Online Operational Planning in a Multi-Provider Ambulatory Allied Healthcare Organisation:

Developing an online appointment scheduling support procedure employing predictive modelling of location occupancy.

Public version: References to the company, the specific field of work and some confidential information are hidden from the reader

Milan Westerink

A thesis presented for the graduation of
Industrial Engineering and Management (Msc)

May 2021

University of Twente
1st supervisor:
Dr.ir. A.G. Leeftink
2nd supervisor:
Prof.dr.ir. E.W. Hans

Company X
Supervisor:
Supervisor X

UNIVERSITY OF TWENTE.

Management summary

In this thesis we propose a novel dynamic scheduling procedure for the company that we refer to as "Company X" due to privacy restrictions, which is an ambulatory allied healthcare organisation operating out of the Netherlands. The organisation has approximately 300 locations and 350 employees, of which about 100 treat patients. They are the market leader in their field and their collaboration with affiliated companies in different parts of the supply chain make it a unique organisational context. The organisation is experiencing rapid growth, which inevitably entails challenges. The problem experienced by Company X that this research attempts to solve is the following:

"The access time for new patients is higher than the desired 3 weeks in 35.7% (2019) of the cases, whereas Company X wants this to be at most 20%."

We found the following core problems to be the cause of the problem mentioned above:

- Schedulers often do not ask patients all questions needed to adequately classify what type of appointment they need;
- There is no information on expected future demand;
- There are no guidelines defining which time slots to book for an appointment;
- There is no information about (expected) occupancy per location;
- There are no guidelines on desired occupancies at given times.

These problems can be aggregated to the problem of inadequate scheduling support provided by the scheduling system. In consultation with Company X we decided that a new scheduling support tool is the preferred solution.

The resulting research goal is as follows:

"How can we develop a scheduling support prototype that suggests the most preferred sessions and time slots to schedule the appointment in, to ensure that 80% of new patients get an appointment within 3 weeks?".

To reach this goal we propose a novel two-stage decision support procedure. The proposed scheduling procedure is applicable to multi-provider online (i.e. dynamic) scheduling systems with patient pooling. We assume deterministic processing times of appointments for assessment of performance measures. The first stage consists of a top-k session (unique combination of date, practitioner and location) selection procedure, whereas the second stage is a complete ranking procedure of all available appointment options in the session selection from the first stage. We refer to these stages as Stage 1 and Stage 2, respectively. In both stages we use a method called "Technique for Order of Preference by Similarity to Ideal Solution" (TOPSIS) to rank options. We obtained weight sets for these Multi-Criteria Decision Analysis methods, with a direct scaled weighting procedure. Each stage has a unique weight set for its TOPSIS procedure, as we consider different criteria per stage. To demonstrate the procedure we construct a prototype tool. Company X will not use the front-end of our prototype, since they want the front-end to meet requirements that are yet to be fully defined. The main contribution of our research is therefore the back-end of the proposed tool.

One of the criteria we consider in both stages is the expected utilization of the session in question on the appointment date. We define the utilization of a session as the total appointment time scheduled divided by the initially available session time. To predict how this utilization develops per session in the month to come, we construct a predictive model that predicts the probability of an above-average utilization $P(utilization > 0.922)$, where 92.2% is the average utilization of 2019. We mainly base our research on data of 2019, because the Covid-19 pandemic corrupted the validity of 2020 data. We construct the proposed predictive model using binary logistic regression. The dependent variable of this model, $P(utilization > 0.922)$, functions as an indicator of how crowded a location will be on the appointment date.

We use three methods to evaluate the performance of the proposed method. The first method is a real-life instance simulation. We reconstruct the schedule as it was on 1st July 2019. Then we reschedule all appointments that came in the following month using the proposed procedure and compare the performance on the KPIs of this simulated month to that of the historic data. The second method is a theoretical instance simulation to evaluate the robustness of the tool. In this method we simulate the same month as in the real-life instance twice. However, we respectively use a much larger and much lower number of appointments. In all instance simulations we have a simple representation of patient choice. We do this by randomly picking from the top 5 suggested appointments. The third evaluative method is Data Envelopment Analysis (DEA) to evaluate the effectiveness of the ranking procedure.

Table 1 shows the Key Performance Indicators (KPIs) that we consider and the current performance compared to that of the real-life instance simulation. Mind that the simulation is only for a month and that the system does not settle into an equilibrium in this amount of time, resulting in a lower utilization because the new scheduling procedure schedules some appointments further into the future. We define the fragmentation index as the number of gaps in the schedule of a session (breaks do not count as gaps). In our research we divide patients into three categories: New Patients (NPs), Recurring Patients (RPs) and Periodic Patients (PP). We consider different KPIs per patient type as Table 1 reflects.

KPIs	**Current performance**	**Real-life instance simulation**
Company X perspective		
Utilization of session time [a]	91.6%	90.9%
Variance of utilization per practitioner workday	1.20%	1.03%
Fragmentation index	1.26	1.15
Patient perspective		
NP access time (days)	20.7	14.8
NP access time violations	35.7%	15.9%
RP/PP days deviation from target date [b]	-	3.36
Distance to the appointment location (km)	3.61	3.60

[a] In equilibrium this value is equal to that of the current situation, given the number of sessions stays the same.
[b] Currently, schedulers do not document the target.

Table 1: Key Performance Indicators

The real-life instance simulation shows that we can reduce the percentage of NPs with an access time (day of scheduling until day of appointment) of over 3 weeks from 35.7% (July 2019) to 15.9% (margin of error 1.9%) by using the proposed scheduling tool. This means we meet the goal of 20% that we

had set. We also see that the proposed method spreads appointments more, instead of myopically scheduling as soon as possible subject to ill-considered filters. This shows itself in a reduction of the variance of utilizations from 1.20% to 1.03% (margin of error of 0.07%). The fragmentation of the proposed scheduling procedure reduced from 1.26 in real life to 1.15 (margin of error of 0.056) in the simulation. This improved spread of appointments and lower fragmentation allows for higher average utilizations, which means less sessions are needed to accommodate the same number of appointments, thus reducing costs. Furthermore, the proposed scheduling method ensures the quality of appointments, since we always show the most preferred options considering the patient's needs and Company X's needs. The fact that the patient gets a high-quality appointment instead of one that is myopically selected with incomplete information, means that patients are more likely to be satisfied with their appointment. This improved patient satisfaction means that patients are less likely to cancel their appointment, less likely to not show up for their appointment and more likely to return for another appointment, leading to a higher revenue. The extent of these additional revenues and reduced costs are not yet known, since they are dependent on factors that Company X does not measure, like patient choice, scheduler behaviour and patient satisfaction. The theoretical instance simulation and DEA showed that the proposed scheduling tool is robust and effective.

Our contribution to literature is threefold:

- This research serves to fill a gap in literature concerning multi-location online appointment scheduling systems with patient pooling.
- We propose a multi-attribute ranking-based approach to appointment scheduling, which is not yet present in literature.
- Our research taps into a yet largely unexplored area of combining patient choice and provider choice in a mediating intelligent scheduling support tool.

Acknowledgements

Before you lies the master's thesis, which preludes my graduation of the master Industrial Engineering and Management at the University of Twente. This master's assignment was issued by the company that we refer to as Company X, due to privacy restrictions. The accommodating department was the financial administrations department. The six months I spent working on this thesis would have been a lot harder if not for the help I have received. Therefore, I would like to take this opportunity to express my gratitude to those that have aided me.

I would like to thank my supervisor at Company X, for always showing support and interest throughout my research. Apart from making me feel welcome, he has always made time to ensure my research was going well and that I had all the information I needed. Furthermore, I want to express my gratitude for all people in the financial department with whom I shared an office, for making me feel welcome right from the start. Your friendliness contributed to me accepting a job at Company X after my graduation. I am excited to commence this new chapter of my life with Company X and look forward to being your colleague.

My gratitude also goes out to my university supervisor, Gréanne Leeftink, for guiding me throughout my research. Your feedback has greatly helped me delimitate/structure my research and you have continuously steered me in the right direction from the moment you introduced me to the company. I also want to thank my second supervisor, Erwin Hans, for helping me with validating my research and advising me on how to transfer the message of my thesis to the readers. To both of my university supervisors I want to say that I have very much enjoyed our collaboration. Furthermore, your help greatly contributed to the overall quality of my thesis and as a result I feel proud of the work I have delivered. Therefore, I thank you.

In conclusion, I would like to thank my family and friends. The support (and sometimes distractions) that they gave me have not only been invaluable in the last six months, but also made my time as a student very enjoyable. I hope all of you enjoy reading this thesis.

Glossary

Access Time Number of days between the patient's appointment request and the date that the appointment takes place. 3

Action Problem A discrepancy between the norm and reality, as perceived by the problem-owner [19]. 3

Bayesian Inference Bayesian inference is a method of statistical inference that treats model parameters as random variables rather than as constants [41]. 35

Confidence Interval Proposes a range of plausible values for an unknown parameter (for example, the mean). A 95% Confidence Interval (95%-CI) means that the proportion of confidence intervals (based on random samples of the same population with the same size) that contain the true value of the unknown population parameter is 95%. 105

Cross-Validation A statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model [42]. 8

Deterministic Not influenced by randomness. 5

Fragmentation The occurrence of several unused spaces in the schedule without enough room to allocate appointments to. 6

Frequentist Inference Frequentist inference is a method of statistical inference that treats model parameters as constants [41]. 35

Heuristics A method used to solve a problem by trial and error when an algorithmic approach is impractical. Heuristics often have an intuitive justification [53]. 5

Knowledge Problem A description of the research population, the variables and, if necessary, the relationships that need to be investigated [19]. 7

Logistic regression Logistic regression is a statistical modeling method that takes the natural logarithm of the odds as a regression function of the predictors [28]. 48

Machine learning The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience [36]. 48

Multiple linear regression In linear regression the dependent variable is modeled as a linear function of a set of regression parameters and a random error [57]. Linear regression can be split into simple linear regression and multiple linear regression. The first uses one predictor variable and the second uses multiple. 48

Offline scheduling An algorithm is called an offline algorithm if it processes its input as one unit and the whole input data are available at the moment of the start of execution of this algorithm [13]. 5

Online scheduling An algorithm is called an online algorithm if it processes its input piece by piece and only a part of the input is available at the moment of the start of execution of this algorithm [13]. Online scheduling is also sometimes referred to as "dynamic", "sequential", "myopic and sequential", and "rolling horizon" scheduling [34]. 5

Patient Choice Analysis Analysis of patients' perceptions of the acuity of their need, time-of-day preferences, and degrees of loyalty toward their designated primary-care provider and the corresponding effects on the schedule [17]. 8

Pre-emption If a job is preemptable, then the execution of this job can be interrupted at any time without any cost, and resumed at a later time on the machine on which it was executed before the preemption, or on another one. Otherwise, the job is non-preemptable [13]. 9

Probability Theory The branch of mathematics concerned with probability. 8

Problem Cluster A model used to map different problems and their mutual relationships. A problem cluster serves as a means of structuring the problem context, and is used to identify the core problem [19]. 3

Reservation Levels The proportion of a future day that should be reserved for potential incoming appointments. 8

Scheduling A term used in planning and control to indicate the detailed timetable of what work should be done, when it should be done and where it should be done [45]. 2

Session A working day of a practitioner at a given location in the schedule. If a practitioner works at two locations on a day, we describe this as two sessions. 3

Stochastic Influenced by randomness. 5

Contents

Chapter 1

Introduction

Company X is an ambulatory allied healthcare organisation that specialises in -confidential-. They do so in partnership with many other closely related specialties to attempt to provide the optimal care and service for a varying patient base expressing issues with -confidential-. Company X is the leading organisation in their field in the Netherlands and has seen substantial growth in the last decade. -confidential-. From 2012 to now they have grown from 63 locations to over 300 (as depicted in Figure 1.1) and are currently the national leader in their field. This growth entails an increasingly intricate web of operational challenges.

Figure 1.1: Locations projected onto map of the Netherlands (instance of 04-12-2020)

One of the organisational aspects where they notice these operational challenges is the scheduling of appointments. With increasing number of locations, employees and patients, it is getting more difficult to ensure efficient and effective schedules. In this section we lay a solid foundation for the remainder of this research. We do this by identifying what we need to research and determining how we should structure the research. First, we discuss the motivation for this research in Section 1.1. We determine which problems Company X experiences (consciously and unconsciously) in Section 1.2. In this section we also identify and describe what the core problem is. In Section 1.3 we formulate our research goals and research questions. Based on this, we decide which research methods are suitable.

1.1 Research motivation

In preliminary talks, Company X expressed their feeling that the operational planning process could benefit from some external advice. They felt that the planning department was held back by the inability to look ahead at how the schedules would develop. Early **scheduling** activities would often get in the way of later additions to the schedule and this resulted in the fact that new patients are often seen too late, which they had raised as their main concern. Also, as an organisation that rapidly grew in the last decade, it is getting increasingly hard to come up with ways to optimize these operational aspects, as complexity grows. Additional expertise in planning and scheduling processes, can therefore be helpful to Company X concerning these issues.

1.2 Problem identification

In this section we identify the problems that Company X is experiencing with regards to their scheduling system. From this we extract a main problem that we will focus on. In Section 1.2.1 we first describe the problem context at Company X. After that we state all problems in a **problem cluster**, to give a visual representation of how the problems are related. Then, in Section 1.2.2, we analyse the core problem, which can be deduced from the problem cluster.

1.2.1 Problem description

Company X currently has about 300 locations (varies over time), which they rent in varying frequencies. Often, a location is rented for either a morning or an afternoon. Some locations are available the entire week where others might only be available one morning/afternoon every two weeks. There are over 100 practitioners available (varies over time), which have varying working hours. Some only work 2 or 3 days per week and others work 5 days per week. Each practitioner usually has a standard schedule for every two weeks as to which location they will be occupying. This schedule stays the same until Company X decides to restructure or the scheduler requests a schedule change. The Company X organisation is spread throughout the Netherlands and is divided into regions that have their own responsibilities. Each region has a maximum of 10 practitioners and a maximum of 25 locations. This setup of staff and locations gives an idea of the complexity of the situation.

The planning department must deal with this complexity and currently spends a lot of time trying to find reasonable schedules, which balance workloads for each **session**. Note that we use the terms session and location throughout this document. We define a session as a working day of a practitioner at a given location in the schedule. If a practitioner works at two locations on a day, we describe this as two sessions. A location can accommodate multiple practitioners and thus multiple sessions at once. It is important to recognize this distinction when we use these terms throughout the document. Some locations have multiple schedules for different appointment types (e.g. separate schedule for diabetics). We denote each of these schedules as a separate location, since Company X does this as well. As of now, some locations are overloaded and others have many gaps in the schedule. This leads to high **access times** at some locations, which is undesirable. In Dutch Healthcare there are access time regulations, some of which are required by law and others required by insurers. These regulations are called “Treeknormen”, or Treek standards in English. However, for the type of allied healthcare that Company X performs no Treek standards are applicable, so there are no universal guidelines as to how long access times can be. Company X does have its own standard on access times for new patients. They want the access time for new patients to be under 3 weeks, but in 2019, 35.7% of new patients had an access time of more than 3 weeks. Sometimes the access time even surpasses 10 weeks. We look at the 2019 data because, due to the covid-19 pandemic, the 2020 data is not representative. Instead we look at the data of 2019 and before, while considering growth and other changes. Company X has indicated that these long access times are the main problem. In consultation with Company X we formulate the **action problem** as follows:

“The access time for new patients is higher than the desired 3 weeks in 35.7% (2019) of the cases, whereas Company X wants this to be at most 20%.”

The norm in this problem is formulated as a service level instead of a solid threshold to account for outliers, but note that this does not mean that if a patient cannot be seen within the 3 weeks, he/she can be postponed forever as he/she does not meet the target anyway.

The planning department experiences more issues than just this one action problem, most of which are related to each other. We want to get to the root of the issues that are experienced within the planning department and we do this by creating a problem cluster (see Figure 1.2). In this figure we visualize which problems there are and how they relate to one another in the problem context. By doing this we can show which problems are the core problems and can thus be seen as the main cause of other issues. This can be multiple problems as well as a single problem. In this problem cluster we do not just include problems experienced by the planning department, but also those experienced by other stakeholders. These include the service department, the patients and the employees.

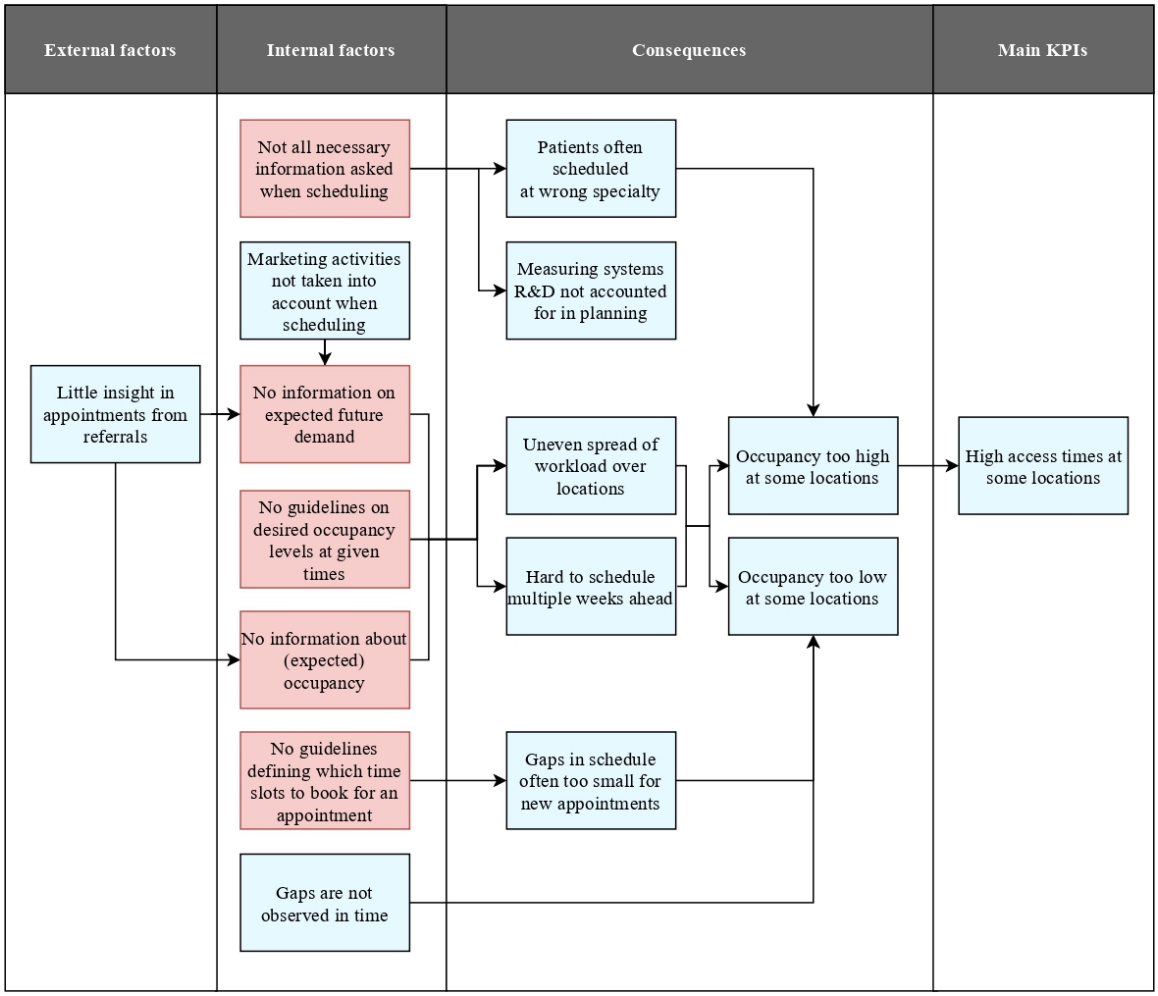

Figure 1.2: Problem cluster depicting all scheduling-related problems experienced by Company X

1.2.2 Core problems

From Figure 1.2 we deduce that we have a set of 7 problems that are not caused by other problems and that can be influenced by Company X. Two of them (light blue) we do not identify as core problems. The fact that marketing activity is not taken into account is just one of many influences on the expected future demand, so we take this into account when solving the core problem of having no information on expected demand. Furthermore, the fact that gaps are not observed in time is not a planning problem but a control issue and therefore lies out of the scope of our project. Also, the

number of gaps is affected by the problem that there are no guidelines about which time slots to book, so this problem is indirectly addressed. This leaves us with 5 core problems (marked red in Figure 1.2):

- Schedulers often do not ask patients all questions needed to adequately classify what type of appointment they need;
- There is no information on expected future demand;
- There are no guidelines defining which time slots to book for an appointment;
- There is no information about (expected) occupancy per location;
- There are no guidelines on desired occupancies at given times.

Note that we use the words occupancy and utilization interchangeably throughout this paper. We define both as the proportion of the schedule that is filled with appointments. The reason why we choose multiple core problems to solve is because these problems can be aggregated to the problem of inadequate scheduling support provided by the scheduling system. Consequentially, they have a collective solution, which entails redesigning the current scheduling tool.

Company X's scheduling activities are currently in the form of an online-offline hybrid, though they mainly perform **online scheduling**. Only twice a week during meetings some **offline scheduling**/rescheduling activities take place. When scheduling appointments we make a distinction between recurring patients and new patients. With a recurring patient, the treating practitioner schedules a new appointment at the end of the previous one. New patients get an appointment by calling the service department, where an employee will then ask some questions to determine which specialty and location/session they should go to. In this research the focus is on the access time for new patients. Furthermore, we focus on the online scheduling procedure and less on the corrective offline scheduling that is done twice a week. We want to make these corrective actions less time-consuming, by decreasing the number of issues in the schedule.

This scheduling problem can be classified as a $P|online - r_j|L_{max}$ problem [15] without loss of generality. This means we have m parallel servers, representing the number of practitioners; we have online scheduling where jobs are released as soon as they are created; and as objective we try to minimize the maximum lateness. In this problem, we define the maximum lateness as $[max(access\ time) - 21]^+$ for each day. This is a basic classification of our problem, though the main objective is obscured by conflicting objectives (e.g. distance and utilization), which we discuss in Section 2). It should be noted that the number of practitioners (servers) varies when Company X hires new practitioners and when practitioners quit or retire. Some additional constraints should be added to model the availability of employees and locations along with some other aspects that are not inherent to this type of problem. To solve instances of such a problem we need a clairvoyant online algorithm [9], which means we assume knowledge of the processing time as soon as the job is created. Company X has standard processing times for each appointment type and since our main issue is the day on which the new patient is scheduled, and not something like overtime, intraday schedules are constructed to have a minimal amount of (unintentionally) scheduled idle time. Therefore, we assume **deterministic** processing times, while in reality **stochasticity** is involved. We will however provide information to Company X about which intraday scheduling **heuristics** could be useful to reduce overtime and intra-day waiting time.

Employees schedule using a scheduling wizard, which shows at what location and time they can schedule the patient. To give the reader an idea of what this looks like, we depict the wizard in Appendix C (wizard is in Dutch). This wizard is purely based on the earliest availability within a given distance from the location to the patient. However, the wizard disregards which location is busiest, how many appointments are expected to still come in, or where/when marketing activity takes place. Ultimately, this leads to some locations suddenly becoming over-occupied, which in turn leads to high access times. The access time is a key performance indicator (KPI) in determining the quality of our schedule. From now on we define access time as the number of days before a patient can get an appointment. We focus on the access time of new patients, since the practitioners schedule recurring patients further into the future and this form of waiting time is not part of the problem. However, we do look at the way we schedule recurring patients, as their appointments can cause higher access times for new patients if not scheduled intelligently.

Company X aspires to have access times for new patients lower than 3 weeks. Many locations usually have longer access times (some even over 10 weeks) and on the other hand many locations have access times of less than a few days. The first indicates a location is overloaded and the second often means that a location is underutilized. Apart from that, unwanted gaps in the schedule often appear at the expense of utilization. This often occurs due to **fragmentation** of the schedule caused by scheduling appointments with an undesirable amount of time (15 or 45 minutes) in between. These gaps are all reported and every week the planning department tries to move things around to fill these gaps. This is very time-consuming and does not always work well since it is hard to move appointments around at such a late stage.

By steering the scheduling of appointments into carefully selected time slots, at carefully selected locations, the aforementioned problems could be diminished. As mentioned earlier, the scheduling wizard, which is used by the planning department, service department and practitioners, now only shows the available time intervals sorted on the earliest availability within a given distance to the patient. ***We want to go to a situation where information is available about the expected occupancy on the locations at the given times and we also want the wizard to give suggestions as to which location and corresponding time slots would be best to use.*** In the remainder of this project plan we state how we plan to solve the core problems and their resulting issues.

1.3 Research plan

Before we can solve the problem that the Company X planning department is experiencing, there is a long road to be travelled. To outline our research, we need to formulate a research goal, research questions about the knowledge that must be acquired to achieve this goal, and what we need to do to obtain this knowledge. This secures that the project is headed in the right direction and that we work concisely, towards an end-product. In Section 1.3.1 we determine our research goal. In Section 1.3.2 we determine what research questions we need to answer to acquire the knowledge needed to achieve our research goal. Then in Section 1.3.3 we define the research methods that we use to acquire the information.

1.3.1 Research goal

In consultation with Company X we decided that a planning support tool is the preferred solution to the problem that we attempt to solve. However, before we can do this and solve the core problem, we need to obtain more knowledge. We can obtain this knowledge by defining a research goal in the form of a **knowledge problem**. The problem being that there is a lack of knowledge that should be clearly specified in order for us to acquire it. We translate our core problem to our research goal in the following way:

"How can we develop a scheduling support prototype that suggests the most preferred sessions and time slots to schedule the appointment in, to ensure that 80% of new patients get an appointment within 3 weeks?".

This research goal is the main knowledge problem that we need to solve. To further explain, the end-product of this case study is a scheduling support tool, which shows to the person scheduling, where and how late the best appointment opportunities are, based on information about the patient, locations and employees. This depends on a lot of variables, with the main two being the (expected) occupancy of the surrounding sessions and the appointment type.

The reason that we aim to develop a scheduling support tool instead of a scheduling algorithm, is that a scheduling algorithm takes away the human aspect of scheduling. The human aspect of scheduling the appointments is needed to provide good service to patients. Appointment scheduling is a game in which patients negotiate with schedulers for time slots that satisfy each party's particular needs [21]. An appointment's time and place might be good for one person, whereas they might not be perceived as good by another, so there is no standardized way of providing the best service. That is why the scheduling decision should still be an interaction between an employee and a patient. The aim is therefore to inform a scheduler of the best options and the consequences when scheduling undesirably.

1.3.2 Research questions

To achieve our research goal, we need to acquire some additional knowledge. We can bridge this knowledge gap by answering the five research questions below. These research questions, which we refer to as RQ1, RQ2, RQ3, RQ4 and RQ5 respectively from now on, each give us a specific piece of the puzzle that is our knowledge problem. In Appendix A we divide these research questions into more specific sub-questions.

Research question 1: *What does the current scheduling process look like and how does it perform on the relevant KPIs?*

RQ1 gives us an understanding as to how scheduling is currently performed with a higher attention to detail than what we have done so far. We plan to answer RQ1 in the form of a flow chart (or if necessary multiple) to depict the process. Furthermore, we need to perform a data analysis to accurately understand and describe the current performance. The flow chart needs to be an exact representation of the current scheduling process and needs to be validated by employees from varying departments at Company X. We carefully construct the chart based on many impressions gained from within various departments. This flow chart, and corresponding performance analysis, can help us understand where improvements can be made and what information could be useful to support the scheduling process. Also, we could identify problem areas and address them.

Research question 2: *How can we make a predictive model of the occupancy of a location and validate it?*

RQ2 provides us with the information that our support tool needs to determine which sessions have a high risk of being over-occupied or under-utilized. To answer RQ2 we need to perform an extensive data analysis. In the data analysis we aim to uncover information about appointment arrivals, scheduling behaviour, no-show/cancellations, frequency of distinct appointment attribute combinations and more. The data is available to the researcher through the database and can also be acquired through the online agenda and various reports. We will also perform a literature review that should provide information on RQ2 as well as on RQ3. We aim to find how to predict future occupancy, preferably related to appointment scheduling. Furthermore, we will use demand forecasting methods and **Probability Theory** to make predictions about which sessions will be busiest. We also need to validate that all the used data is valid and that the model output is accurate as well. We answer how we should do this when answering this RQ, though we will most likely do this by **Cross-Validation**.

Research question 3: *How can we create and validate a scheduling support tool that can adequately inform schedulers on which time slots and sessions to choose, utilizing our predictive model (RQ2)?*

RQ3 treats the design and creation of our final deliverable. RQ3 requires thorough communication between Company X personnel and the researcher, as well as a literature review to provide us with techniques and functionalities that can be of help. We plan to provide the schedulers with multiple forms of descriptive information, which helps them make an educated decision on where and when to schedule the patients. For this, we will combine methods discussed in literature, as well as methods that we devise ourselves and tailor to the situation. When we know exactly what the tool should include and what it should look like, we need to know what the best way is to model each function. This research question considers not only exact science, but also psychology. Furthermore, we look into the possibility of using **Patient Choice Analysis**[17] and **reservation levels** [14]. Finally, we need to carefully devise how we should validate that this tool works the way it is intended to.

Research question 4: *What are the resulting benefits of the scheduling support tool over the current situation?*

RQ4 tells us how the tool performs compared to the current situation. We need to test the performance of the proposed scheduling options, while considering that this also depends on the schedulers' behaviour. This needs to be done without significantly disrupting current operations. We need to ask ourselves how this can be realized. Can we test the model in real life scheduling situations? Can we run a pilot? Can we test it in comparison with historic data? Furthermore, we need not only ascertain quantitative performance measures, but also qualitative measures like ease of use and potential employee resistance. This should all be done before considering implementation, since early flaws can have a rippling effect throughout the system, which can be very laborious to iron out when implemented.

Research question 5: *How can our scheduling support tool be integrated in the existing IT systems at Company X?*

RQ5 directly relates to the implementation of our scheduling support tool. The result of RQ4 is a prototype scheduling support tool that is not yet implemented in the Company X IT systems. To provide a roadmap towards implementation in the IT systems of Company X, we make a detailed implementation plan to ensure a structured and seamless integration of the scheduling support

tool. The implementation plan should give a detailed description of the tool, a step-by-step module-based plan for implementation and guidelines for security and aftercare. We cooperate with the IT department to completely understand the old scheduling wizard and IT system to get knowledge of what steps should be taken to implement the model. Communication with the IT department is essential throughout the development of the implementation plan to ascertain feasibility of each aspect.

1.3.3 Research design

At our disposal we have all information from the database and the help of the employees at Company X. Mainly the IT department, planning department and service department can provide useful information. As mentioned earlier, this scheduling problem can be seen as a $P|online - r_j|L_max$ [15], which has m servers, equal to the number of practitioners, online scheduling and maximum lateness is minimized. Also, jobs are **non-preemptive**. Additional constraints for our problem include the availability of locations and employees, having the right specialty for the patient, distance penalty between patient and location, etc. The objective function should include the minimization of the number of new patients that are not seen within 3 weeks and should include the distance of patients to the location in some way.

Due to the use of online scheduling, each job must be scheduled immediately when it comes in. The sequential nature of this process can be formulated as a multi-stage stochastic program with stages representing each customer request [11]. In each stage, the direct outcome after scheduling is known, but the resulting schedule at the day of the appointment involves a lot of stochasticity. A schedule is influenced by cancellations, no-shows, new patients, recurring patients, sick employees and scheduling behaviour. This stochasticity is the part that we want to better understand and provide information about to the schedulers. This stochasticity is not to be confused with the stochasticity of processing times, which we do not consider as there is no information about this and because our focus is not on intra-day optimization. We aspire to redesign the scheduling wizard in a way that the required information is clear to the scheduler so a well-informed scheduling decision can be made.

Most literature on online scheduling propose optimization-based methods. The issue with this is that it neglects the human aspect of scheduling that lies in negotiation between patient and scheduler. We want to maintain this aspect to ensure high quality of service and patient satisfaction and retention. Thus, our scheduling tool will be based on a top-k ranking with time slot suggestions, supported by some additional features to support the scheduler when he/she wishes to schedule time slots that were not suggested. Ranking algorithms are scarce in appointment scheduling literature. Jie et al. [31] do propose a method that considers patient-provider mutual preference to guide the appointment scheduling process by means of schedule defragmentation. The proposed method ranks all possible appointment slots from lowest fragmentation level to highest and returns the list to the patient, which is encouraged to select a time slot with a higher preference for the care provider. While fragmentation is one of the attributes that we want to base our time slot suggestions on, we also have other attributes that must be considered. Furthermore, a complete ranking is too computationally intensive for our online multi-server problem, which is why we use a top-k ranking. To the author's knowledge, top-k multi-attribute ranking-based approaches to appointment scheduling are not yet present in the literature.

To redesign the scheduling wizard into our desired end-product, we decide to take the following steps:

Step 1: Data analysis, to understand occupancy development over sessions.

Step 2: Establish which functions should be incorporated.

Step 3: Design and validate model that predicts occupancy per session.

Step 4: Redesign scheduling wizard such that it communicates the best scheduling options and all additional desired support information and validate that it works as intended.

Step 5: Assess performance of the new scheduling support tool and compare to the old situation.

These steps should result in the deliverables listed in Appendix B

Chapter 2

Problem context

In this chapter we analyze the current situation and its performance. We discuss the resources that Company X has at its disposal and its implications for this research in Section 2.1. In this section we also describe the case mix that Company X encounters. In Section 2.2 we describe the current scheduling process in detail and several factors that influence the process as well. Then, we assess the performance of the current situation in Section 2.3. After we have a complete picture of the current context, we establish what we want the post-implementation context to look like in Section 2.4. Finally, we conclude this chapter in Section 2.5. From now on we make a distinction between three patient types: New patients (NP), Recurring patients (RP) and Periodic patients (PP). We make this distinction because their desired access times are determined differently for each type. We elaborate on the reason for this distinction in Section 2.1.

Apart from literature, our sources of information come from within Company X. We gain qualitative information using unstructured interviews, observations and service department phone call recordings. The quantitative data that we use comes from the in-house data system, the database and reports that were readily available (upon request). We describe the used data sources, our data processing procedures and data validation procedures in detail in Appendix H. We do this to give the reader an understanding of what we base our research on and for reproducibility in further research. Most of the information based on quantitive data in this chapter, is based on data from 2019 (unless stated otherwise). This is due to unrepresentative data in 2020 caused by the covid-19 pandemic. In Appendix D we show why we label the 2020 data as unrepresentative.

2.1 Resources and case mix

Before we can envision what our process looks like, we need to understand its building blocks. Therefore, we first discuss what our resources and patient group looks like. We analyze our resources in Section 2.1.1. Then, we analyze our case mix in Section 2.1.2.

2.1.1 Resources

We divide Company X's resources in two categories: employees and locations. Of course there are more resources (e.g. money and materials), but from an operational perspective employees and locations are most important for our research. Company X currently has over 300 active locations, but this number is increasing. There are currently 117 active practitioners (the most relevant employee group for our research), but this number is growing as well.

Locations

The locations are subdivided into 17 regions. There are 3 additional regions: -*classified*-. These however, are irrelevant for our research. These 3 regions do not utilize the scheduling wizard and

Region	N.o. locations	N.o. sessions	Avg. sessions per location	Proportion of total appointments
Almelo	17	687	40.41	5.68%
Amsterdam	25	984	39.36	8.29%
Brabant	1	27	27.00	0.20%
Den Haag	21	460	21.90	3.99%
Drenthe	12	512	42.67	4.09%
Enschede	24	1531	63.79	12.00%
Friesland	11	320	29.09	2.48%
Gelderland	37	1500	40.54	12.88%
Groningen	1	1	1.00	0.00%
Hengelo	43	2814	65.44	21.04%
Leiden	9	338	37.56	2.45%
Polder	17	615	36.18	5.71%
Rotterdam	38	820	21.58	7.26%
Utrecht	13	527	40.54	3.95%
Vriezenveen	10	426	42.60	3.22%
Wierden	12	847	70.58	6.05%
Zoetermeer	4	88	22.00	0.71%
Total	**295**	**12497**	**37.78**	**-**

Table 2.1: Information about regions (from appointment data and generated session data in 2019)

thus we leave them out of the scope of this research. From now on we refer to these 3 regions as artificial regions, since they do not represent a tangible region. In the remainder of this research, when we refer to regions, we do not include these artificial regions.

Each region is responsible for its own performance and is represented by a region leader. In Table 2.1 we show some descriptive metrics per region. What we see is that regions that were added later, mostly non-eastern regions, have a lower number of sessions per location. This happens because Company X's value proposition is based on location. They want to be as close to the client as possible to attract patients. For this purpose they often rent locations for just 1 day every 1 or 2 weeks. A problem that is inherent to this strategy is that a location's capacity is quickly filled, which results in high local waiting times. This mainly happens in the big cities, where travel distance has a much higher relative impact than in less densely populated areas. Thus, one can imagine that appointment negotiations pose different challenges for an appointment in a big city than for appointments in a rural area.

Many locations can only accommodate specific appointments types. This is often in some way described in the name of the location, but is not fully standardized. To classify what types of appointments each location can accommodate, we propose a set of location types (later also referred to as session type when talking about sessions at a location). Table 2.2 shows the proposed set of location types that we use throughout this research. In this table we also state which classification rules we use to classify a location using its name. Sometimes multiple classification rules apply to a location. Therefore, we also state a priority, where a lower number designates a higher priority. For example, when a location has "DM" and "Huisbezoek" or "HB" in its name, we designate it as a "Home visits" location, because "Home visits" is a more important distinction (in our opinion). In future research Company X could think about simply designating a location with multiple types. However, in the remainder of this research we use 1 type per location for simplicity. Unfortunately, which appointments types each location type can accommodate is not clear. It would be advisable to

Location type	Classification rule	Priority
Regular	If none of the below applies	9
DM	Contains "DM"	7
Sport	Contains "Sport"	6
Reumatism	Contains "Reuma"	5
Child	Contains "Kind"	4
Laser	Contains "Laser"	3
Home visits	Contains "Huisbezoek" or "HB"	2
External	Contains "ziekenhuis", "zorginsteling", *-classified-*, "poli", "MST", "ZGT" or "AMC"	1
Other	Contains "[" and none of the keywords above	8

Table 2.2: Location types and classification rule applied to location name

Profession[a]	N.o. Employees	Percentage of total appointments	NP (% per profession)	RP (% per profession)	PP (% per profession)
Profession 1	95	93.3%	13.3%	62.6%	24.2%
Profession 2	6	5.7%	0.3%	1.3%	98.3%
Profession 3	2	0.1%	13.3%	56.7%	30.0%
Profession 4	1	0.2%	3.9%	28.5%	67.6%
Profession 5	1	0.4%	0.0%	100.0%	0.0%
Profession 6	2	0.3%	2.4%	33.3%	64.3%

[a]Professions are classified to ensure anonymity of Company X

Table 2.3: Information per profession type (from appointment data of 2019)

define this in further research. In the current situation this is not yet easy to do, since appointment types are often used interchangeably.

Employees

From unstructured interviews with the supervisor at Company X we know that we should take into account the difference between professions. We extracted a list of professions from the IT system (to see how, we refer to Appendix H). We found that there are 6 profession types that we have schedules and appointments for. We state the number of employees and proportion of appointments (per patient type NP, RP, PP) per profession in 2019 in Table 2.3. We see that by far most employees (that treat patients) are of the profession "Profession 1", which can do nearly any activity. We also see that for PP appointments practitioners of "Profession 2" perform a substantial number of appointments. "Profession 2" only performs standard periodic care appointments. Students can perform simple appointments, often under supervision of "Profession 1" (depending on the stage of their studies). "Profession 4" only performs simpler appointments to alleviate the schedule of "Profession 1". "Profession 5" performs appointments concerning laser therapy. "Profession 6" also sometimes performs simple appointments (mainly the more standardized PP appointments). Unfortunately, we are currently not able to establish which profession types can perform which appointment types. This is because schedulers often use these appointment types interchangeably. Just as we advise to make a possible selection of appointments types per location type, we advise the same for professions. This requires some standardization and communication throughout the organization.

It should be noted that we often use "practitioner" throughout this thesis as a collective term to

refer to any employee that treats patients. This means that, when we mention a practitioner, this does not necessarily refer to a specific profession. When we refer to a profession we specify this.

2.1.2 Case mix

The case mix at Company X is described by the patient types and the appointment types. Each appointment type belongs to one patient type. Some appointment types can only be performed at some locations. This is due to equipment that is only available at some locations. Furthermore, some activities are only performable for certain professions, though Company X has not documented an exact division. The tool used for scheduling should account for all these constraints.

Patient types

We make a distinction between New Patients (NP), Recurring Patients (RP) and Periodic Patients (PP). Company X already used the term new patient, but we created the other two classifications for this research specifically. These patient types account for 12%, 59% and 29% of total appointments, respectively (in 2019). We make this distinction since all three have different planning horizons. Company X wants to schedule NPs within 3 weeks; RPs are planned a certain period after the previous appointment, which can for instance be a month for a short-term checkup (shortly after receiving care) or a year for a yearly checkup; and PPs' appointments need to be scheduled pre-defined periods apart. Therefore, different scheduling challenges arise for each patient type. We subdivide appointment types per patient type to determine to what patient type each appointment belongs. We discuss this in more detail in Section 2.1.2.

New Patients
NPs form our focus group, since it is their access time that we are trying to decrease. We aim to do so by scheduling NPs, RPs and PPs more intelligently. New patients are not necessarily patients that have never had an appointment at Company X before. Patients that have not had an appointment in the last 5 years are also classified as an NP by Company X. Patients usually get an appointment over the phone. Patients either call Company X because they want an appointment, or Company X calls patients if their general practitioner referred them to Company X. We depict the scheduling process for NPs in Appendix E.

In Figure 2.1 and Figure 2.2 we show the distribution of performed NP appointments over months and days of the week respectively. Figure 2.1 shows that at the end of the year the most NPs are scheduled. This happens because a lot of people have appointments for Company X's field of allied healthcare in their supplementary health insurance package. At the end of the year people still have those left and want to use them. Figure 2.2 shows that most NPs are scheduled on Wednesdays and that there are hardly any appointments on Saturdays and Sundays. This is logical since only very sporadically some locations have sessions on Saturday. We also notice that fewer appointments are scheduled on Friday, which happens because many practitioners do not work on Fridays or only do a morning session.

Recurring Patients
RPs account for the greatest number of appointments (59%) of all three patient types. These appointments need to be scheduled far into the future. How far depends on what the last activity was. The most frequently occurring RP appointments are: a -*classified*- pickup after a consult, a short-term check-up after a pickup, and a yearly checkup after a consult or check-up. Scheduling for

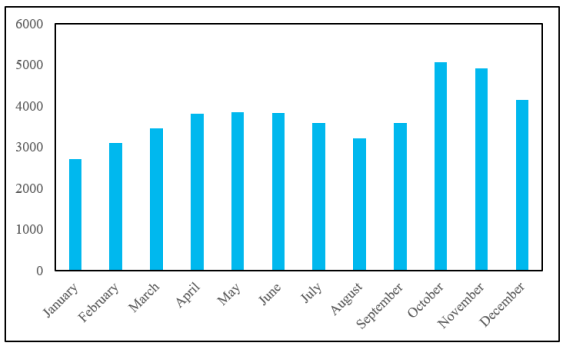

Figure 2.1: Aggregated number of NP appointments performed per month from 01-01-2016 to 31-12-2019

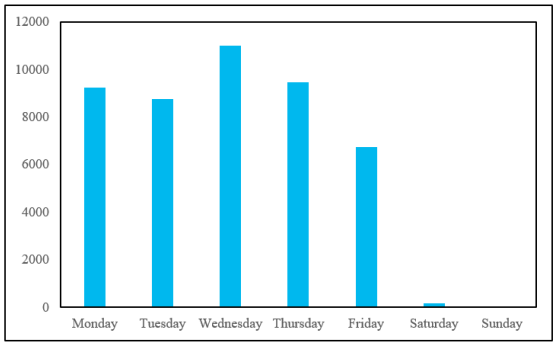

Figure 2.2: Number of NP appointments performed per day of the week from 01-01-2016 to 31-12-2019

RPs and PPs differs from NPs, since they are often scheduled by practitioners at the end of the last appointment and only some times by employees of the service department, as opposed to NPs, which are always scheduled by the service department. Furthermore, the desired access time is based on the foregoing activity, instead of having a standard of 3 weeks. The access time is also of a different form with RPs and PPs. With NPs the desired access time is everything below a maximum of 3 weeks and preferably as low as possible. Whereas, with RPs and PPs, the desired access time is a certain number of days later and not too many days less or more than that. Also, RP appointments are often less urgent than NP appointments, because they have often already been helped. E.g. an NP with issues is more urgent than a patient that has a check-up, which does not necessarily have issues. Access time requirements are therefore generally more important for NPs.

In Figure 2.3 and Figure 2.4 we show the distribution of performed appointments over months and days of the week respectively. We take the average over multiple years, to counteract monthly seasonality. We see very similar distributions as compared to the distributions of NPs. Figure 2.3 shows that at the end of the year the most RPs are scheduled. Again, this happens because a lot of people have a number of appointments per year for Company X's field of allied healthcare in their supplementary health insurance package. If they still have any left by the end of the year they often use them before the year ends. The distribution of appointments over days of the week is also similar to that of NPs.

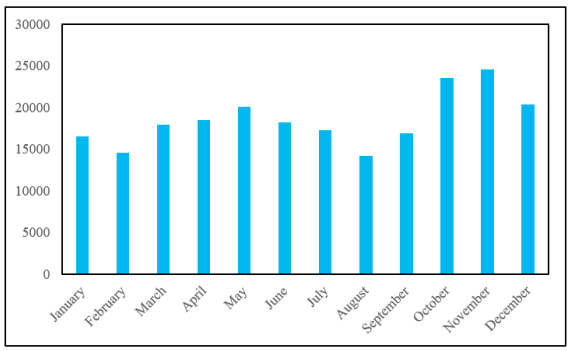

Figure 2.3: Aggregated number of RP appointments performed per month from 01-01-2016 to 31-12-2019

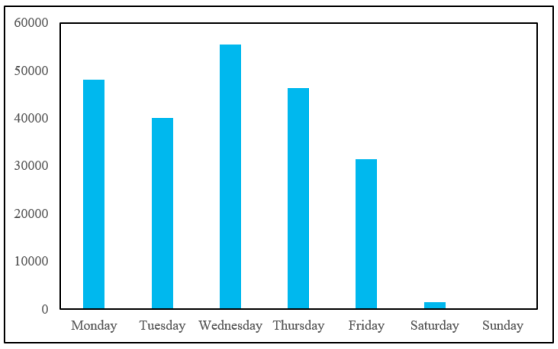

Figure 2.4: Number of RP appointments performed per day of the week from 01-01-2016 to 31-12-2019

Periodic Patients

The third patient group is "periodic patients" (PPs), which accounts for 29% for appointments (in 2019). The majority of PPs are diabetics that need periodic care to check their state. Diabetics often show neuropathy and can therefore not always perceive issues before it is too late. Additionally, they have reduced blood flow (angiopathy), meaning wounds heal very slowly. Another issue originating from diabetes is decreased joint mobility, which can cause issues as well. This also contributes to wound and pressure point forming. Company X decides how frequently a patient should be treated, based on their care profile. The care profile ranges from 1 to 4, with 4 requiring the most attention. Apart from diabetics, PPs mostly consist of elderly patients. Elderly also have increased risk of pressure points and when they cannot tend to pressure points, calluses, etc. themselves, issues quickly arise. Therefore, periodic care is necessary for this patient group as well. The desired access time for a PP is defined by the severity of their issues. For diabetics this is determined using their care profile, but for other PPs this assessment must be performed in collaboration between the patient and the care provider. The patients then periodically return to receive care.

In Figure 2.5 and Figure 2.6 we show the distribution of performed appointments over months and days of the week respectively. We notice that distribution of appointments per month is very different from that of NPs and RPs. For PPs we see a lot more appointments at the start of the year. This happens firstly because periodic diabetic patients need to be seen at the start of the year to determine the care plan for the rest of the year. Secondly, this happens because periodic appointments are insured by the basic health insurance, which means they do not have a specific number of appointments that they can use. How many appointments the patient gets is purely based on the care plan and is covered by the insurance. Figure 2.6 shows that the most used day of the week for PPs is Thursday, whereas for NPs and RPs this is Wednesday.

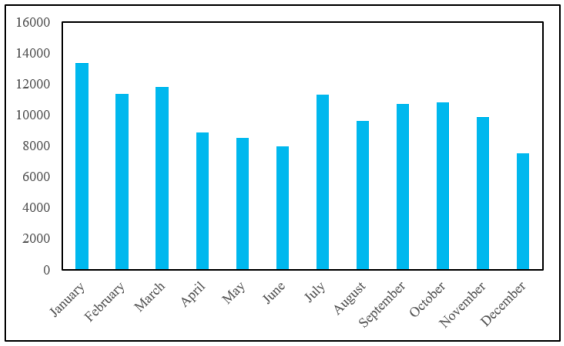

Figure 2.5: Number of PP appointments performed per month added together from 01-01-2016 to 31-12-2019

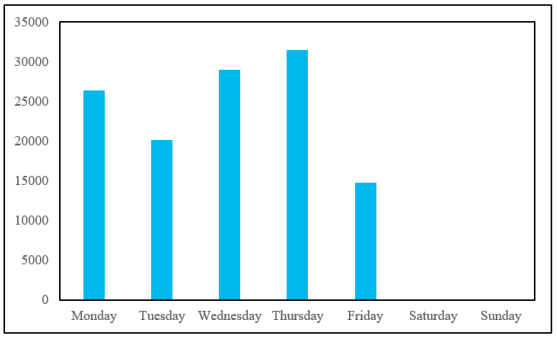

Figure 2.6: Aggregated number of PP appointments performed per day of the week from 01-01-2016 to 31-12-2019

Appointment types

From 2016 till 2019 we see 51 appointment types, called activities in our data set. In Appendix F we depict all of the activities with their corresponding relevant metrics. We extracted the realized care path per patient and counted the frequency of each realized care path from 2016 till 2019. By considering the frequency of each realized care path it becomes easier for us to predict how many appointments will come in for the coming months.

A normal treatment usually consists of at most 3 appointments. A patient's realized care path can be much longer and can include multiple treatments. We want to look at the most frequent realized

care paths of length at most 3 to see what the most frequent treatments are. We describe how we do this in more detail in Appendix H. We explain the algorithm we use for this in Appendix G. The most frequent treatments that we see are:

1. A standard NP treatment, with a consult, after which he/she becomes an RP and gets a -*classified*- pickup and a checkup.
2. An RP treatment with a yearly consult which is followed by a pickup.
3. An RP treatment with a new -*classified*- request consult followed by a pickup. This usually happens if a patient wants new -*classified*- and they do not want to wait until the yearly consult, or they do not have a yearly consult planned.
4. An RP treatment with a normal consult which is followed by a pickup. This usually happens if the patient did not have a yearly consult planned and they want a new appointment.
5. An RP that only gets one normal consult. This often happens when the practitioner lets the patient know that -*classified*- or further treatment are not the solution, or when the patient does not want another appointment after the first (e.g. due to cost). This also happens when the patient wants a follow-up appointment, while they do not have a checkup scheduled.
6. A diabetic PP that gets a periodic treatment.
7. A diabetic PP that gets a periodic check (usually less severe than periodic treatment).
8. A non-diabetic PP that gets a periodic treatment.

	Realized care path	**Frequency**	**Number of appointments**	**Percentage of appointments**
1.	A standard NP path, with a consult, a -*classified*- pickup and a checkup.	17164	51492	13.2%
2.	An RP path with a yearly consult which is followed by a pickup.	9097	18194	4.7%
3.	An RP path with a new -*classified*- request consult followed by a pickup	7505	15010	3.8%
4.	An RP path with a normal consult which is followed by a pickup.	5112	10224	2.6%
5.	An RP that only gets one normal consult	4275	4275	1.1%
6.	A diabetic PP that gets a periodic treatment.	3147	9172	2.3%
7.	A diabetic PP that gets a periodic check	7567	20098	5.1%
8.	A non-diabetic PP that gets a periodic treatment.	9374	27210	7.0%
	Other	-	235410	60.2%
	Total number of appointments (patient-related)	**-**	**391085**	**100.0%**

Table 2.4: Most frequent realized care paths and their metrics (01-01-2016 to 31-12-2019)

In Table 2.4 we show how often each of these treatments occur. The vast majority of treatments that we do not specifically mention above (labeled as other in Table 2.4) are variations on the treatments that we mentioned, varying on one or multiple of the 51 different appointment types.

We also extracted the dates of each appointment in the realized care path. We use this to extract the

average number of days between appointments in the treatments (in the same algorithm mentioned above, see Appendix G). Taking into account the time in between appointments, the most frequent realized care paths that we listed above look like the ones depicted in Figure 2.7.

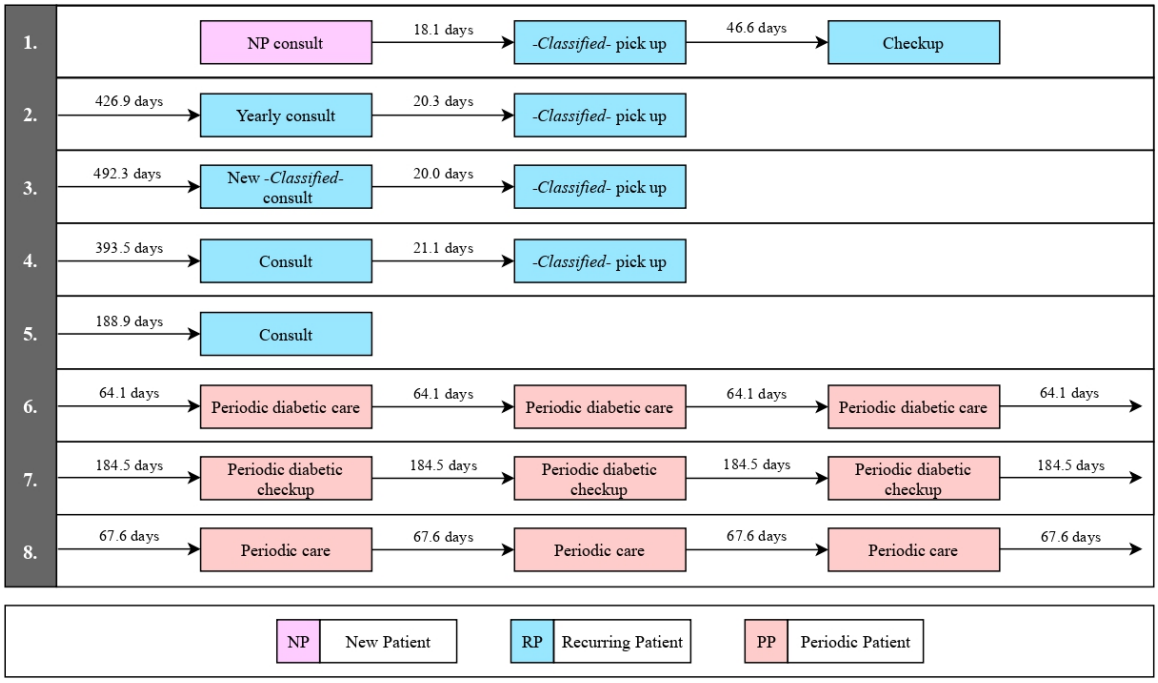

Figure 2.7: A depiction of the most frequent realized care paths from 01-01-2016 to 31-12-2019 (numbered as in Table 2.4), with arrows representing the average time between the previous appointment and the next.

2.2 Current scheduling process

The current scheduling process leaves a lot of decision making to the patient. This is beneficial for patient satisfaction since they can get an appointment wherever they want on a mutually agreed time. However, leaving the decision making to the patient also leads to a lot of inefficiencies for Company X. Appointment scheduling is a game in which patients negotiate with schedulers for time slots that satisfy each party's particular needs [21], but currently Company X's needs are under-represented. Often the planner only considers a single location, while a patient might be able to receive treatment sooner elsewhere. Furthermore, the current way of scheduling does not accommodate an even spread of appointments over locations.

In this section we describe what the current scheduling process looks like. It is important to note that there are currently not many strict rules or guidelines for schedulers, which means every scheduler schedules differently. The description we provide of the current situation describes the scheduling procedure that we mostly observe, which can often cause issues. This does not mean that each scheduler, if any, is complicit to these issues since no strict guidelines are in place. In Figure 2.8 we depict the current scheduling process schematically. In Appendix E we depict the current scheduling process per patient type (NP, RP, PP) in more detail.

As mentioned before, we make a distinction between New Patients (NP), Recurring Patients (RP) and Periodic Patients (PP). These patient types account for 12%, 59% and 29% of total appointments,

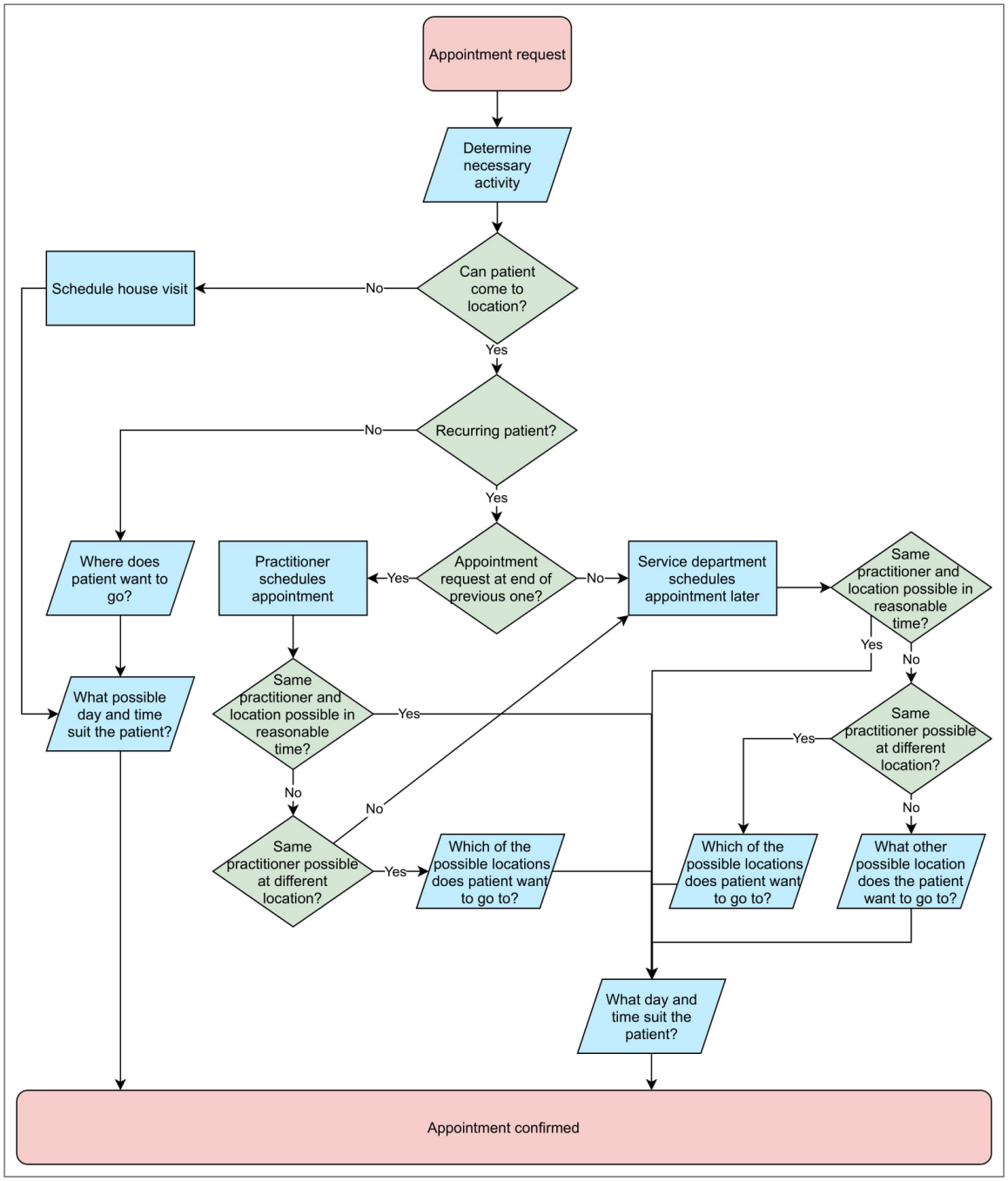

Figure 2.8: Flow chart of current scheduling process

respectively (in 2019). We describe the scheduling process per patient type in the following sections.

2.2.1 New patients

When scheduling NPs, we see that the scheduling habits slightly differ based on the specialty that the patient should go to. What stands out is that the decision making is often left to the patient. With new patients the scheduler will try to find an appointment as soon as possible, preferably within 3

weeks, although schedulers do not actively consider this threshold that Company X has established. The scheduler selects the patient if he/she is already in the system or creates a new patient profile. The scheduler then needs to determine the type of activity that needs to be performed at which specialty and select that in a drop-down window in the scheduling wizard. Then the wizard provides a list with earliest time slots big enough for the standard time set for that activity within a given radius. Usually the scheduler asks the patient where they want the appointment to take place. The patient often does not know the options and the scheduler mentions some locations close to the patient, of which the patient picks one. Then the scheduler tells the patient which time slots are available at that location. If the patient does not agree with any of the provided time slots, the scheduler looks at other locations for an option.

2.2.2 Recurring patients

Much like the current situation for scheduling NPs, Company X's preferences are not taken into account when scheduling RPs. The scheduler, which can be the practitioner or a service department employee, usually does not consider a different location or practitioner when scheduling a patient unless it is too busy to schedule the patient within a reasonable time at the same location with the same practitioner. They do not usually consider a different practitioner because Company X wants the same practitioner to perform a complete treatment path, excluding treatments for PPs. However, a new treatment for an RP can be performed by a different practitioner than the previous treatment. If the practitioner's sessions at the same location are filled, the scheduler informs the patient on other locations that are possible with the same practitioner, or if the practitioner has no room in his/her schedule in reasonable time at any location, the scheduler does consider other practitioners. In Appendix E we schematically show what the process of scheduling RPs looks like.

When the RP did not get a next appointment at the end of his/her last treatment (e.g. yearly checkup), someone at the service department will schedule an appointment for this patient, should they request one. In this case, the scheduler schedules the patient as soon as possible. The scheduler does not make a distinction between NPs and RPs in this case. Exceptions are wound treatments, -classified- and infection. In the first case the patient should be treated within 24 hours (can also be on the next day), whereas the latter two require same-day treatment. The required (access) time in between appointments depends on the diagnosis and the stage of the treatment that the patient is in.

2.2.3 Periodic patients

The scheduling procedure for PPs is the same as it is for RPs, apart from the determination of the desired access time. Usually we see that the patient schedules a new appointment at the end of the previous one with the same practitioner at the same location. Just as with RPs this can contribute to unequal distribution of appointments over sessions. Company X has indicated that having the same practitioner for a PP each appointment is not necessary. There are patients that specifically prefer to have the same practitioner and location, in which case Company X will oblige and schedule the patient where they please. On the other hand there are patients that do not mind going to another location/practitioner. This group of patients is where Company X can improve their process by looking at other locations/practitioners to get the most jointly preferable time slot. We see that in 2019 in 83% of the cases a PP is scheduled at the same practitioner. In 88% of the cases a patient is scheduled at the same location, which shows the myopia of the current way of scheduling.

2.3 Current performance

We know what the current process looks like, but we have not yet established how we compare the performance of the current situation to the situation after implementation of our scheduling tool. We need to determine which KPIs we have to keep track of based on the value proposition and problem context sketched by Company X and the researcher. These KPIs need to be quantified for the current and future situation and this quantification should adequately represent the performance. Also our problems should show in the current performance of our KPIs. In this section we first determine which KPIs we use and underpin why we use them. Finally, we assess the performance on these KPIs for the current situation and confirm that the experienced problems show in the performance.

2.3.1 Key performance indicators

We decide to divide the KPIs that we use into KPIs from the perspective of Company X and KPIs from the perspective of the patient. The reason we do this is to see whether the negligence of Company X's interests, which we observed in the scheduling process, is also represented by an imbalance in performance on these two perspectives. Table 2.5 shows what KPIs we use to measure performance of the scheduling process. We explain how we define these KPIs and why we use them below.

Company X perspective
Utilization
Variance of utilization per practitioner workday
Fragmentation index
Patient perspective
NP access time
NP access time violations
Absolute deviation of RP/PP access times from the target date
Distance from patient to the appointment location

Table 2.5: Key Performance Indicators

KPIs: Company X perspective

Out of our 7 KPIs, 3 are important from a Company X perspective and not as much from a patient perspective. These are utilization, variance of utilization and fragmentation index. Below we discuss these KPIs related to Company X's perspective.

Utilization

For Company X it is important that resources are used well, the resources being the employees and locations. We want an even spread of appointment load over practitioners, so a practitioner on one hand does not have too much idle time and on the other hand does not have an overcrowded schedule, which is detrimental to quality of service and might even lead to a loss of patients due to long access times (number of days from day of scheduling to day of appointment). Therefore, we use utilization per practitioner per day as a KPI. Keeping track of utilization per practitioner (and with that their locations) shows where peaks in demand occur and where we could offload any additional demand. To calculate the utilization we divide the amount of appointment time (in minutes) for a practitioner's day by the amount of session time (in minutes) for this practitioner that day. We deem some sessions invalid, which we exclude from our aggregate utilization. To see how and why we deem

some sessions invalid we refer to Appendix K. In Appendix K we also briefly discuss how we get the amount of appointment time and amount of session time. Furthermore, in Appendix H we describe how we generated the schedules.

Variance of utilization

To show the spread of appointment load over practitioner workdays we use the variation of utilizations. We want to spread appointment load as evenly over sessions as possible, since this allows for optimal utilization of resources. High variation of utilization means that all sessions need to be able to withstand peaks, and for this more capacity is needed. To calculate this we take all the valid utilizations of the practitioners' workdays and calculate the sample variance. Sample size depends on the number of available observations (workdays), but the estimate of the population variance gets more accurate with a higher number of observations.

Fragmentation index

Company X indicated that they suffer from fragmented schedules, which is a common problem in online scheduling environments. Fragmentation of schedules causes increased idle time (and thus decreased utilization). Apart from the fact that idle time is a waste of resources, when present in the form of hard-to-fill slots, it leads to less available time in a session. This causes appointments that could otherwise be scheduled on a certain day to be scheduled on a later date. The result is more access time violations, and in the worst case, loss of patients due to long access times. We cannot measure this fragmentation by idle time, because the total idle time is also often defined by the demand being too low for a full schedule. Fragmentation itself is hard to measure, since it is not directly quantifiable, like for instance overtime or utilization. Jie et al. [31] measure fragmentation using a fragmentation index, which they calculate as the number of open blocks. An open block is defined as one isolated open time slot or multiple adjacent open time slots. E.g in Figure 2.9 (A) there are 6 open blocks and in (B) 2 open blocks, thus we have a fragmentation index of 6 in (A) and 2 in (B). For our research we use the same fragmentation index measure as KPI to track how well our schedules are defragmented. We use the average fragmentation index to describe our overall performance on unworkable schedule gap prevention.

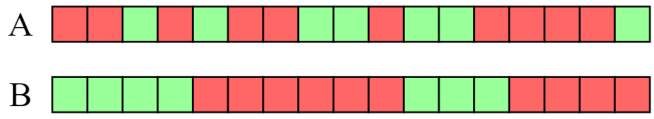

Figure 2.9: High fragmentation schedule (A) vs. low fragmentation schedule (B)

KPIs: Patient perspective

Apart from representing Company X's perspective in the KPIs we also need to represent the patient perspective. As mentioned before, the focus of our research is not on intra-day performance, so we will not be assessing performance in waiting time (in waiting room), quality of treatment, etc. The important measurable variables from a patient's perspective in the scope of our research are the compliance to access time requirements and the distance that patients need to travel. Furthermore, in each aspect of our research we try to only make choices that are not detrimental to patient satisfaction. Measuring patient satisfaction, however, is outside the scope of this research.

NP access time
To measure access time requirements compliance we use different metrics for NPs as opposed RPs and PPs. This is logical, since NP access time requirements are different from those of RPs and PPs. All NP access times are required to be within 21 days. A KPI that we use for NPs is the average access time. For this we take the difference (in days) of the day of scheduling and the day of the appointment.

NP access time violations
The second KPI we use for NPs is the number of access time violations. By counting the number of violations we can calculate a service level of patients being treated in time. If the access time of an NP is larger than 21 days, then we treat it as an access time violation.

Absolute deviation of RP/PP access times from the target date
The most prominent factor for defining the desired access time for RP/PP appointments is the practitioner's diagnosis. It is also based on what stage of his/her treatment the patient is in. For both RPs and PPs we establish a target date. The target date is equal to the current day plus the desired access time that the practitioner determines based on the diagnoses and stage of treatment. We use the absolute number of days deviation from the target date divided by the desired access time (in days) as a KPI. We divide by the desired access time, because deviating a day from the target date is relatively worse if the desired access time is shorter.

Distance from patient to the location
Finally, for patient satisfaction it is important that patients can go to a location that is close to them. From the schedulers' phone call recordings and appointment cancellation data we learned that patients often opt not to make an appointment because the distance between them and the suggested location is too large for them. Therefore, we use the distance from the patient to the location where they have an appointment as a KPI as well. We use distance instead of travel time, since we do not have access to travel time information (e.g. google maps routing API). Apart from this reason, it is also much more computationally demanding to determine travel time than it is to determine a Euclidian distance. Therefore, we calculate the distance between the patient and the location as the crow flies. For this purpose the database registers geographical coordinates for each new zip codes it encounters and stores it. Since the earth's surface is not a flat plane it would not strictly be correct to use the pythagorean theorem. Instead, we use the haversine method. When applied to our data the pythagorean method would produce estimates that deviate up to 3% (average of 0.9%) from the value given by the haversine method. Whereas the pythagorean method assumes a flat plane, the haversine method assumes the earth is spherical. In reality the earth is actually very slightly ellipsoidal. However, the ellipsoidal effect would not make a significant difference on the distances we calculate. The haversine method is described by the following three functions:

$$a = sin^2(\Delta\phi/2) + cos(\phi_1)cos(\phi_2)sin^2(\Delta\lambda/2)$$

$$c = 2 * arctan2(\sqrt{a}, \sqrt{1-a})$$

$$distance = Rc$$

With $\phi = latitude$ (expressed in radians), $\lambda = longitude$ (expressed in radians), $R = radius$ (of the corresponding sphere), which we assume to be 6371km (mean radius of earth).

2.3.2 Current performance on KPIs

Based on the appointment data of 2019 (172368 appointments included) and the schedule data of 2019 (11570 days/12497 sessions included), we assess the performance of the current situation on the defined KPIs (unless we specifically state another time period). See Appendix H for an elaborate explanation of the data gathering and processing, which we used for assessing the current performance.

Utilization and utilization variance

The average utilization of a practitioner's schedule is 91.7% (Standard deviation: 9.58%). The average available minutes per day per practitioner is 381.9 minutes. This means an average idle time 32.0 minutes per day. We show the utilizations per region and the variance within regions in Table 2.6. Note that the sum of sessions over all regions is higher than the number of days that we have utilizations over. This is because some of those utilizations span over two sessions, since practitioners can visit two locations in a day. Apart from the internal factors arising from the scheduling procedure (e.g. month, day, region), there are some external factors that can influence utilizations. We discuss the analysis of these external factors in Appendix I. An external factor of which we could confirm that it has an influence are mailings. Mailings are a marketing activity in which mails go out in batches in a certain region, reminding patients that they still have a reimbursed appointment left from their insurance in this year. From two other external factors, the Google Adwords budget and the wheather, we could not conclude that they have a significant effect on the number of incoming appointments.

Regio	**N.o. sessions**	**Avg. utilization**	**Variance of utilization**	**Avg. fragmentation**
Almelo	687	90.7%	0.96%	1.55
Amsterdam	984	92.0%	1.05%	1.36
Brabant	27	87.3%	1.08%	1.89
Den Haag	460	91.9%	0.66%	1.52
Drenthe	512	88.5%	1.17%	1.81
Enschede	1531	92.6%	0.65%	1.29
Friesland	320	92.7%	0.66%	1.07
Gelderland	1500	93.5%	0.62%	1.21
Groningen	1	58.3%	-	2.00
Hengelo	2814	91.8%	0.89%	1.32
Leiden	338	89.7%	1.21%	1.52
Polder	615	85.9%	1.37%	2.20
Rotterdam	820	91.0%	1.13%	1.53
Utrecht	527	90.6%	0.86%	1.49
Vriezenveen	426	95.0%	0.38%	0.96
Wierden	847	91.6%	0.81%	1.42
Zoetermeer	88	95.4%	0.35%	0.72
Total	**12497**	**91.6%**	**0.91%**	**1.40**

Table 2.6: Utilization and fragmentation per region (n = 11570 days/12497 sessions, period = 2019, source = Company X database/IT system agenda)

The planning department at Company X performs schedule optimisation activities, which influences the utilization and fragmentation. Each week the planning department of Company X gets a report with the amount of free time in the schedules and selects the least utilized schedules to optimize.

They move appointments around on the same day to create bigger spaces in the schedule and move appointments from later days forward to this session. This decreases fragmentation and increases utilization, but only for the more extreme cases. The utilizations and fragmentations we present are post-optimization. Establishing the effect of optimization is outside the scope of this research.

Fragmentation index

The fragmentation index varies from 0 to a maximum observed value of 7. The average fragmentation index in 2019 is 1.40, meaning there is an average of 1.4 unfilled gaps in the schedule. Usually the cause of the gaps is inefficient scheduling, leaving hard-to-fill gaps of 15 minutes. The fragmentation index is influenced by the schedule optimization activities that we mention above. In Table 2.6 we show the average fragmentation index per region.

If this was a single server system, a 1.40 fragmentation index would not necessarily be bad. However, since this is a multi-provider system, appointments can (to some extent) selectively be spread over locations. Thus, often it should be able to fill small gaps by directing shorter appointments to locations that need them. Due to myopic scheduling, patients often go to the same location each time and this selective effect is diminished. Also, when scheduling the scheduler often tells the patient there is room in the schedule between two certain times. Take, for example, an instance where those times are 15:00 and 17:00 and the appointment to be scheduled is a 60-minute appointment. When the patient says he/she would like an appointment at 15:30 (or even worse, 15:15), the scheduler often obliges. This results in fragmentation and thus less space for longer appointments. To the knowledge of the researcher there is no literature on appointment scheduling fragmentation that states any benchmarks that should be met. Thus, the interpretation of this KPI is subjective to Company X's and the researcher's opinion.

NP access time and access time violations

In Figure 2.10 and Figure 2.11 we show the distribution of NP access times using a histogram and box plot respectively. Furthermore, in Table 2.7 we show the average NP access times, the number of access time violations and the service level per region. We define the service level as the proportion of NP appointments that do not have an access time violation. The total average access times for NPs in 2019 is 20.65 days, which is close to the access time violation threshold of 21 days. The number of access time violations was equal to 6916, which is equal to 37.1% of all NP appointments. This number exceeds the 20% that we have set as a goal (80% service level).

We notice that regions with higher proportions of NPs have a higher NP access time. Amsterdam is the region with the highest NP access time. It is interesting to note that the utilization, variance in utilization and fragmentation index in this region are close to average. We hypothesize that in cities like Amsterdam, Den Haag and Rotterdam, patients are much less inclined to travel far for allied healthcare. Company X has noticed that people in these densely populated cities often do not want to leave their part of town for an appointment. Company X has many locations in these cities to be able to cater to the patients' needs. However, most of these locations only have one session per week. When one of the locations is overloaded, the patients close to this location often reject an appointment at another location and opt for the later option at the closer location, which may cause

the high access times in these cities.

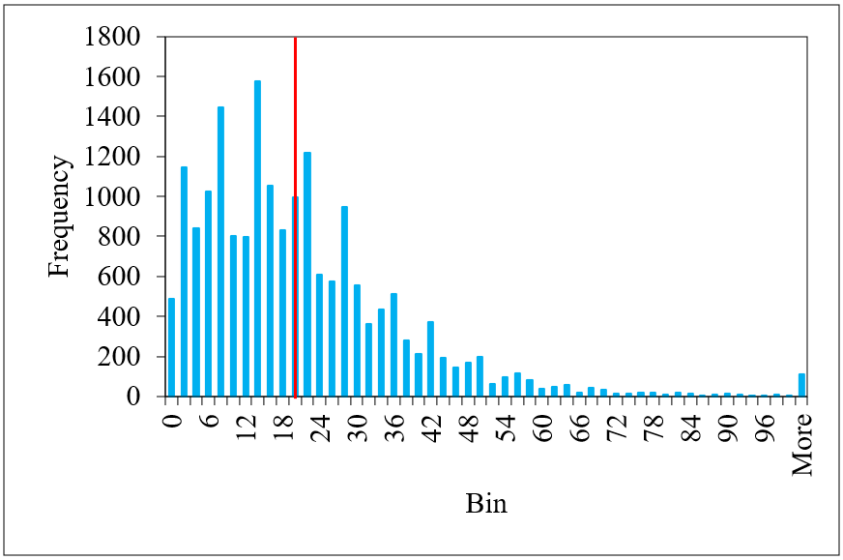

Figure 2.10: Histogram of NP access times in 2019 (bins of size 2; the red line depicts the target of 21 days)

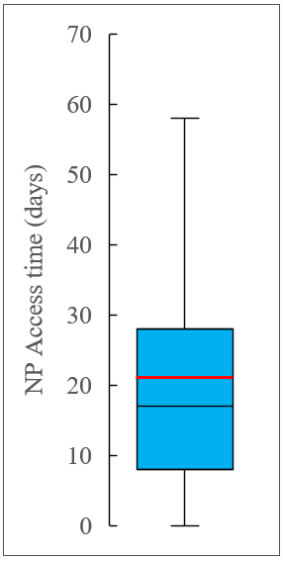

Figure 2.11: Box plot of NP access times in 2019 (the red line depicts the target of 21 days)

Region	N.o. patient appointments	Percentage NP	Average NP access time	Access time violations	Service level
Almelo	8820	9.6%	13.60	132	84.4%
Amsterdam	12841	16.7%	30.20	1410	34.1%
Brabant	250	20.8%	10.21	4	92.3%
Den Haag	6121	21.5%	24.41	604	54.1%
Drenthe	6033	15.9%	17.29	294	69.4%
Enschede	19010	8.7%	17.03	418	74.6%
Friesland	3906	20.3%	26.00	444	44.0%
Gelderland	19403	12.3%	19.63	873	63.3%
Groningen	3	100.0%	2.67	0	100.0%
Hengelo	30738	8.1%	16.34	640	74.2%
Leiden	3634	21.9%	17.76	233	70.8%
Polder	8880	10.9%	15.46	185	81.0%
Rotterdam	10783	19.0%	25.62	905	55.9%
Utrecht	5446	17.1%	18.38	326	65.1%
Vriezenveen	4926	7.8%	18.55	126	67.3%
Wierden	9260	6.7%	16.71	177	71.6%
Zoetermeer	1048	25.3%	27.41	144	45.7%
Total	**151102**	**12.34%**	**20.65**	**6915**	**62.9%**

Table 2.7: NP access times and access time violations per region (n = 151102 appointments, period = 2019, source = IT system agenda)

Another reason for the high access times is fragmentation of the schedules. Schedules often have multiple small gaps in the schedule, which can house 15 or 30-minute appointments, but not the 60-minute NP appointments. This reflects in the fact that regular RP consults have an average access time of 23.5 days. This activity does not have the 21 day threshold like NP appointments have, or

any other predefined access time requirements. Yet, the average access time is very close to that of NP appointments. Since these appointments are 30-minute appointments they can fit into many slots where NP appointments cannot. If the schedules would be less fragmented, less small gaps would occur and more big gaps would arise, which could accommodate NP appointments.

Furthermore, since the locations in cities like Amsterdam, Den Haag and Rotterdam are relatively new, they have a higher proportion of NPs compared to most eastern regions, where the organisation has matured already. NP appointments usually take 60 minutes, whereas almost all other appointments take 30 or 15 minutes. This makes NP appointments much harder to schedule since gaps in the schedule are often smaller than an hour.

PP/RP access times

Unfortunately, the diagnosis of a patient, which is what we base the desired access time on, is currently neither standardized, nor linked to the appointment data. Even if they were, there would be a vast amount of combinations of diagnoses and appointment types, for which we would have to manually appoint a desired access time. Therefore, it is not possible to structurally determine the diagnosis per appointment to measure our current performance. Figure 2.12 shows the average access times per RP activity (only including those relevant to our research) in 2019 and their corresponding interquartile ranges. The figure does not include "Jaarlijkse controle" (i.e. Yearly checkup) (Avg: 177.4, IQ-range: [16, 350]), since including it would significantly stretch the y-axis, which would be detrimental to interpretability of the figure.

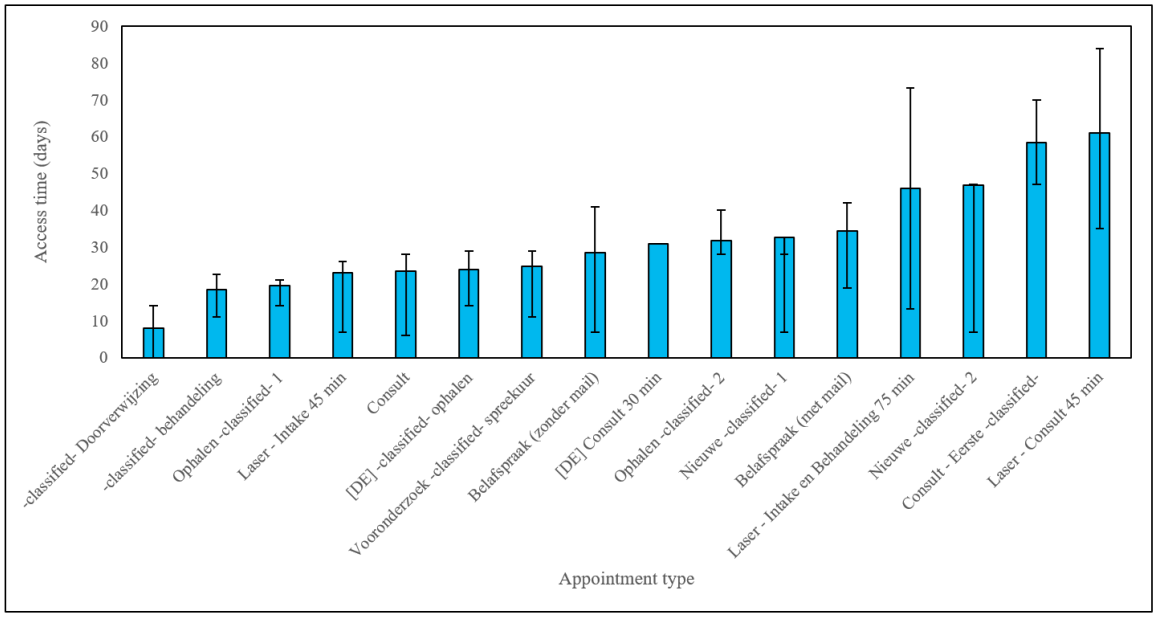

Figure 2.12: Average access times for relevant RP activities (2019) with corresponding interquartile ranges

Note that some interquartile ranges are completely below the average, which is due to a small number of observations substantially increasing the average. A large interquartile range indicates that there is no clear desired access time for this, as access times vary a lot. We see that "Laser - Intake 45 min" and "Consult" have similar access time distributions to NP appointments (Avg: 20.6 days, IQ-range: [8,28]) and thus probably have similar access time requirements. We also notice that "Ophalen

-*classified*- 1", "Ophalen -*classified*- 2" and "Consult - Eerste -*classified*-" have relatively narrow interquartile ranges, meaning that for these appointment types it is more clear what the access time should be than with other types. From unstructered interviews we know that "Ophalen -*classified*- 1" has a standard access time of 2 weeks, which confirms this. Though some characteristics can be inferred from this visualisation, it does not show whether the current performance on this KPI is good or bad.

Distance from patient to location

Finally, we want to know what the current performance is on our distance KPI. In Table 2.8 we depict the average distance travelled from patient to location per region. On average this is 3.45km, which includes only patients with geographic coordinates within the maximum and minimum latitude and longitude of the Netherlands. The reason for this is that we see a lot of potentially faulty coordinates that are hundreds or thousands of kilometers outside of the Netherlands, which could skew the data significantly.

By calculating the distance for each patient to each location we also see that in 2019 a patient lived on average 2.29km from the closest location. The discrepancy between average travel distance and the closest location is mainly due to two reasons. First, new locations open up all the time and a patient might have a preference for the location it has always gone to even though it is not the closest anymore. Second, the closest location is not always the best/an available option (e.g. location with only medical pedicure) and is thus not always chosen. Figure 2.13 shows a comparison of the distributions of distances from the patient to the appointment location and to the closest location. It shows that the medians (1.41 and 1.06 resp.) are close to each other. This indicates that the mean of the distances to the appointment location is significantly skewed by a small number of large distances. Thus, for most patients the distance to the appointment location is smaller than what the mean suggests. This means that Company X schedules patients close to (or at) the closest location rather often. We see that, in regions with a bigger area, the average distance is higher. Examples are Drenthe, Friesland etc., which stretch over entire provinces and have relatively less locations per km^2 (we cannot calculate the N.o. locations per km^2, since we do not have a delimitation of where one region stops and another starts). Regions that have more locations per km^2 usually have a lower average distance. The main location is an exception, since Company X sends a lot of patients from far to the main location. This location has a high capacity, but is also considered a showpiece by Company X, due to its modernity and novel equipment.

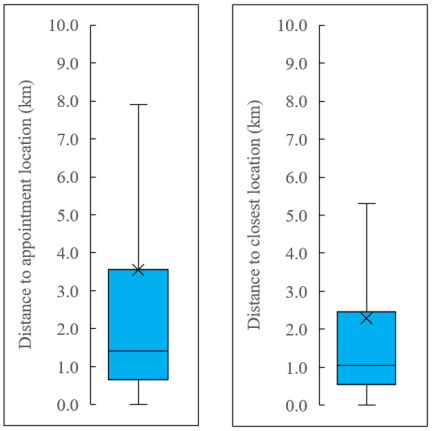

Figure 2.13: Distance from patient to appointment location compared to distance to closest location

Summary of KPI analysis

In conclusion, from Company X perspective we see an average utilization of 91.7%, which means we have 32.0 minutes of idle time on average for each workday. This closely relates to the fragmentation index of 1.40, which means there are on average 1.40 unused gaps in the schedule per workday. If

Regio	N.o. appointments	N.o. locations	N.o. sessions	Average distance (km)
Almelo	8412	17	687	3.12
Amsterdam	11510	25	984	3.04
Brabant	227	1	27	2.41
Den Haag	5779	21	460	2.27
Drenthe	5751	12	512	4.52
Enschede	18352	24	1531	2.25
Friesland	3508	11	320	5.55
Gelderland	18546	37	1500	3.59
Groningen	2	1	1	9.45
Hengelo	29805	43	2814	4.57
Leiden	3447	9	338	5.33
Polder	6949	17	615	4.00
Rotterdam	9761	38	820	2.89
Utrecht	5222	13	527	2.37
Vriezenveen	4709	10	426	2.95
Wierden	8906	12	847	2.44
Zoetermeer	997	4	88	3.04
Average	**8346.06**	**17.35**	**735.12**	**3.45**

Table 2.8: Distance travelled from patient to location (n = 141883 appointments, period = 2019, source = IT system agenda/Company X database)

Company X would solely focus on their own interest's, the utilization would be closer to 100% and the fragmentation would be closer to 0. Furthermore, the standard deviation of the utilization is 9.58%. This shows that there is a significant fluctuation in utilization and thus an uneven spread of workload over sessions.

From a patient perspective, we see that access time for NPs is 20.65 days on average, which is already very close to the threshold of 21 days. A reason for this is that schedulers often ask a patient which location they want to go to, while the patient has no knowledge of which locations are crowded. If the patient chooses a crowded location, it often gets an appointment at this crowded location at a date later than the 21 day threshold. Another reason for high access times is schedule fragmentation, which results in a lot of gaps too small for NP appointments. This fragmentation is induced by the schedulers letting patients pick an appointment time in an open interval instead of suggesting a time that does not further fragment the schedule. We see that the median distance from the patient to the appointment location (1.41 km) is close to the median distance from the patient to the closest location (1.06 km). This is caused by schedulers asking the patient which location they would like to go to, which usually is the closest one. They then often proceed to schedule the appointment at that location without considering others.

The KPIs show that blindly letting patients do the decision-making is detrimental to KPIs from Company X's perspective. Furthermore, when it comes to access time, this way of scheduling is actually detrimental to the patients themselves as well. A well-informed scheduler can actually often pick more preferable slots for the patient than what the patient could suggest without complete information. Currently, patients often seem to simply go to one of the closest locations.

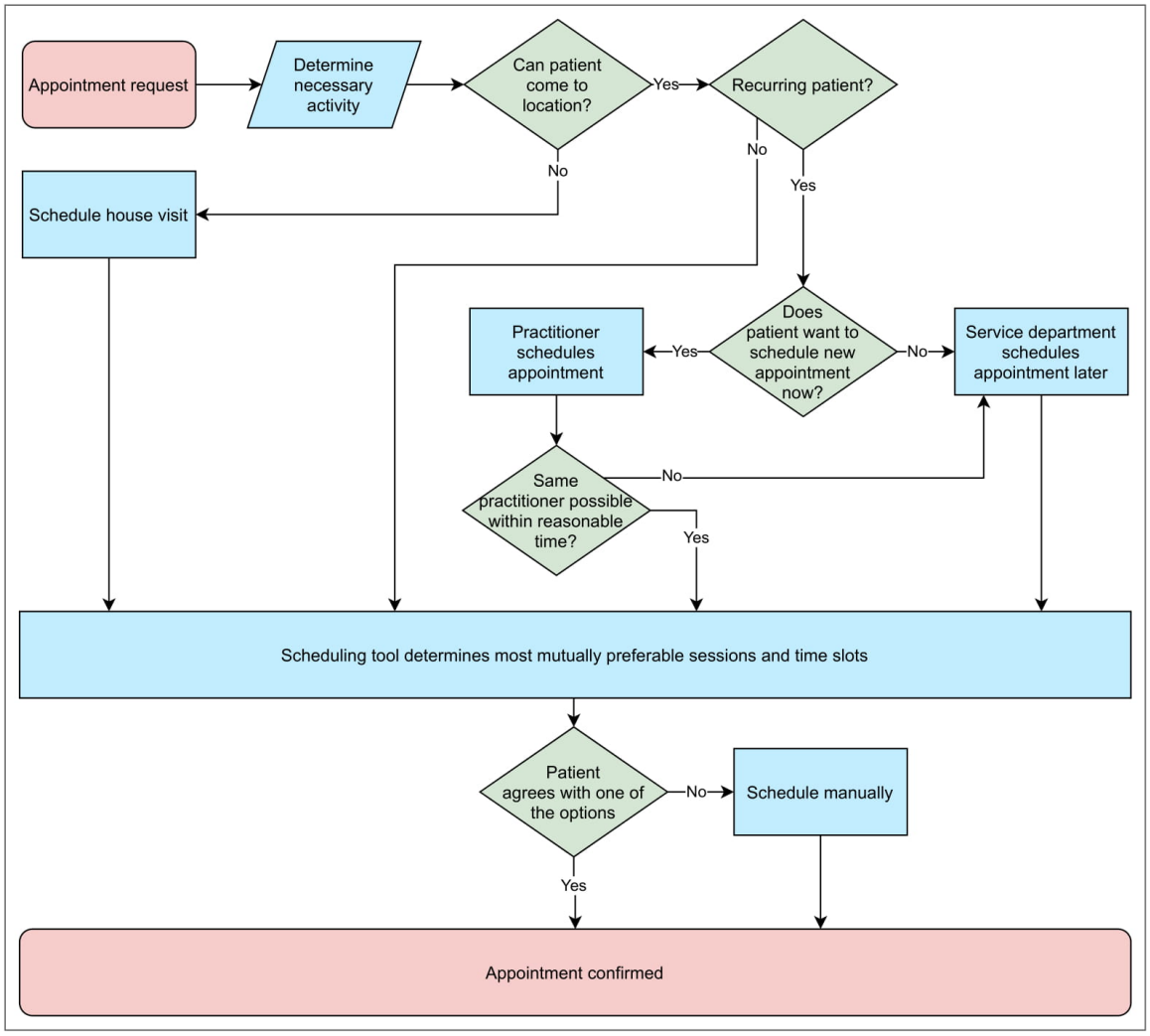

Figure 2.14: Flowchart depicting the desired new scheduling process

2.4 Desired scheduling process

Now we have established what the current situation looks like, we can determine what the new situation should look like. We determine what the new situation should look like based on our interviews with the stakeholders within Company X and what we learned from the data and phone call recordings. The new situation should resemble the one depicted in Figure 2.14. As the flow chart shows, a lot of decision making is transferred from the patient and scheduler to the scheduling wizard to provide the patient with an option they prefer that also considers Company X's interest. In Appendix J we depict the desired scheduling process in more detail, per patient type. In this section we also briefly discuss the desired scheduling process per patient type.

2.4.1 New Patients

When we look at the NP scheduling process, we want to go to a situation where the scheduling tool takes over the finding of sessions and time slots. The tool should suggest the most preferable options for Company X, of which the patient can pick one. If the patient has very specific needs the

scheduler can still schedule manually. Manually in this case means that the scheduler does not pick an appointment from the suggestions and schedules an appointment that suits the patients specific needs. In this case we do still show information on the consequences of scheduling at certain locations and times. This way the scheduler can still take both the preference of the patient and the preference of Company X into account and maybe steer the patient to a slot that is not too detrimental for Company X. Our scheduling will be able to predict overflow and will preventively try to send anyone that is willing to further locations in earlier spots, alleviating the issue of high access times in cities like Amsterdam, Den Haag and Rotterdam.

2.4.2 Recurring Patients

The rigidity that the RP way of scheduling entails, causes an uneven spread of appointments and fragmentation of the schedule. Always scheduling at the same location also causes overloaded locations to stay overloaded. Based on the phone call recordings between the patients and the schedulers, we notice that scheduling at a different location is often no problem for the patient, even less so if the patient can be treated there sooner. Our proposed scheduling tool should state which option is most preferable based on the access time per scenario, the predicted occupancy for the sessions and a possible preference for a practitioner. Furthermore, the suggested time slots should prevent fragmentation of the schedule. As mentioned in 2.3.1, the desired access time for RPs is not documented, nor are there standards documented for this purpose. In our scheduling tool we therefore want a field in which the scheduler chooses the desired access time, or selects "as soon as possible" (mainly for NPs), "within 24 hours" or "today". When scheduling an appointment, the tool will document the desired access time for each appointment, which means we will be able to measure performance in the future.

2.4.3 Periodic Patients

Similar to scheduling NPs and RPs we want our scheduling tool to represent Company X's preference in scheduling options and the scheduler to communicate this preference to the patient, such that together they can find a time slot/location/practitioner combination that suits both Company X and the patient's preferences and needs.

2.5 Conclusion

Company X serves patients in three distinct non-overlapping groups: New Patients (NPs), Recurring Patients (RPs) and Periodic Patients (PPs). In 2019, these groups account for 12%, 59% and 29% of total appointments, respectively. The NP's maximum allowed access time is a standard of 3 weeks; the RP's desired access time is defined by the foregoing appointment; and the PP's desired access time is based on a predefined time in between appointments determined by their care plan.

We decide to measure performance using the following KPIs (current performance measured over 2019):

- Utilization, where the average utilization of a practitioner's workday was 91.7%.
- Variance of utilization per practitioner workday, which was 0.917% (standard deviation: 9.58%).

- Fragmentation index, which was equal to 1.40 (unused gaps per schedule for 1 day, for 1 practitioner).
- NP access time, where the average NP access time was 20.65 days (Variance: 384.5 days, Norm: 21 days).
- NP access time violations, of which there were 6915 (37.1% of all NP appointments).
- Absolute deviation of RP/PP access times from the target date (current day plus desired access time) divided by the desired access time. There is no data to measure performance for this KPI. Our scheduling tool will use a field in which the scheduler states the desired access time and document this. Therefore we will be able to measure performance on this KPI in the future.
- Distance from patient to the location, which was 3.45km on average, though the median is only 1.41km (due to a small number of appointments with very high distances skewing the average).

Figure 2.8 schematically shows the current scheduling process. This process does not take Company X's preference into account. The patient makes most scheduling decisions, which are often not mutually beneficial. We see that KPIs from Company X's perspective suffer from the current way of scheduling. From a patient perspective, the access time suffers from this way of scheduling as well, mostly due to fragmentation and myopically only considering one location. This means 37.1% of NP appointments are not performed in time, which is much higher than our norm of 20%. A well-informed scheduler can actually often pick more preferable slots for the patient than what the patient could suggest without complete information. Currently, patients often seem to simply go to one of the closest locations, as the median distance to the appointment location is close to the median distance from the closest location to the patient.

The utilization of sessions has a standard deviation of 9.58%. To get this down want to incorporate a predictive model on the proposed scheduling tool. This model should predict how busy locations are going to be, so we can spread appointments to other locations before it is too late. Apart from the internal factors arising from the scheduling procedure (e.g. month, day, region), there are some external factors that can influence crowding. An external factor of which we could confirm that it has an influence is mailings. Mailings are a marketing activity in which mails go out in batches in a certain region, reminding patients that they still have a reimbursed appointment left from their insurance in this year. From two other external factors, the Google Adwords budget and the wheather, we could not conclude that they have a significant effect on the number of incoming appointments.

We can conclude that the current way of scheduling is detrimental from the perspective of Company X and often also from the perspective of the patient. In this research we therefore propose a new scheduling tool, which provides the scheduler with appointment options that consider Company X's interests as well as the patient's interests. The patient and scheduler could then negotiate an appointment from these options. Figure 2.14 schematically shows the proposed new scheduling situation.

Chapter 3

Literature

This chapter discussess literature relevant for our research. We divide our efforts in three parts. In Section 3.1 we research predictive modelling methods. In Section 3.2 we review literature that focuses on decision support tools in online appointment scheduling. Finally, we dedicate Section 3.3 to model validation techniques, since this is something that we need to do throughout our research. In Section 3.4 we conclude the chapter.

3.1 Predictive modeling

In this section we research literature to answer our second research question:

"How can we make a predictive model of the occupancy of a location?".

Predictive analysis is a broad term that describes a variety of statistical and analytical techniques used to develop models that predict future events or behaviours [37]. The methods used in Predictive Analytics mainly stem from three data analytic disciplines: classical statistics, Data Mining (DM) and Machine Learning (ML) [2].

The distinction between DM and ML can be labeled as somewhat vague and are often confused. In fact, classical statistics, DM, and ML all overlap in some way. DM concerns finding useful patterns in the data. However, there is a wide variety of definitions and criteria for data mining [26]. DM combines the knowledge of statistics, subject matter expertise, database technologies, and machine learning techniques to extract meaningful and useful information from the data. Predictive Analytics can be seen as a subset of DM, for which classical statistics techniques and ML methods can be used. The main three classes of predictive analytics are Classification, Regression and Time Series methods.

3.1.1 Model requirements

We search the literature for methods that meet the following requirements:

- The method should use multiple predictors and predict one continuous response variable.
- The method should be dynamic as it should be usable for many years to come.
- The method should require little computation time, because the occupancy appointment load estimation should be calculated before scheduling each appointment. This means not too many variables can be considered.
- The method should not be very complex, since we do not want to allocate too much time of our research to building the predictive model and a complex model would be detrimental to computation time and transparency.

Since our predictive model should merely be one of the many factors in determining scheduling options in our scheduling tool, we want to keep it quick and simple. On one hand, because we do not want to allocate too much time of our research to the predictive model, and on the other hand, because the predictive model needs to work very quickly to deliver information almost instantly upon request during scheduling. Therefore, we limit ourselves to basic, widely used methods. We do not discuss all methods we find in literature in much detail to limit the length of the thesis and to keep the literature review concise. If the reader wishes to familiarize him/herself with one of the methods mentioned in the literature review that we do not discuss in much detail, we advise him/her to review literature on this method, either from the references from this document or from other sources.

3.1.2 Methods in literature

Kyriakides & Polycarpou [27] reviewed a number of time series modeling approaches for load forecasting that could also be useful in our research. They summarized the available methods into the following categories: regression models, linear time series-based methods, state-space models, and non-linear time series modeling through machine learning. In the research in [27] the focus is on electric load forecasting as used by electricity providers, but the same methods can easily be extended to regular demand forecasting. For our predictive model we choose not to use state-space and non-linear time series models. From a quick review of literature we believe these model types seem too complex for the simple and fast predictive model that we want. The focus is therefore on regression models and linear time-series based models.

Box & Jenkins [4] formalized a number of time series analysis methods. These methods are Auto-Regression (AR) and Moving Average (MA) in time series and the combined methods of Auto-regressive Moving Average (ARMA) and Auto-regressive Integrated Moving Average (ARIMA). Many forecasting methods in current literature are in some way derived from these methods.

Mirowski et al. [35] apply a Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model, which is an extension on auto-regressive moving average (ARMA) models [16]. They also use a state-space model using Kalman filtering, Holt-Winters double seasonal exponential smoothing, three kernel regression methods and a neural network model. The purpose of the research was short-term load forecasting on aggregates of power meters in a mid-sized city.

Yan & Su [57] extensively discuss linear regression as a measure of prediction. In regression the dependent variable is modeled as a function of independent variables, corresponding regression parameters (coefficients), and a random error term which represents variation in the dependent variable unexplained by the function of the dependent variables and coefficients. In linear regression the dependent variable is modeled as a linear function of a set of regression parameters and a random error. Linear regression can be split into simple linear regression and multiple linear regression. The first uses one predictor variable and the second uses multiple. Simple linear regression does not meet our requirements and will thus not be further discussed in this review. Apart from linear regression there is also non-linear regression, which assumes that the relationship between dependent variable and independent variables is not linear in regression parameters. Nonlinear regression model is more complicated than linear regression model in terms of estimation of model parameters, model selection, model diagnosis, variable selection, outlier detection, or influential observation identification. Therefore we decide that non-linear regression does not meet our low complexity requirement, which means we also do not discuss non-linear regression in the remainder of this literature review.

Tanizaki et al. [47] evaluate four demand forecasting methods: **Bayesian** Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, and the Stepwise Method. The authors use these methods for the purpose of demand forecasting in restaurants using machine learning and statistical analysis.

Demir [8] compares and validates the performance of four predictive analytics methods. Logistic regression, Regression trees, Generalized Additive Methods (GAMs) and Multivariate Adaptive Regression Splines (MARS) models. The author uses this methods in the design of a decision support tool for predicting patients at risk of readmission. This research is therefore mainly focused on classification.

Williams [52] evaluates an Autoregressive Integrated Moving Average method with Explanatory Variables (ARIMAX) for a multivariate vehicular traffic flow prediction. The ARIMAX is an extension to the ARIMA method that incorporates more than one variable in the MA and AR. This way external factors other than time can be incorporated.

Sutthichaimethee & Pruethsan [46] use a method that combines ARIMAX with Vector Autoregressive Moving Average (VARMA) in a method dubbed Vector Autoregressive Moving Average with Explanatory variables (VARIMAX). They use the method to forecast CO2 emission from the energy consumption in the Rubber, Chemical and Petroleum Industries sectors in Thailand. VARIMAX is a vector based approach, which means it can use multiple time series for prediction and also predict vectors instead of a single scalar. The distinction between the ARIMAX and VARMA is that in ARIMAX the explanatory variables may be deterministic or under the experimenter's control.

Evaluation of methods

In Table 3.1 we summarize the predictive modeling methods that we found in literature and we assess for each if they meet our requirements. It is hard to establish how computationally demanding and complex the methods will be for our problem context (third and fourth requirement, as defined in Section 3.1.1). The assessment of those requirements is therefore purely based on our beliefs acquired by qualitative information that we read in literature and might not be completely accurate for each approach. The methods that we list in Table 3.1 may be generalizations or extensions of other models mentioned. In case of generalizations we refer to the basic form (not extensions), e.g. multiple linear regression refers to the non-bayesian (**frequentist**) method.

Methods	Multiple predictors & single continuous response	Dynamic	Low computational effort	Low complexity
Regression				
Multiple linear regression	Yes	Yes	Yes	Yes
Bayesian Linear Regression	Yes	Yes	Yes	Yes
Boosted Decision Tree Regression	Yes	Yes	No	Yes
Decision Forest Regression	Yes	Yes	No	No
Stepwise regression	Yes	Yes	Yes	No
Logistic regression	Yes	Yes	Yes	Yes
Regression trees	Yes	Yes	Yes	Yes
GAMs	Yes	Yes	No	No
MARS	Yes	Yes	No	No
Time series methods				
ARMA	No	No	Yes	Yes
ARIMA	No	Yes	Yes	Yes
SARIMA	No	Yes	Yes	Yes
ARIMAX	Yes	Yes	Yes	Yes
VARIMAX	Yes	Yes	Yes	Yes

Table 3.1: Predictive Analytics methods found in literature

We see six methods that meet all of our requirements: Multiple linear regression, Bayesian linear regression, Logistic regression, Regression trees, ARIMAX and VARIMAX. Before we go into further detail with these methods we eliminate Bayesian linear regression, regression trees, and VARIMAX. We eliminate Bayesian linear regression simply because we adopt a frequentist point of view. We eliminate Regression trees because they seemingly do not lend themselves well for our prediction problem. Being able to result only in the values at the leaves of trees seems counter-intuitive for a prediction that is closer to forecasting and should result in a continuous variable based on time series data and other external factors. VARIMAX describes the relations between all variables in the vector, which is slightly exorbitant since we are only interested in the effect of multiple predictor variables on a single response variable. VARIMAX, or VARMA for that matter, is more suitable when we want to predict multiple values based on their correlations. ARIMAX is therefore more suited for our problem. This leaves us with Multiple linear regression, Logistic regression and ARIMAX.

Our model should give us an indication of how high the future occupancy will be on a location based on time series processes. These processes are influenced by external factors, which are not necessarily time series-based (e.g. constructing time series of current utilization for each location requires too much computing power). This is quite scarcely discussed in literature. The predictive model should reflect the number of appointments that are expected to come in from now to a given point in the future. The time series describing incoming appointments can be thought of as being composed of five components, namely level, trend, seasonal variations, cyclical movements and irregular random fluctuations [44]. However, the occupancy of a location on a given number of days into the future is also highly state-dependent. Factors like the current utilization on the given number of days ahead significantly influence the expected future utilization. Furthermore, as we show in Appendix I, mailings influence occupancy as well.

It is hard to incorporate external factors in an a straightforward ARIMA based model. Therefore, we opt to evaluate both a multiple linear regression model and a Logistic regression model. In both these models we incorporate elements from the ARIMA model. Traditional ARIMA models can all be expressed as a regression where the previous time periods/moving average etc. are factors. What this means is that we can add additional factors representing exogenous variables or to represent periodicity. Specifically, we will use a Multiple linear regression and a Logistic regression, in which we represent the decomposed time series elements as predictors. For instance, we could use month of the year as a 12-way categorical predictor and marketing activity as a boolean predictor indicating if there will be marketing events (e.g. a mailing) leading up to the prediction date. We can include the MA aspect by updating weights for predictors based on new observations. The updating procedure we will use is simply retraining the model with a moving time-interval. Other options include training partial regression models over new data and using for instance a gradient descent approach. However, since our model only takes a few seconds to train, there is no need for more intricate procedures.

3.2 Decision support tools in online appointment scheduling

Before we start building our final deliverable we need to know exactly what functionalities it should feature. To determine which functions are necessary we first research what common challenges there are in online appointment scheduling, which is also commonly referred to as sequential, rolling-horizon, or dynamic scheduling. We try to focus on literature on multi-provider systems, as is the situation at Company X. However, there is not a vast amount of literature on this subject. In this section we gather information which helps us answer our third research question:

"How can we create and validate a scheduling support tool that can adequately inform schedulers on which time slots and locations to choose?"

We discuss the validation of the predictive model as well as that of our scheduling tool in Section 3.3. Literature should provide us with several ideas on how to help schedulers make the right decisions. For example, this could include heuristics, algorithms, or visual aids that inform schedulers on what to do. This should give us information on what we show the scheduler and what should happen in the back-end of the tool. We first discuss a literature gap that we experienced. Then, we research decision making methodologies to base the suggestion-generation in our tool on and corresponding criteria weighting procedures. Finally, we discuss methods for performance visualization in decision support systems, which we aim to use in our scheduling tool.

3.2.1 Literature gap

Most literature discusses a single server or a single location. Literature about multi-provider systems is far less frequent. Multi-server models are rich in the queueing literature, but as pointed out by Erdogan and Denton [10], the appointment scheduling problem differs from a typical queueing model in two main aspects. First, the steady state assumptions in queueing models do not hold in the transient clinical environment. Second, for most queueing models, it is assumed that patients arrive stochastically rather than according to a schedule. Therefore queuing literature also does not provide much information relevant to us. Company X has locations all throughout the country and a lot of them are close together which means they can exchange/pool patients. Enschede, for example, has 36 locations, 16 of which are duplicates of a different type, e.g. diabetics location on same physical location as regular location. There is a central scheduling system for all locations in the country, which can allocate patients to any one of these locations. This is actually a rare situation in healthcare, or in any industry for that matter.

Apart from having a fairly unique problem context, most similar ambulatory care services negotiate appointment scheduling with the patient by means of schedulers or directly by the caregiver. As a means of service, this process is not often guided by intelligent decision support tools. Our research taps into a yet largely unexplored area of combining patient choice and provider choice in a mediating intelligent scheduling support tool.

Another limitation in literature on online scheduling is that it mostly proposes optimization-based single-solution methods [48] [56]. The issue with this is that it neglects the human aspect of scheduling that lies in negotiation between patient and scheduler. In [57] the negotiation is actually considered, but only by providing the patient with the possibility to refuse the suggested interval of time and reiterating, listing the previously suggested solution as tabu. Furthermore, they consider the human aspect by considering patient fairness, which is outside the scope of our research. As mentioned in Chapter 1, we want to maintain the human aspect to ensure high quality of service and patient satisfaction and retention. This means we should consider patient preference. However, this patient preference should be interpreted by the scheduler, which is where most patient choice methods deviate from what we want to achieve. We want to incorporate some specific clearly-defined patient preference functionality in our tool (e.g. requesting specific practitioner), but leave the more subtle preferences to the scheduler (e.g. trade-off between distance to location and access time). Thus, we want our scheduling tool to provide a top-k ranking with time slot suggestions, based on a selection algorithm that takes into account the KPIs from a patient perspective and from Company X's perspective. This so-called Top-k retrieval problem means obtaining the k best results according to a ranking function [32]. The ranking function should incorporate the suggestions' effect on our KPIs. From the resulting suggestions, the scheduler selects the one that he/she considers to be most preferable for both parties and negotiates another suggestion upon refusal of a suggestion. To the author's knowledge, top-k multi-attribute ranking-based approaches to appointment scheduling are not yet present in literature, nor is providing additional information to the scheduler on the non-transparent effects of the scheduling decision (which we discuss in Section 3.2.5).

3.2.2 Multi-attribute ranking methods

In Section 1.2.2 we classified the problem as a $P|online - r_j|L_{max}$ problem in essence. This means we have m parallel servers, representing the number of practitioners; we have online scheduling where

jobs are released as soon as they are created; and as objective we try to minimize the maximum lateness. Here we defined the maximum lateness for a day as $[max(access\,time) - 21]^+$. However, this objective is obscured by other conflicting factors that are the KPIs in Section 2.3.1. This can be represented by penalizing bad performance on those conflicting factors, or by seeing the problem as a multi-objective problem. Penalties in objective functions are used mostly when a variable crosses a certain threshold, whereas multi-objective problems intend to maximize/minimize multiple objectives. In our case a multi-objective problem is more fitting, seeing as most criteria are not linked to a threshold value.

In literature one can find extensive research on optimization techniques such as mathematical programming. These methods could be applied to our context as well. A linear program would be able to find the optimal solution given an objective function. Using a Tabu list in an iterative algorithm, we would be able to find a top-k ranking. However, in our context, the strength of linear programming methods is diminished by the fact that we have multiple conflicting objectives (trade-offs). This means that we are not looking for a strictly optimal solution, but the most satisfying solution [55]. To get a viable solution, the conflicting objectives would be weighted/normalized (or both), which makes this a subjective method, rather than an objective optimization. Furthermore, for ranking we would have to solve a great number of consecutive linear programs, which in our solution space is not sufficiently computable. There are mathematical programming methods, such as Approximate Dynamic Programming (ADP), used to handle large solution spaces. ADP methods have been applied in online scheduling [50], but to the author's knowledge only for single-provider contexts. When $k = 30$, we would need to solve this same problem 30 times over a multi-provider context, further intensifying the "curse of dimensionality" that ADP attempts to overcome. We presume that running such an algorithm in the limited time available for online scheduling, is not currently feasible without inaccurate approximation (for our context). This, in combination with the fact that the resulting solution would ultimately be subjective, suggests that mathematical programming is not an appropriate solution method in this context.

Often, the aforementioned "curse of dimensionality", related to large solution spaces, is handled using heuristic methods. Heuristic methods are practical problem solving methods that do not have a guarantee of an optimal solution. The fact that the optimal solution in our context is subjective (due to conflicting objectives), means that there is inherently no guarantee that the optimal solution is perceived as the best solution by the decision makers. Therefore, both a heuristic method and an exact method, would not have a guarantee of a best-perceived solution. Thus, in our context, a heuristic method can solve the computation time problematic while resulting a solution that is not necessarily perceived as better or worse than that of an exact method. However, not many heuristic methods are useful for conflicting objectives. A field of research that is concerned with multi-criteria heuristic methods for decision making problems is Multi-Criteria Decision Analysis (MCDA). We discuss this field of research in detail in the remainder of this section.

3.2.3 Multiple-Criteria Decision Analysis

Our ranking function is a pivotal aspect of the proposed scheduling tool and should be well-substantiated. Therefore, we search literature about the configuration of a ranking function and the corresponding weight tuning. Multiple-Criteria Decision Analysis (MCDA), which is a sub-discipline of Operations Research, is an activity which helps making decisions mainly in terms of choosing, ranking or sorting the actions [12]. MCDA is also often synonymously referred to as Multiple-Criteria

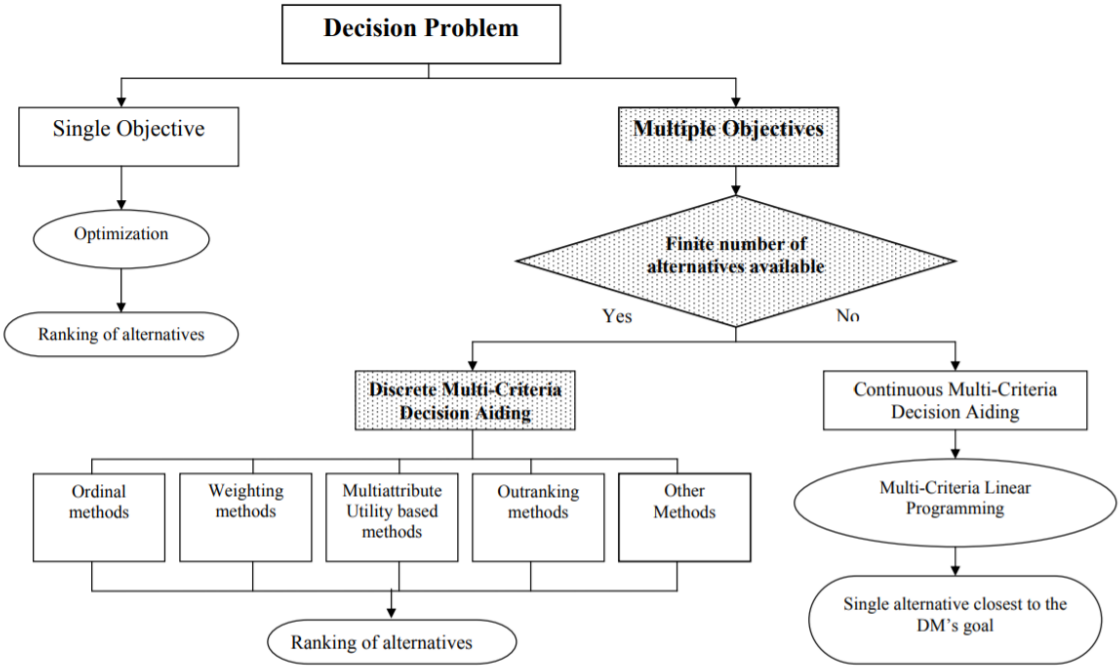

Figure 3.1: Decision Aiding Methods classified by Kodikara [25]

Decision Making (MCDM) or Multiple-Attribute Decision Making (MADM). Usually in MCDA there are three steps to go through: generation of alternatives, determination of criteria and alternative selection. The former two naturally flow from our research thus far. Alternatives correspond to possible combinations of time, location and practitioner represented by an appointment. Criteria correspond to the KPIs that we substantiated in Section 2.3. This leaves us with the selection method, which in literature is often what is referred to when speaking of an MCDA method, as they represent the most discussed part of the three MCDA steps. Figure 3.1 depicts a classification of decision making methods. [43] distinguishes four main decision problematics: selection, sorting, ranking and description. These problematics represent possible goals of an MCDA procedure. For explanation of each problematic we refer to [43]. As we stated before, our problematic is ranking. However, the number of possible alternatives is vast, since Company X plans some sessions years ahead. In Section 4.2.2, we therefore introduce a two-stage ranking method. In the first stage we first filter sessions, after which we obtain a top-k ranking of the filtered sessions. Then, in the second stage, the solution space is small enough to enumerate and rank all appointments options in the top-k session selection.

There are many different MCDA methods, each of which can result in different outcomes. Selecting the right method is therefore a separate preliminary decision process on its own. For this purpose, Wątróbski et al. [54] propose a generalised framework for multi-criteria method selection. We utilize this framework for our MCDA method selection. In the proposed framework, they show that each decision-making problem can be described by the decision maker using the maximum of nine descriptors belonging to the set $\tilde{c} \subseteq c$. At the first level of the hierarchy, the decision maker only defines the general descriptors of the decision problem:

- c_1 – whether different weights of the individual criteria will be taken into account in the decision problem; possible values are: 0 – no, 1 – yes;

- c_2 – on what scale the criterial performance of the variants will be compared; possible values are: 1 – qualitative, 2 – quantitative, 3 – relative;
- c_3 – whether the decision problem is characterized by uncertainty; possible values are: 0 – no, 1 – yes;
- c_4 – what the decision problematic is; possible values are: 1 – selection, 2 – classification, 3 – ranking + selection, 4 – classification + selection.

The knowledge about the decision problem can then be clarified by the decision makers. While we can assume that c_2 is fully defined, the rest of the descriptors of the decision problem on the second level of the proposed hierarchy are presented as follows:

- $c_{1.1}$ – if weights are used, what will their type be? Possible values are: 1 – qualitative, 2 – quantitative, 3 – relative, or 0 if $c_1 = 0$;
- $c_{3.1}$ – if the problem is characterized by uncertainty, which uncertainty aspect does it concern? Possible values are: 1 – input data uncertainty, 2 – DM's preference uncertainty, 3 – both, or 0 if $c_3 = 0$;
- $c_{4.1}$ – if the problematic of ranking is considered, what kind of variants' ranking is expected? Possible values are: 1 – partial ranking, 2 – complete ranking, or 0 if $c_4 = 0$.

The third level of the descriptors' hierarchy refers only to $c_{3.1}$ and addresses data or preference uncertainty in the decision problem:

- $c_{3.1.1}$ – if the uncertainty concerns the data, does it refer to the weights of criteria or to the variants' criterial performance? Possible values are: 1 – criteria, 2 – variants, 3 – both, or 0 if $c_{3.1} = 0$;
- $c_{3.1.2}$ –if the uncertainty concerns the DM's preferences, what thresholds will be used in the decision problem? Possible values are: 1 – indifference, 2 – preference, 3 – both, or 0 if $c_{3.1} = 0$.

Figure 3.2 shows what subsets of MCDA methods are applicable to a given $\tilde{c}$ where $m_{i.j.k.l}$ denotes the characteristics for method i corresponding to $c_{j.k.l}$. The author uses these different notations to be able to apply the framework to a subset $\tilde{c}$ such that $dim(c) \leq dim(m)$ when not all criteria are known. In our case all criteria are known and our values for $\tilde{c}$ are $c_1 = 1$, $c_2 = 2$, $c_3 = 0$, $c_4 = 3$, $c_{1.1} = 2$, $c_{3.1} = 0$, $c_{4.1} = 2$, $c_{3.1.1} = 0$, $c_{3.1.2} = 0$. In our case the possible methods are EVAMIX, MAUT, MAVT, SAW, SMART, TOPSIS, UTA and VIKOR. For detailed explanation about the various MCDA methods we refer to [54], where they provide references for each of the 56 described methods.

EVAMIX uses pairwise comparison of alternatives [49], which requires too much computation power for our problem, due to the large number of alternatives that we have. Therefore, we will not use EVAMIX. MAUT (multi-attribute utility theory) is an extension to MAVT (multi-attribute value theory) that handles contexts where "risks" or "uncertainties" have a significant role in the definition and assessment of alternatives. In MAUT the preferences for each attribute are represented by (marginal) utility functions [12]. These marginal functions are aggregated in a unique overarching utility function. With MAVT this function overarching utility function is replaced by a value function, which is a weighted sum of functions over each individual attribute. The problem with MAUT is that it requires stronger assumptions to ensure additivity and is very difficult to apply [22]. It is

MCDA method properties		m_{i1}	m_{i2}	m_{i3}	m_{i4}	$m_{i1.1}$	$m_{i3.1}$	$m_{i4.1}$	$m_{i3.1.1}$	$m_{i3.1.2}$	Subset of MCDA methods	
											Names	Abbreviations
Rules	R_1	0	1	0	1	0	0	0	0	0	Maximax, Maximin, MIN_MAX[1]	$\{M_X, M_N, E_M\}$
	R_2	0	1	1	4	0	1	0	2	0	FuzzyMIN_MAX[1]	$\{E_F\}$
	R_3	0	1	1	3	0	2	1	0	3	ELECTRE IV	$\{E_4\}$
	R_4	0	2	0	1	0	0	0	0	0	Goal Programming	$\{G_P\}$
	R_5	0	2	1	3	0	1	1	2	0	NAIADE I	$\{N_1\}$
	R_6	0	2	1	3	0	1	2	2	0	COMET, NAIADE II	$\{C_T, N_2\}$
	R_7	1	1	0	1	1	0	0	0	0	ARGUS, Lexicographic method	$\{A_C, L_M\}$
	R_{11}	1	1	0	1	2	0	0	0	0	ELECTRE I	$\{E_1\}$
	R_8	1	1	0	3	1	0	1	0	0	QUALIFLEX, REGIME	$\{Q_F, R_C\}$
	R_{12}	1	1	0	3	2	0	1	0	0	ELECTRE II	$\{E_2\}$
	R_{13}	1	2	0	3	2	0	1	0	0	IDRA, MAPPAC, PACMAN, PRAGMA	$\{I_D, M_P, P_C, P_C\}$
	R_{14}	1	2	0	3	2	0	2	0	0	EVAMIX, MAUT, MAVT, SAW, SMART, TOPSIS, UTA, VIKOR	$\{E_V, M_U, M_V, S_A, S_M, T_P, U_T, V_K\}$
	R_{16}	1	2	0	3	3	0	2	0	0	AHP + TOPSIS, AHP + VIKOR	$\{A_H + T_P, A_H + V_K\}$
	R_{15}	1	2	1	1	2	1	0	2	0	Maximin fuzzy method	$\{M_F\}$
	R_{17}	1	2	1	1	2	2	0	0	1	TACTIC	$\{T_C\}$
	R_{18}	1	2	1	1	2	2	0	0	3	ELECTRE IS	$\{E_S\}$
	R_{19}	1	2	1	2	2	2	0	0	3	ELECTRE TRI	$\{E_T\}$
	R_9	1	2	1	3	1	2	1	0	1	ORESTE	$\{O_R\}$
	R_{10}	1	2	1	3	1	2	1	0	3	MELCHIOR	$\{M_C\}$
	R_{16}	1	2	1	3	2	1	2	3	0	Fuzzy SAW, Fuzzy TOPSIS, Fuzzy VIKOR	$\{S_F, T_F, V_F\}$
	R_{20}	1	2	1	3	2	2	1	0	3	ELECTRE III, PROMETHEE I	$\{E_3, P_1\}$
	R_{21}	1	2	1	3	2	2	2	0	3	PROMETHEE II	$\{P_2\}$
	R_{22}	1	2	1	3	2	3	1	2	3	PAMSSEM I	$\{P_{A1}\}$
	R_{24}	1	2	1	3	2	3	1	3	3	Fuzzy PROMETHEE I	$\{P_{1F}\}$
	R_{23}	1	2	1	3	2	3	2	2	3	PAMSSEM II	$\{P_{A2}\}$
	R_{25}	1	2	1	3	2	3	2	3	3	Fuzzy PROMETHEE II	$\{P_{2F}\}$
	R_{27}	1	2	1	3	3	1	2	1	0	Fuzzy AHP + TOPSIS	$\{A_F + T_P\}$
	R_{28}	1	2	1	3	3	1	2	2	0	AHP + fuzzy TOPSIS	$\{A_H + T_F\}$
	R_{29}	1	2	1	3	3	1	2	3	0	Fuzzy AHP + fuzzy TOPSIS, Fuzzy ANP + fuzzy TOPSIS	$\{A_F + T_F, A_{NF} + T_F\}$
	R_{30}	1	3	0	3	3	0	2	0	0	AHP, ANP, MACBETH, DEMATEL, REMBRANDT	$\{A_H, A_N, M_B, D_M, R_M\}$
	R_{31}	1	3	1	3	3	1	2	3	0	Fuzzy AHP, Fuzzy ANP	$\{A_F, A_{NF}\}$

MIN_MAX[1] - methods of extracting the minimum and maximum values of the attribute.

Figure 3.2: The rules of selecting a suitable MCDA method [54]

hard to accurately express each attribute and their aggregation function as a utility function, due to the non-linearity associated with utility theory. Consequently, we exclude MAUT from consideration. SMART is aimed at qualitative attributes, but can also be used on converted quantitative attributes. The main advantage of this method is that it handles multi-level attributes, whereas other methods do not. However, our problem context does not require multi-level attributes. Thus, we also exclude SMART from our possible MCDA methods. UTA(UTilité Additive) is a method that assumes the axiomatic underlying of MAUT, as it assumes the existence of an additive utility function. It assesses a model based on a set of given utility functions, which we do not have. Determining utility functions for each criteria, is outside the scope of this research. Thus, we exclude UTA from consideration as well.

SAW, VIKOR and TOPSIS are applicable methods to our research problem. In Figure 3.1 SAW would classify as a weighting method and the latter two would classify as "other methods". VIKOR and TOPSIS can be classified as methods involving distance from an ideal alternative (distance methods) [25]. SAW (Simple additive weighting) means calculating a weighted sum of normalized values of all attributes to determine the ranking, which one of the most simple, yet intuitive ways to get a ranking. In [51], the authors compare various MCDA methods on a data set with a known ideal ranking. They conclude that the TOPSIS method achieves higher accuracy than SAW, when comparing the resulting rankings to the ideal ranking. For this reason, we opt to use TOPSIS over SAW. VIKOR is closely related to TOPSIS as they both treat preference as geometric distance from the ideal [38]. However, with VIKOR one has the possibility to use different weights on the distance to the ideal and the distance to the negative-ideal. We choose not to use the VIKOR method, since we believe determining weights on such an abstract concept will not be benificial to transparancy of the method. We need this transparancy to articulate the scheduling tool to the decision makers and lower the threshold to implementation. Furthermore, we believe that the resulting weights might not be well-substantiated, since the decision maker (Company X) does not have knowledge of the used methodology. The argumentation above leads us to TOPSIS as our MCDA method of choice.

The TOPSIS method consists of the following 6 steps:

1. Perform a vector normalization on the decision matrix (Equation 3.1)

$$n_{ij} = \frac{r_{ij}}{\sqrt{\sum_{i=1}^{I}(r_{ij})^2}} \tag{3.1}$$

Where:
n_{ij} denotes the normalized performance score of alternative i on attribute j;
r_{ij} denotes the performance of alternative i on attribute j before normalization.
We do this for each combination of i and j.

2. Form a weighted normalized decision matrix (Equation 3.2)

$$V = N\mathbf{w} \tag{3.2}$$

Where:
V is the weighted normalized decision matrix;
N is the normalized decision matrix from step 1;
$\mathbf{w}$ is the vector of attribute weights $\{w_1, w_2, ..., w_J\}^T$.

3. Determine the positive ideal solution (PIS) and negative ideal solution (NIS) (Equation 3.3 and 3.4)

$$PIS = \mathbf{v}^+ = \{v_1^+, v_2^+, ..., v_J^+\} \tag{3.3}$$

$$NIS = \mathbf{v}^- = \{v_1^-, v_2^-, ..., v_J^-\} \tag{3.4}$$

Where:
v_j^+ is the best value on attribute j over all alternatives i (i.e. lowest value for cost attributes and highest value for benefit attributes) and v_j^- is the worst value on attribute j over all alternatives i (vice versa).

4. Calculate the Euclidian distances from each alternative to the PIS and NIS in J-dimensional Euclidian space. We do this using a generalization of the Pythagorean theorem to higher dimensions (Equation 3.5 and 3.6).

$$d_i^+ = \sqrt{\sum_{j=1}^{J}(v_{ij} - v_j^+)^2} \tag{3.5}$$

$$d_i^- = \sqrt{\sum_{j=1}^{J}(v_{ij} - v_j^-)^2} \tag{3.6}$$

Where:
d_i^+ is the distance from alternative i to the PIS and d_i^- is the distance from alternative i to the NIS.

5. Calculate the closeness coefficient (CC) of each alternative i (Equation 3.7).

$$CC_i = \frac{d_i^-}{d_i^+ + d_i^-} \tag{3.7}$$

6. Rank all alternatives i by their corresponding closeness coefficient (CC_i).

3.2.4 Weighting methods in multiple-criteria decision analysis

TOPSIS does not have an intrinsically associated weighting method for determining criteria weights. We consult literature to find a weighting method that fits our preferences. Again, we refer to the sources we mention for elaboration on the methods stated in this section. Weighting methods can generally be divided into compositional and decompositional methods [20]. Compositional methods perform weighting separately from scoring, whereas decompositional methods do both in the same procedure. Our decision model will rank a number of available appointments based on their performance on some KPIs, which represent the attributes. Scoring alternatives on attributes is therefore not necessary. Consequently, we exclude decompositional methods from the possibilities. Marsh et al. [33] define a typology in which they describe a number of weighting methods. The types of compositional weighting methods they discuss are ranking, direct rating, pairwise comparison and swing weighting. Of these types, ranking (sorting on most important criteria, then on second most important, etc.) is overly simplistic for our purpose, which is why we do not use this weighting type. In direct rating, the decision makers assign a weight directly, usually by allocating 100 points over attributes or using a 0-100 scale for each attribute. In pairwise comparison, the attributes are compared to each other attribute and a measure of preference is assigned for each combination. Swing weighting assigns weights by first ranking the preference of a "swing" from the worst to the best observed value for each of n attributes. The decision maker then assigns a rating on each of $n + 1$ ranked alternatives (including worst-case, with lowest observed value on all attributes) from 0 to 100, where the highest ranked option gets a rating of 100 and the worst-case scenario gets a rating of 0.

Aspects we need to keep in mind when selecting the weighting procedure include transparency and the reliability of weights. Transparency is increasingly important when the decision maker has more knowledge of the trade-off that the attributes entail. Should the decision makers initially not have a clear idea on how the weights should be allocated, then a pairwise or swing weighting method might be a better solution, as it distills weights from simple comparisons. These less transparent methods will exert a lower cognitive load on the decision maker. Our decision makers (controller and practitioner at Company X) however, have substantial knowledge of the underlying trade-off of our attributes. Consequently, the results of a pairwise comparison or swing weighting procedure might not reflect the weighting that they already have in mind. Therefore, a direct rating procedure seems most qualified at translating the decision makers' preferences. Usually, one of two direct ratings methods is chosen: direct scaled rating and point allocation. Bottomley et al. [3] compare these two methods and conclude that direct scaled rating provides higher reliability of weights than point allocation. In conclusion, this means that we will use direct scaled weighting.

3.2.5 GUI design in decision support systems

Should the patient state preferences that do not seem represented by the available suggestions, then we encourage the scheduler to filter on location, practitioner and date accordingly. When the scheduler applies heavy filtering, suggestions might be far from optimal. Therefore, we want to give them an indication of the quality of the resulting suggestions. Clarkson et al. [6] use color coding in their decision support tool to visualize whether a solution is good or not. We will implement the same principle in our scheduling tool. The color coding should visualize the quality of the solutions on performance measures that are not apparent to the scheduler. In Chapter 4.1 we decide to use the

probability of an above average utilization as the response variable for our predictive model. This could be a good variable to base our color coding on, since it shows the quality of a suggestion. Also, the fragmentation delta might be a good performance measure to consider for our color coding. To the author's knowledge, providing additional information to the scheduler on the non-transparent effects of the scheduling decisions is not yet present in literature.

3.3 Validation methods

To make our research viable, we need to confirm that all our research methods perform as intended. For this purpose, we search literature for validation methods for our predictive model and for our scheduling tool. In Section 3.3.1 we discuss validation methods that we can apply to our predictive model. In Section 3.3.2 we discuss validation methods that we can apply to our scheduling tool.

3.3.1 Predictive model validation

Current approaches towards model validation and assessment of predictability include graphical evidence of a good match between the predicted and observed values, analysis of parameter uncertainty and sensitivity, and analysis of model associated uncertainty [23]. Graphical representations of model fitting often include a histogram of the residuals, a normal probability plot of residuals, a residuals versus fits plot and a residuals versus order plot, which we will all evaluate. Analysis of parameter uncertainty and sensitivity can be performed on the standard errors of the coefficients, T-value, P-value and variance inflation factor (VIF) for each predictor. This shows whether predictors have a significant effect on the response variable and if there is no severe multicollinearity (correlation between predictors). Most represented in literature is analysis of model associated uncertainty. The most frequent solution seems to be (k-fold) cross-validation [5] [40] [30] [24]. Cross-validation means splitting the training data in a training set and a test set for validation. In a typical k-fold cross-validation procedure for a linear model, the data set is randomly and evenly split into k parts (if possible). A candidate model is built based on $k-1$ parts of the data set, called a training set. Prediction accuracy of this candidate model is then evaluated on a test set containing the data in the hold-out part. By respectively using each of the k parts as the test set and repeating the model building and evaluation procedure, we choose the model with the smallest cross-validation score (typically, the mean squared prediction error MSPE) as the 'optimal' model. Given p independent variables, there are a total of $2p-1$ possible models. In the k-fold cross-validation procedure, each model is in fact evaluated k times. Therefore, a single 'optimal' model is selected via $k(2p-1)$ times of model evaluation. We consider all methods mentioned above during validation of our predictive model. Reason for this is the considerable representation of these methods in literature and applicability in our context.

3.3.2 Scheduling tool validation

As discussed before, the type of scheduling tool that we aim to develop is not frequently represented in literature (if at all). Therefore, it proves difficult to find validation techniques aimed at our context. Thus, to try to find validation methods we look at sources aimed at other research subjects as well. One of those other sources is decision support tool validation, which is related to our subject, since our tool is essentially a specialized decision support tool. Furthermore, we search literature about scheduling algorithms in general. However, we are not able to find applicable validation methods

specifically related to these areas. To validate our tool we will therefore use the more general method of face validation [29]. If the model's results are consistent with perceived system behavior, then the model is said to have face validity. For our research face validation mainly consists of evaluating performance on KPIs of the top-k suggestions compared to historic performance.

There is no standardized way of determining how well our tool performs, which is partially due to the fact that our solution is subjective (because of conflicting objectives). However, it is possible to see if our weights result in an efficient ranking. In this context, "efficient" refers to the ability to use the inputs and produce outputs in optimal proportions given those inputs [39]. In a good ranking, higher ranked options should theoretically be more efficient. Data Envelopment Analysis (DEA) is a method that can evaluate efficiency for Decision-Making Units (DMUs), which in our case are appointments options. DEA produces an econometric frontier, which in our case is a best-practice frontier (different name for different applications, e.g. in production it is a production frontier). An example of such a frontier
is depicted in Figure 3.3. In this Figure y_1 and y_2 represent outputs and x represents an input variable. The efficiency frontier consists of the set of outer points such that we get a concave (in case of maximization) line. Each point represents a DMU, which in our case are appointments options (possible decisions). Points on this efficiency frontier are efficient points. Any options set of inputs that does not result in a point on the line is considered inefficient. This means it does not optimally use its resources. DEA suggests a translation of inputs that would be necessary for the option to be considered efficient (e.g. R to R'). By doing this for each option we get an average translation of the inputs such that the efficiency is optimized. In our case this would show whether our set of weights results in efficient solutions. DEA is completely non-parametric, which means no assumptions are necessary and subjectivity is eliminated. In our case there are more inputs/outputs, which means we construct this frontier over more dimensions than the two-dimensional case in the figure.

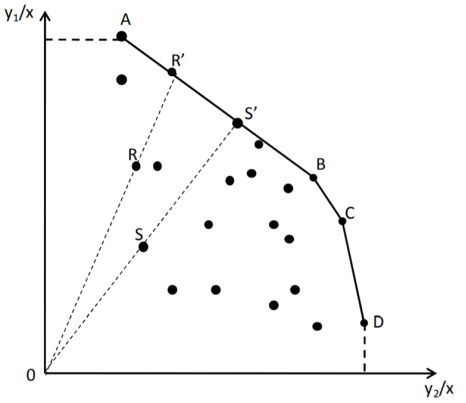

Figure 3.3: DEA frontier example

DEA can either be input-oriented or output-oriented. The input-oriented DEA minimizes inputs and the output-oriented DEA maximizes outputs. Since in our case lower values on the criteria are preferable, we use an input-oriented DEA model where our criteria are input variables. The way one can see this is that we "pay" with our criteria (e.g. some distance and some fragmentation) to get to our output of 1 appointment and we want to minimize the use of our criteria to get to that output of 1 appointment. Zhu and Cook [58] designate three models as the three basic DEA models: CCR (Cooper, Charnes and Rhodes), BCC (Banker, Charnes and Cooper) and additive models (by Charnes et al.). The most basic DEA method is CCR [7] (as developed by Charles, Cooper and Rhodes). This is the standard method when applicable. However, as we discuss in Section 4.2.2, we have a partially negative variable with the fragmentation delta. CCR can only handle non-negative values. Furthermore, it is not translation invariant, which means we cannot translate the variable to be non-negative without changing the optimal solution of the linear programming model that DEA solves for each DMU. BCC is partially translation invariant, as it is translation invariant for inputs when doing an output-oriented BCC and vice versa. Unfortunately, we need to translate an input of

an input-oriented DEA to be non-negative. Additive models are translation invariant for both input and output, since the efficiency evaluation does not depend on the origin of the coordinate system when this model is used. Therefore, we use the Slack Based Measure model, which is an additive model augmented with a slack based measure to evaluate the efficiency [7]. The efficiency of CCR and BCC is always as great as or greater than the efficiency of SBM, since SBM uses a different definition of inefficiency (mix inefficiency vs technical inefficiency for CCR).

3.4 Conclusion

In Section 3.1 we showed various predictive analytics methodologies. We concluded that we will consider a multiple linear regression model and a binary logistic regression model. In Section 3.2 we searched literature regarding several aspects of our scheduling tool. We first discussed the literature gap we observed in this research area in Section 3.2.1. Then, we searched literature on methods we can use for multi-attribute ranking in Section 3.2.2. We concluded that Multiple-Criteria Decision Analysis (MCDA) was a promising field of research for our problem, which we dedicate Section 3.2.3 to. Here, we substantiate our choice for a TOPSIS procedure. Corresponding to this MCDA method we searched literature for an attribute weighting procedure in Section 3.2.4. We opt for a direct scaled rating (performed by stakeholders) to produce criteria weights. Finally, we look at literature on GUI design of decision support systems in Section 3.2.5 and find a color scaling method that we can apply as well. In Section 3.3 we discuss validation methods for our predictive model and scheduling tool. We found a number of validation methods for predictive models, including k-fold cross-validation, graphical representation of fit between the predicted and observed values, and analysis of parameter uncertainty and sensitivity. Clear-cut validation methods applicable to our scheduling are not available in literature. To validate our model we will therefore resort to face validation, including evaluation of performance of the top-k suggestions compared to historic performance. Also, we find DEA to be a method that can show the quality of our solution method by evaluating the efficiency of the solution.

Chapter 4

Solution design

In this chapter we discuss our solution design, with the solutions being our predictive model and our scheduling tool. In Section 4.1 we discuss the choices we make for a predictive model, which we want to use to predict how busy locations will be. This is important to be able to spread appointments in an early stage. Also, we discuss how we constructed this predictive model, how it works, and what role it has in our scheduling tool, in detail. Moreover, we discuss its performance and validity. In Section 4.2 we discuss what choices we make for our scheduling tool and why. Furthermore, we discuss in detail how we construct the scheduling tool and how it should work. We conclude the chapter in Section 4.3.

4.1 Predictive model

We want to achieve a better spread of appointments over sessions. Doing so allows for a higher possible average utilization. The KPI we use for this is the variance of utilization. However, it remains a challenge to bring this variance down. When a future session currently already has a high utilization, it is often already to late to spread appointments evenly. This is because appointments that we can schedule further ahead are usually easier appointments to spread evenly. With NPs and other more urgent appointments we cannot really consider the spread of appointments. This is because we do not want to make concessions on these appointments, since they are imperative for patient satisfaction and retention. We therefore want appointments that are not that urgent to be spread over locations in such a way that we have room for NPs at most locations on the short-term.

To be able to this we want to predict how crowded locations are going to be on a given date. For this purpose we want to include a predictive model in the proposed scheduling tool. This predictive model should describe how the occupancy of sessions develop over time. In our Literature section (Section 3) we searched literature for various predictive analytics methods that we could apply to obtain such a model. From these options we deemed **Multiple linear regression** and **Logistic regression** as the most suitable methods.

Linear regression is a supervised **machine learning** method. In linear regression the dependent variable is modeled as a linear function of a set of regression parameters (predictors) and a random error [57]. Linear regression can be split into simple linear regression and multiple linear regression. The first uses one predictor variable and the second uses multiple. Logistic regression on the other hand takes the natural logarithm of the odds as a regression function of the predictors and translates the predicted variable back to a value between 0 and 1 (probability). In this section we first discuss the components that we want to use as model parameters. Additionally, we explain how we construct the models and discuss (and substantiate) the choices we make in detail. Finally, we discuss the performance and validate the chosen model. In Section 4.2 we discuss how we utilize this predictive model in the scheduling tool.

4.1.1 Model components

For our predictive model we first need to determine what variable we want to predict. We refer to this variable as the dependent variable. We did already decide that we want to predict future occupancy, but this can be expressed in many different ways and on many different levels (e.g. location, region, practitioner). After we have established exactly what we want to predict, we need to determine the variables that we use to predict this dependent variable.

Dependent variable

The dependent variable, which is the variable that we predict, needs to be related to the occupancy of a region, location, practitioner or single session. When deciding what we need to predict, we first look if we want to predict at region, location, practitioner or session level. Before making a choice lets repeat our research goal:

"How can we develop a scheduling support prototype that suggests the most preferred sessions and time slots to schedule the appointment in, to ensure that 80% of new patients get an appointment within 3 weeks?"

Our research goal states our focus on ensuring an appointment for NPs within 3 weeks. To achieve this, there needs to be space for an NP appointment at a location in the vicinity of the patient (vicinity is dependent on patient preference) within 21 days (access time violation threshold). We want to identify which locations are at risk of having a full schedule too soon. This way we can redirect appointments in time and prevent unavailability of schedule space to accommodate NPs. Occupancy per region would be too general for what we want to achieve, which is to know how crowded a schedule is going to be. Predicting the occupancy of a practitioner might not be accurate since the occupation in the morning and afternoon can be very different if the practitioner works two locations that day. We could opt for session occupancy, but we do not have data for this. Generating it would mean we would have to spend a substantial amount of time to come up with a data set and algorithm that can do this, which we cannot justify given the limited time that we have for this research. Therefore, *we decide that the dependent variable of our predictive model should be related to location occupancy(/utilization).*

We utilize our predictive model when scheduling an appointment in the scheduling tool. It should give an indication of what the occupancy of a location is for the day on which we want to schedule an appointment. For the dependent variable, we have the choice to predict the occupancy itself or some variable derived from it. We evaluate two different dependent variables. The first is the occupancy for a location on the appointment date. The second is the probability that the occupancy of a location (on the appointment date) is above average (92.2%). In Section 4.1.2 we compare a model with the first dependent variable and a model with the second dependent variable. The reason for evaluating the second dependent variable is because, as we show in Section 4.1.2, the second dependent variable results a model that more realistically represents its underlying dependent variable.

Predictors

Above we determined that we want to predict location occupancy on the appointment date, at the moment that we are scheduling an appointment. This is determined by a number of internal factors that arise from the scheduling process, like the region, month, and day of the week, which we discussed in 2.1. However, there are also a number of external factors that influence the utilization of a session.

In Appendix I we analyzed three possible external factors: Wheather, Company X's Google Adwords budget, and mailings. Only, for mailings we were able to conclude that they have a significant effect. For more details on these external factors and the corresponding analysis we refer to Appendix I. The combination of this external factor with the internal factors from the scheduling process, results in the following set of predictors that we choose to consider:

- Region of the location that we want to schedule an appointment at
- Year of appointment date
- Month of appointment date
- Day of week of appointment date
- If mailing (marketing activity) in corresponding region occurs in month leading up to appointment date, then number of days between mailing date and appointment date, otherwise 0
- Number of days between date of scheduling and date of appointment
- Current utilization of appointment date (as observed on scheduling date)
- Average current utilization during access time
- Working time in the last 4 weeks at the selected location

We expect these variables to be the main influences on the utilization.

Region
In Chapter 2 we established that the utilization differs based on the region. For this reason we include it in our model. We exclude the artificial regions, as they are not relevant for our scheduling tool. As of now, the region "Groningen" does not have much data to base a prediction on and may significantly differ from future expectations. This should remedy itself over time, since after implementation a regression coefficient updating procedure should be in place.

Time series variables
We include time series elements like seasonal indices and trend using the day of the week, month and year of the appointment date. These variables can significantly influence the utilization. In Section 2.1.2 we show the effect of these seasonal factors on the number of appointments.

Mailing in preceding 30 days
In Appendix I we showed that mailings have a significant effect on the number of incoming appointments. From unstructured interviews with the marketing department we know that the mailings target RPs and that conversions to appointments mainly result in new *-classified-* appointments. We see that the average access time for new *-classified-* appointments that come in on the day after a mailing (this is when the biggest peak occurs) is 20.5 days. Therefore, we use the occurrence of a mailing in the month leading up to the appointment as a variable as well. If a mailing (marketing activity) occurs in the corresponding region, in the month leading up to appointment date, then we use the number of days between the mailing date and appointment date, otherwise we use 0. We did not perform a sensitivity analysis concerning the period over which a mailing influences utilization, since it is out of the scope of this research. However, this could be beneficial to do in future research to improve the predictive model. The period of 30 days we use is selected intuitively, based on the information about new *-classified-* appointments' access times and our analysis of mailing effects. It

is notable that the average access time for new -*classified*- appointments normally (in 2016 to 2019) is 44.5 days, as Appendix F shows. The reason that the access time is this much lower on the day after a mailing, is because practitioners often schedule new -*classified*- appointments ahead at the end of an appointment. We see that they often do this half a year, a year or sometimes even two years ahead. When someone requests a new -*classified*- appointment after a mailing, this is not the case. Therefore, the access time distribution for this type of appointment is significantly different on the day after a mailing.

Days between scheduling date and appointment date
The days between the scheduling date and the appointment date should also influence the realized utilization. This variable depicts the amount of time left to fill the schedule. The more days there are left until the appointment, the more we expect the utilization to increase.

Current utilization on appointment date
The current utilization on the appointment date describes the current state of the schedule. When constructing our model we evaluate if the dependent variable is truly dependent on all predictors. Current utilization and the number of days between the scheduling date and the appointment date could form a powerful interaction in predicting the dependent variable. Therefore, we also include the variable "Current utilization * Number of days ahead" as an interaction variables. This variable is equal to the product of the two variables, which is the standard way of including interaction between two variables in regression [18].

Average current utilization during access time
Not only the utilization on the appointment date is important to indicate crowding. The utilization in the time between scheduling and the appointment is also an indicator for expected utilization. If the schedule for the location is full during access time, we expect a higher utilization than when the sessions leading up to the appointments session are empty. Just as with the current utilization for the appointment date, this variable might also have a relevant interaction with the number of days between the scheduling date and the appointment date. We therefore add an interaction variable for these two variables as well.

Working time in 4 weeks preceding appointment
To predict how full a schedule is going to be, it might be a good idea to know how much time we normally have to fill. We add "Working time in 4 weeks preceding appointment" (including the appointment date) as a variable, to show how much treatment time we have available at the location. Whether a certain utilization is good or bad might depend on this, so we add an interaction between these two variables as well. We use a multiple of 2, since we have recurring schedules spanning over 2 weeks (in the vast majority of the cases). We use 4 weeks rather than 2 to reduce the effect of the erratic occurence of extra workdays or less workdays.

4.1.2 Model construction

We determined the basic components for our regression model: our predictors and dependent variables. We need to translate these into a model. The way we construct our model is by transforming our available data into training data for our model. From the data we have (raw data and previously generated analytic data) we should generate data containing our model parameters. This data set will serve as training data to create our model. We designed an algorithm that evaluates the appointment data and extracts the corresponding data from data sets that we generated before,

about locations' utilization, fragmentation, practitioners and available time per day. Appendix L describes this algorithm and other data selection/analysis steps used for constructing the model.

The newly generated training data serves to train a regression model for both our dependent variables. We do this in a statistics package called Minitab. The regression "learns" the relation of each predictor to the dependent variable using the training data. For multiple linear regression this is a linear relation, but for logistic regression this is a linear relation with the log-odds of the dependent variable (further elaboration in Appendix M). Additionally, the regression procedure evaluates the significance of the relation of each predictor to the dependent variable and some metrics about the fit of the model. Relevant outcomes of fitting these regression models include regression coefficients for each predictor and P-values per predictor. In multiple linear regression the coefficients describe the linear effect of the predictor on the dependent variable. For logistic regression the coefficients describe the linear relationship between the predictor variables and the log-odds (also called logit) of the event that $P(Utilization > 0.922)$ is true. The P-values show whether the effect of the predictor on the dependent variable is significant. Appendix M shows the complete regression summaries with interpretation and a more detailed explanation of the methods.

We evaluate three different regression models. The first model (1) is a multiple linear regression model that predicts the location occupancy. The second model (2) is a binary logistic regression model that predicts the probability of a utilization that is above average, $P(utilization > 0.922)$. In Appendix M we explain why the average of 92.2% differs from the one of 91.6% mentioned before. The third model (3) is a multiple linear regression model that predicts the probability of a utilization that is above average, $P(utilization > 0.922)$. First, we evaluated all models using all predictors and only excluded predictors that were not statistically significant. Though, when we implemented the model in our scheduling tool, we noticed that including the "working time in the preceding 4 weeks" demanded too much computational power (50% increase in computation time), for only a very small improvement in model performance (0.02% increase in correct predictions, where correct means a predicted probability of more than 0.5 and with an above average final utilization and vice versa). Therefore, we exclude this predictor from the model.

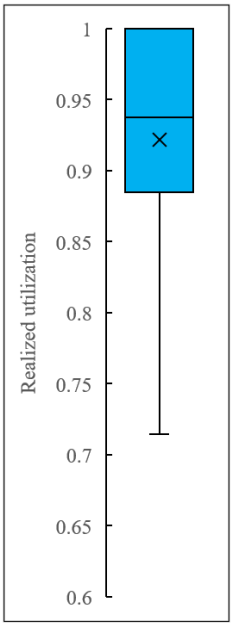

Figure 4.1: Boxplot of realized utilizations

Figures 4.2 and 4.3 show the average predicted values against the realized utilization. These figures have realized utilizations from 0.7 to 1, since (as Figure 4.1 shows) utilizations outside this range are outliers. We see that Model 1 (Figure 4.2) is very inaccurate when the utilization on the appointment date is much lower than the average. The reason for this is that the process that we are trying to predict is not organic. When the occupancy for a location is very low, the planning department often interferes and tries to move appointments to these sessions. Therefore, even when a location's schedule is very empty a few days in advance, it is usually relatively full when the session takes place. This causes our model to structurally predict a utilization close to the average.

Due to this unpredictable human interference it is not possible to predict historical utilizations accurately, which is why we cannot conclude that this model accurately predicts the utilizations of locations. The model should show whether a location will have a relatively high/low utilization. Thus, we need an indication of whether it thinks that the targeted location will be crowded or not.

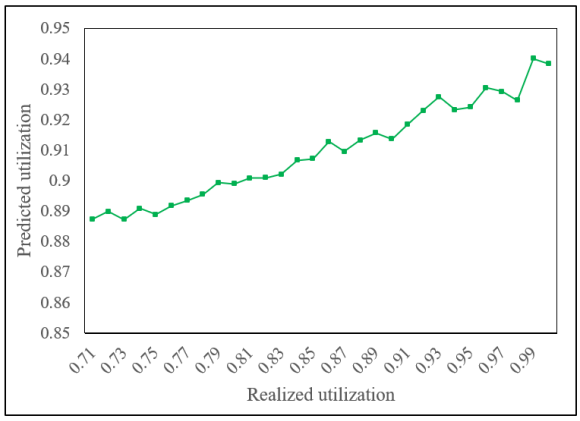

Figure 4.2: Predicted occupancy vs realized utilization

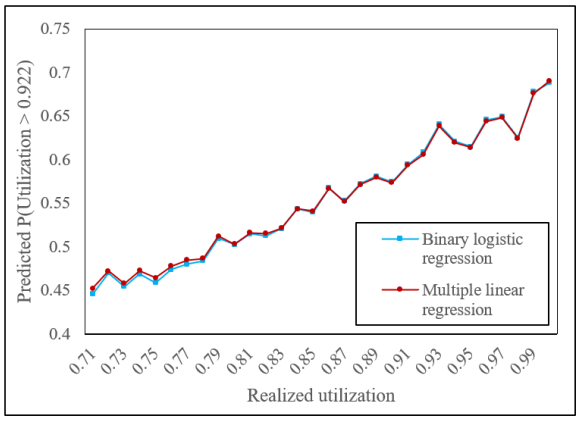

Figure 4.3: P(utilization > 0.922) vs realized utilization

The resulting value does give us that, but labeling it as the predicted future occupancy might not be the way to go. For this reason we construct models that predict the probability that the realized utilization will be above average. When the probability is high, it is not preferable to further increase the probability of crowding at the location with another appointment and when the probability is low we should direct appointments towards this location. We show the predicted probability $P(utilization > 0.922)$ for both Model 2 (Binary logistic regression) and Model 3 (Multiple linear regression) in Figure 4.3. We still see that with lower utilization the model expects higher utilizations, which is what usually happens due to the human interference. Though, this model gives a much clearer difference in predicted value over the same range of realized utilizations. This suggests this dependent variable results is a better indication of whether a location will be crowded or not. Furthermore, we see that Model 2 is slightly better than Model 3, since it has slightly lower predictions in the lower range and slightly higher predictions in the higher range.

4.1.3 Model analysis

The constructed predictive model serves as an indication of whether a location will be crowded on a given day. In this section we describe the model performance and whether or not it gives a valid indication of the crowding possibility. As discussed in Appendix M, there is a statistically significant lack-of-fit. That is, the model does not seem to accurately describe the actual probability of an above average utilization. The reason is that the process itself is non-organic and somewhat random. The human interference in the process (e.g. schedulers quickly fill low utilization sessions nearing due date) makes that the model usually expects utilizations to be near-average. Whenever low utilizations slip through it is unexpected, thus resulting in a lack-of-fit. Figure 4.3 shows this tendency towards the middle, by showing moderate probabilities even for more extreme realized utilizations. Figure 4.4 also shows this in its frequency of the predictions. Figure 4.5 shows the proportion of times that the prediction was correct at a given predicted $P(utility > 0.922)$, where correct means $(P(utilization > 0.922) > 0.5)$ and $(utilization > 0.922)$ or vice versa.

To show how the model reacts to changes in predictor values we use a 2-level full factorial experiment design for the 4 independent continuous predictor variables that we included. For all experiments we set the region, month and day equal to 1 (Appendix M shows what this means). We set the low and high levels for predictors to their first and third quartile values. These quartile values give

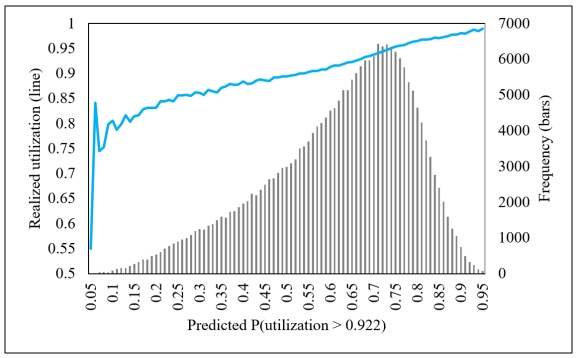

Figure 4.4: Realized utilization vs average predicted P(utilization $>$ 0.922) and frequency of predicted P(utilization $>$ 0.922) (bins of 0.01)

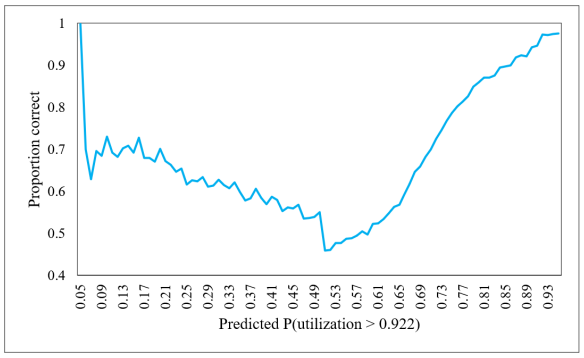

Figure 4.5: Proportion correct vs predicted P(utilization $>$ 0.922)

more realistic appointment options than for instance taking the maximum and minimum of each factor. Only for mailing we use a max-min leveling (0 to 30), since the first and third quartile are both 0 for this predictor. Table 4.1 shows the results of swinging from first to third quartile for each independent continuous predictor (complete results in Appendix N). The effects of the predictors seem as expected, which advocates the use of this variable as an indicator of the probability that a location will be busy or not.

Term	1^{st} **quartile**	3^{d} **quartile**	**Effect of swing (1^{st} quartile to 3^{d} quartile) on prediction**
Mailing[a]	0	30	0.0168
N.o. days ahead	7	21	0.1878
Current utilization	0.4643	0.8333	0.2402
Access time utilization	0.6360	0.8488	0.0445

[a]Min-Max values instead of quartile values for mailing, since quartile values are both 0

Table 4.1: 2^k-factorial experiment results

The experiments in Table 4.1 show the effect of swings in predictor variables and the results do not show any apparent flaws. Also, in Figures 4.3, 4.4 and 4.5 we show that on average our predictive model is a good indicator of future utilization at a location. Still, this does not say much about the reliability of individual predictions. To test this, we grab a small random batch of appointments and perform an elementary analysis to see if the predicted values are logical relative to the other appointments. Table 4.2 shows 10 random historical appointments from 2017 to 2020 and the corresponding predicted P(utilization $>$ 0.922). Relatively high values are red and relatively low values are green. Overall, we see one peculiarity. Row 1 and Row 5 have similar prediction values, though Row 1 seems like it should be higher than Row 5, since it has a higher current utilization and it is the same number of days ahead. The utilization during the access time is slightly lower for Row 1, but this does not account for such a discrepancy. The first reason for the similar probability is that Row 5 is an appointment from the region "Gelderland" (third highest regression coefficient), whereas Row 1 is from the region "Drenthe" (second lowest regression coefficient). The second reason is that Row 5 is from July (second highest regression coefficient), whereas Row 1 is from February (third lowest regression coefficient). All other rows seem logically balanced, which advocates the

Row	Marketing	N.o. days ahead	Current utilization of session	Access time utilization of location	Predicted P(utilization >0.922)
1	0	6	78.1%	78.1%	55.0%
2	0	27	31.3%	31.3%	57.2%
3	0	9	93.9%	89.5%	81.9%
4	0	10	81.3%	81.3%	63.5%
5	0	6	56.3%	84.9%	56.9%
6	21	12	82.4%	86.1%	73.3%
7	0	14	28.1%	65.8%	35.0%
8	5	27	73.5%	51.5%	74.7%
9	0	1	95.8%	95.8%	68.8%
10	0	16	48.4%	81.7%	55.7%

Table 4.2: 10 random historical appointments with the corresponding prediction for $P(utilization > 0.922)$

reliability of the model as an indicator of crowding. From this experiment and our findings in the 2-factor full factorial experiment we can conclude that the predictive model is a sufficient indicator of $P(utilization > 0.922)$.

4.1.4 Updating procedure

This predictive model is based on data of the last 4 years. However, it is unknown how representative the model will be over the years, as model coefficients might need to change. Therefore, we use an updating procedure to adjust our variables to observed changes in the process. The updating procedure we use is simply retraining the model with a moving time-interval. Other options include training partial regression models over new data and, for instance using a gradient descent approach. However, since our model only takes a few seconds to train, there is no need for more intricate procedures. The retraining frequency is a trade-off between working time for exporting the data set to fit the model to and keeping the model up-to-date, which Company X should assess for themselves. Though, we do recommend to update the model at least once a year, since a lot can change in a year and over a year we know that changes are not due to difference in seasonal factors per region.

4.2 Scheduling tool

In this section we describe a scheduling tool with the functionalities that should be included after implementation at Company X. Our main contribution does not lie in the front-end of the tool, as the front-end is subject to strict requirements still being worked out by the quality management department. Our main contribution does lie in the back-end of our tool, which utilizes the crowding prediction model that we proposed in Section 4.1. The scheduling tool we propose, runs a two-stage decision support procedure. The first stage consists of a top-k session selection procedure, whereas the second stage is a complete ranking procedure of all available appointment options in the session selection from the first stage.

Before constructing the tool, we make decisions about what the tool should look like and what functionalities it should have. First, we discuss the front-end of our tool in Section 4.2.1. We discuss what input we need from the scheduler and how that reflects in our GUI. In Section 4.2.2 we discuss the back-end of our tool. Here we introduce what multi-criteria decision making procedures we use

and how we apply them. Some of the decisions that we mention in this section, we already discussed in the literature review. We repeat them here to get a clear picture of the most relevant decision making regarding the proposed scheduling tool.

4.2.1 Front-end of scheduling tool

The tool design is inspired by the current scheduling tool to reduce resistance to implementation. Therefore, the proposed tool uses similar input data to that of the current scheduling tool. Before the scheduler uses the current scheduling tool, he/she selects the patient for whom he/she is scheduling an appointment, or creates a new patient profile in case of a new patient. The current scheduling tool uses this patient data of the selected patient as part of the input data and so does the proposed tool. Since our scheduling tool cannot be linked directly to the database yet, we include a field where the scheduler states the patient's postal code. We need this postal code to retrieve the patient's geographical coordinates, which in turn we need to calculate the distance from the patient to a given session. After implementation this postal code should be pre-filled, but in this demonstrational tool (pre-implementation) the scheduler should provide a postal code to represent the patient's location.

Apart from selecting the patient info, the scheduler is tasked with indicating the activity type, the appointment duration and the required access time. Of the required input data for the proposed tool, only the required access time is something that the current tool does not include in the same way (currently the scheduler can only include a specific date range for this purpose). From unstructured interviews with a practitioner, who serves as advisor to the management team, we know that schedulers are capable of making these assessments. In the proposed tool, the scheduler also has the option to state a specific date range, practitioner and/or location. Furthermore, the scheduler can indicate whether to include specific schedule types (e.g. diabetes, sport, or other location types). This is all that the scheduler has to/can do in the front-end. This set of data is what we refer to as the input data in the remainder of this section. Appendix O shows what the GUI for our preliminary tool looks like. All input fields either are dropdowns or have input validation, which signals the scheduler when input is faulty. This accommodates easy and properly-formatted input.

4.2.2 Back-end of scheduling tool

TOPSIS

In Section 3.2 we searched literature for multi-criteria decision making methods and substantiate our choice for a method called "Technique for Order of Preference by Similarity to Ideal Solution" (TOPSIS), with a direct scaled weighting procedure. We discuss our weighting procedure at the end of this section. For an elaborate description of the TOPSIS method we refer to Chapter 3. The TOPSIS method results in a ranking of appointment options. To get this ranking, the criteria values for all options are first normalized using vector normalization, i.e. for each criterion the vector of all values is transformed to a vector with magnitude equal to 1 (unit vector), keeping the same direction. Then, each value is multiplied by its criterion weight. Over these weighted and normalized values a Positive (PIS) and Negative Ideal Solution (NIS) are determined. The PIS is the set of the best observed values out of all options for each criterion, whereas the NIS is the set of the worst observed values. Then, the euclidian distance from each option to the PIS and NIS is calculated. As ranking variable, TOPSIS uses the Closeness Coefficient (CC), which considers both the distance to the PIS

and NIS. The Closeness Coefficient for alternative m is defined as follows:

$$CC_m = \frac{d_m^-}{d_m^+ + d_m^-} \tag{4.1}$$

Where:

d_m^+ = distance from alternative m to the PIS;
d_m^- = distance from alternative m to the NIS.

Two-stage decision support procedure

By calculating the Closeness Coefficient for each option we can rank all appointment options on this value. However, the number of appointment options is vast, as sessions are defined for many years to come. This solution space is incomputable. Therefore, we first need to reduce the set of sessions that we use to generate possible appointment options over. For this purpose, we choose to develop a two-stage decision support procedure. In Figure 4.6 we schematically show what these two stages look like. The first stage consists of a top-k session selection procedure, whereas the second stage is a complete ranking procedure of all remaining appointment options in this session selection. For the remainder of this thesis we will refer to these stages as Stage 1 and Stage 2, respectively. In both stages we use a TOPSIS methodology for ranking of options, though with different criteria and weight sets. In Stage 1 we only select the top-k options from our TOPSIS procedure to pass to Stage 2. Stage 2 results a complete ranking of appointments options.

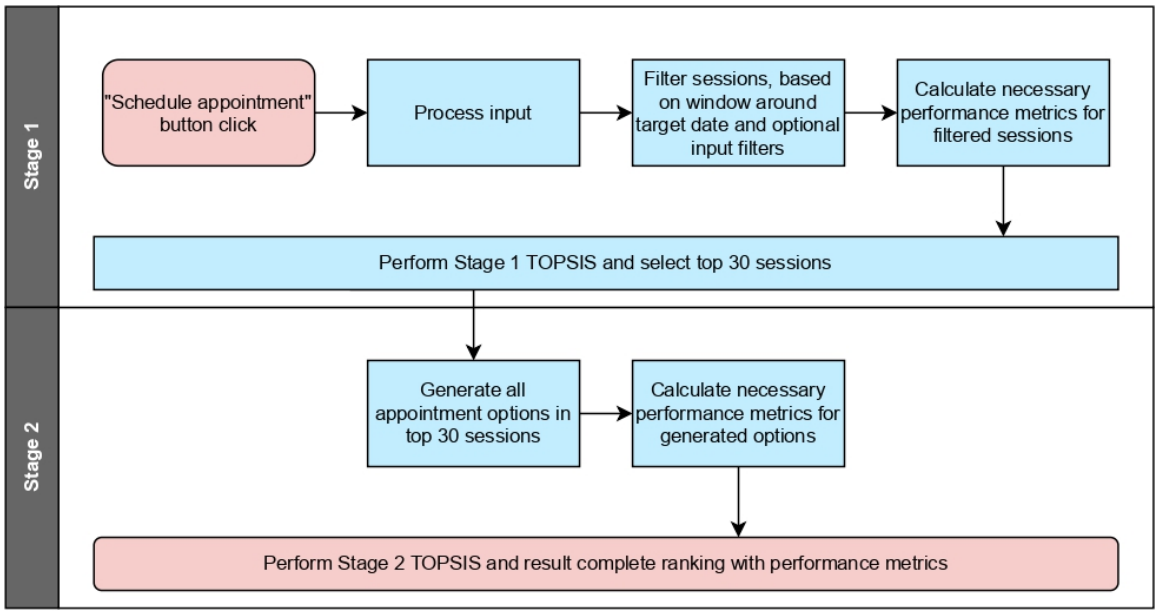

Figure 4.6: Block diagram of 2-stage decision making procedure

Which set of criteria c_{hij} we use for the TOPSIS procedure, depends on the parameter h, denoting whether the required access time is greater than 30 days or not, and the stage i. The value that an option has on a given c_{hij} is before normalization and weighting. What values it can take is therefore completely dependent on the underlying, e.g. distance can theoretically be between 0 km and approximately 20 038 km (half the circumference of the earth). We formally define h, i and j as follows:

$$h = \begin{cases} 1, & \text{if } \textit{required access time} \leq 30 \textit{ days} \\ 2, & \text{otherwise} \end{cases}$$

$$i = \begin{cases} 1, & \text{if } \textit{tool in Stage } 1 \\ 2, & \text{if } \textit{tool in Stage } 2 \end{cases}$$

$j \in \{1, 2, ..., J\}, \;\; \textit{with } J = \textit{number of criteria used for the given combination of } h \textit{ and } i.$

We only use our predictive model when the required access time is less than or equal to 30 days (i.e. $h = 1$). This is because the predictive model is constructed for predictions up to 30 days ahead. For appointments further into the future we use the current utilization of the schedule as a criterion. The other factor, on which the criteria we use depends, is the stage $i \in \{1, 2\}$ that the tool is in. Table 4.3 shows what criteria we use for each h and i. The Stage 2 TOPSIS uses more criteria, because it also includes criteria calculated on appointment level. Stage 1 can only take into account criteria that can be calculated on session level. Consequently, the two stages also use different weight sets. As Table 4.3 shows, h and i result in separate 4 criteria sets and thus 4 different weight sets. Appendix P provides a more complete explanation of the two stages and their criteria.

Required access time $\leq$ 30 days ($h = 1$)		Required access time $>$ 30 days ($h = 2$)	
Stage 1			
j	Criteria c_{11j}	j	Criteria c_{21j}
1.	Distance	1.	Distance
2.	Predicted $P(utilization > 0.922)$	2.	Utilization
3.	Days deviation from target	3.	Days deviation from target
4.	Appointment same type as session	4.	Appointment same type as session
Stage 2			
j	Criteria c_{12j}	j	Criteria c_{22j}
1.	Distance	1.	Distance
2.	Predicted $P(utilization > 0.922)$	2.	Utilization
3.	Days deviation from target	3.	Days deviation from target
4.	Fragmentation index delta	4.	Fragmentation index delta
5.	Appointment same type as session	5.	Appointment same type as session
6.	Access time violation days [a]		

[a] Only non-zero when the patient to schedule is NP and access time $>$ 21 days

Table 4.3: Decision criteria c_{hij} classified by required access time h and stage i, with corresponding weight sets

Weighting procedure

To weight these criteria c_{hij} we use a direct scaled weighting procedure. This means we present the criteria to stakeholders, which rate the importance of each criterion from 0 to 100. However, our weighting procedure slightly defers from a standard direct weighting procedure. We give the stakeholders an initial set of weights, intuitively selected by the researcher as a viable set, with its given results for a given appointment configuration. Then, we let the stakeholders adjust the weighting and show them the results. We reiterate until the stakeholders are satisfied. This method is more

transparent and should provide better perception of the effects of the weighting for the stakeholders than a normal direct rating. We have to perform this procedure separately for each combination of h and i, since both values of h and both stages i feature different criteria. This results in 4 distinct weight sets, with weights w_{hij} corresponding to each c_{hij}. We determine these weights in Section 5.1. The stakeholders involved in this procedure are the supervisor at Company X (controller) and a practitioner that is currently active as main body of the quality management department. The resulting weight sets are balanced weight sets. The proposed tool can easily be altered to adopt multiple weight sets for different preferences. The scheduler could then select which suits the needs of the patient best. Examples would be a weight set to prioritize close locations and a weight set to prioritize appointments with low access times. Furthermore, Company X might want to use different weight sets for different patient types. It would however take time to define and test more weight sets, which we cannot do within the scope of this research. We recommend Company X to further research this opportunity.

4.3 Conclusion

The predictive model that we decide to use in our proposed scheduling tool is a binary logstic regression model in which we predict the probability of above average utilization on a location, denoted $P(utilization > 0.922)$. In Appendix M we substantiate why we take 92.2% as the average. The predictors that we use in this predictive model are:

- Region of the location that we want to schedule an appointment at
- Year of appointment date
- Month of appointment date
- Day of week of appointment date
- If mailing (marketing activity) in corresponding region occurs in month leading up to appointment date, then number of days between mailing date and appointment date, otherwise 0
- Number of days between date of scheduling and date of appointment
- Current utilization of appointment date (as observed on scheduling date)
- Working time in the last 4 weeks at the selected location

Due to human interference the underlying utilizations of the predictive model evolve inorganically, since low utilizations are often quickly filled in the days leading up to the session. Therefore, the model has a tendency towards the average utilizations. However, we do establish that the predictive model is still a good indicator of whether or not a sessions is at risk of crowding or of being underutilized.

The front-end of our demonstrational tool is similar to the one that Company X currently uses. The only additional input we need from the scheduler is the required access time, which a practitioner, active as advisor to the management team, stated the schedulers would be able to assess. The fact that both inputs of the current tool and the proposed tool are similar is beneficial to preventing resistance to implementation. However, our main contribution is the back-end of the proposed solution. We developed a two-stage decision support procedure, which results in a ranking of appointment options within reasonable time, considering that this is an online scheduling problem. The first stage of the

procedure consists of a top-k session selection procedure, which limits the solution space. The second stage is a complete ranking procedure of all available appointment options in the session selection from the first stage. For the remainder of this thesis we will refer to these stages as Stage 1 and Stage 2, respectively. For both stages we use a TOPSIS methodology for ranking of options, though with different criteria and weight sets. To weight criteria for the TOPSIS method, we use an iterative direct scaled weighting. In this weighting we present an initial viable balanced weight set and let stakeholders adjust weights after which we show the new results. We iteratively do this until the stakeholders are satisfied.

Chapter 5

Experimental design and results

This chapter discusses the performance of the proposed scheduling tool compared to the current performance. We determine the configuration of the proposed tool that we use for evaluation in Section 5.1. To evaluate the performance of the proposed tool, we use simulation. We recreate an instance from the past and let the proposed tool schedule a given set of appointments. This way, we can compare performance of the proposed tool to historical data. We discuss this simulation in Section 5.2. In Section 5.3 we introduce a real-life instance (recreation of schedule at a point in the past) simulation in which the proposed tool schedules the appointments that originally came in, in the month following the instance. To evaluate the performance in more irregular contexts, we also perform the same simulation for some theoretical instances, in Section 5.4. This assesses the robustness of the tool. Finally, we discuss the use of Data Envelopment Analyses (DEA) in Section 5.5, to assess the efficiency of our solutions. In Section 5.6 we conclude this chapter.

5.1 Tool configuration

Before we evaluate the proposed tool we have to determine what parameter configuration we use. We selected this configuration in consultation with Company X. We also discuss how we configure the simulation, which we use for evaluation of the proposed scheduling tool.

The first parameter we determined is the number of sessions k in our Stage 1 selection, to generate appointment options over in Stage 2. We set $k = 30$ sessions. The reason for this number is that in areas with a lot of sessions, this allows the tool to select all sessions we feel should be considered. If k would be any lower, some good sessions might be overlooked. If k would be higher, our computation time would get too high.

We also selected the weight sets for each set of criteria c_{hij}. Table 5.1 shows the weights w_{hij}, corresponding to each criterion c_{hij}, that resulted from the iterative direct scaled weighting procedure with the stakeholders, which we discussed in Section 4.2.2. The Stage 1 weight sets remained unchanged from the initially proposed sets (intuitively selected by the researcher), as the stakeholders agreed that this set already resulted an adequate selection of sessions. The Stage 2 weight sets required more precise calibration (i.e. shifting of weights w_{hij} from some criteria c_{hij} to others), since they should result in a precise ranking of options, as opposed to the rough selection of Stage 1. In Section 5.5 we therefore evaluate the robustness of the Stage 2 weight sets.

5.2 Simulation

The simulation that we use to evaluate the tool uses a given instance of the schedule (all sessions with corresponding schedule and some additional metrics) and a set of appointments to schedule. The simulation iterates over each appointment and uses the background procedure of the proposed

Required access time $\leq$ 30 days ($h = 1$)			Required access time $>$ 30 days ($h = 2$)		
Stage 1					
j	Criteria c_{11j}	w_{11j}	j	Criteria c_{21j}	w_{21j}
1.	Distance	0.83	1.	Distance	0.83
2.	Predicted $P(utilization > 0.922)$	0.10	2.	Utilization	0.10
3.	Days deviation from target	0.05	3.	Days deviation from target	0.05
4.	Appointment same type as session	0.02	4.	Appointment same type as session	0.02
Stage 2					
j	Criteria c_{12j}	w_{12j}	j	Criteria c_{22j}	w_{22j}
1.	Distance	0.30	1.	Distance	0.33
2.	Predicted $P(utilization > 0.922)$	0.33	2.	Utilization	0.37
3.	Days deviation from target	0.08	3.	Days deviation from target	0.10
4.	Fragmentation index delta	0.09	4.	Fragmentation index delta	0.10
5.	Appointment same type as session	0.10	5.	Appointment same type as session	0.10
6.	Access time violation days [a]	0.10			

[a]Only non-zero when the patient to schedule is NP and access time $>$ 21 days

Table 5.1: Decision criteria c_{hij} classified by required access time h and stage i, with corresponding weights w_{hij}

tool to schedule them one-by-one. After scheduling one of the appointments, the simulation adds the appointment to the instance before scheduling the next appointment. This results in a realistic development of the schedule. For a more detailed explanation of the simulation we refer to Appendix Q.

5.2.1 Simulation configuration

The proposed scheduling tool cannot be linked to the database until implementation. Also, we want to be able to compare performance to historical data. For testing and development we therefore use the instance of 01-07-2019 (at the start of the day, before any new appointments come in). We chose this month because July was close to average in terms of number of appointments in 2019, which makes it good month to use for evaluation (2020 was not representative due to the Covid-19 pandemic). We recreate the schedule as it was at the beginning of this day. This serves as the base point of the simulations that we run. We also use this instance for our DEA in Section 5.5.

We include a simple element of patient choice in the simulation by randomly picking one of the top 5 appointments, unless there are less than 5 options. Then we pick the top ranked one. We intuitively chose the number 5, since in the implemented version Company X does not want to initially show more than 5 appointment options to the scheduler (they are currently considering a front-end which only shows 3). We denote what patient choice (1 to 5) we get for each appointment and from the results this allows us to see what the average effect is when a patient chooses different options out of the top 5. Also, excluding patient choice would mean that we overestimate the abilities of the proposed tool, since in reality the patient will not always pick the top-ranked appointment. Only after implementation we can know how the choice of options by patients in the proposed scheduling tool is distributed, so we use this simplification to at least simulate patient choice to some extent. It is harder to simulate filter choices insisted by the patient (e.g. if the patient is adamant on a specific location), so we do not include it in our simulation. In Appendix U we do simulate the current scheduling tool and show the discrepancy between the options that this tool shows and the original performance.

Additionally we argue that the discrepancy between the proposed tool simulation and the actual expected results is much smaller, since no myopic filtering is needed to get valid appointment options on screen. We validate the use of this simple representation of patient choice in Appendix Q.

5.2.2 Simulation simplifications and assumptions

The simulation of the proposed tool has the following simplifications and assumptions:

Simplification 1: Schedule appointments at same session type.
We force the appointment to be scheduled in the same session type as the type they were originally scheduled in. The reason is that Company X has not sufficiently documented which appointments should go where. There are activity types destined for certain schedule types, but they are often overlooked. Therefore, a lot of appointments that were scheduled in regular sessions could also have been scheduled in non-regular schedules. The other way around, some appointments scheduled in non-regular sessions could also have been done in regular sessions. However, we cannot label which appointments we can schedule at both types and which have to stay in the same type.

Simplification 2: Leave out appointments originally scheduled in non-existent sessions.
The instance that the tool uses does not include appointments that were previously scheduled somewhere while there was no session on at that location on that day. The assessment of current performance also did not consider these appointments. These appointments are often artificial.

Simplification 3: Schedule all non-patient related activities first.
To make sure the non-patient related activities can be scheduled in the same place, we first schedule all non-patient related activities. In reality non-patient related activities are scheduled somewhere where there are no appointments and also very often outside session times. They often take up large blocks of time, sometimes larger than the session itself. Allowing to schedule these activities anywhere would give lots of problems for these reasons and thus we choose to schedule them first in their original place.

Simplification 4: Schedule appointments of certain session types at the same time and session.
The proposed tool will not influence schedules of the type "House visits", "External" and "Other". Therefore, we schedule appointments that were originally scheduled in these sessions, at the same time and sessions as before.

Simplification 5: We use a simple representation of patient choice in the simulation.
If we do not include patient choice in our simulation and simply pick the top-ranked option, we overestimate the performance of the proposed tool. However, we do not have any information on patient choice at Company X, let alone for the yet to be implemented tool. To still include some form of patient choice, the simulation randomly picks one of the top 5 appointment options. We intuitively chose the number 5, since in the implemented version Company X does not want to initially show more than 5 appointment options to the scheduler (they are currently considering a front-end which only shows 3).

Simplification 6: We do not simulate the patient choice in additional filters.
The patient to be scheduled sometimes insists on going to a specific location/practitioner or having an appointment at a specific time. We have no information on how often this happens and which filters would be insisted upon. Thus, we cannot simulate it. We do elaborate on their effect on the

eventual performance in Appendix U and Q.

Assumption 1: Difference in effect of patient choice in reality and simulation is not large enough to invalidate the simulation.
We assume that our simplification of patient choice will not make the results of our simulation invalid. To assure this we run our simulation while mimicking the current scheduling procedure. This means selecting sessions based on a distance filter, which is set to 10 km by default, and ranking appointment options solely on access time. We show the results from this simulation in Appendix U. In Appendix U we also discuss the discrepancy between the simulation of the current scheduling tool and the original performance, and how this discrepancy translates to the proposed scheduling tool.

Assumption 2: Appointments must be scheduled at least 30 minutes into the future.
We assume that appointments can only be scheduled in the future. Additionally, we only generate appointment options that are at least half an hour later than the time of scheduling.

Assumption 3: The desired access time in weeks for RP/PP appointments is equal to number of days originally scheduled ahead divided by 7, rounded to the nearest integer.
For NP appointments, we know that the desired access time is 0 weeks (as soon as possible). For RP/PP appointments this varies per appointment, based on the treatment and diagnosis. Since the desired access time is not documented, we do not know where they should have been scheduled. We therefore assume that the desired access time in weeks is the number of days that the appointment was originally scheduled ahead, divide it by 7, and round it to the nearest integer. Thus, the proposed scheduling tool will try to find a desirable appointment option close to when the RP/PP was originally scheduled.

5.3 Real-life instance simulation

The first evaluative method we use is a real-life instance simulation. For this simulation we recreate the schedule as it was at some point in the past. Then, we let the proposed tool schedule the appointments that were originally created (not necessarily performed) in the following month and compare the performance on our KPIs for this month to the real-life performance.

5.3.1 Real-life instance experiment design

This simulation shows how well the proposed tool would perform when facing the same situation as was the case in reality. For comparison, we also run the appointments of July 2019 at exactly the same time and place as they were originally scheduled. This way, we obtain the original instance at the end of the month. We calculate all performance measures in the same way as we did in our data analysis in Section 2 and make the same exclusions (e.g. outliers and sessions without appointments). As mentioned before, we simulate patient choice by randomly picking one of the the top-5 ranked options. In Appendix R we show the effect of our simulated patient choice on the KPIs in this real-life instance simulation.

5.3.2 Real-life instance experiment results

Tables 5.2 and 5.3 show the results of the real-life instance simulation. Table 5.2 shows the results of the performance measures from Company X's perspective. Here we show the difference between

Context	Utilization	Fragmentation	Variance of utilization
Original	91.6%	1.26	1.20%
Tool simulation	90.9%	1.15	1.03%

Table 5.2: Results of real-life instance simulation, average performance from Company X's perspective in July 2019

Context	NP access time (days)	NP access time violations	RP/PP days deviation from target [a]	Distance (km)
Original	20.7	35.7%	-	3.61
Tool simulation	14.8	15.9%	3.36	3.60

[a]Target based on original instance (see assumption 3 in Section 5.2.2)

Table 5.3: Results of real-life instance simulation, average performance from patient perspective in July 2019

the instance at the end of July 2019 from the original historical data and the new instance created by the proposed tool. Keep in mind that these values are still influenced by appointments in July 2019 that were scheduled before the simulated month starts. Only 32% of appointments in this month is scheduled by the simulated tool. The others were already in the instance, since they were scheduled in July before the simulated month commenced. We therefore expect observed changes in fragmentation and variance of utilization to be larger if the proposed scheduling tool reaches an equilibrium (i.e. all appointments in the evaluated period were scheduled by the proposed tool). The utilization will remain at about 91.6%, since we still have the same ratio of sessions and appointments. The fragmentation and variance of utilization are both lower in the simulation than in the original instance, which shows that the proposed tool performs as desired. The fragmentation decreased from 1.26 to 1.15 (margin of error of 0.056) and the variance of the utilization decreased from 1.20% to 1.03% (margin of error of 0.07%). Note that the average utilization is lower for the simulated instance. This is because the proposed tool spreads appointments, rather than myopically scheduling as soon as possible. Therefore, the simulation scheduled some appointments that were originally scheduled in July, in August. Should the tool be employed for a longer period, the utilizations of the original and the simulation would quickly converge. It should be noted that we calculated these KPIs per sessions rather than per workday. Therefore the fragmentation index is lower and the variance of utilization is higher than those in Section 2.3. The fragmentation is lower per session, since there are on average more gaps in a workday than in a session, because a workday can consist of two sessions. The variance per session is higher per sessions, because a workday can consist of two sessions which has a mediating effect on the variance.

Table 5.3 shows the results of performance measures from the patient's perspective. The outcome of the simulation of the proposed scheduling tool shows better performance than the historical performance in July 2019 on each of the KPIs. The reason why the proposed tool can perform better on each KPI, is twofold. First, schedulers currently let the patient suggest where to go and when to schedule. This often leads to overlooking the best options for the patient, since the patient does not have complete knowledge of the situation, and selecting an appointment option that would be dominated by the unconsidered options (inefficient solutions). This is why we can perform better on the patient KPIs. Second, the current scheduling tool does not consider Company X's interests, so the fact that the proposed scheduling tool does is why we can also perform better on KPI's from Company X's perspective. The main KPI, NP access time violations, decreases from 35.7% to 15.9%

(margin of error 1.9%). The average absolute number of days deviation from the target for RPs/PPs cannot be calculated for the original instance, since it is not known what the target was. In the simulated instance we set the target (in weeks ahead) for RPs/PPs (target for NPs is 0 days) to the number of days originally scheduled ahead, divided by 7, rounded to the nearest integer (see assumption 3 in 5.2.2). Therefore, we can calculate this KPI for the simulated instance, but not for the original instance, and thus cannot compare them on this KPI. The distance also decreases a little, although we chose a low weight for this criterion. We chose a low weight, because in the current situation, the performance on distance was already good. The distance was already low, because patients often choose one of the closest locations. The reason why the simulated performance on distance is still better than the historical performance, is that the proposed scheduling tool takes distance into account, whereas the current tool does not. The current scheduling tool only sets a filter to 10 km (along with a date filter) and does not even show what the distance is between the patient and the location. Also, the fact that the patient often picks a dominated appointment (worse on every KPI than another option) due to the current way of scheduling, makes the overall quality of appointments worse.

5.4 Theoretical instance simulation

The real-life instance simulation shows how the proposed tool performs in an average real-life situation. However, it does not show how it performs when the context is out of the ordinary. Using theoretical instances can show how robust the proposed tool is. The theoretical instances we evaluate are an instance where abnormally many appointments come in and an instance where abnormally few appointment come in, considering the available capacity. This shows us whether the proposed tool is still reliable in more extreme contexts. The reason why we evaluate these two theoretical instances is because this realistically could happen. An example of another theoretical instances that could happen, and is interesting to simulate, is an instance with abnormally high number of appointments from a certain region. Other instances where we alter other parameters could also be interesting. However, we choose to only do the instances with overcapacity and undercapacity, due to time restrictions (a simulation run of an average month takes between 8 and 10 hours).

5.4.1 Theoretical instance experiment design

We want to show how the proposed tool reacts to a situation when very few (Instance 1)/many (Instance 2) appointments come in. In Instance 1 we run the real-life instance we used before, in which we randomly deleted 50% of the appointments in the appointment set. In Instance 2 we add a random sample of 50% the appointments in August 2019 to the appointment set of the month before. We exclude non-patient related appointments and appointments from session types "External", "Home visit" and "Other". We schedule these types of appointments at the same time and place (substantiated in Appendix Q) as where they were originally scheduled, which might collide with these types of appointments from the month before. We bring the appointments, included in the sample, forward by a month to be included in our instance. Instance 1 represents a low (appointment) demand context and Instance 2 represents a high demand context. In reality, the number of appointments that we have in our theoretical instances do occur, but then it is compensated by a higher or lower number of sessions. In these theoretical instances we use the same number of sessions as July 2019 with a 50% increase/decrease of appointments, so we simulate a month with undercapacity/overcapacity.

5.4.2 Theoretical instance experiment results

Tables 5.4 and 5.5 show the results of the theoretical instance simulations. In the crowded instance the access time violations and distance are higher than in the original instance, but this is to be expected. From a patient perspective the low demand case is better for each KPI, which makes sense since there is more room for them to come sooner at closest locations etc.

Demand	Nr. of appointments	Utilization	Fragmentation index	Variance of utilization	Predicted $P(utilization > 0.922)$ [a]
Low	6793	78.1%	1.994	2.23%	43.3%
Regular [b]	13486	91.5%	1.261	1.20%	58.1%
High	18226	94.8%	0.769	0.83%	66.3%

[a]This is not one of the main KPIs, as discussed in Section 2.3.1, but nevertheless an interesting KPI to assess the robustness of the predictive model
[b]Original instance

Table 5.4: Results of theoretical instance simulations, average performance from Company X's perspective

Demand	Nr. of appointments	NP access time (days)	Access time violations	RP/PP days deviation from target [a]	Distance (km)
Low	6793	11.36	4.2%	3.090	3.805
Regular [b]	13486	20.69	35.7%	-	3.614
High	18226	19.46	44.9%	4.048	5.184

[a]Target based on original instance (see assumption 3 in Section 5.2.2)
[b]Original instance

Table 5.5: Results of theoretical instance simulations, average performance from patient perspective

The utilization varies as expected, low for the low birth rate (of appointments) case and high for the high birth rate case. In the high birth rate (crowded) case we have 20 minutes of idle time per sessions on average. The reason why this is not lower is threefold. First, the start of the month is not affected by the crowding. Second, sessions of type "external", "home visits" and "other", keep exactly the same schedules (Appendix Q explains why). Third, there are not a lot of 15-minute appointments to fill gaps. Therefore, the fragmentation index in the crowded instance is still closer to 1 than it is to 0. However, we do see that in the crowded case the proposed scheduling tool manages to get the fragmentation index down by 39% compared to the original instance. We also manage to spread appointments better than in the original instance. It is notable that the proposed tool gets better at spreading appointments and avoiding fragmentation with increasing number of appointments. This is partially caused by the predictive model, which is trained by the original situation where they align capacity with demand, and thus, where such under-utilization does not happen often (if ever). In the under-utilized instance, the predictive model might expect a certain session to be filled in the access time for an appointment, which could very well not be the case. This also shows in the fact that the average predicted $P(Utilization > 0.922)$ is still at 43.3%, while the utilization is much lower than 92.2%, at 78.1%. However, all values do seem rather stable. We do not see any values that vary disproportionally. In the crowded instance the access time violations and distance are higher than in the original instance, which is logical when the schedules are more filled. We can conclude that the proposed scheduling tool is robust. On the other hand, the predictive model is not as robust, and therefore has room for improvement.

5.5 Data Envelopment Analysis

To see if our weights result in an efficient solution (i.e. the resulting appointment options) we use Data Envelopment Analysis (DEA). In this context, "efficient" refers to the ability to use the inputs and produce outputs in optimal proportions given those inputs [39]. This section discusses how we use DEA as evaluative method. In Section 3.3.2 we discuss DEA in more detail.

5.5.1 Data Envelopment Analysis experiment design

Stage 1 of our scheduling tool is concerned with filtering sessions. As long as the best sessions are included in this stage, the efficiency of the options is not a prominent issue, which is why we focus the DEA on Stage 2. We run two DEAs on the results of Stage 2 of our two-stage TOPSIS procedure. The first DEA (DEA 1) is on the resulting appointment options for a random sample of 5 NP appointments. We limit the sample size to 5 because processing DEA results from existing DEA solvers takes a lot of time. If we wanted a bigger sample size we should have coded our own DEA solver to perform multiple DEAs consecutively and store the results. However, due to time restrictions we cannot justify this. DEA shows the efficiency of each Decision Making Unit (DMU). In our case a DMU is an appointment option. To see how robust our solutions are we compare the ranking resulting from our 2-stage TOPSIS procedure to the corresponding efficiencies in the DEA. If the efficiency of the highest ranked options is low, the solution is not robust. This would mean that we could have gotten higher output with the given input. To also evaluate the robustness when the required access time is over 4 weeks (no predictive model and NP access time violations), we do the same for 5 random appointments that have a required access time of 7 weeks (DEA 2). The average non-NP access time is 49.5 days, hence the 7 weeks. Apart from seeing how robust the proposed tool is with the different criteria, another purpose of this second DEA is to see how robust the proposed tool is when the schedules are still relatively empty. Apart from resulting DMU efficiencies, DEA also shows optimal weight distribution per DMU to make the DMU as efficient as possible. By taking the average of these weights we get a weight set that results in optimal average efficiency.

5.5.2 Data Envelopment Analysis experiment results

Figures 5.1 and 5.2 show the efficiency for the top 50 ranked appointments for DEA 1 and DEA 2, respectively. The higher the required access time, the bigger the window in which we can schedule, thus the more appointment options there are. Therefore, we evaluate the first 50 options for DEA 1 (number of options between 62 and 286) and the first 150 options for DEA 2 (number of options between 482 and 705). The figures show that the tool (in these cases) always results an efficient option first and then, as we move lower in rank, results options with a lower efficiency. This advocates the robustness of the ranking obtained through our two-stage TOPSIS procedure.

Table 5.6 shows the weight sets to obtain optimal efficiency for a subjective decision making process, as obtained from the DEA procedure. We do not make changes to the weight set we currently use, since this weight set should articulate Company X's preferences rather than having maximum efficiency. For instance, the days deviation from target is not that important to Company X, since we already filtered the options to be in a certain window and they are rather indifferent to which day we schedule in that window (though closer to the target is more preferable). Yet, the weight set proposed for maximum efficiency has a high weight for this criterion. However, the resulting weight set is a good place to start when generalizing the proposed tool to a context outside Company X,

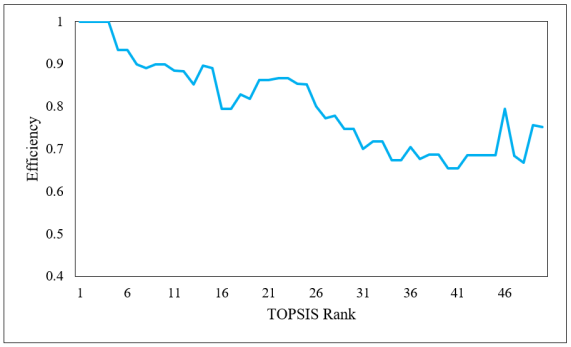

Figure 5.1: Average efficiency of top 50 appointment options for 5 random NP appointments, according to DEA

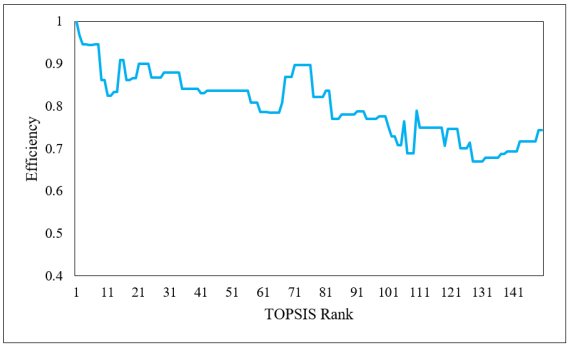

Figure 5.2: Average efficiency of top 150 appointment options for 5 random appointments with required access time 7 weeks, according to DEA

as it is the most efficient weight set from an unbiased perspective. When generalizing the proposed tool to a context other than the Company X organisation, we advise decision makers to start with the weight sets obtained from the DEA. Then, they should incrementally apply shifts in weights, where they prioritize criteria which they feel are most important in their context. It is advisable to tune weights with a toy instance (maybe only 1 appointment), before going to large scale instance testing. We run our real-life instance simulation again with the weight set obtained from the DEA. The results of this simulation are in Appendix T.

Required access time $\leq$ 30 days ($h = 1$)			Required access time $>$ 30 days ($h = 2$)		
Stage 2					
j	Criteria c_{12j}	w_{12j}	j	Criteria c_{22j}	w_{22j}
1.	Distance	0.307	1.	Distance	0.270
2.	Predicted $P(utilization > 0.922)$	0.242	2.	Utilization	0.089
3.	Days deviation from target	0.192	3.	Days deviation from target	0.498
4.	Fragmentation index delta	0.091	4.	Fragmentation index delta	0.138
5.	Appointment same type as session	0.047	5.	Appointment same type as session	0.005
6.	Access time violation days [a]	0.122			

[a]Only non-zero when the patient to schedule is NP and access time $>$ 21 days

Table 5.6: Weights w_{hij} on the criteria c_{hij} for optimal efficiency in Stage 2, as obtained from DEA

5.6 Conclusion

In this chapter we discussed the results of our proposed scheduling tool. To compare the performance of the proposed tool to the current performance, we use three evaluation methods: real-life instance simulation, theoretical instance simulation and Data Envelopment Analysis (DEA). In the real-life instance simulation we rescheduled a month of appointments from historical data (Section 5.3), including an element of patient choice. The results show that we improve on all criteria. The real-life instance simulation shows a reduction of the main KPI, access time violations, from 35.7% (July 2019) to 15.9%. The average fragmentation index for the corresponding month went from 1.26 to 1.15 and the variance in utilization went from 1.20% to 1.03%.

The theoretical instance simulation in Section 5.4 shows how robust the proposed scheduling tool is in overcrowded/underutilized contexts. In the overcrowded instance we schedule 50% more (regular)

appointments and in the underutilized instance we schedule 50% less appointments than in the real-life instance simulation. The performance varies as one would expect. We see that the proposed tool gets stronger at spreading workload and avoiding fragmentation, with a greater number of appointments. We can conclude that the proposed tool is robust. The predictive model, on the other hand is less robust as it is trained for the current ratio of capacity and demand.

Finally, we evaluate the robustness of the resulting appointment suggestions. More specifically, we determine the efficiency of the resulting appointments option ranking in a number of DEA procedures over the rankings for 5 random NP appointments and the ranking of 5 random appointments with access time of 7 weeks. The DEA method we used is an input-oriented Slack-Based Measure (SBM) method. This showed that the highest ranked options show perfect efficiency, which decreases as we move lower in rank. This advocates the robustness of our method. Furthermore, from these DEAs we got suggested weight sets for optimal efficiency. Since we tailored our weights to Company X's preferences, we do not change our weights. However, when generalizing the proposed tool to contexts outside Company X, these weight sets are a good place to start.

Chapter 6

Discussion

In this chapter we discuss the results of our research. We first describe what conclusions we infer from this research, in Section 6.1. After this, we discuss the limitations that we encountered, which should be taken into account, in Section 6.2. Then, we discuss the main assumptions necessary throughout the research, in Section 6.3. Finally, we discuss our recommendations and what future research is advisable in Section 6.4.

6.1 Conclusions

In Section 1 of this research, we state the **research goal**:

"How can we develop a scheduling support prototype that suggests the most preferred sessions and time slots to schedule the appointment in, to ensure that 80% of new patients get an appointment within 3 weeks?".

To achieve this goal we broke the problem into smaller pieces by introducing some research questions. In this section we concisely summarize our research by answering these questions one by one. We should mention that for readers interested in the results, Research Questions 4 is most interesting as it summarizes the findings from the experiments and states the benefits of the newly proposed scheduling tool over the current one. It also includes the conclusion of whether we have reached the research goal and the assessment of the overall quality of the solution.

Research question 1:

What does the current scheduling process look like and how does it perform on the relevant KPIs?

The current scheduling process suffers from a myopic scheduling tool and overall procedure. Currently, Company X pre-processes the sessions and generates all available gaps in the schedule. When they schedule an appointment in the gap, they adjust this in the system. By default, the current scheduling tool retrieves these gaps and filters them, such that we only have gaps from sessions in the next 6 weeks, within 10 km from the patient. These are then sorted by lowest access time, which is the time between scheduling and the appointment option, after which they are presented to the scheduler. Additionally, there are a lot of optional filters that the scheduler can use (e.g. specific location, practitioner, time window, etc.). Since the current scheduling tool does not take other KPIs than access time into account, the top options are often not good options (e.g. large distance). Therefore, the scheduler usually asks the patient where they want to go and then tells the patient what the first option at that location is. However, the patient does not have complete information. There could have been closer locations, locations with a shorter access time, or both. Apart from the fact that the current scheduling procedure often does not get the most beneficial appointment for the patient, Company X's interests are completely ignored.

KPIs	Performance
Company X perspective	
Utilization	91.7%
Variance of utilization per practitioner workday	0.917%
Fragmentation index	1.40
Patient perspective	
NP access time (days)	20.65
NP access time violations	37.1%
Absolute deviation of RP/PP access times from the target date (days)[a]	-
Distance from patient to the appointment location (km)	3.45

[a]This KPI cannot be calculated over historical data since there is no target date documented

Table 6.1: Key Performance Indicators and corresponding performance in 2019

The KPIs that we decide to use, and the corresponding performance over 2019 (2020 data is corrupted due to the covid-19 pandemic), are in Table 6.1. From Company X's perspective, we see an average utilization of 91.7%, which means we have 32.0 minutes of idle time on average for each workday. This closely relates to the fragmentation index of 1.40, which means there are on average 1.40 unused gaps in the schedule per workday. If Company X would solely focus on their own interest's, the utilization would be close to 100% and the fragmentation would be close to 0. Furthermore, the standard deviation of the utilization is 9.58% (Variance of 0.917%). This shows that there is a significant fluctuation in utilization and thus an uneven spread of workload over sessions.

From a patient perspective, we see that access time for NPs is 20.65 days on average, which is already very close to the allowed threshold of 21 days, as stated in our research goal. A reason for this is that, as mentioned above, schedulers often ask a patient which location they want to go to. The patient usually does not know all locations and what the implication of this location is on their access time. If an NP chooses a crowded location, he/she often gets an appointment at this crowded location with an access time far over the 21 day threshold. Another reason for high NP access times is schedule fragmentation, which results in a lot of gaps too small for NP appointments. This fragmentation is induced by the schedulers letting patients pick an appointment time in an open interval (a gap in the schedule) instead of suggesting a time that does not further fragment the schedule. The average distance from the patient to the appointment location is 3.45 km. For reference, on average the closest location to a patient is 2.29 km away. However, a small number of appointments with high distances are skewing the average. We see that the median distance from the patient to the appointment location (1.41 km) is close to the median distance from the patient to the closest location (1.06 km). This is caused by schedulers asking the patient which location they would like to go to, which usually is the one they feel is closest. They then often proceed to schedule the appointment at that location without considering others.

In conclusion, the KPIs show that blindly letting patients do the decision-making is detrimental to KPIs from Company X's perspective. Furthermore, when it comes to access time, this way of scheduling is actually detrimental to the patients themselves as well. A well-informed scheduler can actually often pick more preferable slots for the patient than what the patient could suggest without complete information.

Research question 2:

How can we make a predictive model of the occupancy of a location and validate it?

To give an indication of the expected occupancy of a given session, we decide to use a binary logistic regression model. This model predicts the probability that a session's utilization will be above average (i.e. utilization > 0.922), also referred to as, $P(utilization > 0.922)$. Appendix M explains why we use 92.2% instead of the 91.7% stated above at Research Question 1. The predictor variables it uses are:

1. Region of the location that we want to schedule an appointment at
2. Month of the appointment date
3. Day of the week of the appointment date
4. If a mailing (marketing activity) in the corresponding region occurs in the month leading up to the appointment date, then the number of days between the mailing date and the appointment date, otherwise 0
5. Number of days between the date of scheduling and the appointment date (today = 0; tomorrow = 1)
6. Current utilization on the appointment date (as observed at moment of scheduling)
7. Average current utilization of sessions at the same location during access time (as observed at moment of scheduling)
8. (5) multiplied by (6)
9. (5) multiplied by (7)

Due to human interference the underlying utilizations of the predictive model evolve inorganically, since sessions with low utilizations are often quickly filled in the days leading up to the session. Therefore, the model has a tendency towards the average utilizations, which results in a lack-of-fit. However, we do establish that the predictive model is still a good indicator of whether or not a session is at risk of crowding, or of being underutilized. A 2-level full factorial experiment design shows that the effects of swinging predictors from their first-quartile value to their third-quartile value, are in line with what we see in reality. This advocates the use of this variable as an indicator of the probability of above-average utilization. A random sample of appointments also showed that the predicted $P(utilization > 0.922)$ varies as desired for each appointment.

Research question 3:

How can we create and validate a scheduling support tool that can adequately inform schedulers on which time slots and sessions to choose, utilizing our predictive model (RQ2)?

Our main contribution is the back-end of the proposed scheduling tool. We developed a two-stage decision support procedure, which results in a ranking of appointment options within reasonable time. This is necessary considering that this is an online scheduling problem. The first stage of the procedure consists of a top-k session selection procedure, which limits the solution space. The second stage is a complete ranking procedure of all available appointment options in the session selection from the first stage. For both stages we use a TOPSIS methodology for ranking of options, though with different criteria and weight sets. One of the criteria that we use is the value for $P(utilization > 0.922)$ that

we obtain from our predictive model. The criteria sets differ per stage and also depend on whether we schedule more than 30 days ahead or not. The predictive model for $P(utilization > 0.922)$ is trained for 30 days into the future. If we schedule further into the future, we therefore use the current utilization. Table 6.2 shows the criteria we use for the TOPSIS procedure per stage.

Required access time $\leq$ 30 days ($h = 1$)		Required access time $>$ 30 days ($h = 2$)	
Stage 1			
j	Criteria c_{11j}	j	Criteria c_{21j}
1.	Distance	1.	Distance
2.	Predicted $P(utilization > 0.922)$	2.	Utilization
3.	Days deviation from target	3.	Days deviation from target
4.	Appointment same type as session	4.	Appointment same type as session
Stage 2			
j	Criteria c_{12j}	j	Criteria c_{22j}
1.	Distance	1.	Distance
2.	Predicted $P(utilization > 0.922)$	2.	Utilization
3.	Days deviation from target	3.	Days deviation from target
4.	Fragmentation index delta	4.	Fragmentation index delta
5.	Appointment same type as session	5.	Appointment same type as session
6.	Access time violation days [a]		

[a] Only non-zero when the patient to schedule is NP and access time $>$ 21 days

Table 6.2: Decision criteria c_{hij} classified by required access time h and stage i, with corresponding weight sets

The procedure can use multiple weight sets for different situations. For this research we use 1 balanced weight set per scenario (e.g. Stage 1, less than 30 days ahead). To validate the scheduling tool, we use simulation and DEA. We simulate a real-life instance and some theoretical instances. In these instances we recreate the schedule at some point in the past and schedule a given set of appointments using the proposed scheduling tool. In the real-life instance we schedule all appointments from the month following the original instance, based on the date they came in (not when they were performed). We compare the performance in this simulated month to the performance that we get when we schedule the appointments in the exact same time and place as they were originally scheduled. Also, we compare performance measures per appointment (e.g. distance and access time). In the theoretical instance we do the same, but we use two different sets of appointments: one high-demand set and one low-demand set. We use DEA to make sure our solutions are robust. This means that our highest-ranked options should be efficient and our lowest-ranked options should not be efficient. In this context, "efficient" refers to the ability to use the inputs and produce outputs in optimal proportions given those inputs [39].

Research question 4:

What are the resulting benefits of the scheduling support tool over the current situation?

The proposed scheduling tool actively spreads appointments over locations and prevents fragmentation of the schedule. A good spread of appointments means that there are more locations with room in the schedule to accommodate NP appointments within 3 weeks. Less fragmentation means less small, hard-to-fill gaps occur in the schedule and that we get more gaps large enough for an NP appointment within the 3-week threshold. Therefore, NP appointments can often be scheduled sooner. From simulating the use of the proposed scheduling tool in a real-life instance of a month, we see

improved performance from the patient's perspective as well as from Company X's perspective. For instance, using the balanced weight set that we propose, the standard deviation of session utilizations decreased from 10.9% (in the original instance of this month) to 10.1% and the fragmentation index decreased from 1.26 gaps to 1.15. This is only after one month that is still heavily influenced by the appointments that were scheduled in this period before the start of the simulation. We also see that for the appointments that we scheduled in the simulation, the percentage of access time violations is 15.6%, whereas in the original instance it was 35.7%. This is well below the 20% that we stated in our goal.

Also, the proposed scheduling procedure allows a change of the scheduling procedure for schedulers. The current scheduling tool shows gaps within 10 km sorted by increasing access time. Since only access time is taken into account, the top options (i.e. the options visible to the scheduler) are often not good options on other important performance measures like distance. Since schedulers know that the the options they see are not necessarily good options, and because it is hard to see which options are good options, schedulers usually ask patients where they would like to go, to filter the options. This myopically restricts the solution space and has a high probability of pruning the most beneficial solutions. The patient does not have complete information and thus often blindly picks the location he/she thinks is closest, or the one that they were at previously. This could be a very busy location and might actually not even be the closest location. The quality of appointments is therefore often far from optimal. After asking to pick a location, the scheduler also often lets the patient suggest a time instead of suggesting a preferable time himself/herself. This results in the aforementioned small, hard-to-fill gaps in the schedule on many occasions, which is often lost time for Company X. It also thwarts the scheduling of NP appointments (longer than RP/PP appointments), by not having many 60-minute gaps.

The proposed scheduling tool immediately shows good solutions at the top that balance distance and access time, which means the scheduler can simply suggest the top option (and then another upon rejection). This prevents that schedulers need to ask the patient where to go to filter options. Of course additional filtering remains an option, but this should only be used when a patient insists on a certain location, practitioner, etc. Eventually, this leads to appointments where the most beneficial option from the patient's, but also Company X's perspective are chosen. Therefore, the quality of appointments in the proposed scheduling procedure is better than with the current scheduling procedure.

Additionally, the proposed scheduling tool already boasts functionalities that are useful for Company X in the future. These functionalities include prioritizing sessions of a certain type (for which they first need to standardize these types) and using standardized ranges to search appointments in (desired access time), which is useful for standardization of activities. Company X can use these functionalities in any way (and to any extent, using the weighting) they desire. Also, weights for the TOPSIS decision-making procedures can easily be altered and extra weight sets for different scenarios can be added. Company X can therefore align the proposed scheduling tool to the strategic preferences whenever/however they want.

The decreased fragmentation and better spread of appointments means that the average utilization can grow, thus better utilizing resources and reducing costs for additional sessions. It also means that there are more short-term options for NPs at more locations. Furthermore, the proposed scheduling tool actively tries to optimize the quality of the appointment for the patient, increasing patient

satisfaction. This means they are less likely to cancel the appointment, less likely to not show up, and more likely to return for another appointment, generating additional revenue. The extent of these additional revenues and reduced costs will become apparent after implementation. They cannot be quantified beforehand, since they depend on factors that are not currently measured, like patient satisfaction, patient choice and scheduler behaviour.

Research question 5:

How can our scheduling support tool be integrated in the existing IT systems at Company X?

To make sure the transition to the proposed scheduling tool happens smoothly, we created an implementation plan in Appendix S. This implementation plan is aimed at the implementation of the proposed tool in the information system at Company X, by the IT department of Company X and might not be as useful in other contexts. During this research we did not only consider the technical implications and changes, but also a possible resistance to implementation. Therefore, the proposed scheduling tool runs on similar input data and data structures. This means the IT department does not have to make a lot of changes to the database for this procedure to work. To reduce resistance to implementation we also took into account the performance (time from button click to appointment) and the ability to use the procedure with a front-end of Company X's choice. The proposed scheduling tool uses the same input that the current front-end uses with only one additional input variable, which is the desired access time. However, using this variable takes away the necessity of entering a start and end date as range for appointments, which means that in most cases there is one less input box to be filled. As long as the front-end that Company X chooses (which Company X has not decided on yet) provides the back-end with the same input data and the desired access time, then the implementation plan provides enough information for implementation.

6.2 Limitations

The limitations we encountered were mainly data-related. In the data that was available we encountered a lot of misclassification and misdocumentation. Examples are a lot of appointment types that are used interchangeably and a lot of wrong appointment types for certain schedule types. This ultimately is the result of a deeper issue, which is lack of structured and concise classification. A lack of standardization in processes within Company X contributes to this issue. This does not only apply to the scheduling process, but also other processes.

Another limitation of the data was that Company X currently does not document whether a patient instigates a reschedule or if Company X does. This means we cannot recalculate the access times after rescheduling. This can result in an overestimation of access times when the patient wants to move the appointment to a later time. However, it can also result in underestimation if a scheduler brings the appointment forward following schedule optimization (offline scheduling activities to fill gaps). We assume the effects of these over/underestimations to be equal before and after implementation (see Assumptions).

The final mentionable limitation of the data is that it does not show if an access time violation is due to inability to schedule within the desired access time or if it is the patient's choice. It might be that Company X is able to schedule the patient within 3 weeks, but that the patient prefers to have an appointment on a later date (e.g. if the patient is on vacation). This might mean that we overestimate the number of access time violations.

6.3 Assumptions

In this section we discuss the relevant assumptions that we made throughout the research. We list and explain these assumptions below.

1. *The pre-implementation access time under/overestimation is equal to the post-implementation access time under/overestimation.* Therefore, we assume the difference in access time before and after implementation to be valid. The under/overestimation is mainly caused by frequent access time reduction/elongation due to schedule optimization/reschedules. The access time can not be recalculated in these cases due to a lack of reschedule trigger information. Another cause of access time under/overestimation is misclassification and misdocumentation of appointments, the effect of which we also assume to be equal before and after implementation.

2. *Each practitioner should be allocated an equal workload.* In reality some practitioners might work slightly slower, e.g. due to less experience. However, this is how Company X currently does it as well, except for certain professions. Company X usually schedules appointments for these professions specifically for the employee using a practitioner-specific filter, which is also possible in the proposed tool.

3. *An access time violation is always due to the inability to schedule within the desired access time.* In reality it is sometimes the patient's choice to wait longer than the predefined desired access time for that activity. This contributes to overestimation of the access time, but as mentioned before, we asssume that pre-implementation access time overestimation is equal to the post-implementation access time overestimation. Therefore we assume the difference in access time performance before and after implementation to be valid.

4. *The weights translating the decision maker's desired balance of KPIs is known.* For this purpose we let stakeholders weight the tool in consultation with the researcher. When generalizing the tool to other contexts, the decision makers should give significant attention to weight selection. One could use the optimal efficiency weight set of the DEA as a starting set and incrementally shift the weights to their preference.

5. *Configuring the proposed tool using the instance of 01-07-2019, does not influence the validity of the proposed tool in the present.* We constructed the proposed tool to use a list of sessions that is available in the database, which is not different now from how it was then (apart from having different sessions in it of course). Criteria values are also not distributed differently, so we can assume the weights are still balanced. For instance, if the average distance would have been twice as large and the other criteria would have remained equal, we would have wanted to change the configuration by shifting the weights. However, in this case, there are no substantial changes in criteria values that disrupt the balance of our weights. On the other hand, the number of sessions does increase each year (1065 sessions in July 2019 vs 1266 in July 2020). Therefore, it might be good to check if $k = 30$ in Stage 1 still does not exclude sessions that should be included in Company X's opinion. However, we intuitively chose k large enough to make sure we include all sessions that we deemed worthy of evaluation in areas with the most sessions. Thus, leaving $k = 30$, will likely still not often exclude sessions of which the appointments options would have been ranked high after Stage 2.

6.4 Recommendations and future research

In this section we discuss the recommendations that we have for Company X. Also we discuss the future research that we recommend for further improvement of the stated methods. We first list the recommendations and below that we list the recommended future research.

Recommendations

1. *Implement the proposed scheduling tool.* This thesis substantiates the benefits of using the newly proposed scheduling procedure. The research shows promising results and the proposed scheduling tool allows for a less myopic, less restrictive, and better-informed scheduling procedure. We therefore recommend Company X to implement the proposed scheduling procedure in the near future. In doing so they have the liberty to configure the scheduling tool to their liking, though we advise not to deviate too far from the proposed parameter settings, as the evaluation of the tool shows these settings to be effective. However, should Company X decide to use Stage 1 to select gaps in the schedule instead of sessions (which they are considering since they have a pre-processed table for this in the database), it is wise to increase the number of options k in Stage 1.

2. *Standardize more in the data structure.* Currently a lot of information is reported in the form of a story (log), but for data analysis this is not useful. Also reduce the appointment types by only using a minimal number of types and maybe splitting out information like patient type (DM, NP, etc.). Similarly, we recommend to introduce schedule type classification. The classification set we propose in Section 2.1.1 is our suggestion, but Company X might want to specify more.

3. *Start storing data on whether the patient rescheduled or if Company X did.* This is important for analytic purposes. Currently it is not possible to accurately determine the access time after a reschedule, since we do not know if the patient wanted the appointment to be later or if Company X needed to reschedule the appointment. For the same purpose it is a good idea to also store from when to when the appointment was moved. This would be to prevent access time underestimation (when moving the appointment to an earlier date during schedule optimization) as well as overestimation. Some schedulers note this already, but it is not standardized and often not done by a lot of other schedulers.

4. *Make an ERD of database that shows relations between data tables.* Something that impeded on our data analysis processes, was the fact that a lot of information was hard to find. Most of the information was available in the database, but it was unclear where. The current entity-relationship diagram (ERD) is only a grid of all tables with all their columns, without clear relational information. We advise an ERD per subject (i.e. the first name before "_" in a table name), for example patient or employee. In these ERDs relations should be indicated with lines. Then, one ERD should show the relations between each subject, if possible. Also, it is possible to add a comment to tables, which might be good idea to do for all tables, as currently the function of some tables is unclear. This will support any further researchers and employees with database access.

Future research

1. Some performance gain is possible in the proposed tool. We designed the tool to calculate performance measures itself to see how quick our tool performs with the current data structure. It would be significantly faster to not calculate utilization and other KPIs for sessions in the algorithm, but by updating the values in the instance after scheduling an appointment and during pre-processing. Pre-processing often happens on a daily basis (usually at night) in an automatic procedure. This makes sure all tables are up to date. In reality, this requires that the IT department adds a number of columns to the database table "index_agenda_tijblokken". The columns that they would have to add (for this performance gain) is the fragmentation index, utilization, schedule type and predicted $P(Utilization > 0.922)$ corresponding to the "index_agenda_dagdata_id" per session. The researcher and IT department discussed these changes and concluded that this would not be an issue. Scheduling an appointment in the demonstrational tool takes approximately 2.4 seconds on average, with observations between 0 and 6 seconds. This means running the simulation that we used for evaluation in Chapter 5 takes between 8 and 10 hours, which is why we limited the simulation runs to July 2019. We expect the proposed scheduling tool to be significantly faster than this demonstrational tool after implementation. First of all, as discussed above, a lot of the performance measures that we need will be calculated in pre-processing instead of during the scheduling procedure. Second, the database holds a pre-processed table with all available gaps in the schedule, which means we do not have to loop over the entire schedule for a day during the appointment generation in Stage 2. Finally, the implemented tool will be programmed in C#, which is known to be much faster than Python (the programming language we used for the demonstrational tool). We elaborate on this in the implementation plan in Appendix S.
2. The proposed predictive model could be made more accurate. Due to time restrictions we had to keep this model simple. Therefore we opted for binary logistic regression, which assumes linear relations of each variable to the log-odds of the predicted value. In reality these relations are non-linear. If one would research these non-linear relations, it might be possible to construct a more accurate predictive model. For the time being though, the proposed predictive model is a viable indication of how crowded a location will be up to a month into the future.
3. To further optimize the weights to Company X's preferences it is wise to research the performance per patient type/session type/required access time/etc. If there is any imbalance this can be fixed by shifting weights accordingly. Also, new weight sets could be instantiated for these specific contexts.

6.5 Generalization of proposed tool to other contexts

This research considers Company X as the target context. However, the tool is generalizable to other contexts. The proposed tool can be used in most multi-provider online appointment scheduling systems where patient pooling is allowed. This is not limited to healthcare, but other online appointment scheduling systems in large organisations as well. Smaller organisation might benefit from more exact methods as the solution space would be better calculable, though the proposed model would still work for smaller organisation with some tuning. Think of a reduction of number of sessions (Stage 1, top-k session ranking) to consider. In much larger organisations runtime might become a problem as the solution space grows. Stronger filtering of the initial session set (before top-k) can counter this

problem of dimensionality. Currently we only use a date window filter for sessions (to not exclude sessions that might be necessary for emergency situations), but a geographical filter would be an easy way to limit the number of sessions to evaluate (Stage 1, top-k session ranking). The criteria we use are tailored to the wishes that Company X has, but for implementation in other contexts it is possible to add or remove criteria and adjust the weights according to the wishes of the decision maker(s). The main challenge when implementing this tool outside Company X, is that the decision makers' database might need to be altered to accommodate the proposed tool.

6.6 Contribution to literature

We consider the contribution of this research to literature to be threefold. First, this research serves to fill a gap in literature concerning multi-location online appointment scheduling systems with patient pooling. To the author's knowledge there is currently no literature on systems of this type. Existing solutions are patient allocation methods and single-location dynamic scheduling procedures. Generalizing single-location solutions proposed in literature to multiple locations is possible, but does not seem feasible within reasonable time. Also, this literature does not suggest methods of incorporating additional attributes into the decision process. They often propose optimization based methods, which usually do not incorporate multiple conflicting objectives. Another issue with optimization-based methods is that they neglect the human aspect of scheduling that lies in negotiation between patient and scheduler.

The proposed two-stage decision-making procedure reduces the solution space intelligently (Stage 1) and ranks the remaining appointment options, considering multiple locations and providers (Stage 2). To be able to obtain ranked appointments, based on multiple conflicting objectives, we use the Multi-Criteria Decision Making method TOPSIS in both stages. To the author's knowledge, multi-attribute ranking-based approaches to appointment scheduling are not yet present in literature, which is the second contribution of this research to literature.

The criteria we use for these stages represent the organisation's preferences as well as the patient's preferences. The result is a set of mutually preferable appointment options, which the scheduler suggests while negotiating an appointment with the patient. Our research therefore taps into a yet largely unexplored area of combining patient choice and provider choice in a mediating intelligent scheduling support tool. This is our third and final contribution to literature.

Bibliography

[1] Daniel Berrar. *Cross-Validation*. 01 2018. ISBN 9780128096338. doi: 10.1016/B978-0-12-809633-8.20349-X.

[2] R Blackburn, K Lurz, B Priese, G Rainer, and I Darkow. A predictive analytics approach for demand forecasting in the process industry. *International Transactions in Operational Research*, 22(3):407–428, 2015.

[3] Paul A. Bottomley, John R. Doyle, and Rodney H. Green. Testing the reliability of weight elicitation methods: Direct rating versus point allocation. *Journal of Marketing Research*, 37(4): 508–513, 2000. ISSN 00222437.

[4] G.E.P. Box, G.M. Jenkins, and H. Day. *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day, 1976. ISBN 9780816211043.

[5] Michael W Browne. Cross-validation methods. *Journal of Mathematical Psychology*, 44(1): 108–132, 2000. ISSN 0022-2496. doi: https://doi.org/10.1006/jmps.1999.1279.

[6] Edward Clarkson, Jason Zutty, and Mehul Raval. A visual decision support tool for appendectomy care. *Journal of Medical Systems*, 42, 02 2018. doi: 10.1007/s10916-018-0906-9.

[7] W. Cooper, L. Seiford, and K. Tone. Data envelopment analysis: A comprehensive text with models, applications, references and dea-solver software. 1999.

[8] E. Demir. A decision support tool for predicting patients at risk of readmission: a comparison of classification trees, logistic regression, generalized additive models, and multivariate adaptive regression splines. *Decision Sciences*, 45(5):849–880, 2014.

[9] Maciej Drozdowski. *Scheduling for Parallel Processing*. Springer-Verlag London, 01 2009. doi: 10.1007/978-1-84882-310-5.

[10] S. A. Erdogan and B. T. Denton. Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORM J. Computing*, 25(1):116–132, 2013.

[11] S. Ayca Erdogan, Alexander Gose, and Brian T. Denton. Online appointment sequencing and scheduling. *IIE Transactions*, 47(11):1267–1286, 2015. doi: 10.1080/0740817X.2015.1011355. URL `https://doi.org/10.1080/0740817X.2015.1011355`.

[12] José Figueira, Salvatore Greco, and Matthias Ehrgott. Multiple criteria decision analysis, state of the art surveys. *Multiple Criteria Decision Analysis: State of the Art Surveys*, 78, 01 2005. doi: 10.1007/b100605.

[13] Stanisław Gawiejnowicz. *Time-Dependent Scheduling*. Springer-Verlag Berlin Heidelberg, 01 2008. ISBN 978-3-540-69445-8. doi: 10.1007/978-3-540-69446-5.

[14] Yigal Gerchak, Diwakar Gupta, and Mordechai Henig. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42(3):321–334, 1996. ISSN 00251909, 15265501. URL `http://www.jstor.org/stable/2634346`.

[15] R.L. Graham, E.L. Lawler, J.K. Lenstra, and A.H.G. Rinnooy Kan. Optimization and approximation in deterministic sequencing and scheduling : a survey. *Annals of Discrete Mathematics*, 5:287–326, 1979. ISSN 0167-5060. doi: 10.1016/S0167-5060(08)70356-X.

[16] George Gross and Francisco D. Galiana. Short-term load forecasting. *Proceedings of the IEEE*, 75(12):1558–1573, 1987.

[17] Diwakar Gupta and Lei Wang. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, 56:576–592, 06 2008. doi: 10.1287/opre.1080.0542.

[18] M. Hardy. *Regression with dummy variables*. 1993.

[19] Hans Heerkens and Arnold van Winden. *Solving Managerial Problems Systematically*. Noordhoff Uitgevers, 2017. ISBN 9789001887957. Translated into English by Jan-Willem Tjooitink.

[20] Roland Helm, Armin Scholl, Laura Manthey, and Michael Steiner. Measuring customer preferences in new product development: Comparing compositional and decompositional methods. *International Journal of Product Development - Int J Prod Dev*, 1, 01 2004. doi: 10.1504/IJPD.2004.004888.

[21] Rainer Herrler, Christian Heine, and Franziska Klgl. Appointment scheduling among agents: A case study in designing suitable interaction protocols. *AMCIS 2002 Proceedings*, 199, 07 2002.

[22] M. V. Herwijnen. Multiple - attribute value theory (mavt). 2010.

[23] Ayaz Hyder, D. Buckeridge, and B. Leung. Predictive validation of an influenza spread model. *PLoS ONE*, 8, 2013.

[24] Yoonsuh Jung and Jianhua Hu. A k-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics*, 27(2):167–179, 2015. doi: 10.1080/10485252.2015.1010532.

[25] Prashanthi Nirmala Kodikara. Multi-objective optimal operation of urban water supply systems. 2008.

[26] V. Kotu and B. Deshpande. *Predictive Analytics and Data Mining : Concepts and Practice with Rapidminer*. Amsterdam: Elsevier, 2014.

[27] E. Kyriakides and M. Polycarpou. *Short term electric load forecasting: A tutorial*, volume 35 of *Studies in Computational Intelligence*. 2006.

[28] Michael P. LaValley. Logistic regression. *Circulation*, 117(18):2395–2399, 2008. doi: 10.1161/CIRCULATIONAHA.106.682658.

[29] A. Law and W. D. Kelton. Simulation modeling and analysis. 1982.

[30] Denis Heng-Yan Leung. Cross-validation in nonparametric regression with outliers. *The Annals of Statistics*, 33(5):2291–2310, 2005. ISSN 00905364.

[31] Jie Lian, Kori Distefano, Somer Shields, Cindy Heinichen, Melissa Giampietri, and Lian Wang. Clinical appointment process: Improvement through schedule defragmentation. *Engineering in Medicine and Biology Magazine, IEEE*, 29:127 – 134, 05 2010. doi: 10.1109/MEMB.2009.935718.

[32] Nicolás Madrid and Pavel Rusnok. A top-k retrieval algorithm based on a decomposition of ranking functions. *Information Sciences*, 474, 09 2018. doi: 10.1016/j.ins.2018.09.014.

[33] Kevin Marsh, Maarten IJzerman, Praveen Thokala, Rob Baltussen, Meindert Boysen, Zoltán Kaló, Thomas Lönngren, Filip Mussen, Stuart Peacock, John Watkins, and Nancy Devlin. Multiple criteria decision analysis for health care decision making—emerging good practices: Report 2 of the ispor mcda emerging good practices task force. *Value in Health*, 19(2):125–137, 2016. ISSN 1098-3015. doi: https://doi.org/10.1016/j.jval.2015.12.016.

[34] William P Millhiser and Emre A Veral. A decision support system for real-time scheduling of multiple patient classes in outpatient services. *Health care management science*, 22(1):180—195, March 2019. ISSN 1386-9620. doi: 10.1007/s10729-018-9430-1. URL `https://doi.org/10.1007/s10729-018-9430-1`.

[35] P Mirowski, S Chen, T. K. Ho, and C.-N. Yu. Demand forecasting in smart grids. *Bell Labs Technical Journal*, 18(4), 2014.

[36] T.M. Mitchell. *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, 1997. ISBN 9780071154673.

[37] Charles Nyce. Predictive analytics white paper. 01 2007.

[38] Serafim Opricovic and Gwo-Hshiung Tzeng. Compromise solution by mcdm methods: A comparative analysis of vikor and topsis. *European Journal of Operational Research*, 156(2):445–455, 2004. ISSN 0377-2217. doi: https://doi.org/10.1016/S0377-2217(03)00020-1.

[39] C. Paço and J. M. C. Pérez. The use of dea (data envelopment analysis) methodology to evaluate the impact of ict on productivity in the hotel sector. 2013.

[40] Richard R. Picard and R. Dennis Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984. ISSN 01621459.

[41] B. Puza. *Bayesian Methods for Statistical Analysis*. ANU Press, 2015.

[42] Payam Refaeilzadeh, Lei Tang, and Huan Liu. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. doi: 10.1007/978-0-387-39940-9_565. URL `https://doi.org/10.1007/978-0-387-39940-9_565`.

[43] Bernard Roy. *Paradigms and Challenges*, pages 3–24. Springer New York, New York, NY, 2005.

[44] E. Silver, D. Pyke, and D. Thomas. *Inventory and Production Management in Supply Chains*. Boca Raton : CRC Press, 2016. doi: 10.1201/9781315374406.

[45] Nigel Slack, Alistair Brandon-Jones, and Robert Johnston. *Operations Management (7th edition)*. Pearson, 2013.

[46] Pruethsan Sutthichaimethee. Varimax model to forecast the emission of carbon dioxide from energy consumption in rubber and petroleum industries sectors in thailand. *Journal of Ecological Engineering*, 18(3):112–117, 2017. doi: 10.12911/22998993/70200. URL `http://dx.doi.org/10.12911/22998993/70200`.

[47] T. Tanizaki, T. Hoshino, T. Shimmura, and T. Takenaka. Demand forecasting in restaurants using machine learning and statistical analysis. In *Procedia CIRP*, volume 79, pages 679–683, 2019.

[48] I. B. Vermeulen, S. M. Bohte, P. A. N. Bosman, S. G. Elkhuizen, P. J. M. Bakker, and J. A. La Poutré. Optimization of online patient scheduling with urgencies and preferences. 2009.

[49] H Voogd. Multicriteria evaluation with mixed qualitative and quantitative data. *Environment and Planning B: Planning and Design*, 9(2):221–236, 1982. doi: 10.1068/b090221.

[50] Jin Wang and Richard Y.K. Fung. Adaptive dynamic programming algorithms for sequential appointment scheduling with patient preferences. *Artificial Intelligence in Medicine*, 63(1):33–40, 2015. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2014.12.002.

[51] M Widianta, Taufiq Rizaldi, Dwi Setyohadi, and Hendra Riskiawan. Comparison of multi-criteria decision support methods (ahp, topsis, saw & promenthee) for employee placement. *Journal of Physics: Conference Series*, 953:012116, 01 2018. doi: 10.1088/1742-6596/953/1/012116.

[52] Billy Williams. Multivariate vehicular traffic flow prediction: Evaluation of arimax modeling. *Transportation Research Record*, 1776:194–200, 01 2001. doi: 10.3141/1776-25.

[53] W.L. Winston and J.B. Goldberg. *Operations Research: Applications and Algorithms*. Operations Research: Applications and Algorithms. Thomson Brooks/Cole, 2004. ISBN 9780534423629. URL `https://books.google.nl/books?id=tg5DAQAAIAAJ`.

[54] Jarosław Wątróbski, Jarosław Jankowski, Paweł Ziemba, Artur Karczmarczyk, and Magdalena Zioło. Generalised framework for multi-criteria method selection. *Omega*, 86:107–124, 2019. ISSN 0305-0483. doi: https://doi.org/10.1016/j.omega.2018.07.004.

[55] Lihong Xu, Qingsong Hu, Haigen Hu, and E. Goodman. Conflicting multi-objective compatible optimization control. 2010.

[56] Chongjun Yan, Jiafu Tang, Bowen Jiang, and Richard Y.K. Fung. Sequential appointment scheduling considering patient choice and service fairness. *International Journal of Production Research*, 53(24):7376–7395, 2015. doi: 10.1080/00207543.2015.1081426.

[57] Xin Yan and Xiao Gang Su. *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Co., Inc., USA, 2009. ISBN 9789812834102.

[58] Joe Zhu and W. Cook. Modeling data irregularities and structural complexities in data envelopment analysis. 2007.

Appendices

Appendix A

Research questions with subquestions

Research question 1: What does the current scheduling process look like and how does it perform on the relevant KPIs?

- What are the KPIs?
- What patient care paths are there?
- Which questions does the service department ask new patients to determine where and when they should be scheduled?
- How do schedulers generally decide where and when to schedule patients (not just service desk and not just new patients)?
- What is the current performance on the KPIs?

Research question 2: How can we make a predictive model of the occupancy of a location?

- What variables influence the occupancy and how are they related?
- How far into the future should we be able to make accurate predictions?
- What type of model could do this?
- How can we validate the model?

Research question 3: How can we create and validate a scheduling support tool that can adequately inform schedulers on which time slots and sessions to choose, utilizing our predictive model (RQ2)?

- Which visual aids could we add?
- What online scheduling heuristics/algorithms could be of use (incorporating our occupancy model)?
- Which of these methods should we apply at Company X?
- What data do we need to feed the tool?
- How can we model the desired support functionalities?
- What should the lay-out be?
- How do we validate the model?

Research question 4: What are the resulting benefits of the scheduling support tool over the current situation?

- How can we test the model without disrupting the running IT systems at Company X?
- How can test the performance of the model, taking into account that a big part of performance is dependent on the schedulers' behaviour?

- How does the new scheduling procedure's performance compare to that of the old one?

Research question 5: How can our scheduling support tool be integrated in the existing IT systems at Company X?

- What does the structure of the IT system look like?
- How can we subdivide our prototype into comprehendible modules for step-by-step implementation, such that it fits in the current IT system structure?
- Is it possible to alter the current scheduling wizard such that it displays the desired scheduling support information?
- How can we clearly and concisely communicate how steps should be performed?
- What guidelines should be instated to ascertain security and aftercare?

Appendix B

List of deliverables at the end of master's thesis period

Deliverables:

1. Flow chart visualising scheduling process
2. Predictive model describing future occupancy
3. Planning support tool prototype
4. Summary of planning support tool functionality
5. Implementation plan
6. Master's thesis

Appendix C

An image of the current scheduling wizard

Before getting the screen below the scheduler selects a patient or adds a new one. The scheduler then selects an activity, which has a standard duration. Though this duration can be altered. Furthermore, the scheduler selects a time window, a distance from the patient (postal code is pre-filled). Optionally, the scheduler can fill in a specific region or location. In the list at the bottom, the scheduler can add filters as well.

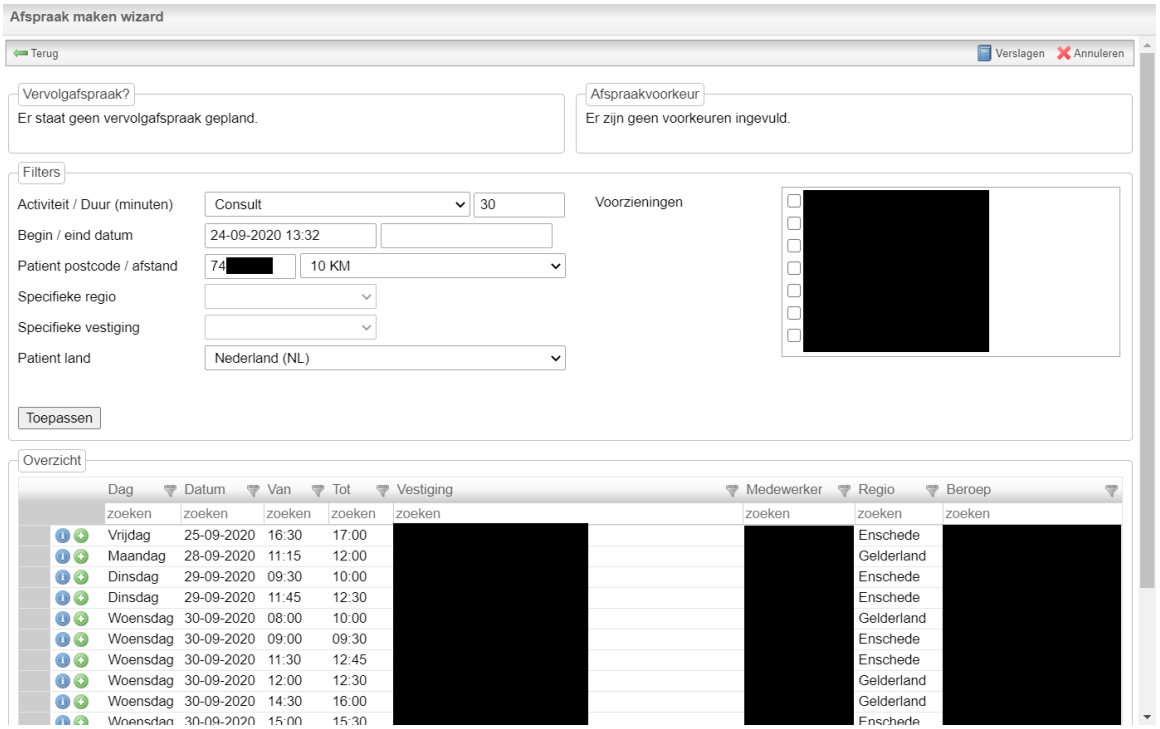

Figure C.1: Image of the current scheduling wizard

Appendix D

Analysis of the implications of the covid-19 pandemic on the 2020 data

To analyse the ramifications of the Covid-19 pandemic on the course of events within Company X in 2020, we analysed the number of completed appointments, number of incoming appointments, number of incoming NP appointments and the average NP access time per month from January to September for each of the last 5 years. We use the data from January to September because that is all we have from 2020 at the moment of this analysis. The following 4 figures best describe the effect of the Covid-19 pandemic. The figures clearly show that the 2020 data is not representative.

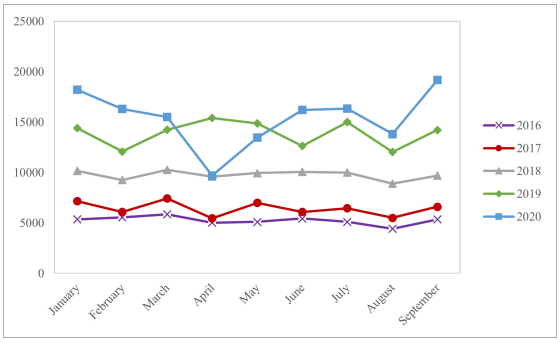

Figure D.1: Number of completed appointments from January to September

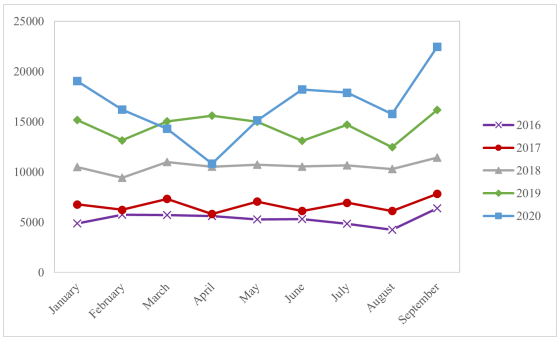

Figure D.2: Number of incoming appointments from January to September

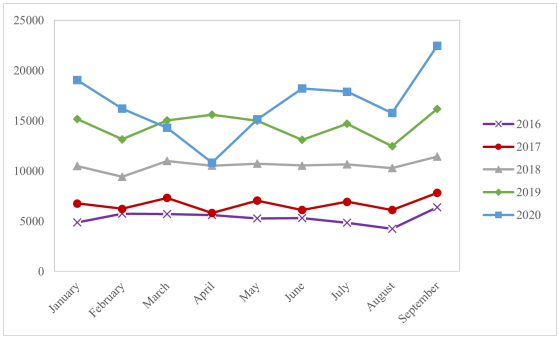

Figure D.3: Number of incoming NPs from January to September

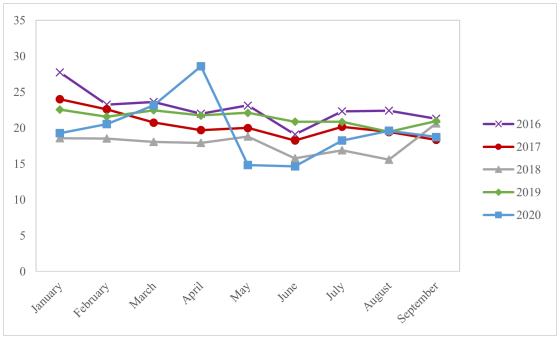

Figure D.4: Average access time for incoming NPs from January to September

Appendix E

Flow charts describing the current scheduling process per patient type

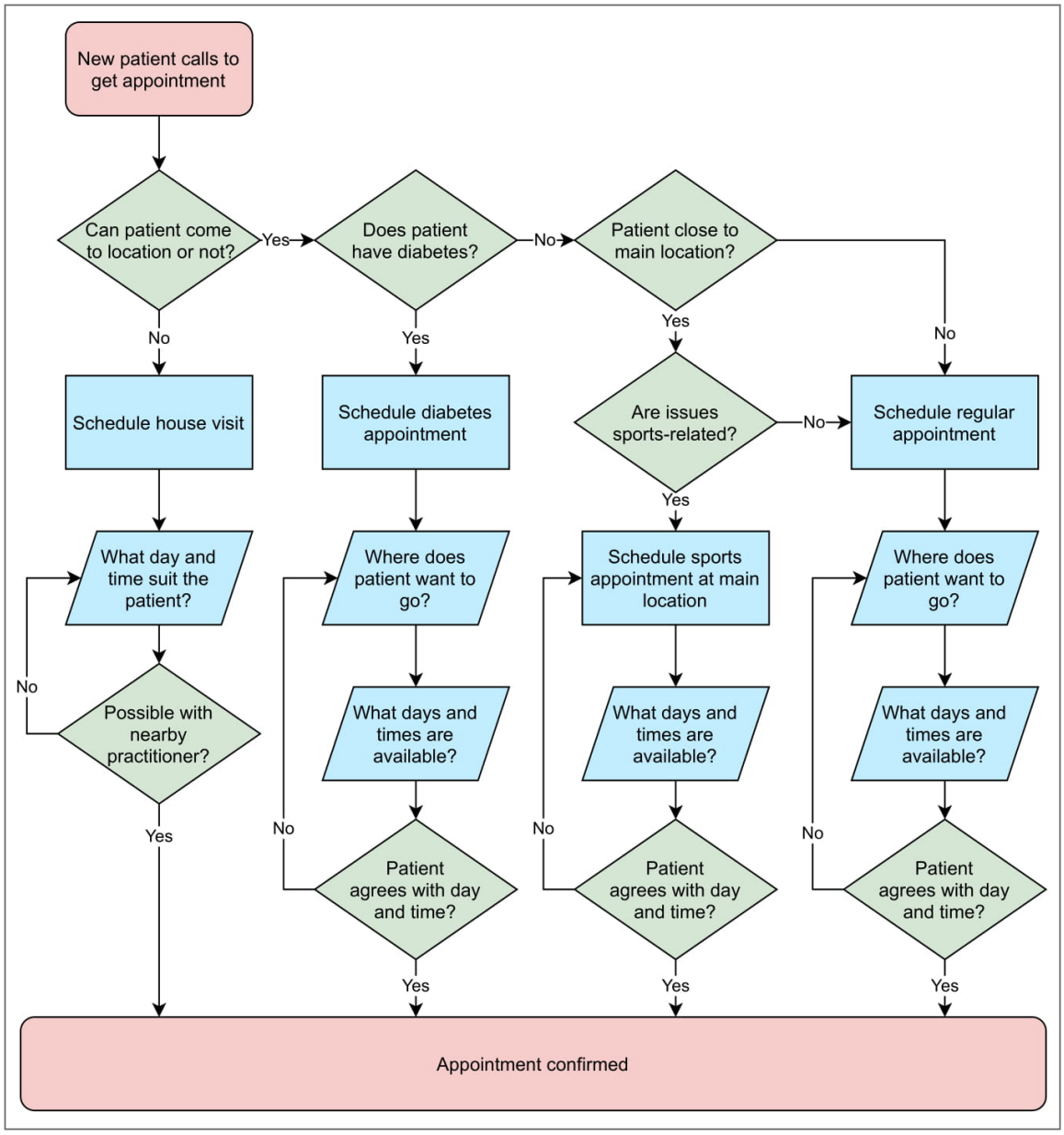

Figure E.1: Flow chart of scheduling new patients in current situation

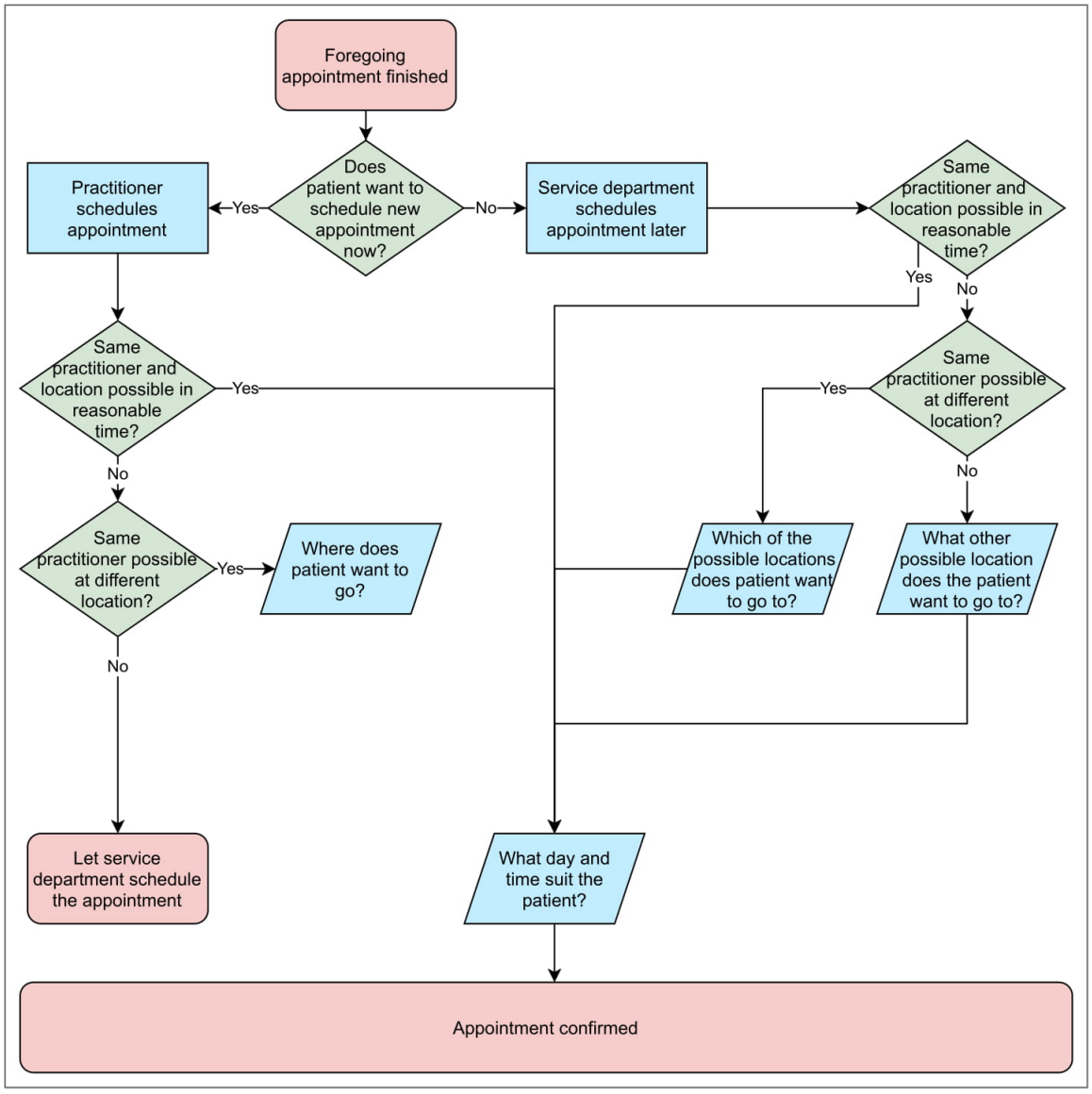

Figure E.2: Flow chart of scheduling recurring patients in current situation

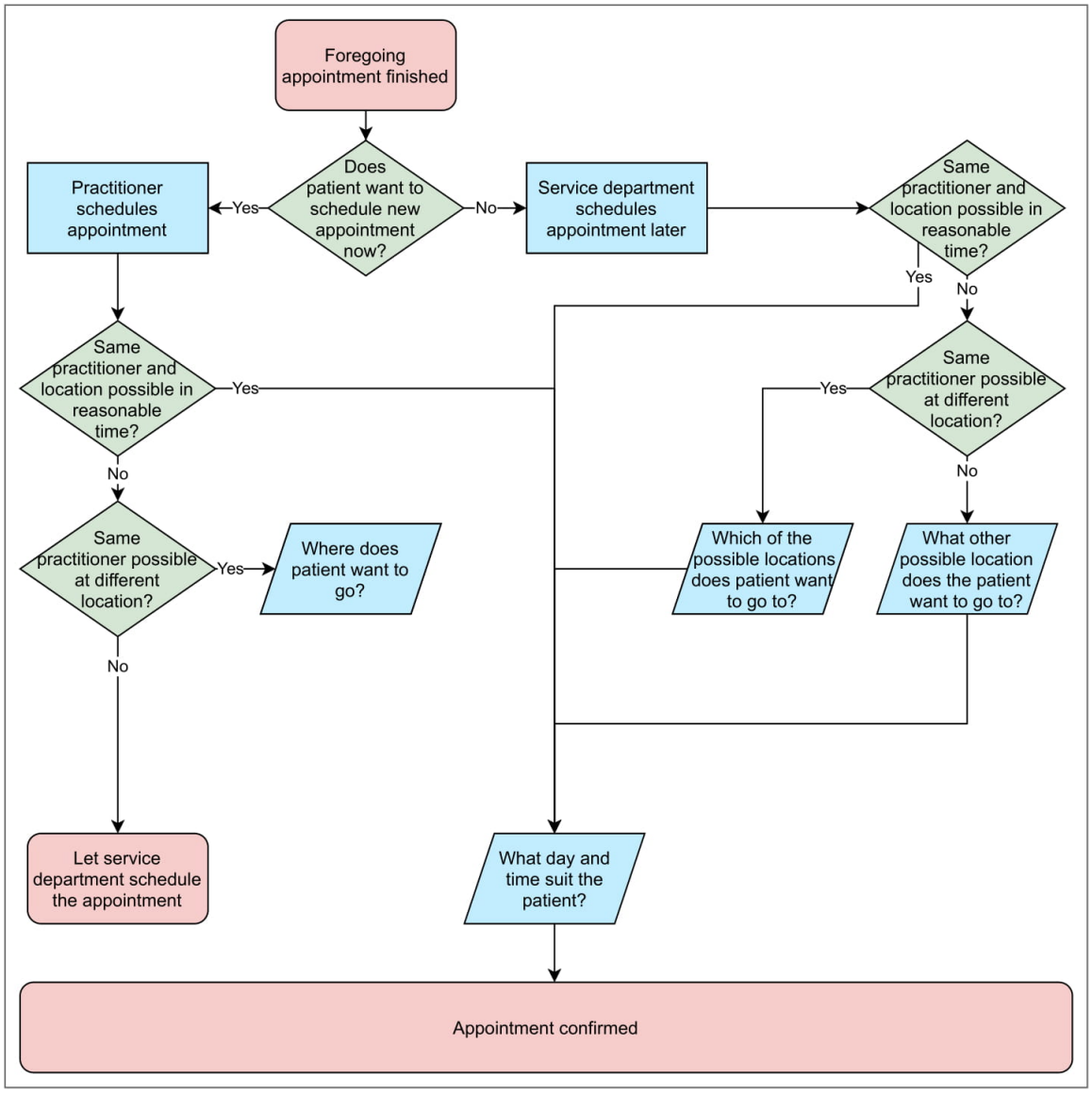

Figure E.3: Flow chart of scheduling periodic patients in current situation (same as RP)

Appendix F

Table showing metrics for all activities

-Classified-

Appendix G

Treatment extraction algorithm description

Our algorithm for extracting treatments from the care paths works as follows. We loop over each appointment of each care path and evaluate the selected appointment and its two successors. These successors can also be empty if at the end of a care path. We only count combinations of appointments with less than half a year between each of them. If the first successor is more than half a year after the selected appointment, we consider the single appointment as a separate treatment of length 1. If the first successor is less than half a year after the selected appointment, but the second successor is more than half a year after the first successor, then we consider the selected appointment and its first successor as a treatment of length 2. We do not evaluate the last appointment in each care path as this, in combination with the two empty slots after the appointment, would always be counted as a separate treatment of length 1 and overestimate the frequency of treatments of length 1 (care path is cut off at 31-12-2019).

If we encounter a treatment that we already saw, we add to the frequency of that combination and skip to the first appointment after the treatment. If we did not encounter the combination of appointments yet, we add it to the treatments list as long as the list is not already 500 entries long. We choose a list length of 500 because we notice that at this length the frequencies of the later entries stays low. This means the more common combinations were already included earlier. Also, the computation power needed for this length is acceptable, which is why we do not lower the list length. For the PP care paths we alter the maximum time in between appointments to 400 days. We do this because we want to include PPs that get a yearly treatment (so we use a little over a year).

While we register each treatment we also add the number of days between the appointment preceding the treatment (if there is one) and the first appointment of the treatment (and keep track of the number of times this is registered). We do the same for the number of days between the first and second appointment and the time between the second and third appointment (again only if possible). At the end of the algorithm we write the frequencies and the averages of these times to the sheet per treatment.

Appendix H

Data sources and processing

Apart from literature, our sources of information come from within Company X. We gain qualitative information using unstructured interviews, observations and service department phone call recordings. The quantitative data that we use comes from the IT system agenda, the database and reports that were readily available (upon request). We describe the data sources to give the reader an understanding of what we base our research on. Furthermore, we need to confirm that all data, original and newly extracted, is valid. In this appendix we describe our data sources and some of the data processing steps that we consider as valuable for reproduction in further research.

H.1 Qualitative data

The qualitative data that we obtain defines our understanding of the current scheduling process and how it can best be improved. Mostly we shape our view of the process from observing the current appointment scheduling tool and the appointment data, but also simply by asking questions within Company X by means of unstructured interviews. Unstructured interviews are also very important in establishing the constraints of our research. For example, patients have to be scheduled at the same practitioners in many RP appointments, which is the kind of information that we acquire with unstructured interviews. The supervisor at Company X is the most direct source of information, which means questions are usually answered by him if possible. For shaping our understanding of the scheduling process and the current scheduling tool we performed unstructured interviews with the service department (schedulers), a practitioner, the marketing department, the IT department, a management team advisor, controllers, quality management and the R&D department.

Our unstructured interviews with the service department and practitioner were part of observation moments that we discuss below. Unfortunately, due to the covid-19 pandemic, contact with patients and employees has to be kept to a minimum, which gives a high intuitive resistance to observation in person. This means we tried to limit the number of observation periods to what was strictly necessary. During the observation period at the service department we tried to obtain information to help shape a picture of the current scheduling process and how the current scheduling wizard contributes to it. The observation period at the practitioner helped us gain an elementary understanding of what the process looks and showed us to some extent how the practitioner schedules a new appointment for the patient at the end of the appointment. We had an unstructured interviews with the marketing department to acquire the marketing activity data that we need to analyze its effect on demand and an explanation on how to interpret this data. We had unstructured interviews with the IT department to get access to the database and to get an explanation on how and when we could extract information from it. Furthermore, we had contact for additional questions and information for the implementation plan. We can only extract large amounts of data from the database outside of office hours, since large extractions temporarily delay any other requests. We regularly have meetings with the supervisor and a practitioner that is active as advisor to the management team once every 2-3 weeks (though

this varies throughout the research) to discuss our progress and the validity when possible. In these meetings we make sure the researcher and Company X agree on each aspect of the research and that the research is headed in the right direction according to both parties. Unstructured interviews with a controller often occur in the form of simple questions about qualitative aspects of the process. The supervisor at Company X is a controller himself, so usually these questions can be answered by him, but sometimes we are referred to others, e.g., someone from quality management. At the beginning of the research we had an unstructured interview with someone at the R&D department to determine their interests concerning the scheduling tool. They state that in the current scheduling tool, their research equipment is not accounted for. For instance, a lot of different parties use pressure plate equipment for measurements and currently it is not registered at which appointments they are used. Often that means that research equipment is not available because it is used elsewhere. After consultation with the supervisor at Company X, we came to the conclusion that this problem, mentioned by R&D, is outside the scope of our research. However, we do advise Company X to further look into this problem.

We observe the process in many ways. Simply seeing how the current scheduling tool works, gives us a lot of information. Furthermore, we establish how schedulers work by observing them for some time, after due consultation. For this purpose we had two observation moments. At the first observation moment the researcher observed the work of a practitioner for about an hour and a half, in which we also got a look at how a practitioner schedules the patient. The way a practitioner schedules only differs from how a scheduler schedules in that practitioners only schedule appointments for themselves instead of considering other practitioners. With the second observation moment, the researcher observed a scheduler for a period of 4 hours to see and hear the scheduling process happen over the phone while observing everything the scheduler did on screen. We also have access to phone call recordings from appointment negotiations between the scheduler and patient. This is important to consider as well, since having an observer watching might change a scheduler's behaviour. Apart from that, it simply gives a lot of additional information on top of the 4 hours of information that we have from the observation moment. The phone call recordings give us a good view of the scheduling process without any external influences, yet are limited by the fact that we cannot see what happens on the scheduler's computer screen. The combination of both gives us a more complete understanding of the current scheduling behaviour. We listened to a random selection of 58 phone call recordings provided by the supervisor at Company X from August till October, 2020. We assume that the schedulers' behavior in these 3 months is representative for our research apart from the fact that some questions related to the covid-19 pandemic occur. 26 of the phone calls were incoming calls (patient calls Company X) and the 32 remaining calls are outgoing calls where Company X calls the patient. The latter happens if an NP is referred to Company X by an external party (usually general practitioner). An outgoing call also occurs when an RP or PP did not want to schedule an appointment directly at the end of the previous appointment and does not call for a new appointment him/herself (assuming that they should still receive care).

Qualitative data is also leading in how we can use quantitative data. For instance, information came up during the research that some quantitative data was invalid due to a flaw in documentation. This allowed us to rid ourselves of the compromised data and secure validity of the research. There are many more examples like this. We keep collecting qualitative data throughout the research to make sure that we optimally understand the current situation and corresponding quantitative data that we use for analysis and base our model/tool on.

H.2 Quantitative data

The quantitative data that we use comes from the IT system agenda, the database and reports that were readily available (upon request). In this section we describe what data we use, how we (pre)processed it, how we use it and how we validate it.

H.2.1 General information and limitations

The data of 2020 is significantly influenced by the Covid-19 pandemic. The pandemic has had its ramifications for Company X by taking away many appointments from March and April, some of which were lost, whereas others were reallocated to the months that followed. Furthermore, since the pandemic, Company X sends the *-classified-* through post with explanation instead of a pick-up appointment. They normally perform these pick-up appointments so they can explain to the patient how the *-classified-* should be used and how it will help, which they ought to be necessary for a good service. We analyzed the 2020 data and concluded that the 2020 data is not representative, which is why most of our research is based on the data from 01-01-2016 to 31-12-2019. Appendix D contains more detailed information on the 2020 data based on the analysis. Leaving out the 2020 data, means that 444642 appointments remain to be used for analysis. When we do use 2020 data we specify this and if necessary exclude the significantly affected months.

H.2.2 IT system at Company X

One of the most significant data sources that we have is the in-house IT system, from which we exported all appointment data from 01-01-16 to 30-06-22 (the day of export was 02-10-20, so any date later consists of appointments that were scheduled before, or on, the day of export). This consists of 629711 schedule entries, including non-patient related activities, and 561994 excluding non-patient related activities. The latter set can be classified as patient appointments.

Access times

From the appointment data we extracted the date of creation of the appointments using the log entries included in the appointment data. From this we could deduce the access times for patients, which is one of our main KPIs. We deleted a lot of appointment data due to invalidity. We describe what data and why in the anomalies section. Furthermore, we exclude the regions *-classified-*, which are artificial regions that only apply to specific locations. *-classified-*. While appointments can take place in *-classified-*, our scheduling tool will not have anything to do with them. Corresponding schedules are not meant to be optimally utilized, since they are used as a service that is performed only a few times per day. When assessing other KPIs we also do not include these regions.

Employee data

From the Company X database we extracted all employee data, but the employee types we found were not in line with the aforementioned ones. It turns out that the employee type column (medewerker_soort_id) in the employee table is simply for rights in the IT system. In the employee table there is a column called profession ID (medewerker_beroep_id), which is what we needed. However, there is no description available in the database. Profession descriptions were only available per employee in a part of the IT system that the researcher has no access to. By looking at the page source code

from the IT system on the account of the supervisor at Company X (who did have access to this part of the IT system) we found out that the list of descriptions was typed in the source code instead of having a table in the database like one would expect. We extracted this list with 41 professions from the source code and found that there are only 6 profession types used for employees that we have schedules and appointments for. The other professions do not treat patients or are used to describe relations (e.g., physicians). We state the number of employees and proportion of appointments (per patient type NP, RP, PP) per profession (for the 6 professions that treat patients) in 2019 in Section 2.1.1.

Realized care paths and treatments

To analyse what the realized care path of each patient looks like we extracted all activities and their dates for each patient in chronological order from the appointment data. We extracted the realized care path per patient and counted the frequency of each realized care path from 2016 till 2019. We limit the realized care paths to a maximum of 20 entries, since we notice that patients with more than 20 appointments are PPs from which the rest of the realized care path can be logically deduced based on the foregoing appointments. There are some RPs that have up to 20 appointments, but not more than that. We limit the realized care path lengths to prevent overflow (some patients have over a 100 appointments), since a VBA (the programming language we use for data analysis) array cannot hold as much data as what we would need if we did not limit the dimensions.

A normal treatment usually consists of at most 3 appointments. A patient's realized care path can be much longer and can include multiple treatments. We want to look at the most frequent realized care paths of length at most 3 to see what the most frequent treatments are. We describe the algorithm we use for this in Appendix G. In Section 2.1.2 we show how often each of these realized care paths occur. It should be noted that for the PPs' realized care paths we count each occurrence of 2 or 3 successive appointments of this type as a treatment. This is just to keep the number of occurrences of this treatment comparable with the other treatments. The vast majority of treatments that we do not specifically mention in Section 2.1.2 (labeled as other in Figure 2.4) are variations on the treatments that we mentioned, varying on one or multiple of the 51 different appointment types. As mentioned before we also extracted the dates of each appointment in the realized care path. We use this to extract the average number of days between appointments in the treatments (in the same algorithm mentioned above, see Appendix G). Taking into account the time in between appointments, we depict the most frequent realized care paths in Figure 2.7 in Section 2.1.2.

H.2.3 Marketing data

Based on the appointment data we can also make a demand forecast of new patients coming in. Any time series can be thought of as being composed of five components, namely level, trend, seasonal variations, cyclical movements and irregular random fluctuations [44]. In our case the timing and content of marketing activities also influence the number of new patients coming in. We therefore also have access to historic data about the marketing activities at our disposal. Therefore, we can analyse the effect of the marketing activity on the number of new patients.

Anomalies

An anomaly we observe in the data is that some appointments have an unrealistically high access time. We trace nearly all of them to events where 5 locations in the region of Rotterdam were taken over by Company X and their clients were also adopted into the system. Nearly all appointment data in 2019 for these locations is compromised, which is why we exclude this data from our data set. We can justify leaving this data out since it is not representative for the current situation and leaving it in would cause a biased comparison between the current situation and the situation after implementation. Other positive outliers are expectedly due to false classification and documenting, which we cannot confirm as false data. However, after implementation of our scheduling tool these misclassifications and misdocumentations should occur at the same frequency, so there should be no significant bias.

What we also see is that there are some negative access times in our data. This happens when someone schedules an appointment that supposedly happened in the past. However, often these appointments are fictitious. This often happens to administrate that diabetic patients were with Company X in that quarter of the year after a takeover of a location (and thus its patients) takes place. Company X gets subsidized for each diabetic patient per quarter, no matter if they have an appointment or not. Therefore, when the diabetic patient did not have an appointment since the takeover, they need to administrate that the diabetic patient was with them in the foregoing quarter, for which they sometimes use a fictitious appointment scheduled in the past. They schedule in the past so the appointment does not fill the future schedule. With non-diabetic NPs something similar often happens. In those cases fictitious appointments are often scheduled in the past to register patients for an invoice if patients are seen without an appointment, e.g. when working in a healthcare center sometimes other care providers ask practitioners to see a patient in between appointments. For this reason negative access times can be seen as invalid and we therefore remove them from our data set. Without the negative access times and invalid locations we have 439033 valid appointments from 01-01-2016 to 31-12-2019.

Figure H.1 shows how the NP access times in 2019 are distributed using a box plot. For RPs and PPs a box plot (or other graphs) would be less informative, since those appointments do not all have the same access time requirements. We see a lot of upward outliers, while we have already filtered some sources of upward outliers (invalid locations discussed in Section H.2). However, we do not take these outliers out of the data set, since there is no proof of invalidity. We expect most of these outliers to be due to NPs postponing their appointments to a much later date. Unfortunately, we cannot establish if this is true, since we are not able to see who instigates a reschedule (which we describe below in the limitations section). Should some of these outliers actually be invalid, then we assume the effect of inclusion of invalid data entries to be equal before and after implementation of our scheduling tool and therefore that the measured impact of our scheduling tool remains valid.

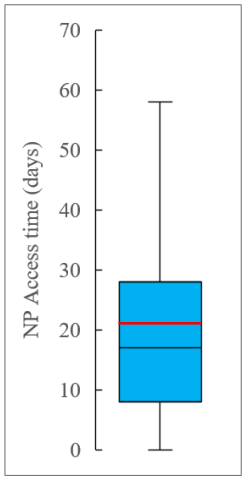

Figure H.1: Box plot of NP access times in 2019 (red line depicts the threshold of 21 days)

Another anomaly we see is that a lot of activities (appointment types) are left blank in the data set. We notice that this only happens in 2016 and that it happens 388 times. Analysing these entries together with the supervisor at Company X showed that these entries were non-patient related activities.

However, the name for this activity was changed in 2016, which is why the
activities are blank. Therefore, we enter "non-patient related" ("niet patiëntgebonden" in Dutch) as the activity for all blanks.

Limitations

A limitation of our data set that cannot be solved is that appointment reschedules are not structurally labelled as being instigated by the patient or Company X. It is sometimes noted in the comment column of an appointment entry, but more often it is not. This means we do not know if a patient asked if their appointment could be moved or if Company X wanted to move the appointment. In the first case, an access time violation after the reschedule would not be Company X's fault, as the patient previously agreed to the appointment time. In this case the access time for the patient should be recalculated as the number of days between the reschedule call and the new appointment date. However, in the second case (Company X reschedules) the access time should be recalculated as the difference between the day that the appointment was originally created (or the date of an earlier reschedule triggered by the patient) and the new appointment date. Since there is no information on what party triggered the reschedule, we leave the access time as the difference between the date that the appointment was created and the date that the appointment eventually takes place. This means we often overestimate the access time, but we assume this overestimation is equal before and after implementation of our scheduling tool. Therefore, the difference in access time in the current and new situation should still be valid. However, when assessing the performance of the new scheduling tool (before implementation), this behavior cannot be taken into account and thus will result in a bias. We advise Company X to start storing whether the patient rescheduled or if Company X did in the database so this behavior can be analysed in the future.

Similarly the data does not show if an access time violation is due to inability to schedule within the desired access time or the patient's choice. We assume that an access time violation is always due to the inability to schedule within the desired access time, while in reality it is sometimes the patient's choice to wait longer than the predefined desired access time for that activity. This contributes to overestimation of the access time, but as mentioned before, we assume that this overestimation is equal before and after implementation of our scheduling tool and that therefore the difference in access time in the current and new situation is valid.

H.2.4 Company X database

Another major data source is the Company X database. We could only acquire data from this database after consultation outside of working hours while on the Company X network. Extracting large amounts of data from this database is very time-consuming, which is why we only extracted the necessary data from 2019, since that is the most recent year that is valid (due to covid-19). The acquired data include location data, employee data, region data and schedule data. The schedule data was mainly what we needed the database for. This allows us to construct the schedules for all employees which we can use for analysis of, for instance, utilizations.

Schedule generation

There are no standard schedules, so for each employee a specific schedule is constructed. The database stores this information rather inefficiently. Each employee has a schedule that cycles for two weeks.

For each location that the employee is active at they have a separate schedule and often multiple for the same location. Additionally there are schedules for single days that they do something different than their normal schedule and the schedules often change. Therefore, in 2019, each practitioner had an average of 56.6 schedules and it is not uncommon for a practitioner to have over 100. There is a separate table in the database that states for each schedule which days they are active but also entries specifying each day that they are not active, which is excessive/redundant. Only a small portion of those entries, which specify each day that the schedules are not active, are useful. Unfortunately, since we cannot distinguish a useful entry from a useless one, this means we cannot bin any of the redundant entries, which is a vast amount. 64.4% of all entries specify that the employee will not be working on that day, and the majority of that information is useless. Since the database is designed to hold 2-week-schedules schedules have 14 day entries. Only single day schedules (used to overrule the normal schedules for 1 day) have 1 entry specifying the new schedule and, if at a different location than the normal schedule, 1 additional entry blocking the normal schedule. The schedule that blocks the current schedule is formatted in the exact same way as the redundant entries that state that patients are not working. There are 62099 day schedule entries. We use a couple of conditions to filter our data:

- We only use schedules that were active in 2019, since it is the most recent valid year.
- We only use practitioner schedules, so not office employee information.
- We only use schedules that are not set to inactive.
- We only use schedules at locations that are also present in appointment data.
- We filter all schedules of employee-location combinations that are not present in the appointment data.

This leaves us with 10162 schedule data entries. We use this data to reconstruct the schedules that the employees had in 2019. We end up with 11697 day schedules that are valid and longer than 0 minutes. For each practitioner schedule we note the opening/closing time, the location, time in between sessions(if applicable), all breaks, schedule time and schedule time subtracting break times. In Section 2.3 we use our appointment data on those schedules to gather information on Company X's current performance.

Just as the schedules are stored inefficiently, the breaks in the schedule are as well. Each day schedule has 9 breaks, some of which are filled and most of which are not. Usually there are only 1 or 2 breaks in a schedule and the other entries are just breaks from 00:00:00 to 00:00:00. This results in an enormous number of entries that is useless. However, for the same reason as with the schedules we cannot filter these entries out, because some of those entries actually state that there will be no break on that day. We have 345713 entries in this table. Just as with the schedule entries we distill active practitioner break entries from 2019 and we filter out the combinations of employees and locations that are not in the appointment data. Also some schedules no longer exist but still have breaks, so we filter out those as well. This brings the number of data entries down to 78963.

When we calculate the amount of time in a workday we take the time between the opening and closing time of the session(s) on that day and subtract the time from breaks for the practitioner for that day and if applicable the times in between sessions. However, there are schedule breaks and employee breaks. Employee breaks are often events like vacations, leaves of abscence, which can last entire days or pieces of days. These employee breaks can overlap schedule breaks, but needs to be

subtracted by the amount of break time, time between sessions and time outside opening/closing time that it overlaps.

Location data

The database registers geographical coordinates for all new postal codes it encounters and stores them. For each appointment in 2019 we determine the latitude and longitude for the patient and the location at which they are treated. We want to get the distance from the patient to the location. We notice a lot of erroneous geographical coordinates resulting in unrealistically large distances (while corresponding to a dutch zip code). We filter data rows that have a latitude or longitude higher or lower than the maximum or minimum latitude and longitude in the Netherlands. Another error we see is that many patients do not have longitudes and latitudes in the database, which occurs very often. We filter these out as well, since they tell us nothing. After filtering all this data we go from 172368 entries to 146519. From the data entries that were deleted 24971 were because the patient had no latitude or longitude. Usually this means the corresponding postal code does not exist or the appointment was non-patient related. Finally, we ignore the house visits and nursing home visits (another 2777 data entries). Still we see some values that are not realistic, so we compare the number difference in postal code divided by the distance (in km) for each appointment and manually filter the locations and patients for which the coordinates are clearly wrong. Finally, we exclude the artificial regions *-classified-*, as we do with all KPIs. This brings us to 141883 valid appointments to determine our distance KPI.

Limitations

Apart from the redundant data, another limitation of the database is that there is no way of knowing when schedules went inactive, since this is not logged. Therefore, we cannot include schedules that were running in 2019 but have been marked as inactive since. We could evaluate if there were appointments scheduled on that location for those days, but it could for instance be that a schedule from 8:00 to 17:00 was set to inactive and replaced by one from 8:30 to 17:00. If we would then perform the data analysis all metrics would corrupted. Therefore we delete the inactive schedules from our data, which means we might lose some data that was valid.

H.3 Validity

We ensure validity of the data by constantly validating newly generated data. We do so by sampling a couple of data entries in the new set and looking up the information in the database to see if all attributes check out. Furthermore, whenever we see values that seem abnormal, we do this same check and discuss the validity with the supervisor. If there are suspicions of invalidity we consult with the supervisor how to deal with the issue. We list the major data issues/invalidity concerns in the Limitations section (6.2) of the Discussion chapter (Chapter 6). This way we maintain a valid data set throughout the research.

Appendix I

External factors analysis

The predictive model that we want to use should describe how the occupancy of sessions develop over time. This is determined by a number of internal factors, like the region, month, and day of the week, which we discussed in 2.1. However, there are also a number of external factors that influence the utilization of a session. Company X has mentioned that they experience crowding as a result of marketing activities. We want to know how much of an effect this has on new appointments. Furthermore, we research if weather conditions have an influence on the number of new appointments. In Appendix I we show the analysis of these external effects in more detail, whereas below we discuss the findings.

Adwords

In the marketing data that the marketing department has provided, we see the Adwords budget for each week from week 17 in 2017 to week 40 in 2020. We use the 2017 to 2019 data for our analysis. Before we can assess if the adwords costs influence the number of incoming NPs we need to detrend and deseasonalize the data. We performed a regression with the observation (costs for that week) as dependent variable and the observation number (observation 1 is week 17 of 2017, observation 140 is week 52 of 2019) as explanatory variable. We then took the residuals as observations to detrend the data. After that we deseasonalized these residuals based on the monthly seasonal index. Then we performed a regression with these values as dependent variable and the Adwords costs as explanatory variable. The P-value was equal to 0.80, which means with level of significance 95% we cannot conclude that the Adwords costs have significant influence on the number of scheduled appointments in the same week. This seems strange since one would expect a higher number of new appointments with higher costs. The problem lies in the way the Adwords costs are calculated. With Adwords the client pays per click, so the more clicks you get, the more you pay. The Adwords costs do grow over the years, but this mainly happens because the business grows and thus the number of clicks. After detrending the data the adwords costs are rather stable and thus it is hard to discern an effect.

Mailings

The dates and regions of mailings can be found in the comments row for each week in the marketing data. However we had to manually read through the comments and register each mailing since no standard way of reporting was used. After registering each of them we get a total of 65 mailings. Only in the 2019 and 2020 data there was information about mailings. For this analysis we do include the 2020 data since there is not enough data to analyse otherwise. We compare the week before each mailing (excluding the mailing day) with the 2 weeks following the mailing (including the mailing day), without taking into account weekend days since these often have 0 appointments. As long as we do not include mailings where those three weeks overlap with new covid-19 measures in the Netherlands, the comparison should be valid. A lot of the mailings were done on Tuesday or Thursday, which means the pattern in differences in appointments was influenced by this. We therefore normalized the number of appointments per day using the seasonality in weekdays.

If we compare the week before the mailing compared to the first week after the mailing we see an average increase in appointments scheduled of 10.2% (95% **confidence interval** is [7.13%, 13.3%]) in the corresponding region. On the first day after the mailing the number of appointments scheduled increases by 31.4% (95% CI: [21.5%, 42.2%]). The average width of 95% confidence intervals over all 14 days on a daily level is 18.1%, which is caused by a rather high standard deviation of observations and a low number of available observations. Therefore, with a level of significance of 5%, we can say that on the first, second, fourth and fifth day after the mailing there is a significant increase in the number of incoming appointments. Thus, mailings have a significant effect on the number of new appointments. We depict the average increase for 13 days after the mailing as a percentage of the normalized average of the week before the mailing in percentages in Figure I.1.

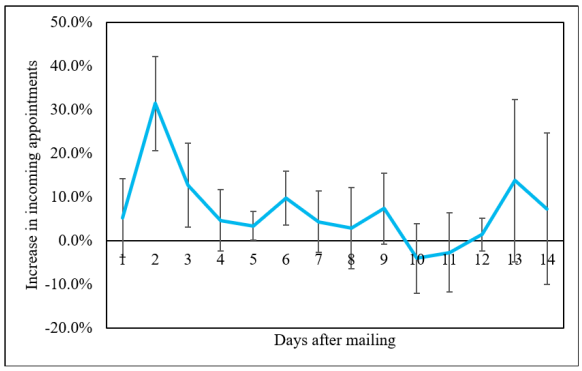

Figure I.1: Percentage increase in incoming appointments in 2 weeks after mailing (on day 1) compared to the average of the week before, with error bars representing 95% confidence interval

Weather conditions

For sake of completeness we decided to evaluate the effect of weather conditions on the number of new appointments. For instance, when it is cold muscles tend to be more tense, which might lead to issues. This could result in a patient calling for an appointment. The weather conditions might have an influence on scheduling behaviour in many other ways. To find out if we might need to take this into account for our predictive model, we obtained the weather data for 2016 to 2020 from the KNMI (the Royal Netherlands Meteorological Institute) database. We used the data from the region Twente (as defined by KNMI), since this region is where most of Company X's appointments are from. We used regression to determine if the number of scheduled appointments is influenced by the maximum temperature in a day or the number of hours of rain in a day. Both variables had a P-value of far over 5%, which means we cannot conclude that these weather conditions have a significant effect on the number of appointments. Since this is a simple linear regression it could be that there is no linear dependency, but there might a non-linear dependency. However, visual analysis of the number of appointments per day against the maximum temperature shows no dependency whatsoever. The same goes for the number of hours of rain.

Appendix J

Flow charts describing the desired scheduling process per patient type

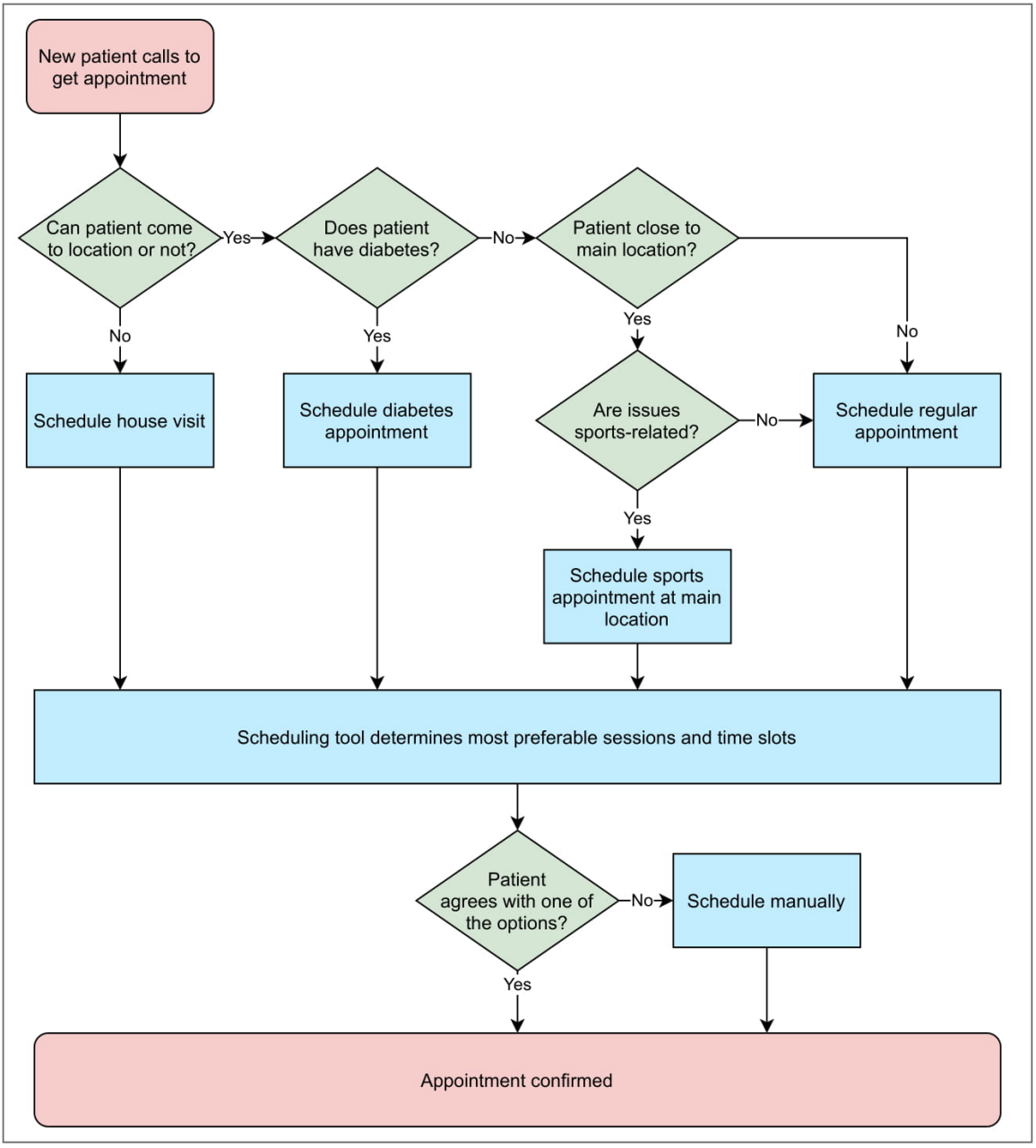

Figure J.1: Flow chart of scheduling new patients with new scheduling tool

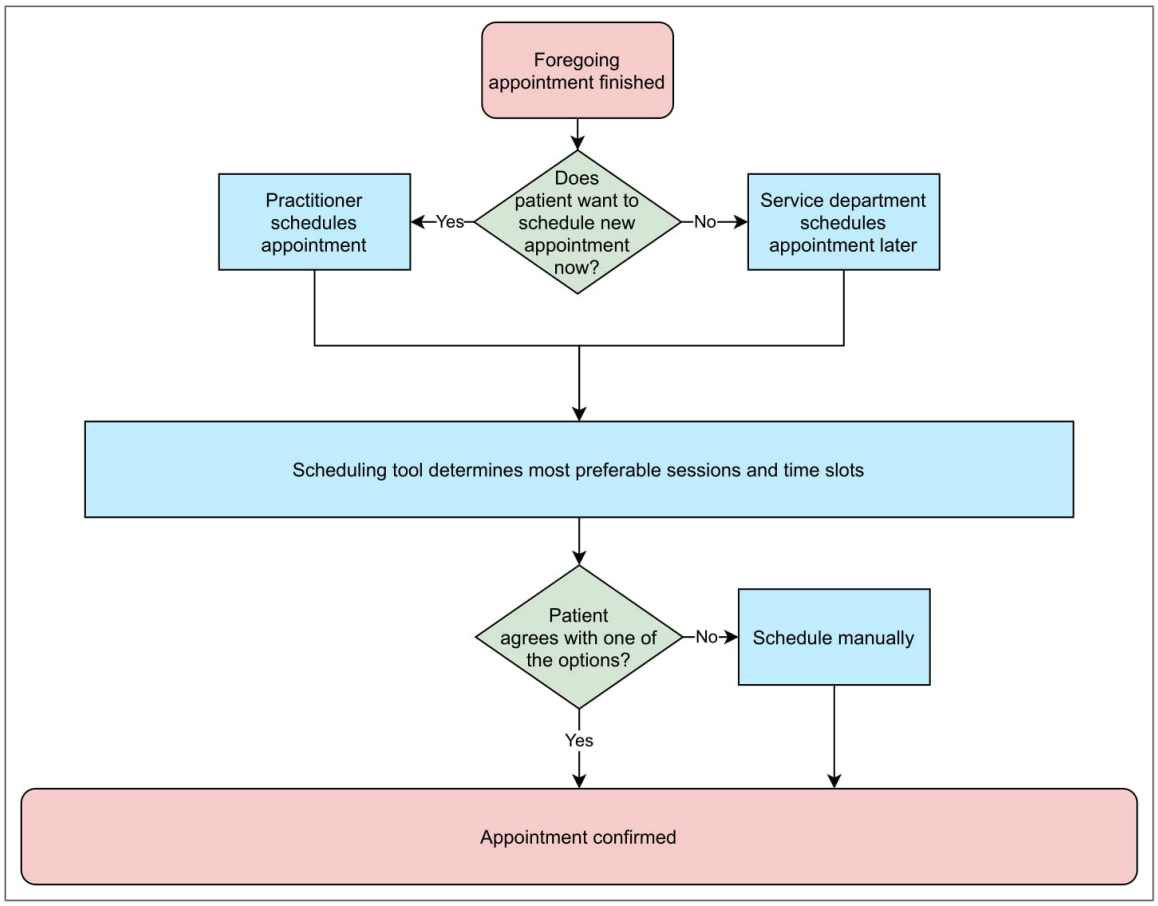

Figure J.2: Flow chart of scheduling recurring patients with new scheduling tool

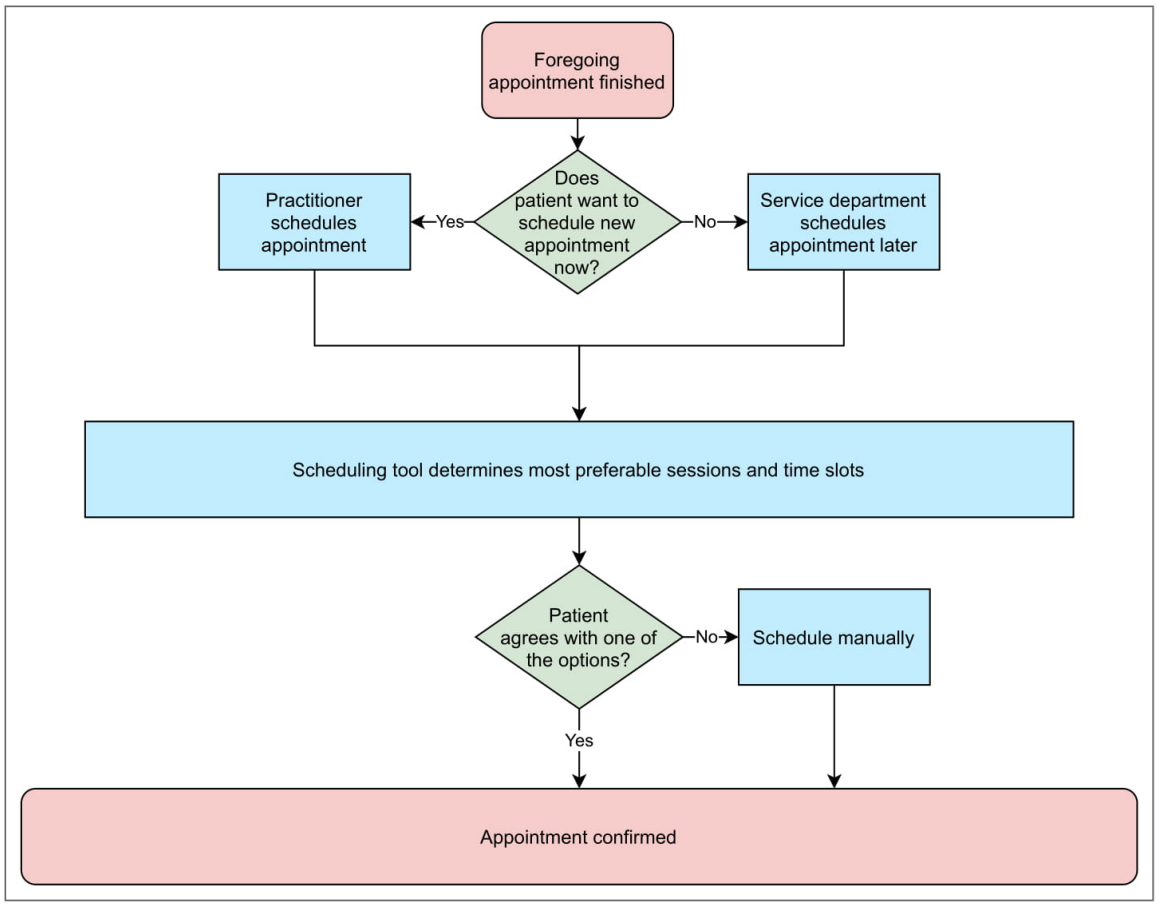

Figure J.3: Flow chart of scheduling periodic patients with new scheduling tool (same as RP)

Appendix K

Calculation of utilizations and fragmentations

To calculate the utilization we divide the amount of appointment time (in minutes) for a practitioner's day by the amount of session time (in minutes) for this practitioner that day. We calculate the fragmentation index by counting the number of gaps in the schedule. The appointment time is including non-patient related activities. However, if there are only non-patient related activities on that day, then we deem it invalid. We delete employees that have only non-patient related appointments and also delete those appointments from the appointment data (necessary for the algorithm to work). Appointment time that is processed outside session hours are not counted as contributing to utilization, otherwise we would get utilizations of over 100%. Also, non-patient related activities often range outside session times while that time should not be counted as utilization. Therefore, we subtract appointment time outside opening/closing time from the appointment duration when calculating utilization. There are some requirements that the schedule for that day has to meet. If the schedule for a certain practitioner on a day does not meet these requirements, we deem the workday invalid for our KPI calculation. These requirements are:

- Two appointments should not overlap. This usually happens when a non-patient related appointment overlaps a patient-related appointment. In this case we cannot assess if the non-patient related appointment should be considered as utilizing the session time or not. When non-patient related activities do not overlap with other activities, we assume that the non-patient related activity is utilizing the sessions minutes.
- A practitioner should have appointments at only two locations on one day. More than that should not happen. For home visits a separate schedule ("location") is used and not the patient's location, which means we do not deem these sessions invalid.
- Breaks in the schedule should not overlap. As explained in I the way the schedules are build and stored is equivocal. Sometimes this results in breaks that should have been overwritten but are not, which is due to an anomaly on a higher level than the data that the researcher has access to.
- A break should not overlap with the time in between two sessions of a practitioner on a day (provided that the practitioner works on two locations that day).

We use the appointment data in combination with the schedules to assess the utilization and fragmentation. After this cross-examination we end up with 10988 valid practitioner workdays, which contain 12497 sessions.

We also generate the utilizations, fragmentation, practitioners and amount of available time per location based on the practitioners' information that we generated. In 94.0% of the cases there is only 1 practitioner at a location on a given day. In 5.8% of the cases there are 2 practitioners and in 0.2% there are 3. When this happens we calculate the location's utilization by averaging the utilizations of the practitioners that are working on the location on that day. Having a weighted

average (based on session duration) would require a chain of algorithms to be altered which would take a substantial amount of time. Furthermore, we expect that there is a positive correlation between practitioner's utilizations at the same location. Given the limited time we have for our research, we feel we can justify this simplification.

Appendix L

Data analysis for predictive model

In this Appendix we describe the data analysis we performed in the process of developing our predictive model. We first discuss a simplification that we needed to make. After that we discuss the algorithm we created to generate training data.

L.1 Simplifications

When we use our predictive model we want to obtain an indication of how crowded a certain session will be. However, we do not have utilizations per session, we have them per practitioner workday and per location, which can both consist of multiple sessions. Generating the utilization per session would take a substantial amount of time, since we would have to develop an algorithm for this. We cannot justify developing algorithms for each small data analysis step, given the limited amount of time that we have, so we opt not to do that in this case. However, we do have utilizations per location and in 94.0% of the cases there is only 1 practitioner at a location on a given day. In 5.8% of the cases there are 2 practitioners and in 0.2% there are 3. In case there are multiple practitioners we simplify by assuming an equal utilization for each session at the location for that day. We expect that this does not significantly affect our model due to the low frequency of occurrence and due to the fact that we expect utilizations of sessions at the same location to be positively correlated (if one session has a high utilization, the other is likely to have a high utilization as well).

Another simplification is that we predict occupancy per region and equally allocate it over the number of locations in this region, while in reality demand might unequally allocated over smaller geographical instances. For instance, one city might have a relatively higher demand than another city in the same region. The reason why we aggregate per region is firstly because the number of classes (regions in our case) would become very big, which is detrimental to performance and complicates data warehousing. Secondly, it might result in overfitting of the model.

L.2 Generating training data

To construct a predictive model we need training data. We use the statistics package Minitab to construct our models. To construct regression models in Minitab, the training data needs to be in a format where each row describes an observation and each column describes a predictor or response value. The moment that the predictive model should be put to use is when the scheduler uses our scheduling tool to schedule an appointment. That is why we use each appointment as an observation moment. The predictor variables that we evaluate are as follows:

- Region of the location that we want to schedule an appointment at
- Year of appointment date
- Month of appointment date

- Day of week of appointment date
- Occurrence of mailing (marketing activity) in corresponding region in month leading up to appointment date
- Number of days between date of scheduling and date of appointment
- Current utilization of appointment date (as observed on scheduling date)
- Utilization during access time
- Working time in the last 4 weeks at the selected location

.

L.2.1 Training data generation algorithm

The algorithm we constructed evaluates all appointments from January 2017 to November 2020. We leave out the 2016 data because we want to base our model on relatively recent data. However, we do want to have multiple years, to get a representative number for our seasonalities and trend. We exclude non-patient related appointments and appointments that took place in March, April and May of 2020 (the covid-19 pandemic had a significant impact in these months). The appointments should be in chronological order (considering the date on which the scheduling took place) to keep track of utilization on the appointment date at the time of scheduling. The algorithm also uses all mailing data and checks for each appointment if there was a mailing in the corresponding region in the 30 days leading up to the appointment. We use the appointment data and generate the variables above for each appointment. For each appointment we update the number of appointment minutes. When evaluating a session the algorithm also looks at how many work days the location had in the last 4 weeks and what the average utilization is on the location in the access time (including the appointment date). To do so we keep track of cumulative appointment minutes for each session. We have the number of session minutes for each session, which we use to calculate the current utilization, working time in the 4 weeks leading up to the appointment and the utilization during the access time.

Appendix M

Regression summaries

At the end of this appendix we have the regression summaries of the multiple linear regression model that predicts occupancy (1), the binary logistic regression model that predicts $P(Utilization > 0.922)$ (2) and the multiple linear regression model that predicts $P(Utilization > 0.922)$ (3), respectively. With $P(Utilization > 0.922)$ we mean the probability that the final realized utilization is higher than the average of 92.2%. This 92.2% is different from the 91.7% that we stated as average utilization in Section 2.3 for two reasons. The first reason is that the 92.2% average is the average over all observed final utilizations in our training data, whereas the 91.7% is the average of all session utilizations. In the first case we get a higher average, because the sessions with a higher utilization have more appointments on average and are thus represented more often. The second reason that there is a difference is that 91.7% is the average over 2019 and 92.2% is the average over 2017 to 2020 (excluding March, April and May of 2020, because of the covid-19 pandemic). Nevertheless, whether 91.7% or 92.2% was the threshold does not make a difference, since the model should merely indicate which session has a higher probability of being crowded compared to another. In both cases the model would have done this equally well. We performed the regressions using a statistics package called Minitab. In the regression we have continuous variables and categorical variables. The categorical variables are region, month and day of the week. We numbered the regions as follows:

Number	Region
1	Almelo
2	Amsterdam
3	Drenthe
4	Enschede
5	Gelderland
6	Hengelo
7	Rotterdam
8	Utrecht
9	Vriezenveen
10	Wierden
11	Polder
12	Friesland
13	Den Haag
14	Leiden
15	Zoetermeer
16	Brabant
17	Groningen

Minitab splits categorical variables into binary variables for each category. For instance, a training data entry with Region Rotterdam, which is Region 7, will have $Region(x) = 1$ for $x = 7$ and $Region(x) = 0$ for $x \neq 7$. Furthermore, the first term called "Constant" is the intercept included in the model. One can interpret the intercept as the value the regression formula would take if each variable would be equal to zero. This logically follows from the regression model, which is of the

form:

$$y = a + b_1x_1 + b_2x_2 + ... + b_nx_n, \qquad i \in \{1, ..., n\} \tag{M.1}$$

Here, y denotes the predicted value of the dependent variable, a denotes the intercept, b_i denotes the coefficient (denoted "Coef" in the regression summary) of variable i and x_i denotes the value of variable i. a is equal to the coefficient of the Constant term. In the binary logistic regression model (2) the predicted variable is the log-odds (also called logit) of the event that ($Utilization > 0.922$) is true. The log-odds is $log(\frac{p}{1-p})$, where $p = P(Utilization > 0.922)$ and the logarithm can have any base that the researcher prefers. Minitab uses a natural logarithm by default, so we use e (Euler's number) as the base. To get $P(Utilization > 0.922)$ from the log-odds we just have to apply the following mapping to our log-odds: $\frac{e^y}{1+e^y}$.

For our multiple linear regression models we use 10-fold cross-validation. From [1]: *"In k-fold cross-validation, the available learning set is partitioned into k disjoint subsets of approximately equal size. Here, "fold" refers to the number of resulting subsets. This partitioning is performed by randomly sampling cases from the learning set without replacement. The model is trained using k - 1 subsets, which, together, represent the training set. Then, the model is applied to the remaining subset, which is denoted as the validation set, and the performance is measured. This procedure is repeated until each of the k subsets has served as validation set. The average of the k performance measurements on the k validation sets is the cross-validated performance."*. Minitab does not support k-fold cross-validation on binary logistic regression. Therefore, we use a split sample method in which we split the data set in 30% validation and 70% training data. This prevents over-fitting of the model.

In the regression summaries below we have from left to right, the terms (name of predictor variable), coefficients, the standard errors of the coefficients, T-value, P-value and variance inflation factor. The coefficients represent the linear relation between the predictors and the dependent variable in our regression model. The SE coefficient measures the precision of the estimate of the coefficient. The smaller the standard error, the more precise the estimate. Note that we do not include "working time in the preceding 4 weeks" as predictor in the regressions summaries. This variable was previously included, but when constructing the model we noticed that generating the number of work days in the preceding four weeks is too computationally demanding to use in the scheduling tool (50% increase in computation time). Furthermore, the performance of the model was nearly equal when leaving the variable out. Should Company X have this variable readily available, we recommend to include it in the model, but as they currently do not have this and since it is so detrimental to the performance of the scheduling tool, we do not include it in our predictive model.

Dividing the coefficient by its standard error calculates a T-value. If the P-value associated with this T-statistic is less than our significance level (5%), we can conclude that the coefficient is statistically significant. We see that, for the continuous terms (top 9 terms, including constant), most P-values are below 5%, which means that they have a statistically significant relation with the dependent variable. Only for (2), we have a P-value above 5% for the interaction variable "Number of days ahead * Utilization during access time", and for (3), we have a P-value above 5% for working time. We do also see some more variables that have P-values greater than 5%. However, these are binary variables that correspond to categorical variables. Each of the categorical variables also has binary variables with a P-value below 5%. This shows that the categorical variables themselves have a statistically significant relation with the dependent variable. Note that each categorical variables' first category has a coefficient equal to 0, since all other categories are assessed relevant to the first

category. Also note that for day 7 (Sunday), the coefficient is also 0. This is due to the fact that there are no sessions on Sundays in our training data. To make sure the model does not fail if ever a Sunday sessions does come up, we set the coefficient equal to 0.

The VIF describes how much multicollinearity (correlation between predictors) is present in our model. Multicollinearity increases the variance of the regression coefficients, making it difficult to evaluate the individual impact that each of the correlated predictors has on the dependent variable. It is important to note that the VIF increases significantly when one includes interaction variables, which is logical since the interaction term is clearly related to its underlying variables. The researcher should therefore calculate the VIF over the model without interaction variables, or he/she should center the underlying variables by subtracting the mean. VIF $= 1$ means there is no multicollinearity between the predictor and the other predictors. $1 <$ VIF ≤ 5 means there is moderate multicollinearity. VIF > 5 suggests that the regression coefficient is poorly estimated due to severe multicollinearity. When we perform a regression without the interaction variables we have no predictors with a VIF $>$ 5, meaning there is no severe multicollinearity between independent predictors.

We do see a statistically significant lack-of-fit for all these models, which means it does not accurately describe the functional relationship between the experimental factors and the response variable. This is due to the fact that this is an inorganic process caused by human interference. It is infeasible for us to capture this interference in variables. However, since this regression model is only meant to be a simple estimator of our utilization, we can tolerate this lack-of-fit. To not get a lack-of-fit we would have to construct a much more intricate model, which is out of the scope of this research. Our model should merely give an indication of whether a location has a high probability of crowding relative to another location, which it does adequately, as we show in Section 4.1.

The model we eventually use is the second model (2), binary logistic regression predicting P(Utilitization > 0.922). The argumentation for this is in 4.1.2. From this model we exclude the predictors "Year" and the interaction variable "Number of days ahead * Utilization during access time". We exclude the year because we observed a marginal negative coefficient for this predictor. The reason for this is that Company X added a lot of locations to cope with the growth and apparently slightly decreased the utilization on average (also due to low utilization in 2020). It is however not logical to assume that utilization will decrease each year. The other predictor was statistically insignificant, as the regression summary shows.

Apart from excluding the predictors mentioned above, we also choose to exclude "Working time in the preceding 4 weeks". The reason is that finding all sessions in the preceding 4 weeks on a given location is very computationally demanding. This makes the computation time for finding suggestions approximately 50% longer. This is something we experienced during construction of the tool. During construction of the predictive model we did notice that whatever predictors we used, binary logistic regression performed slightly better than multiple linear regression. This is to be expected, since binary logistic regression is meant for predicting a probability of a success, which is what we are doing.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.71484	0.00167	429.31	0	
Marketing	0.000089	0.000031	2.87	0.004	1.17
N.o. days ahead	0.004909	0.000083	59.35	0	19.37
Current utilization on appointm	-0.84212	0.0027	-311.52	0	15.68
Utilization during access time	0.07109	0.00331	21.45	0	12.32
Days ahead * Curr util	-0.00285	0.000145	-19.67	0	23.5
Days ahead * Access util	-0.00058	0.000187	-3.13	0.002	49.31
Regio					
1	0				
2	-0.01008	0.000866	-11.63	0	2.13
3	-0.0114	0.00102	-11.17	0	1.62
4	-0.00405	0.000793	-5.11	0	2.65
5	0.008681	0.000781	11.12	0	2.81
6	0.004692	0.000729	6.44	0	3.67
7	-0.01292	0.000928	-13.93	0	1.85
8	-0.00568	0.001	-5.68	0	1.65
9	0.02311	0.00109	21.27	0	1.5
10	0.00501	0.000882	5.68	0	2.05
11	-0.01495	0.000965	-15.49	0	1.73
12	-0.02373	0.00117	-20.3	0	1.41
13	0.00395	0.00127	3.11	0.002	1.33
14	-0.00942	0.0012	-7.87	0	1.38
15	0.0327	0.00235	13.89	0	1.08
16	-0.01422	0.00276	-5.15	0	1.06
17	-0.09881	0.00595	-16.6	0	1.02
Month					
1	0				
2	0.009463	0.000803	11.79	0	1.96
3	0.009393	0.000883	10.64	0	1.68
4	0.017159	0.00086	19.94	0	1.75
5	0.015567	0.000844	18.45	0	1.81
6	0.010591	0.000785	13.49	0	2.07
7	0.017376	0.000781	22.26	0	2.15
8	0.012088	0.000821	14.72	0	1.9
9	0.010517	0.000796	13.21	0	2.27
10	0.007688	0.000768	10.01	0	2.29
11	0.0129	0.000786	16.4	0	2.07
12	0.019037	0.000822	23.15	0	1.89
Day					
1	0				
2	0.002056	0.000507	4.06	0	1.6
3	0.006828	0.000481	14.21	0	1.68
4	0.006219	0.000494	12.58	0	1.65
5	0.010778	0.000562	19.19	0	1.46
6	0.00999	0.00283	3.53	0	1.03
7	0				

Table M.1: Regression summary of multiple linear regression model predicting utilization

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-3.6057	0.06	-60.08	0	
Marketing	0.00248	0.00105	2.35	0.019	1.17
N.o. days ahead	0.08538	0.00299	28.52	0	21.07
Current utilization on appointm	3.4985	0.0942	37.13	0	15.63
Utilization during access time	0.81	0.119	6.82	0	12.22
Days ahead * Curr util	-0.05175	0.0051	-10.14	0	21.17
Days ahead * Access util	0.00834	0.00677	1.23	0.218	50.86
Regio					
1	0				
2	-0.0794	0.0294	-2.7	0.007	2.08
3	-0.1886	0.0341	-5.53	0	1.63
4	-0.0068	0.0267	-0.25	0.8	2.62
5	0.3025	0.0267	11.34	0	2.62
6	0.0829	0.0245	3.38	0.001	3.6
7	-0.1653	0.0312	-5.3	0	1.85
8	-0.1688	0.0336	-5.02	0	1.65
9	0.8597	0.0404	21.27	0	1.37
10	0.2628	0.0301	8.74	0	1.99
11	-0.1518	0.0324	-4.68	0	1.72
12	-0.3099	0.0392	-7.9	0	1.41
13	0.1581	0.044	3.59	0	1.3
14	-0.0623	0.04	-1.56	0.12	1.39
15	1.296	0.103	12.56	0	1.05
16	-0.1165	0.0896	-1.3	0.194	1.06
17	-1.745	0.251	-6.94	0	1.02
Month					
1	0				
2	0.1834	0.0273	6.73	0	1.96
3	0.2059	0.0302	6.81	0	1.67
4	0.3994	0.0298	13.4	0	1.7
5	0.3583	0.029	12.37	0	1.78
6	0.2577	0.0267	9.65	0	2.06
7	0.4094	0.0268	15.3	0	2.12
8	0.3243	0.028	11.58	0	1.89
9	0.2366	0.0271	8.74	0	2.27
10	0.1715	0.0261	6.57	0	2.28
11	0.2541	0.0268	9.49	0	2.06
12	0.4117	0.0282	14.61	0	1.86
Day					
1	0				
2	0.0492	0.0172	2.86	0.004	1.58
3	0.1825	0.0164	11.15	0	1.66
4	0.1379	0.0168	8.19	0	1.63
5	0.3452	0.0196	17.65	0	1.42
6	0.698	0.104	6.7	0	1.03
7	0				

Table M.2: Regression summary of binary logistic regression model predicting P(Utilization $>$ 0.922)

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.2736	0.0101	-27.01	0	
Marketing	0.000536	0.000186	2.89	0.004	1.17
N.o. days ahead	0.018494	0.000514	35.99	0	20.7
Current utilization on appointm	0.7977	0.0166	48.06	0	16.36
Utilization during access time	0.1507	0.0206	7.33	0	12.58
Days ahead * Curr util	-0.01393	0.000893	-15.61	0	24.71
Days ahead * Access util	0.0038	0.00117	3.24	0.001	53.92
Regio					
1	0				
2	-0.02205	0.00521	-4.24	0	2.14
3	-0.03954	0.00613	-6.45	0	1.62
4	-0.00094	0.00477	-0.2	0.843	2.65
5	0.05672	0.00469	12.08	0	2.81
6	0.01339	0.00438	3.06	0.002	3.67
7	-0.03532	0.00558	-6.33	0	1.85
8	-0.03345	0.00602	-5.56	0	1.65
9	0.15775	0.00653	24.16	0	1.5
10	0.05254	0.0053	9.91	0	2.05
11	-0.03699	0.0058	-6.38	0	1.73
12	-0.07303	0.00703	-10.39	0	1.41
13	0.03465	0.00762	4.55	0	1.33
14	-0.01189	0.00719	-1.65	0.098	1.38
15	0.2175	0.0142	15.37	0	1.08
16	-0.0343	0.0166	-2.06	0.039	1.06
17	-0.3326	0.0358	-9.3	0	1.02
Month					
1	0				
2	0.03177	0.00483	6.58	0	1.96
3	0.04142	0.00531	7.8	0	1.68
4	0.08344	0.00517	16.13	0	1.75
5	0.07602	0.00507	14.98	0	1.81
6	0.05151	0.00471	10.93	0	2.07
7	0.08324	0.00469	17.74	0	2.15
8	0.06975	0.00494	14.13	0	1.9
9	0.04621	0.00478	9.66	0	2.27
10	0.03457	0.00462	7.49	0	2.29
11	0.05352	0.00473	11.32	0	2.07
12	0.08555	0.00494	17.31	0	1.89
Day					
1	0				
2	0.0095	0.00304	3.12	0.002	1.6
3	0.03617	0.00289	12.52	0	1.68
4	0.02928	0.00297	9.85	0	1.65
5	0.07041	0.00338	20.86	0	1.46
6	0.1358	0.017	7.98	0	1.03
7	0				

Table M.3: Regression summary of multiple linear regression model predicting P(Utilization $>$ 0.922)

Appendix N

2-level full factorial experiments

Table N.1 shows the results of the 2-level full factorial experiment of the 5 continuous predictors in our predictive model. We define the predictors as follows:

- A = Number of days between mailing and appointment date if less than or equal to 30 days, otherwise 0
- B = Number of days scheduling ahead
- C = Current utilization on appointment date
- D = Current average utilization during access time
- E = Working time in 4 weeks preceding appointment date (in minutes)

Term	Effect	Coef
Constant	0.5273	*
A	0.016752	0.008376
B	0.18784	0.09392
C	0.2402	0.1201
D	0.04445	0.02222
A*B	-0.000071	-0.000036
A*C	-0.000240	-0.000120
A*D	-0.000106	-0.000053
B*C	-0.03532	-0.01766
B*D	0.002619	0.001309
C*D	-0.000911	-0.000455
A*B*C	-0.001599	-0.000800
A*B*D	-0.000279	-0.000139
A*C*D	-0.000362	-0.000181
B*C*D	-0.004302	-0.002151
A*B*C*D	0.000050	0.000025

Table N.1: Summary of 2-level full factorial experiment with low = 1^{st} quartile and high = 3^d quartile

Appendix O

Scheduling tool GUI

Figure O.1: Company X proposed scheduling tool front-end

Appendix P

Tool construction

We program the proposed scheduling tool in Python using its de facto standard GUI package Tkinter. Since we cannot link our tool directly to the database yet, we mimic an instance from historical data for easy performance comparison. Also, we let the tool run on data (non-visible data, not input data) of a similar format to what the current wizard uses. This data includes sessions by denoting the location, date and employee and the schedule in JSON format. In this JSON format the keys are slot IDs with as corresponding value 1 or 0. This binary format shows whether the slot is vacant (1) or not (0). This way of storing this information allows quick processing and little storage space. For this reason, and to prevent resistance to implementation, we use this same data format. All other data required to run our model is stored in the same path as the tool. Upon startup we load this data into the tool which takes approximately 2 seconds.

After entering the input information, the tool will provide appointment suggestions upon request. As mentioned before, this process is a 2-stage decision making procedure. The first stage consists of a top-k session selection procedure, whereas the second stage is a complete ranking procedure of all available appointment options in the session selection from the first stage. Figure P.1 shows a schematic representation of this 2-stage decision making procedure. We discuss the construction of both stages separately below.

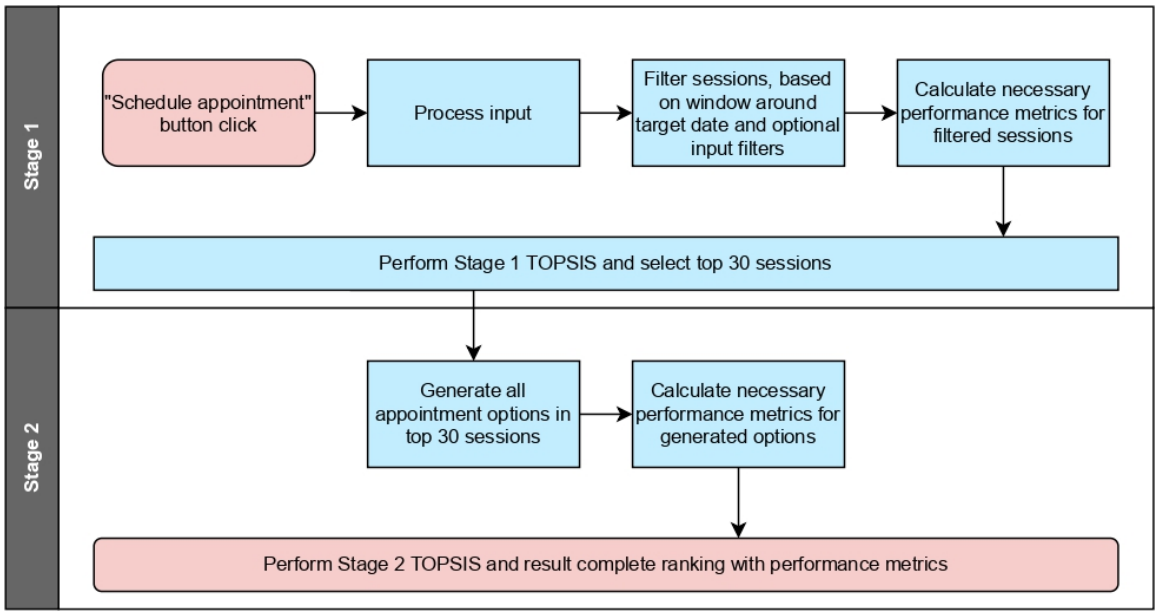

Figure P.1: Block diagram of 2-stage decision making procedure

Stage 1: Top-k session selection

Both stages use a TOPSIS multi-criteria decision making procedure with different criteria and weights. The criteria c_{hij} we choose to use for the Stage 1 TOPSIS procedure are in Table P.1 with the corresponding weights w_{hij}. For explanation on the parameters h, i and j, we refer to section 4.2.2. We introduced the first 3 criteria in Section 2.3 (c_{h11}/c_{h13}) and Section 4.1 (c_{h12}). The tool uses c_{h14}

Required access time $\leq$ 30 days ($h = 1$)			Required access time $>$ 30 days ($h = 2$)		
Stage 1					
j	Criteria c_{11j}	w_{11j}	j	Criteria c_{21j}	w_{21j}
1.	Distance	0.83	1.	Distance	0.83
2.	Predicted $P(utilization > 0.922)$	0.10	2.	Utilization	0.10
3.	Days deviation from target	0.05	3.	Days deviation from target	0.05
4.	Appointment same type as session	0.02	4.	Appointment same type as session	0.02

Table P.1: Stage 1 decision criteria c_{hij} classified by the required access time parameter h, with corresponding weights w_{hij}

to give preference to sessions that have the same type as the appointment, e.g. diabetic appointment in a diabetic session, which is not compulsory (but is preferred). The reason why Section 2.3 does not include this performance measure, is because currently there is no standardized association of appointments to these schedule types, nor are the schedule types standardized. However, we do want to give the scheduler the option to give higher preference to schedules of which we do know have the same type.

We use our crowding prediction model only when the required access time is less than or equal to 30 days. This is because the predictive model is only aimed at predictions up to 30 days ahead. For appointments further into the future we use the current utilization of the schedule as a criterion. When the scheduler selects an access time of 4 weeks, the tool will take sessions between 20 and 36 days ahead into consideration. In this case we assume validity of our model for this extended period and also use the predictive model for sessions between 30 and 36 days ahead. We do this to prevent unfair comparison due to the different decision criteria.

The tool filters all sessions on session date and all optional filters that the scheduler can use in the front-end. Over these sessions it calculates all performance measures needed for the predictive model to work. Then it applies the TOPSIS method to produce the top-k sessions, where we have chosen $k = 30$. By trial and error we saw that this includes all sessions that we feel should be considered for the areas that have the most sessions, thus it is also enough for areas with less sessions. In future research one could consider a different k for areas with more sessions as compared to those with less. However, this would require additional research and methods to indicate when and how to apply this, which is outside the scope of this research.

Stage 2: Appointment options ranking

The criteria we choose to use for stage 2 are in Table P.2. For each of the 30 sessions, the tool generates all possible appointment options. For each of these options it deduces whether the appointment results in a fragmentation index delta (c_{h24}) of -1, 0 or 1. A delta of -1 happens when the appointment directly precedes another appointment or break and directly follows another appointment or break (we lose a gap). When only one of the two is the case the increase is 0 (still a gap). When neither are true the increase is 1 (1 gap is split into 2 gaps). Also, the tool documents for each option the

Required access time $\leq$ 30 days ($h = 1$)			Required access time $>$ 30 days ($h = 2$)		
Stage 2					
j	Criteria c_{12j}	w_{12j}	j	Criteria c_{22j}	w_{22j}
1.	Distance	0.38	1.	Distance	0.39
2.	Predicted $P(utilization > 0.922)$	0.29	2.	Utilization	0.30
3.	Days deviation from target	0.07	3.	Days deviation from target	0.08
4.	Fragmentation index delta	0.11	4.	Fragmentation index delta	0.13
5.	Appointment same type as session	0.10	5.	Appointment same type as session	0.10
6.	Access time violation days [a]	0.05			

[a]Only non-zero when the patient to schedule is NP and access time $>$ 21 days

Table P.2: Stage 2 decision criteria c_{hij} classified by required access time parameter h, with corresponding weights w_{hij}

number of slots still left open and the number of access time violation days. "Access time violation days" is equal to $max\{access\,time - 21\,days, 0\}$ if the patient is an NP and 0 otherwise.

All other criteria are equal to those of their corresponding session in Stage 1. We then apply TOPSIS and write the top 10 appointment options of our ranking to the screen in the scheduling tool. The scheduler can select one of these 10 appointments or choose to see the next 10 options, unless no options are left. When selecting one of the options the selected row indicates how high the probability of above average utilization is by its color (red is high P(utilization $>$ 0.922), yellow is medium, green is low). This way the scheduler can see when the option he/she is selecting is detrimental for overall performance.

For the tool to work with the "index-agenda-dagdata table", which Company X now uses, we need to expand it to hold the inital number of open slots and the current number of open slots (both for utilization). Furthermore, to make the tool work fast, we advise to also put the necessary data corresponding to the schedules (e.g. employee and location) in this same table. Not doing this would significantly slow down the tool. A lot of performance gain is also possible by including the decision criteria in this table and updating it after scheduling an appointment.

Appendix Q

Simulation

The simulation that we use to evaluate the tool uses the instance data to represent the current schedule (i.e. all sessions and their schedules in a JSON format) and a set of appointments to schedule. The simulation iterates over each appointment and uses the background procedure of the proposed tool to schedule them one-by-one. After scheduling one of the appointments, the simulation adds the appointment to the instance before scheduling the next appointment. This results in a realistic development of the schedule. From the solution (ranking of appointment options) of an appointment the simulation randomly picks one of the top 5 appointments (see simplifications). The selected appointment option with corresponding metrics is added to the list of all selected options. From this we can deduce the KPIs from a patient perspective, i.e. distance, NP access time, access time violations and the predicted $P(utilization > 0.922)$. At the end of the simulation we calculate the performance measures from Company X's perspective over all sessions.

It takes 2.4 seconds on average to schedule an appointment, with an upper bound of approximately 6 seconds (obtained by scheduling an appointment with a solution space as large as possible). This means simulating a month of appointments regularly takes between 8 and 10 hours. After implementation at Company X, we expect the tool to be faster. First of all, the database holds a pre-processed field with all available gaps in the schedule, which means we do not have to loop over the entire schedule for a day during the appointment generation in Stage 2. Second, a lot of the performance measures that we need will be calculated in pre-processing instead of during the scheduling procedure. Finally, the implemented tool will be programmed in C#, which is known to be much faster than Python (the programming language that we used for our demonstrational tool). The simulation of the proposed tool has the following simplifications and assumptions:

Simplification 1: Schedule appointments at same session type.
We force the appointment to be scheduled in the same session type as the type they were originally scheduled in. The reason is that Company X has not sufficiently documented which appointments should go where. There are activity types destined for certain schedule types, but they are often overlooked. Therefore, a lot of appointments that were scheduled in regular sessions could also have been scheduled in non-regular schedules. The other way around, some appointments scheduled in non-regular sessions could also have been done in regular sessions. However, we cannot label which appointments we can schedule at both types and which have to stay in the same type.

Simplification 2: Leave out appointments originally scheduled in non-existent sessions.
The instance that the tool uses does not include appointments that were previously scheduled somewhere while there was no session on at that location on that day. The assessment of current performance also did not consider these appointments. These appointments are often artificial.

Simplification 3: Schedule all non-patient related activities first.
To make sure the non-patient related activities can be scheduled in the same place, we first schedule all non-patient related activities. In reality non-patient related activities are scheduled somewhere

where there are no appointments and also very often outside session times. They often take up large blocks of time, sometimes larger than the session itself. Allowing to schedule these activities anywhere would give lots of problems for these reasons and thus we choose to schedule them first in their original place.

Simplification 4: Schedule appointments of certain sessions types at same time and session.
The proposed tool will not influence schedules of the type "House visits", "External" and "Other". Therefore, we schedule appointments that were originally scheduled in these sessions, at the same time and sessions as before.

Simplification 5: We use a simple representation of patient choice in the simulation.
If we do not include patient choice in our simulation and simply pick the top-ranked option, we overestimate the performance of the proposed tool. However, we do not have any information on patient choice at Company X, let alone for the yet to be implemented tool. To still include some form of patient choice, the simulation randomly picks one of the top 5 appointment options. We intuitively chose the number 5, since in the implemented version Company X does not want to show more than 5 appointment options to the patients (they are currently considering a front-end which only shows 3).

Simplification 6: We do not simulate the patient choice in additional filters.
The patient to be scheduled sometimes insists on going to a specific location/practitioner or having an appointment at a specific time. We have no information on how often this happens and which filters would be insisted upon. Thus, we cannot simulate it.

Assumption 1: Difference in effect of patient choice in reality and simulation is not large enough to invalidate the simulation.
We assume that our simplification of patient choice will not make the results of our simulation invalid. To assure this we run our simulation while mimicking the current scheduling procedure. This means selecting sessions based on a distance filter, which is set to 10km by default, and ranking appointment options solely on access time. We show the results in Tables Q.1 and Q.2. We see that the appointments randomly picked from the top 5 options in the current scheduling procedure are very unbalanced. Therefore, schedulers ask the patient without knowledge of the situation where he/she would like to go and apply other myopic filters to get to valid appointments. This explains the major discrepancy between what the current tool shows and what the original real-life performance was. The proposed scheduling procedure takes away this necessity for myopic filtering. Thus, the discrepancy between the performance in simulated performance of the proposed scheduling procedure and the real-life performance will be much smaller.

Context	Utilization	Fragmentation	Variance of utilization
Original	91.5%	1.26	1.20%
Current tool simulation	91.6%	1.70	1.01%
Proposed tool simulation	90.9%	1.15	1.03%

Table Q.1: Results of real-life instance simulation of current scheduling tool, average performance from Company X's perspective in July 2019

Context	NP access time (days)	NP access time violations	RP/PP days deviation from target [a]	Distance (km)
Original	20.7	35.7%	-	3.61
Current tool simulation	12.6	12.1%	-	5.88
Proposed tool simulation	14.8	15.9%	3.36	3.60

[a]Target based on original instance (see assumption 3 in Section 5.2.2)

Table Q.2: Results of real-life instance simulation of current scheduling tool, average performance from a patient's perspective in July 2019

Assumption 2: Appointments must be scheduled at least 30 minutes into the future. We assume that appointments can only be scheduled in the future. Additionally, we only generate appointment options that are at least half an hour later than the time of scheduling.

Assumption 3: The desired access time in weeks for RP/PP appointments is equal to number of days originally scheduled ahead divided by 7, rounded to the nearest integer. For NP appointments, we know that the desired access time is 0 weeks (as soon as possible). For RP/PP appointments this varies per appointment, based on the treatment and diagnosis. Since the desired access time is not documented, we do not know where they should have been scheduled. We therefore assume that the desired access time in weeks is the number of days that the appointment was originally scheduled ahead, divide it by 7, and round it to the nearest integer. Thus, the proposed scheduling tool will try to find a desirable appointment option close to when the RP/PP was originally scheduled.

Appendix R

Patient choice performance implications

Patient choice	NP access time (days)	Access time violations	RP/PP days deviation from target	Distance (km)
1	13.29	11.6%	2.514	3.562
2	14.23	11.7%	3.517	3.309
3	15.14	17.8%	3.690	3.642
4	15.73	19.7%	3.698	3.789
5	15.75	18.9%	3.979	3.698

Table R.1: Patient choice performance implications on KPIs from patient perspective, in real-life instance (July 2019) simulation with proposed scheduling tool

Appendix S

Implementation plan

In this implementation plan we discuss which steps the IT department of Company X should take when implementing the proposed scheduling tool. Note that this implementation plan is specifically aimed at Company X and that it might not be very useful when generalizing the proposed tool to other contexts. We refer to the person/people to whom this implementation plan concerns, as the implementer. We assume the implementer is an experienced programmer and that his/her math competencies are at a high enough level to understand what is written in this implementation plan. The changes that we propose in this implementation plan were discussed with the IT department and approved, though some changes are not directly applicable and might be implemented later. The prototype of the proposed tool is a demonstrational tool programmed in Python. This application can be converted to an executable, but direct integration in the Company X IT system is not possible with this format. Therefore, the researcher consulted with the IT department and agreed that the researcher should present the following information to the IT department, so they can implement the proposed tool in the current IT system (using C#):

- Explanation of database changes needed for implementation
- Procedure in pseudocode
- Python code

In this appendix we discuss the first two points of the above. The pseudocode we provide is aimed at implementation, which means it slightly defers from the demonstrational tool that we programmed in python. A lot of calculation of performance measures happens in pre-processing, so we leave it out of the pseudocode and instead describe the calculation of these performance measures in the Section S.1. We do this because the pseudocode should reflect what the procedure should look like after implementation, not what it looked like in the demonstrational tool. When compiling the Python code, first change the working directory to the folder in which we saved the "Company XSchedulingTool.py" file with the necessary data files.

S.1 Database preparation for implementation

We first state which changes we advise in the data structure for the proposed tool to work as intended, in the list below. Most proposed changes are in "index_agenda_dagdata" and "index_agenda_-tijdblokken". The researcher and the IT department agreed that it would be wise to include the performance measures in the database (wherever possible) and calculate them in pre-processing. After scheduling an appointment the measures would then only need to be recalculated for the corresponding session. This will be beneficial to computation time. From the list below *Data requirement 3* and *Data requirement 6* include data that is not readily available. We elaborate on these data requirements below the list.

The additional required data is:

1. Add initial number of open slots and current number of open slots in "index_agenda_dagdata".
2. Add fragmentation index and current utilization in "index_agenda_dagdata".
3. Add date and region of mailings in the database and add a column in "index_agenda_dagdata" denoting how many days before the session date a mailing occurs in the region, which is equal to 0 if there is no mailing in the region in the month preceding the session date.
4. Add region of the session in "index_agenda_dagdata".
5. Add prediction of P(utilization $>$ 0.922) in "index_agenda_dagdata".
6. Add session type in "index_agenda_dagdata".
7. Add utilization, P(utilization $>$ 0.922) and session type in "index_agenda_tijdblokken" corresponding to "index_agenda_dagdata_id"

Data requirement 1
We want to add *Data requirement 1* to the data table to be able to quickly calculate the utilization in *Data requirement 2*. "index_agenda_dagdata" includes a column with a JSON format representing the open time blocks in the schedule. Upon creation of a session schedule the initial number of open slots should be noted and after scheduling an appointment the current number of open slots should be updated.

Data requirement 2
The utilization in *Data requirement 2* is necessary for our predictive model. The fragmentation is not actually necessary for the proposed tool, as we only take into account the fragmentation delta (-1, 0 or 1), which we calculate in *Data requirement 7*. However, the fragmentation index is good to have for analytic purposes. The utilization can easily be calculated by dividing the current number of open slots with the initial number of open slots. To get the fragmentation index we should evaluate the JSON format in "index_agenda_dagdata", and count how many gaps there are. The current pre-processing procedure already does this while generating the available time blocks to schedule in. By simply keeping count of how many time blocks we generate we have the fragmentation index.

Data requirement 3
As mentioned before, *Data requirement 3* is not yet possible to include, because planned mailing dates and regions are not yet included in the database. We advise Company X to include this in the database so the predictive model can utilize this information. The marketing department stated that they usually know when they will do a mailing more than a month in advance. This can therefore be documented before it happens and used for the predictive model. We saw that the number of appointments in the month following a mailing in that region are affected (Appendix I), which is accounted for in the predictive model in *Data requirement 5*. Should the implementer choose to implement the proposed tool before including this mailing data, the coefficient corresponding to the "mailing in preceding month" should be excluded from the function representing the predictive model, which we discuss in *Data requirement 5*. In this case, we get a slight miscalculation of $P(utilization > 0.922)$ when a mailing occurs. However, the tool does counter this by redirecting appointments when a location's utilization is relatively high, since the predictive model would still give a high $P(utilization > 0.922)$. When mailing dates and regions are available, the implementer should include a column in the "index_agenda_dagdata" table that indicates whether or not there was a mailing in the month preceding the appointment, and if so, how many days before. If there

was no mailing the number should be 0.

Data requirement 4

The implementer needs to add the region in *Data requirement 4*. This is necessary for the predictive model, since some regions are more busy than others. In this research we used the alternative numbering for regions denoted in Table S.1, for more intuitive categorical variables. In the instance data table for the demonstrational tool we had a column with the region numbers per sessions, according to the region number in the model. In the demonstrational tool we obtain the region coefficients for the regression function (predictive model) from local table, extracted from a CSV file. In *Data requirement 5* we elaborate on this. After implementation the proposed tool can either get these coefficients from the database or it can simply be hard-coded in the source code. The implementer should change the region number in this coefficient table to the corresponding regio_id in Table S.1.

Region number in model	regio_id	regio_naam
1	2	Almelo
2	3	Amsterdam
3	5	Drenthe
4	6	Enschede
5	8	Gelderland
6	11	Hengelo
7	15	Rotterdam
8	16	Utrecht
9	17	Vriezenveen
10	18	Wierden
11	20	Polder
12	21	Friesland
13	23	Den Haag
14	32	Leiden
15	34	Zoetermeer
16	35	Brabant
17	36	Groningen

Table S.1: Region number in predictive model with corresponding region IDs in database

Data requirement 5

Most changes in the database structure need to be made to accommodate the regression function calculation in *Data requirement 5*. The IT department and researcher agreed that calculation of all these additional columns (including *Data requirement 5*) would best be done in pre-processing and recalculated after scheduling an appointment. We state the regression equation below.

Regression Equation

$P(1) = exp(Y')/(1 + exp(Y'))$

$Y' = -3.6057 + 0.00248a + 0.08538b + 3.4985c + 0.810d - 0.05175e + 0.00834f + 0.0Region_1 - 0.0794Region_2 - 0.1886Region_3 - 0.0068Region_4 + 0.3025Region_5 + 0.0829Region_6 - 0.1653Region_7 - 0.1688Region_8 + 0.8597Region_9 + 0.2628Region_{10} - 0.1518Region_{11} - 0.3099Region_{12} + 0.1581Region_{13} - 0.0623Region_{14} + 1.296Region_{15} - 0.1165Region_{16} - 1.745Region_{17} + 0.0Month_1 + 0.1834Month_2 + 0.2059Month_3 + 0.3994Month_4 + 0.3583Month_5 + 0.2577Month_6 + 0.4094Month_7 + 0.3243Month_8 +$

$0.2366Month_9 + 0.1715Month_{10} + 0.2541Month_{11} + 0.4117Month_{12} + 0.0Day_1 + 0.0492Day_2 + 0.1825Day_3 + 0.1379Day_4 + 0.3452Day_5 + 0.698Day_6 + 0.0Day_7$

where

$P(1)$ = *Probability of an above average utilization (i.e. utilization > 0.922), also referred to as P(utilization > 0.922)*;
a = *N.o. days from most recent mailing in region as seen from appointment date (so mailing date can be in the future), 0 if more than 30*;
b = *N.o. days scheduling ahead*;
c = *Current utilization on appointment date*;
d = *Utilization on corresponding location during access time*;
$e = b * c$;
$f = b * d$.

As discussed above, the region variables correspond to those in Table S.1 and refer to the corresponding regio_id in the database. How these categorical variables work is, if the region of the session is region 1, then the corresponding variable $Region_1 = 1$ and $Region_i = 0,\ \forall i \neq 1$. The categorical month variables range from $Month_1$, which corresponds to January, to $Month_{12}$, which corresponds to December. The day variables range from Day_1, which corresponds to Monday, to Day_7, which corresponds to Sunday. These categorical variables work in the same way as the region variables. d = *utilization during access time* can be calculated by querying sessions with the corresponding location ID, where the date is between (and including) the scheduling date and the appointment date. From this selection we take the average of the utilization.

If the implementer wants to exclude a variable from the predictive model, this is possible, but it should be kept in mind that excluding variables makes the model less accurate. Variables with high possible absolute values and a high absolute coefficient have a higher impact on the solution and are thus more detrimental to exclude. When excluding variables we strongly advise to retrain the predictive model without the variables.

Data requirement 6
We mentioned already that *Data requirement 6* is not currently possible, since Company X does not yet have standardized schedule types (e.g. diabetics). We recommend Company X to implement standardized schedule types as soon as possible, since it would allow for better classification of schedules. It would also allow filtering on schedule type. We proposed a set of schedule types with classification rules in Section 2.1.1, which Company X can use. If Company X wants to adopt a different set, the implementer should replace the schedule types in the tool with the new set. Same as with mailing dates, omitting the session type prioritization is possible, by setting all session types to "Regulier". In this case the scheduler would make this distinction him/herself. The weights can remain equal, since the value for the same type variable will be constant for each option and have no relative impact.

Data requirement 7
In *Data requirement 7* the implementer should include the utilization, P(utilization > 0.922) and session type in "index_agenda_tijdblokken" corresponding to "index_agenda_dagdata_id". In the earlier data requirements that we discussed, the implementer will have calculated these values. During pre-processing and after scheduling an appointment, these values should also be updated.

S.2 Procedure description

Figure S.1 shows the procedure schematically. We briefly explain the two stages of the procedure, after which we state the corresponding pseudocode for the steps A to F.

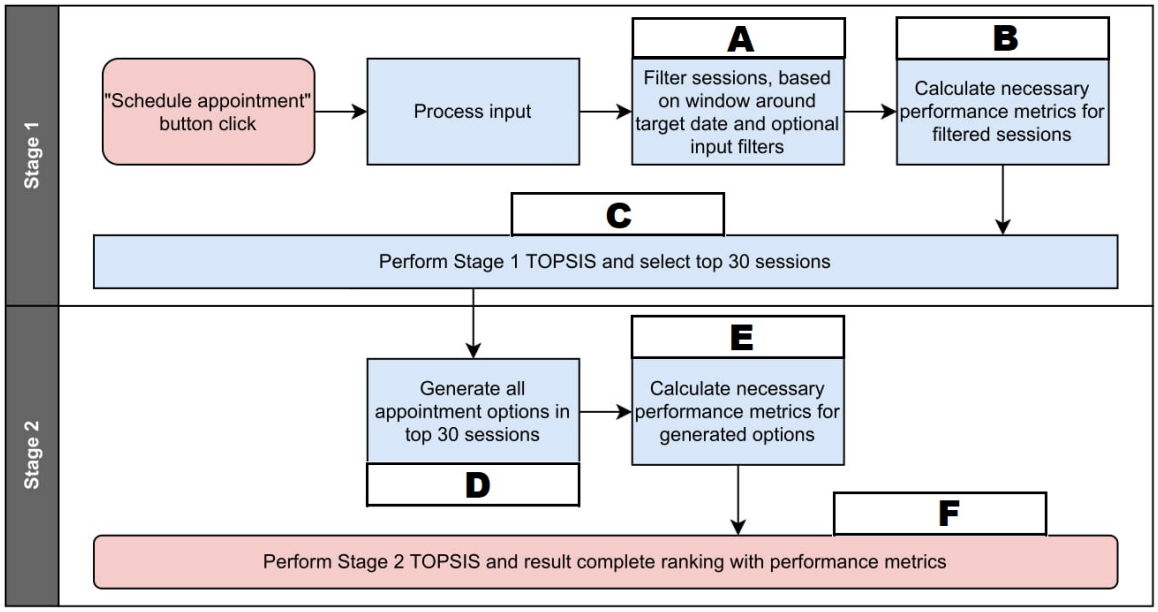

Figure S.1: Block diagram of 2-stage decision making procedure

In Stage 1, we first process the input. The implementer can do this as he/she pleases. The only additional input data, apart from what the current scheduling tool already has, is the required access time and desired schedule type for the appointment to be scheduled. The required access time is the desired number of weeks between scheduling and appointment. Options for this also include today, within 24 hours and as soon as possible (used for NPs). We explain how we handle the required access time input variable in the pseudocode for Step A. The schedule type should be used for the TOPSIS procedures, if available. To see what input variables we have and what formats they take, we refer to the demonstrational tool. After processing the input, the sessions should be filtered to only include sessions in a certain time window around our target date, which is based on the required access time. For these sessions we calculate the necessary metrics (e.g. distance from patient to location). The final step of Stage 1 is to run a TOPSIS procedure over these filtered sessions using the calculated performance measures. We then take the top k sessions and proceed to Stage 2. The implementer can change the number of sessions to select to his/her liking. The researcher experienced that $k = 30$ sessions is a good number to make sure to include the most relevant sessions. However, the IT department at Company X has indicated that they would like to select time blocks in this stage, rather than sessions. This means it might be a good idea to choose a higher number. The data structure of the "index_agenda_tijdblokken" with the proposed changes allows for the use of this table to run the procedure. This way we pick time blocks instead of sessions in Stage 1, based on the same performance measures.

In Stage 2 we start with the best k sessions, or time blocks in case of Company X. From these time blocks we generate all appointment options. Over these appointment options we calculate the necessary performance metrics. The implementer only has to determine the fragmentation index

increase and the access time violation days. We calculate the fragmentation index increase by seeing if the start of the appointment is equal to the start of the time block and if the end of the appointment is equal to the end of the time block. If both are true, the increase is -1 (gap filled), if only one is true, the increase is 0 (still 1 gap), if both are not true, the increase is 1 (1 gap becomes 2 gaps). The access time violation days are always equal to 0 if the patient is not an NP (new patient). When the patient is an NP the variable is equal to $max\{0,\ access\,time - 21\,days\}$. The rest of the performance measures necessary for the TOPSIS procedure can be taken from the corresponding time block values, which are readily available from Stage 1. Finally, we perform our Stage 2 TOPSIS procedure to obtain a full ranking of the appointment options.

In Table S.2 we show the criteria c_{hij} and corresponding weights w_{hij} that we use for the Stage 1 and Stage 2 TOPSIS procedure. Which set of criteria c_{hij} we use depends on parameters h and i. For each combination of h and i, we therefore also have a distinct weight set. We formally define h, i and j as follows:

$$h = \begin{cases} 1, & \text{if } required\,access\,time \leq 30\,days \\ 2, & \text{otherwise} \end{cases}$$

$$i = \begin{cases} 1, & \text{if } tool\,in\,Stage\,1 \\ 2, & \text{if } tool\,in\,Stage\,2 \end{cases}$$

$$j \in \{1, 2, ..., J\},\ \ with\ J = number\ of\ criteria\ used\ for\ the\ given\ combination\ of\ h\ and\ i.$$

$$h = \begin{cases} 1, & \text{if } required\,access\,time \leq 30\,days \\ 2, & \text{otherwise} \end{cases}$$

$$i = \begin{cases} 1, & \text{if } tool\,in\,Stage\,1 \\ 2, & \text{if } tool\,in\,Stage\,2 \end{cases}$$

S.2.1 Procedure steps in pseudocode

Below we show relevant (back-end) procedures steps in pseudocode. We label the pseudocodes A to F, corresponding to the steps in Figure S.1. Some pseudocodes are more detailed then others. We do this because the procedure steps with more detailed pseudocodes require more precise explanation. In procedure steps with less detailed pseudocodes the implementer is free to perform the step according to what philosophy he/she thinks is best. For instance, when calculating the distance from a patient to a location, we use the haversine method, but the IT department already has a method in place for this. Also, some things quite simply do not require as detailed an explanation as others. Throughout the entire procedure we have scheduler's input data available, which is why we omit it from the input per step.

Required access time $\leq$ 30 days ($h = 1$)			Required access time $>$ 30 days ($h = 2$)		
Stage 1					
j	Criteria c_{11j}	w_{11j}	j	Criteria c_{21j}	w_{21j}
1.	Distance	0.83	1.	Distance	0.83
2.	Predicted $P(utilization > 0.922)$	0.10	2.	Utilization	0.10
3.	Days deviation from target	0.05	3.	Days deviation from target	0.05
4.	Appointment same type as session	0.02	4.	Appointment same type as session	0.02
Stage 2					
j	Criteria c_{12j}	w_{12j}	j	Criteria c_{22j}	w_{22j}
1.	Distance	0.30	1.	Distance	0.33
2.	Predicted $P(utilization > 0.922)$	0.33	2.	Utilization	0.37
3.	Days deviation from target	0.08	3.	Days deviation from target	0.10
4.	Fragmentation index delta	0.09	4.	Fragmentation index delta	0.10
5.	Appointment same type as session	0.10	5.	Appointment same type as session	0.10
6.	Access time violation days [a]	0.10			

[a]Only non-zero when the patient to schedule is NP and access time $>$ 21 days

Table S.2: Decision criteria c_{hij} classified by required access time h and stage i, with corresponding weights w_{hij}

Algorithm 1: Pseudocode step A

```
/* Determine the time window for session filtering using the required access
time                                                                    */;
input: "index_agenda_tijdblokken" table ;
if required access time = "Zo snel mogelijk" then
    target number of days ahead := 0;
    minimal number of days ahead := 0;
    maximal number of days ahead := 30;
else if required access time = "Vandaag" then
    target number of days ahead := 0;
    minimal number of days ahead := 0;
    maximal number of days ahead := 0;
else if required access time = "Binnen 24 uur" then
    target number of days ahead := 0;
    minimal number of days ahead := 0;
    maximal number of days ahead := 1;
else if required access time = "1 week" then
    target number of days ahead := 7;
    minimal number of days ahead := target * 0.7;
    maximal number of days ahead := target * 1.6;
else if required access time contains "weken" then
    target number of days ahead := required access time (weeks) * 7;
    minimal number of days ahead := max{target * 0.7, target - 14};
    maximal number of days ahead := max{target * 1.3, target + 14};
else
    Raise error /* Happens when scheduler did not enter the required access time
     */;
end
Filter time blocks on this time window and any additional filters;
Return filtered selection of time blocks
```

Algorithm 2: Pseudocode step B

```
/* Calculate performance measures for each time block                      */;
input: Filtered selection of time blocks (A) and "index_agenda_dagdata" table ;
foreach TimeBlock in Filtered selection of time blocks do
    utilization := Retrieve current utilization of TimeBlock;
    P(utilization > 0.922) := Retrieve predicted P(utilization > 0.922) of TimeBlock;
    days deviation := absolute number of days that the TimeBlock day deviates from the target
     day;
    distance := distance from patient to location of TimeBlock;
    same type := 1 if appointment type is the same as session type of TimeBlock, 0 otherwise;
end
return results
```

Algorithm 3: Pseudocode step C

```
/* Perform TOPSIS, for more detailed information on TOPSIS we refer to the
thesis                                                                    */;
input: Filtered selection of time blocks (A) with corresponding performance measures ;
if target ≤ 30 then
    set criteria and weights corresponding to Stage 1 and Required access time ≤ 30 ;
else
    set criteria and weights corresponding to Stage 1 and Required access time > 30 ;
end
/* First we perform vector normalization                                  */;
foreach criterion in criteria do
    criterion normalization factor := sqrt(sum(each criterion value^2 in A)) ;
end
foreach TimeBlock in A do
    foreach criterion in criteria do
        normalized criterion value := criterion value/criterion normalization factor ;
        /* Now we have normalized values, we apply the weighting          */;
        weighted criterion value := normalized criterion value * criterion weight ;
    end
end
/* We determine the positive (best) and negative (worst) ideal solution (PIS
and NIS). For each criterion a lower value is better, except for the same type
variable (1 if same type, 0 if not), where 1 is the best value and 0 the worst.
A possible PIS is for instance [0.1km, 25%, 0 days, 1 (i.e. appointment type
same as session type)].                                                   */;
PIS := Set of best observed values per criterion in A;
NIS := Set of worst observed values per criterion in A;
/* Now we calculate the distance from each option to the positive and negative
ideal solution using a multi-dimensional Pythagorean theorem              */;
foreach TimeBlock in A do
    distance to PIS :=
      SQRT(SUM((Set of criteria values corresponding to TimeBlock − PIS)^2));
    distance to NIS :=
      SQRT(SUM((Set of criteria values corresponding to TimeBlock − NIS)^2));
    /* We calculate the Closeness Coefficient (CC), which is what we use to rank
    */;
    CC := distance to NIS/(distance to PIS + distance to NIS)
end
Top − k selection := Select top k time blocks based on highest CC
return Top − k selection
```

Algorithm 4: Pseudocode step D+E

```
/* Generate appointment options                                    */;
input: Top − k selection of time blocks (A) with corresponding performance measures ;
initialize empty list of appointment options foreach TimeBlock in A do
    foreach TimeSlot in TimeBlock where begin of TimeSlot to end of TimeBlock is enough
     time for appointment do
        /* fragmentation increase is -1 if begin TimeSlot is begin TimeBlock and
        end TimeSlot is end TimeBlock (gap filled), 0 if one these is true (still
        1 gap), 1 if both are not true (1 gap becomes 2 gaps)          */;
        Calculate fragmentation increase := ;
        /* access time violation days is max{n.o. days TimeBlock is ahead - 21, 0}
        if the patient is an NP (New Patient), 0 otherwise           */;
        Calculate access time violation days ;
        Append appointment option, with these performance measures and those corresponding
         to the TimeBlock, to our appointment option list;
    end
end
return Appointment options list with corresponding performance measures
```

Algorithm 5: Pseudocode step F

```
/* Perform TOPSIS, for more detailed information on TOPSIS we refer to the
thesis                                                                    */;
input: Generated appointment options (A) with corresponding performance measures ;
if target ≤ 30 then
    set criteria and weights corresponding to Stage 2 and Required access time ≤ 30 ;
else
    set criteria and weights corresponding to Stage 2 and Required access time > 30 ;
end
/* First we perform vector normalization                                  */;
foreach criterion in criteria do
    criterion normalization factor := sqrt(sum(each criterion value^2 in A)) ;
end
foreach AppointmentOption in A do
    foreach criterion in criteria do
        normalized criterion value := criterion value/criterion normalization factor ;
        /* Now we have normalized values, we apply the weighting              */;
        weighted criterion value := normalized criterion value * criterion weight ;
    end
end
/* We determine the positive (best) and negative (worst) ideal solution (PIS
and NIS). For each criterion a lower value is better, except for the same type
variable (1 if same type, 0 if not), where 1 is the best value and 0 the worst.
A possible PIS (when target ≤ 30) is for instance [0.1km, 25%, 0 days, -1 (filled
a gap), 1 (appointment type same as session type), 0 days].                */;
PIS := Set of best observed values per criterion in A;
NIS := Set of worst observed values per criterion in A;
/* Now we calculate the distance from each option to the positive and negative
ideal solution using a multi-dimensional Pythagorean theorem               */;
foreach AppointmentOption in A do
    distance to PIS :=
      SQRT(SUM((Set of criteria values corresponding to AppointmentOption − PIS)^2));
    distance to NIS :=
      SQRT(SUM((Set of criteria values corresponding to AppointmentOption − NIS)^2));
    /* We calculate the Closeness Coefficient (CC), which is what we use to rank
    */;
    CC := distance to NIS/(distance to PIS + distance to NIS)
end
sort appointments options based on highest CC
return sorted appointment options
```

Appendix T

Real-life instance simulation with DEA weights

We see that the DEA weightset strongly prefers days deviation from target. This is what results in the decrease in NP access time and thus NP access time violations. We see that the DEA weightset simulation schedules appointments very quickly and thus has more appointments in the month that we evaluate over (July 2019). Therefore the utilization for this month is higher. In Section 5.4 we saw that a higher utilization gives lower fragmentation and lower variance of utilization. We can therefore expect that the fragmentation and variance of the direct weighting method will slightly decrease when the utilization reaches its equilibrium. In consultation with two stakeholders within Company X, we came to the direct weighting weightset. Still, the DEA weightset is a good weightset and might be worth considering. Though, this does rigidly steer towards quick appointments instead of balancing.

Context	Utilization	Fragmentation	Variance of utilization
Direct weighting weightset	90.9%	1.153	1.03%
DEA weightset	92.1%	1.165	0.95%

Table T.1: Results of real-life instance simulation, average performance on KPIs from Company X's perspective, using DEA weights

Context	NP access time (days)	NP access time violations	Predicted $P(Utilization > 0.922)$ [a]	Distance (km)
Direct weighting weightset	14.82	15.9%	57.6%	3.598
DEA weightset	12.73	8.0%	57.1%	3.759

[a] Only appointments with required access time $\leq$ 4 weeks taken into account

Table T.2: Results of real-life instance simulation, average performance on KPIs from a patient's perspective, using DEA weights

Appendix U

Current tool simulation

Tables U.1 and U.2 show the results of running our simulation with a representation of the current scheduling tool. To mimic the current scheduling tool, we select all sessions within 10 km, in the same date ranges as we use for the proposed tool (targets based on original access time and patient type). From these sessions we generate all options, which we sort on lowest deviation from target date. From these options we randomly pick one of the top 5 options, as we did in the other simulations.

What stands out is the imbalance, as the appointment options that the current tool shows have very low access times, but very high distance. As we mentioned in Section 2.2, the fact that the current tool shows options that are not necessarily good, makes that schedulers always ask patients where they want to go, when they want an appointment, etc. The patient however, does not have complete knowledge of the situation, which means they often do not pick the most beneficial appointment for themselves and certainly not for Company X. This is evident from the tables below by looking at the discrepancy in the current tool simulation and the original instance performance.

In the proposed tool the options that we show the scheduler are the options that we believe to be the best for the average patient, while taking into account Company X's preferences. This means that the scheduler does not have to myopically filter the options to get valid options on the screen. They only have to filter when a patient is adamant on a specific filter option (e.g. specific location). Therefore, we can expect the discrepancy between the proposed tool simulation and actual performance to be much smaller than the discrepancy between the current tool simulation and the original performance. We cannot determine exactly how large this discrepancy will be, since this depends on factors that are not currently measured, like patient satisfaction, patient choice and scheduler behaviour.

Context	Utilization	Fragmentation	Variance of utilization
Original performance	91.5%	1.26	1.20%
Proposed tool simulation	90.9%	1.15	1.03%
Current tool simulation	91.6%	1.70	1.01%

Table U.1: Results of real-life instance simulation of current tool, average performance from Company X's perspective in July 2019

Context	NP access time (days)	NP access time violations	RP/PP days deviation from target [a]	Distance (km)
Original performance	20.7	35.7%	-	3.61
Proposed tool simulation	14.8	15.9%	3.36	3.60
Current tool simulation	12.6	12.1%	0.15	5.88

[a] Target based on original instance (see assumption 3 in Section 5.2.2)

Table U.2: Results of real-life instance simulation of current tool, average performance from the patient's perspective in July 2019