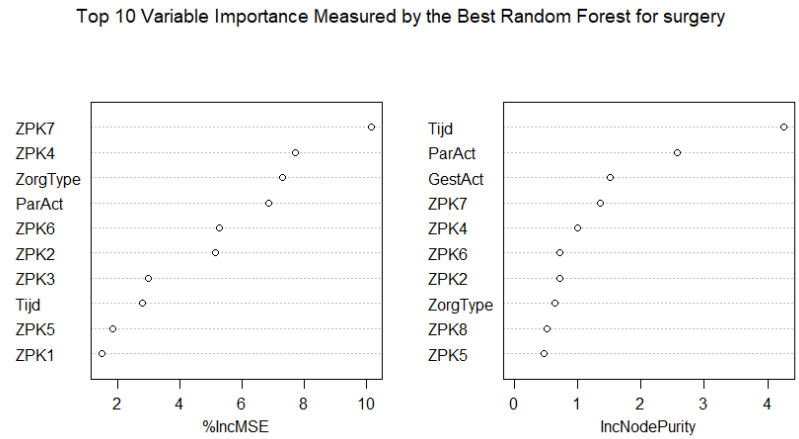

Figure 32: Variable Importance Plot for Surgery

The threshold that is found by the algorithm to achieve the highest specificity is 0.769, the corresponding accuracy is 0.6835, the sensitivity is 0.6765, and the specificity is 0.7273. To increase the accuracy, the threshold should be set at 0.705, then the accuracy and specificity of 0.7595 and 0.3636, respectively, are reached. With this threshold of 0.705, the following confusion matrix with the test data is constructed:

Surgery	Predicted rejected	Predicted approved
Actual rejected	4	7
Actual approved	12	56

The ROC curve is found in figure 33, where the AUC is 0.569.

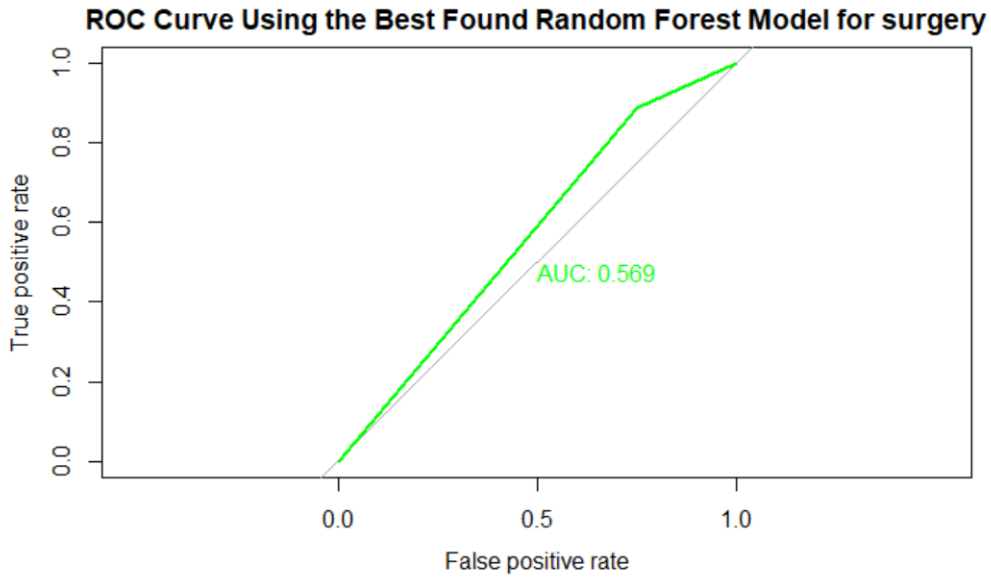

Figure 33: ROC curve for surgery with random forest with threshold 0.705

6.5.4 Internal medicine

Looking at the variable importance plot in figure 34, only ZPK4, ZPK7, and ZPK8 rank in both plots as important.

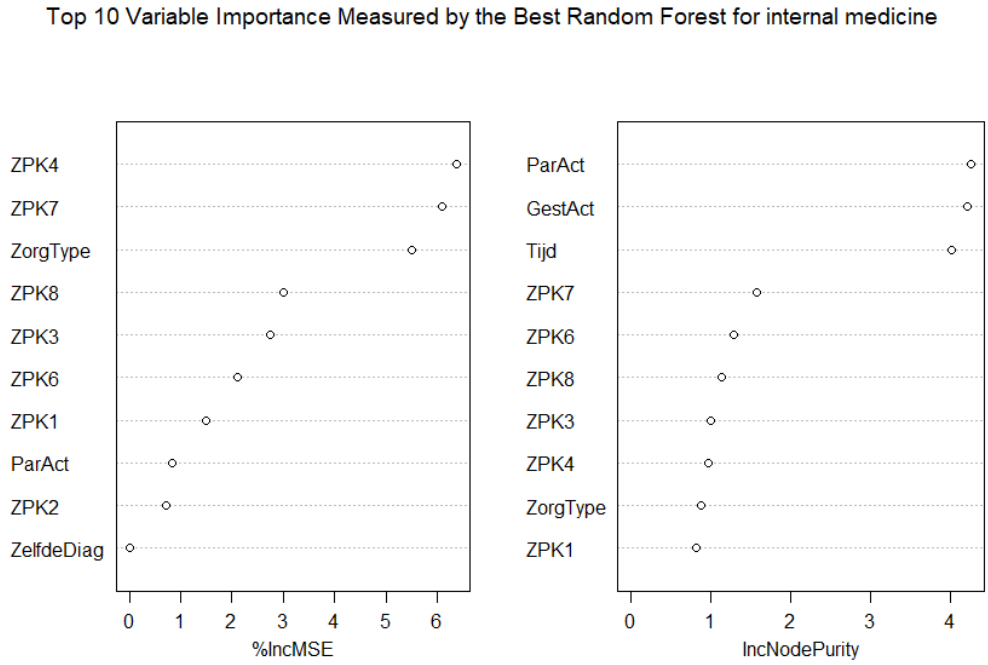

Figure 34: Variable importance plot for internal medicine with random forest, where the top 3 indicate other variables with high importance

The threshold calculated to get the highest sensitivity is 0.878. Using this threshold, an accuracy of 0.7071 is calculated, a sensitivity of 0.6897, and a specificity of 0.8333. When changing the threshold to get an accuracy of at least 0.75, the threshold is decreased to 0.85, and the accuracy belonging to this threshold is 0.7576, with a specificity of 0.6667. Implementing this threshold with the testing data, the following confusion matrix is generated.

Internal medicine	Predicted rejected	Predicted approved
Actual rejected	8	4
Actual approved	20	67

The ROC curve for internal medicine using random forest can be found in figure 35, where the AUC is 0.615.

Figure 35: ROC curve for internal medicine with random forest with a threshold of 0.850

6.5.5 Remaining specialties

The last data set to look into is the remaining specialties. The variable importance plot can be found in figure 36. The number of care activities in the observed sub-trajectory is following both graphs of high importance, also ZPK6 has high importance.

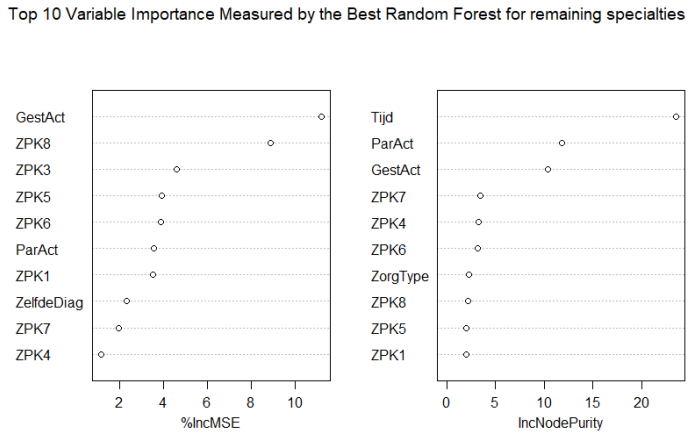

Figure 36: Variable Importance plot for remaining specialties, in the top 3 with the highest importance is both the number of care activities in the observed sub-trajectory

The threshold found for internal medicine in the random forest algorithm to achieve the highest specificity is 0.8483. With this threshold the accuracy is 0.6401, the sensitivity is 0.6235, and the specificity is 0.7381. To compare the model, the accuracy must be at least 0.75. The threshold that returns this is 0.745, then the accuracy is 0.7509, and the specificity is 0.3333. With this threshold, the below confusion matrix is generated.

Remaining specialties	Predicted rejected	Predicted approved
Actual rejected	14	28
Actual approved	44	203

When plotting the ROC curve as shown in figure 37, the AUC is 0.560.

Figure 14: ROC curve for remaining specialties with random forest with a threshold of 0.745

6.6 Comparison between classification tree, logistic regression, and random forest

In this section, we compare the three models with each other. Therefore, table 7 gives an overview of all performance measures. The performance measures are so that the accuracy is close to 0.75, this is done to compare the models. Also, the average is calculated to compare the models' overall sets and to compare the differences between the different data sets over all models.

	Decision tree	Logistic regression	Random forest	*Average*
All specialties	Threshold: 0.76	Threshold: 0.858	Threshold: 0.750	*Threshold: 0.789*
• Accuracy	• 0.7872	• 0.7568	• 0.7553	• *0.7664*
• Sensitivity	• 0.8726	• 0.8272	• 0.8290	• *0.8429*
• Specificity	• 0.2118	• 0.2824	• 0.2588	• *0.2510*
Ophthalmology	Threshold: 0.91	Threshold: 0.880	Threshold: 0.805	*Threshold: 0.865*
• Accuracy	• 0.7668	• 0.7617	• 0.7668	• *0.7651*
• Sensitivity	• 0.7709	• 0.7709	• 0.7821	• *0.7746*
• Specificity	• 0.7143	• 0.6429	• 0.5000	• *0.6201*
Surgery	Threshold: 0.63	Threshold: 0.800	Threshold: 0.705	*Threshold: 0.712*
• Accuracy	• 0.7848	• 0.8101	• 0.7595	• *0.7848*
• Sensitivity	• 0.8971	• 0.9412	• 0.8235	• *0.8873*
• Specificity	• 0.0909	• 0.0000	• 0.3636	• *0.1515*
Internal medicine	Threshold: 0.73	Threshold: 0.850	Threshold: 0.850	*Threshold: 0.770*
• Accuracy	• 0.7879	• 0.7576	• 0.7576	• *0.7677*
• Sensitivity	• 0.8506	• 0.8046	• 0.7701	• *0.8084*
• Specificity	• 0.3333	• 0.4167	• 0.6667	• *0.4722*
Remaining specialties	Threshold: 0.77	Threshold: 0.836	Threshold: 0.745	*Threshold: 0.823*
• Accuracy	• 0.7924	• 0.7543	• 0.7509	• *0.7659*
• Sensitivity	• 0.8866	• 0.8623	• 0.8219	• *0.857*
• Specificity	• 0.2381	• 0.1190	• 0.3333	• *0.2301*
Average	*Threshold: 0.76*	*Threshold: 0.885*	*Threshold: 0.771*	***Threshold: 0.805***
• *Accuracy*	• *0.7838*	• *0.7681*	• *0.7580*	• ***0.7700***
• *Sensitivity*	• *0.8556*	• *0.8412*	• *0.8053*	• ***0.8340***
• *Specificity*	• *0.3176*	• *0.2922*	• *0.4245*	• ***0.3448***

Table 7: All performance measures per model per data set

Table 7 gives an overview of the performance of all the different models. When comparing all these measures, the focus is on the specificity, as the accuracy is at the desired level, and the specificity is most important to Performation. For all specialties, the highest specificity is reached in the random forest model. So if only a model is incorporated that does not divide the different specialties, the random forest is the best model. When looking at the model for ophthalmology, the highest specificity is achieved with the decision tree model. For surgery, the highest specificity is reached with the random forest model.

Predicting the probability of internal medicine cases can best be done with a random forest. For the remaining specialties also the random forest performs best. From this, random forest outperforms the other models, except for ophthalmology. When looking at the averages per data set, ophthalmology has the highest performance. On average the threshold is 0.805, which means that when the probability is below 80.5%, the case should be checked.

6.7 Example

How the above-mentioned models work in practice can best be illustrated with an example from the testing data. An example is:

A patient has two parallel sub-trajectories within ENT (Ear, Nose Throat), and the diagnoses of the two sub-trajectories are different. The observed sub-trajectory has 2 care activities registered, and the parallel sub-trajectory 7 care activities. Both sub-trajectories have a ZPK5, only the observed sub-trajectory has a ZPK1, only the parallel sub-trajectory has a ZPK2 and ZPK6, and neither sub-trajectory has a ZPK3, ZPK4, ZPK7, or ZPK8.

As ENT is one of the remaining specialties, the probability can be calculated using the models for all specialties or the remaining specialties. If we use the classification tree of all specialties, at node 1 the answer is yes, as only the observed sub-trajectory (AG) has a ZPK1. At node 2, the answer is no, as the number of activities is equal to 2, so not less than 2. Therefore the probability using this tree is 0.89. Now let's look at the classification tree for the remaining specialties, Figure 21. At node 1, the answer is yes, as both have a ZPK5 (B). Then we go to node 2, then node 5 as neither has a ZPK3. From node 5, terminal node 11 is reached as the number of care activities in the observed sub-trajectory is equal to 2. So using this tree, the probability is 0.88. For the probability of using logistic regression for all specialties, table 3 is used. The model generates a response of 0.72. Where for the logistic regression for the remaining specialties, the response is 0.65. Lastly looking at the random forest, the response from the all specialties model is 0.48, and for only the remaining specialties this is 0.2805. This case was observed as rejected, so when logistic regression or random forest is used, this case is correctly predicted.

6.8 Threshold analysis

The threshold is chosen such that the accuracy is at least 0.75 and the specificity is as high as possible. In this section, we will give an analysis of how the threshold would change if we set the desired level of accuracy higher and lower. If we look at the random forest model for internal medicine, we see specificity is already high in comparison with the other data sets. In Figure 38 the accuracy and specificity per threshold (with steps of 0.05) are displayed. The dotted line is the desired level of 0.75. As we can see in the graph, the highest accuracy is achieved with a threshold of 0.65, where the specificity is 0.25. At the highest threshold, 0.95, the accuracy is 0.46. So if the desired level of accuracy is set higher, the specificity drops. Both the accuracy and specificity stay around the same values between a threshold of 0.50 and 0.60. From 0.60 to 0.65, the accuracy goes up by 0.03, and the specificity from 0.00 to 0.25. Between 0.65 and 0.85 the accuracy decreases linearly to the threshold, after that, there is an increase in the slope of the decrease. For the specificity, there is no linearity in the slope. It increases with 0.25 between the thresholds 0.60 and 0.65 and between the thresholds 0.80 and 0.85 again an increase of 0.25. If the desired level of 0.75 is more flexible for the hospitals, a higher threshold is chosen with higher specificity. The 0.75 is set in consultation with Performation, however, this is not yet discussed with the hospitals.

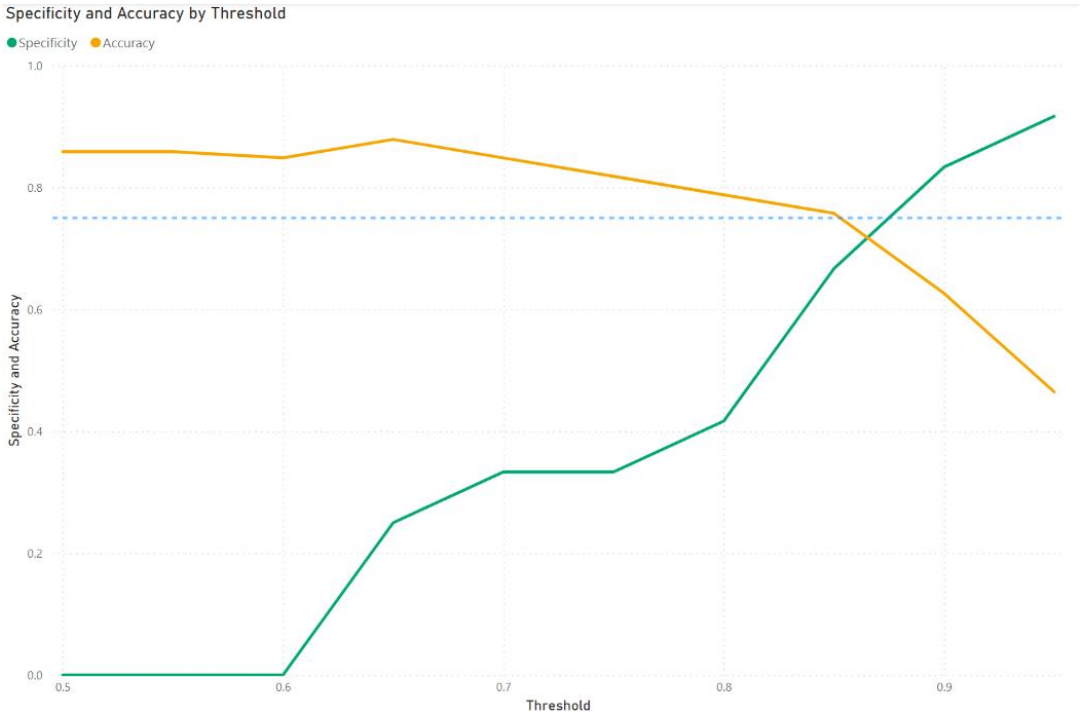

Figure 38: Specificity and accuracy of the random forest for internal medicine for different thresholds

6.9 Sensitivity analysis

The threshold is chosen to achieve the highest possible sensitivity with an accuracy of at least 0.75. This minimal accuracy is set in consultation with Performation, but hospitals can choose to deviate from this and choose a threshold that has a higher specificity. With a higher specificity the workload increases, as more cases have to be checked. To illustrate this, we look at the random forest model for the remaining specialties. This model has a specificity of 0.4245, which is preferred to be higher. In Figure 39 the relationship between the specificity and the number of sub-trajectories to check is given. The number of sub-trajectories to check is given in percentage of the total number. The sub-trajectories that are predicted as 'rejected' are the sub-trajectories that should be checked. This graph gives insight into how a higher specificity influences the workload, which has a linear relationship with the number of sub-trajectories to check. Hospitals have hundreds of parallel sub-trajectories, which can not all be checked because this will take up to much time. With this graph, hospitals can choose what their desired level of specificity is, based on the proportion of sub-trajectories that they can check.

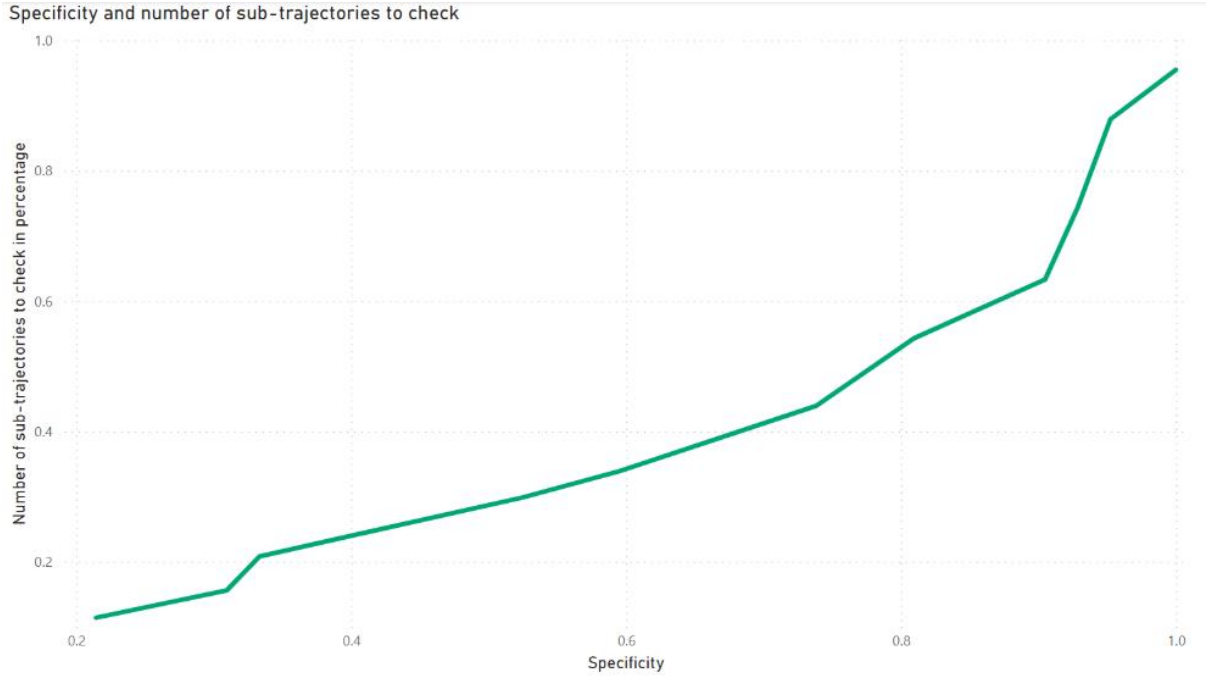

Figure 39: Specificity and number of sub-trajectories to check as a percentage of the total for remaining specialties with random forest

7 Discussion & conclusion

This chapter is divided into 3 sections. Section 7.1 presents the main conclusion of this research. In Section 7.2 the results are discussed and in Section 7.3 future research directions are proposed.

7.1 Conclusion

In this research the following main research question is addressed:

How can statistical learning be used to predict the probability of a parallel sub-trajectory getting wrongfully declared?

To get an answer 7 sub-questions were considered.

1. What are the current challenges for registering parallel sub-trajectories?

The answer to the first sub-question gives insight into the current challenges for registering a parallel sub-trajectory. The difficulties are due to the mismatch between the regulations providing the theoretical guidelines, and the practical cases that have to be checked manually one by one. The challenge lies in that staff needs to be trained to register parallelity correctly. For the doctors, this means only opening a second sub-trajectory when the parallelism holds to all requirements. For the administrative staff, this means that they should check on when the parallelism is registered. A pitfall in wrongly opening an extra sub-trajectory is that it is because at the start it is not sure yet if it is a new care question. Another pitfall in wrongly opening is that the extra sub-trajectory is a complication of the previous care question. Another pitfall lies in that the treatment can be combined with the two care questions, or that the new care question does not involve treatment. The checks on parallel sub-trajectories are on the sample drawn, which can be seen as a lottery. When you have luck, so your sample has fewer wrongly opened parallel sub-trajectories, your costs are lower. When you have bad luck, so your sample has a high number of wrongly opened parallel sub-trajectories, your costs are higher. For self-examination, the sample size is determined by the NZa. For horizontal supervision, this is not the case. Here the hospitals can check as much as they want, but with a minimum that is discussed with the insurers. To lower the impact of the lottery, the sample size should be as big as possible, however, this is not possible as the workload will be too high. Therefore it is important to detect possible wrongfully opened parallel sub-trajectories before the declaration, this leads to a lower chance of mistakes in the sample tests. This detecting right now is done based on the gut feeling of administrative staff on where they think the risk lies for their hospitals.

2. How can we visualize parallel sub-trajectories that have been judged in the sample tests?

The second sub-question concerns how can we visualize parallel sub-trajectories that have been judged in sample tests. The answer to this question is in the form of two dashboards both for horizontal supervision and self-examination. The outcomes can be found in Chapter 3. The most important takeaways from this are that the specialties of surgery, ophthalmology, and internal medicine have the highest number of observations combined in the sample tests. It was already expected that these specialties were the biggest, as there is a lot of parallelity possible and aloud within these specialties. On average, the error margin of horizontal supervision is lower than for self-examination. When looking at the 10 biggest specialties regarding the number of parallel observations for self-examination, the error margins for pediatrics (KIN), neurology (NEU), and urology (URO) are 17.92%, 20.00%, and 24.64% respectively.

3. What literature is available on using statistical learning to predict a binary outcome?

The third question regards the literature review of how statistical learning can be used to predict a binary outcome. This question is answered in Section 4.2. For this thesis, the focus is on supervised learning, and three models are chosen and incorporated based on their appliance and ease of implementation.

4. What models are suitable to use in this research?

This is the answer to the fourth question of what models are suitable for this research, which are decision trees, logistic regression, and random forest. The decision tree is easy to interpret, which makes them especially useful for cases that require transparency and explainability. For the decision tree, the disadvantage that is important to take into account is that this algorithm is prone to overfitting. This disadvantage is where the random forest comes into play. As this algorithm overcomes this by averaging the predictions of multiple trees. However, this algorithm is harder to interpret. Lastly, logistic regression is modeled, for this algorithm the advantages are that it is computationally efficient and provides interpretable results.

5. Which quantitative variables can be used for predicting the outcome?

The fifth question is about what quantitative variables can be used for predicting the outcome. In the electronic patient record, a lot of variables are stored, however, in this research, the focus is on the care type, the overlap time, the diagnosis, the number, and what type (ZPK) of care activities are performed in the sub-trajectories. The importance of the variables can be compared by looking at the variable selection per data set per model. The number of care activities registered in the observed

sub-trajectory is included in most models, followed by the overlap time. The least included variables are the care type, ZPK2, ZPK5, and ZPK8. From these four, the low importance of ZPK5 is interesting. As ZPK5 states whether the sub-trajectory is operative or not. From hospital experience, the presence of ZPK5 is checked when searching for wrongful parallelism. The reason why this might not be a result of the models is that some specialties, internal medicine, for example, are reflective specialties. This means that there are no operative activities, so no ZPK5. It might be that the combination between certain diagnoses and the presence of ZPK5 could make be more important, however, this was not checked during this thesis.

6. What are the results of the models?
The sixth question regards the results of the model. The specificity should be as high as possible with an accuracy of at least 0.75. This is to keep the number of false positives as low as possible, but not classify too many cases wrong, so both the number of false positives and the number of false negatives should be low. From the 3 models, the random forest outperforms the other in most cases. This is because this is a more iterative process, and therefore less sensitive to outliers meaning that random forests mitigate the impact of noisy data by aggregating predictions from multiple trees. The data was also split to improve the performance. First, we take a look at the performance of the models including all specialties. The highest performance is reached with the logistic regression algorithm, so if only one model is implemented into Notiz which should take into account all specialties, the advice is to use a logistic regression model. If multiple models can be implemented into Notiz, with division into the biggest specialties different algorithms that perform best for each specialty can be used. For ophthalmology, this is the decision tree that has the highest specificity. As the decision tree is easier to implement into Notiz, there is no need necessary between link Notiz and R. For surgery the advice is to implement the random forest model based on the surgery training data set, as the performance of this algorithm is the best. It can be discussed whether surgery should have its own algorithm, rather than using the logistic regression of all specialties. However, this is not investigated in this research. For internal medicine, the random forest has the highest specificity, therefore the advice is to implement this into Notiz with a link to R. For the remaining specialties, the highest specificity is achieved with the random forest model.

7. In which way can the model help improve care registrations?
The last sub-question is how the model can improve care registration. The answer to this is that this model can help detect registrations that have a higher risk based on historical observations of being rejected. Performation helps hospitals to get a higher rate of 'first-time-right registration', this starts with the doctors that have to register care. This research helps the administrative controllers to determine which cases should be checked, to detect the registrations that are opened parallel wrongfully. For Performation this research introduces the opportunity to involve a prediction model in Notiz for parallel sub-trajectories at risk of getting wrongfully declared. The further contributions of this research can be found in the practical contributions in Section 7.2.3.

Let's combine the answers to the sub-questions into one that answers the main research question. The most suitable statistical learning method depends on the data that is modeled. For all specialties logistic regression performs best, for ophthalmology this is the decision tree, and for surgery, internal medicine, and the remaining specialties the random forest is best. The advice to Performation is to implement the ophthalmology decision tree, and for the other algorithm import more data to get a higher specificity. The outcome of the model returns a probability of a parallel sub-trajectory is opened correctly. The lower the probability, the higher the chance that this parallel sub-trajectory is wrongfully opened, and will be rejected during checking of sample observations. This probability can be used in the decision making which parallel sub-trajectories should be checked before the declaration. This model can be applied in two ways. On one hand, hospitals can check all the cases with a probability lower than the advised threshold or the chosen threshold by the hospital. The advised threshold is when the accuracy is at least 0.75 and the highest specificity is achieved. The chosen threshold by the hospital can depend on their current error margin, where a lower error margin will probably lead to a lower threshold, or depend on the other checks that are already in place, where more checks will probably lead to a lower threshold. On the other, hospitals can start with the cases with the lowest probability and use their time on the cases with the highest risk.

7.2 Discussion

7.2.1 Limitations

The total available data is 5150 judged observations, with 3070 observations for horizontal supervision and 2080 observations for self-examination. However, when splitting the data into different sets based on specialties, the smallest total data set is 261 observations. Where 182 observations are for the training set and only 79 for the testing set. When looking at the performance measures in Table 7, the performance fluctuates between the different sets. With more available data from self-examination, better estimates and predictions can be made. It is hard to indicate how much more reliable the findings will be if the data sizes increase. However, with more data, improvements in specialty-specific models can likely be made. However, ophthalmology has the most observations, i.e. 596 observations. This is 2.5 times the number of observations than for surgery. Whether the number of observations is related to a better performance of the models is not easy to say, but when adding

more observations for the surgery specialty, hopefully, the performance will go to the same level as the performance of ophthalmology.

In total 3 different models were implemented. Per the data set the performance of the different models will be discussed. In general, the random forests perform best concerning the highest specificity, with an average of 0.4245. This can be due to the multiple subsets that were used. Cross-validation can be a technique that can make the predictions average over multiple folds. For random forests, the complexity is the biggest limitation as this outcome is harder to interpret and implement in the real world. This is because random forest uses multiple trees with varying structures, therefore not one clear decision tree can be constructed. This means that for the implementation in Notiz, the data must be run through the model, instead of implementing into the application itself. So a link need to be established between Notiz and R.

For the decision tree, a general limitation is that it tends to overfit the training data too closely. If we look at the decision tree for ophthalmology, we might conclude that this is happening here. We see that there are a lot of regions, also known as terminal nodes, with less than 5% of the observations. However, the best-performing model is the decision tree for ophthalmology, so even if overfitting might have happened, the performance did not suffer from it. Logistic regression is limited to linear relationships between the input variables and the output variable where this relationship might be more complex. Linear relationships in logistic regression mean that a one-unit change in the value of a predictor variable is associated with a constant change in the log odds of the binary outcome, regardless of the initial value of the predictor variable.

The higher the specificity, the lower the number of false positives, which is the goal of this research. However, accuracy is also important to keep the workload for administrative staff in mind. The desired level of specificity was not specified during this research. When in the future improving this model, or when comparing a new algorithm, a desired level of specificity must be set.

The AUC of the ROC is for all models low, this indicated that the model's predictive power is not much better than random chance. As already found in the literature study, such a model can still be used. As the model is used to indicate parallel sub-trajectories that have a higher risk of being wrongfully opened, so rather as a decision-making tool which sub-trajectories to check, than to predict which sub-trajectories are wrongfully opened. The impact of a false negative is low, as the consequence is that the administrative employee had to put time and effort into checking a case that was wrongfully classified as wrongly opened, however, there are no financial consequences. The impact of a false positive is bigger, as this has financial consequences. However, this financial consequence would also be encountered if no sub-trajectories were checked before the declaration. This model incorporates multiple factors that can contribute to wrongful parallel registration detection. Now this is only done based on assumptions of what might indicate risks. An example of this is that both diagnoses or one can be a general diagnosis and not specific. Even though the AUC is around 0.5, this research still provides insights and patterns in that data that were not previously known. With updating the model when extra data is added, the AUC hopefully also increases. In all the ROC curves there is only one kink in the line encountered. When the ROC curve should have been nearly straight, it indicates that the model's predictive power is similar across different classification thresholds. With only one kink, it typically indicates that the model's predictive performance changes at a specific threshold. When setting the threshold manually, it was already encountered that adjusting just a few 0.001 already could make a big difference in the performance measures. This can be explained that the main focus is on categorical variables and that the probability therefore can only take several different values. If we for example look at the logistic regression model for surgery, only two binary variables (as ZPK6 can only take one value, so these three variables are connected) are taken into account, which leads to only 8 possible outcomes. As the possible outcomes are not evenly distributed, the performance measures are not increasing smoothly when adjusting the threshold. Another possible reason for the shape of the ROC curve is the imbalance between the distribution of positive and negative cases in the data. In our data the number of negative cases is much smaller than the number of positive cases, this can lead to the classifier being biased to the positive class. Therefore, it is important to balance the dataset in future research, so leaving out positive cases. Which case should be left out is a question that should be investigated and answered together with Performation and the hospitals. Some cases are almost always correct, but the specialties that are already more sensitive to mistakes can be chosen to be taken into consideration.

7.2.2 Theoretical contributions

During the literature review, we noticed that there is not much research done on predicting the correctness of care declarations. However, if we take a step back, there have been a lot of studies performed that use statistical learning to predict a binary outcome. Where within the study field, so with the use of a care trajectory, it is more common to predict the probability of whether the patient will die from a specific disease or not. This study shows that it is possible to predict the probability of whether a parallel sub-trajectory is getting wrongfully declared or not,

7.2.3 Practical contributions

For Performation, this research contributes to improving checks on parallel care registration. The implementation of this research can be two-sided. On one hand, Performation can implement that all the parallel sub-trajectories with a high risk (based on the threshold that is chosen by the hospitals, but advised is 0.805 on average) are marked as high risk and these registrations should be checked. On the other hand, Performation can implement the probability calculated by the model for every parallel sub-trajectory in Notiz. Here also the financial impact can be incorporated. With the latter options, hospitals can choose which parallel sub-trajectory they want to check based on the combinations of probability and what the financial impact is when the sub-trajectory would be rejected. The specificity is in most cases not very high, i.e. below 0.5, however, for Performation this is not a dealbreaker. As this model will be implemented as a starting model, where more data can be added to improve the performance. The competitors of Notiz do not include statistical learning to detect wrongfully declared parallel sub-trajectories. So even with a lower specificity, this gives Perfomation a competitive advantage.

Not every hospital would benefit equally from this research, as some hospitals already have a low error margin, therefor the extra time that has to be put into the extra check might not be worth it. The check is for the hospital that does not have a grip on the parallel sub-trajectories, where also the dashboard can come in handy to decide which hospitals could use some extra checks to detect parallelism at risk of getting wrongfully declared. The results of the dashboards are partly already discussed with the hospitals, and most of them asked for a follow-up where their consultant together with the administrative staff will indicate the risks, and make a personal improvement plan. The biggest win is for the hospitals that have higher error margins, as they have bigger financial benefits from detecting wrongfully opened parallel sub-trajectories. For hospitals that have a lower error margin, it is still useful to detect the risk full sub-trajectory as every mistake has its financial impact, so 1 big mistake can result in higher costs than multiple smaller mistakes. Hospitals can also learn from each other, as some hospitals perform better than others. However, it is hard to copy their way of working as better performance can have multiple reasons. The obvious reason is that the doctors are well-informed about the regulations and know how to register parallel care correctly. Another reason is that the administrative staff already performs more checks to indicate risk cases. Parallelism also occurs more frequently for some diagnoses and specialties, and not all hospitals have the same specialties, therefore, the observations per hospital can not always be compared easily. The last possible reason is under-registration. This means that too little care is registered: where the hospital was allowed to register parallel, but chose not to. This leads to a lower error margin, however, financially this is not covering the costs. So a lower error margin does not directly imply completeness and lawfulness of care registrations.

7.3 Further research

The prediction models do not take diagnoses into account, however, this could substantially contribute to the performance of the different models. The diagnoses are not incorporated as there are too many different diagnoses per specialty. To give an idea of this, for horizontal supervision in total 716 different diagnoses are present and for self-examination, this is 577. The diagnoses should be grouped or the number of observations per diagnosis should be increased. The grouping of diagnoses can be based on the type of diagnosis, so general diagnoses can be grouped, but also groups can form diagnoses that are related to each other. An example is to group the diagnoses into general diagnoses, acute diseases, and chronic diseases. If we were to group the diagnoses within specialties, orthopedics is a good example. Here you can group all diagnoses that are connected to for example the left arm, right arm, left leg, and right leg. Another research that can be incorporated is text mining, as the check has to be supported by the report of the care activities. With this research, this is not done as the goal was to look at quantitative variables. Also, other models can be tested to improve the performance, such as support vector machines or artificial neural networks. These machine-learning models were not included in this research. The model can be improved in a way that the accuracy stays at the desired level, but that the specificity for all data sets is at a higher level. However, in these models, the data is imbalanced, which leads to the sensitivity influencing the accuracy bigger than the specificity. As already to be said, to overcome this limation, the data should be balanced, so the number of rejected and approved sub-trajectory should be around the same size.

The outcome, the judgment on whether the observed sub-trajectory is approved or rejected, is based on experts' opinions, thus the human factor might make it hard to find clear relationships. Within the current situation, it isn't known what the total number of parallel sub-trajectories is or what hospitals already do to detect and check wrongful parallel sub-trajectories. This knowledge can be used to explain the differences between the hospitals. If the hospital already does multiple checks to detect possible mistakes, the error margin in the sample test will probably be lower than when no checking was done. What can be done, is also incorporate these prior checks into the data. With this also more rejected cases will be present in the data set. In future research, it is also important to have conversations with the hospitals and how they explain their error margin. By doing this, more insight can be gained into which ways controlling rightful parallelism are resulting in a lower error margin. In the future the dashboards can be extended to give more feedback to the hospitals, so shifting to benchmarking.

Bibliography

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Gelder, C. de. (2022, September 30). *Helft budget ziekenhuizen gaat niet naar patiënten: "Niet slecht, maar simpelweg nodig."* EenVandaag. https://eenvandaag.avrotros.nl/item/helft-budget-ziekenhuizen-gaat-niet-naar-patienten-niet-slecht-maar-simpelweg-nodig/

Hastie, T. and T. R. and F. J. H. and F. J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Horizontaal Toezicht Zorg. (2023). *Een vorm van samenwerking tussen de zorgverzekeraars en een zorgaanbieder*. https://www.horizontaaltoezichtzorg.nl/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition*.

Koen Kuijper. (2022, March 28). *Dit heeft 16 jaar marktwerking in de zorg ons gebracht*. Zorgwijzer. https://www.zorgwijzer.nl/zorgverzekering-2023/dit-heeft-16-jaar-marktwerking-in-de-zorg-ons-gebracht

Koenraadt, B. (2022, September 20). *Miljoenennota 2023: dit zijn de gevolgen voor je zorgverzekering*. Zorgwijzer. https://www.zorgwijzer.nl/zorgverzekering-2023/miljoenennota-2023-dit-zijn-de-gevolgen-voor-je-zorgverzekering

Microsoft. (n.d.). *SQL Server Management Studio (SSMS)*. Retrieved March 7, 2023, from https://learn.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms?view=sql-server-ver16

Nagtegaal, F. (2018, January 9). *Horizontaal Toezicht leidt tot enorme daling herdeclaraties*. NVZ Kennisnet. https://www.nvz-kennisnet.nl/nieuws/811-horizontaal-toezicht-leidt-tot-enorme-daling-herdeclaraties

Nederlandse Zorgautoriteit. (2022). *Regeling medisch-specialistische zorg-NR/REG-2207a*. http://puc.overheid.nl/doc/PUC_652304_22

NZa. (2022). *Regeling medisch-specialistische zorg - NR/REG-2306a*. https://puc.overheid.nl/nza/doc/PUC_720738_22/1/

Performation. (n.d.-a). Internal communication. In *2022*.

Performation. (n.d.-b). *Zorgregistratie en zorgadministratie*. Retrieved September 13, 2022, from https://performation.com/zorgadministratie/

Pfizer. (n.d.). *Uitgaven aan dure geneesmiddelen*. Retrieved April 3, 2023, from https://www.pfizer.nl/zorgthema/uitgaven-aan-dure-geneesmiddelen#:~:text=Volgens%20de%20definitie%20van%20de,per%20pati%C3%ABnt%20per%20jaar%20kost.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*.

Zorgwijzer. (n.d.). Welke zorgverzekeraars zijn er? *Zorgwijzer*. Retrieved March 3, 2023, from https://www.zorgwijzer.nl/faq/welke-zorgverzekeraars-zijn-er

Appendix

List of Abbreviations

NZa = Nederlandse Zorgautoriteit
DBC = Diagnose-Behandelcombinatie
NVZ = Nederlanse Vereniging van Ziekenhuizen
ZN = Zorgverzekeraars Nederland
NFU = Nederlandse Federatie van Universitair Medische Centra
FMS = Federatie Medisch Specialisten
ALL = Allergologie
ANE = Anesthesiologie
CAR = Cardiologie
CHI = Chirurgie
DER = Dermatologie
GER = Klinische geriatrie
GYN = Gynaecologie
INT = Interne geneeskunde
KIN = Kindergeneeskunde
KNO = Keel-Neus-Oorheelkunde
LON = Longgeneeskunde
MDL = Maag-Darm-Levergeneeskunde
NEU = Neurologie
OOG = Oogheelkunde
ORT = Ortopedie
PLA = Plastische chirurgie
PSY = Psychiatrie
REU = Reumatologie
RTH = Radiotherapie
URO = Urologie
AG = Only in observed sub-trajectory
AP = Only in parallel sub-trajectory
B = In both observed as parallel sub-trajectory
GvB = In neither observed nor parallel sub-trajectory
GestAct = Number of activities in the observed sub-trajectory
ParAct = Number of activities in the parallel sub-trajectory
Tijd = Overlap time

R Code

In this part of the Appendix the layout of the code used to get to the results will be explained.

Librarys

```
library(readxl)          # read excel files from finder
library(tidyverse)       # data wrangling and visualization
library(knitr)           # beautifying tables
library(broom)           # for tidy model output
library(effects)         # for probability output and plots
library(rpart)           # decision tree
library(rpart.plot)      # plot of the decision tree
library(ggplot2)         # logistic regression plot
library(caret)           # for confusion matrix
library(pROC)            # AUC under ROC
library(ROCR)            # performance log model
library(caret)           # specifivity and sensitivity
library(randomForest)    # for random forest
```

Splitting the data based on specialties

```
# filter datasets
DfOog <- filter(DfNamen, Specialisme == "OOG")
DfChi <- filter(DfNamen, Specialisme == "CHI")
DfInt <- filter(DfNamen, Specialisme == "INT")


# get data set of other specialties
OverigSpecialisme <- c("KNO", "PLA", "ORT", "URO", "GYN", "DER", "KIN", "MDL",
"CAR", "LON", "REU", "ALL", "PSY", "NEU", "GER", "RTH", "ANE")
DfOverig <- filter(DfNamen, Specialisme %in% OverigSpecialisme)
```

Get training set (example)

```
# get training set
trainTotaal <- sample(1:nrow(DfNamen), size = 0.7*nrow(DfNamen))
DataTotaal <- mutate(DfNamen, Train = if_else(row_number() %in% trainTotaal,
"Train" , "Testing"))
```

Calculate error margins (example)

```
# calculate error margins
errorDfTotaal <- (sum(DataTotaal$IndGoedgekeurd == 0))/nrow(DataTotaal)
```

Decision tree (example)

```
# decision trees
TotaalTree <- rpart(IndGoedgekeurd ~ ZorgType + Tijd + ZelfdeDiag + GestAct +
ParAct + ZPK1 + ZPK2 + ZPK3 + ZPK4 + ZPK5 + ZPK6 + ZPK7 + ZPK8, data =
TrainDataTotaal)
  rpart.plot(TotaalTree, main = "Classification tree for all specialties",
cex=1, extra=101,box.palette="RdBu", shadow.col="gray", nn=TRUE)
```

Logistic regression (example)

```
# Logistic regression model
Totaalfit <- glm(IndGoedgekeurd ~ ZorgType + Tijd + ZelfdeDiag + GestAct +
ZPK1 + ZPK2 + ZPK3 + ZPK4 + ZPK5 + ZPK7 + ZPK8, data=TrainDataTotaal, family =
binomial)
Totaalfit_back <- step(Totaalfit, direction="backward")
```

Random forest (example)

```
# try different values for m. This sequence includes m = 5 which is
approximately the square root of the number of predictors of 29
  for (m in seq(3,29,2)) {
    # fit a random forest model
    randomforest.TotaalFit <- randomForest(IndGoedgekeurd ~ ZorgType + Tijd +
ZelfdeDiag + GestAct + ParAct + ZPK1 + ZPK2 + ZPK3 + ZPK4 + ZPK5 + ZPK6 + ZPK7
+ ZPK8, data = TrainDataTotaal, mtry = m, importance = TRUE)
```