# Demand and Capacity Management for Medical Practices

Zur Erlangung des akademischen Grades eines

**Doktors der Ingenieurwissenschaften**

**Dr.-Ing.**

von der KIT-Fakultät für
Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

**DISSERTATION**

von

**Dipl.-Math. Anne Zander**

| | |
|---|---|
| Tag der mündlichen Prüfung: | 14.09.2021 |
| Hauptreferent: | Prof. Dr. Stefan Nickel |
| Korreferent: | Prof. Dr. Peter Vanberkel |

# Abstract

This thesis on tactical demand and capacity management for medical practices consists of four main parts. In the first part, we analyze the general planning and control decisions that need to be taken by a practice manager when opening and then running a medical practice. We further present a best-case data set containing all relevant information on interactions between patient and practice. We compare several real-world appointment data sets to this best-case data set, commenting on the consequences of not collecting specific data. We discuss the fundamental problem of defining model parameters from data and give recommendations for modelers and practitioners to bridge the gap between theory and practice.

In the second part, we present a flexible analytical queueing model to investigate the relationship between the physician's daily capacity, the panel size, and the distribution of indirect waiting times of patients. Essential features of the basic model are the consideration of queue length-dependent parameters such as the appointment request rate, the no-show probability, and the rescheduling probability. We present several extensions to the basic model, including the consideration of queue length-dependent service times. Finally, we investigate the model behavior by conducting extensive numerical experiments.

In the third part, we propose deterministic integer linear programs that decide on the intake of new patients into panels over time, considering the future panel development. Here, we minimize the deviation between the expected panel workload and the physician's capacity over time. We classify panel patients and define transition probabilities from one class to another from one period to the next. Experiments are conducted with parameters based on real-world data. We use the programs to define upper bounds on the number of patients in a patient class to be accepted in a period through solving the programs several times with different demand inputs. When we use those upper bounds in a stochastic discrete-event environment, the expected differences between workload and capacity can be significantly reduced over time, considering several future periods instead of one in the optimization. Using a detailed classification of new patients decreases the expected differences further.

In the last part, we present further integer linear programs to decide on the intake of new patients. For example, we consider several physicians with overlapping panels and capacities as decision variables. Last but not least, we investigate how the queueing model and the panel management programs could be combined.

# Acknowledgements

I started working on my dissertation with just those four words: "appointment planning in healthcare." My advisor, Prof. Dr. Stefan Nickel, felt this was a promising topic with many possible research directions to investigate. And he was right. At first, I was overwhelmed by the already vast body of literature. However, over the years, I found my personal dissertation research niche, deviating slightly from appointment planning to demand and capacity planning, building models, discarding models, extending models until I was finally satisfied with the results. Thank you, Stefan, for indicating this topic to me that I am still passionate about and supporting me to do very independent research over the years.

I further thank Prof. Dr. Peter Vanberkel for his valuable feedback and input, which led to a significant improvement of this work.

I thank Prof. Dr. med. Frank Peters-Klimm, Prof. Dr. med. Martin Schuster and Prof. Dr. Alexander Alscher for arranging contacts with medical practices, providing much-needed data sets, and sharing a lot of practical wisdom and insights with me.

A big thank you goes to all of my current and former colleagues for creating a welcoming and supporting working atmosphere. I especially thank Dr. Fabian Dunke for his helpful comments on the first version of this dissertation and Dr. Melanie Reuter-Oppermann for pushing me to publish our ideas. I further thank the students David Koza, Marlene Dickmann and Grigory Korzhenevich for their support of my research.

I want to thank my family and friends, who all supported me on my journey. A special thanks goes to Ulrich Tietze and Markus Grotz for reading preliminary versions of this work. However, my biggest thank you goes to my partner Maik Bukowczan. He went through the highs and lows of my thesis work with me and always supported me and believed in me. Thank you!

Anne Zander

*Karlsruhe, July 2021*

# Contents

**Abstract** i

**Acknowledgements** iii

**1 Introduction** 1
- 1.1 Planning for medical practices with panels . . . 2
- 1.2 Focus of this thesis . . . 5

**2 Ambulatory care logistics** 9
- 2.1 Introduction . . . 9
- 2.2 Planning and control decisions . . . 10
- 2.3 Literature reviews . . . 13
  - 2.3.1 Location planning . . . 14
  - 2.3.2 Layout planning . . . 16
  - 2.3.3 Service design planning/case-mix planning/capacity planning . . . 17
  - 2.3.4 Personal/workforce planning . . . 18
  - 2.3.5 Panel and capacity management . . . 18
  - 2.3.6 Appointment planning . . . 20
- 2.4 Demand definition and patient-practice interactions . . . 23
- 2.5 Factors that influence patient demand . . . 25
- 2.6 Best-case data set . . . 26
- 2.7 Real-world data sets . . . 31
- 2.8 Defining model parameters from raw data . . . 32
- 2.9 Conclusion and outlook . . . 34

**3 An analytical queueing model to determine the distribution of indirect waiting times** 37
- 3.1 Introduction . . . 37
- 3.2 Literature review . . . 39
- 3.3 Conceptual and mathematical modeling . . . 43
- 3.4 Basic queueing model . . . 47
  - 3.4.1 Queue length distributions . . . 47
  - 3.4.2 Indirect waiting time distribution . . . 63
  - 3.4.3 Arrival process . . . 66
  - 3.4.4 Embedded Markov chain immediately after departures . . . 69

3.4.5 Transient queueing model behavior . . . . . . . . . . . . . . . . . . . 70
3.4.6 Other performance measures . . . . . . . . . . . . . . . . . . . . 70
3.5 Model extensions . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
3.5.1 Variable appointment offerings . . . . . . . . . . . . . . . . . . . 73
3.5.2 Physicians with panel patients . . . . . . . . . . . . . . . . . . . . . 75
3.5.3 Capacity division . . . . . . . . . . . . . . . . . . . . . . . . . . 82

**4 Queuing model implementation and numerical experiments 89**
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
4.2 Real-world data analysis . . . . . . . . . . . . . . . . . . . . . . . . 90
4.3 Model and simulation implementation . . . . . . . . . . . . . . . . . . 93
4.3.1 Simulation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 95
4.3.2 Simulation output analysis . . . . . . . . . . . . . . . . . . . . . . 96
4.4 Numerical experiments . . . . . . . . . . . . . . . . . . . . . . . . . 98
4.4.1 Basic parameter settings . . . . . . . . . . . . . . . . . . . . . . . 99
4.4.2 Runtime of the model implementation . . . . . . . . . . . . . . . . . 102
4.4.3 Verification of the mathematical model and its implementation . . . . . 102
4.4.4 Validation of simplifications in the basic model . . . . . . . . . . . . 106
4.4.5 Approximation of the arrival process . . . . . . . . . . . . . . . . . 107
4.4.6 Basic model results . . . . . . . . . . . . . . . . . . . . . . . . . . 108
4.4.7 Integrating rescheduling arrivals into the total arrival rate . . . . . . . . 119
4.4.8 Results for linear total arrival rates . . . . . . . . . . . . . . . . . . 121
4.4.9 Results for different arrival distributions . . . . . . . . . . . . . . . . 123
4.4.10 Results for variable appointment offerings . . . . . . . . . . . . . . . 123
4.4.11 Results for different constant individual arrival rates for panel patients . 127
4.4.12 Results for a mix of panel and non-panel patients . . . . . . . . . . . 128
4.4.13 Results for fair capacity division between two independent demand streams . . . . . . . . . . . . . . . . . . . . . . . . . . . . 130
4.5 Application in practice . . . . . . . . . . . . . . . . . . . . . . . . . . 131
4.6 Conclusion and outlook . . . . . . . . . . . . . . . . . . . . . . . . . 133

**5 Managing the intake of new patients into a physician panel over time 137**
5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 137
5.2 Literature review . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 139
5.2.1 Panel sizing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 139
5.2.2 Case-mix and workload distribution . . . . . . . . . . . . . . . . . . 141
5.2.3 Continuity of care . . . . . . . . . . . . . . . . . . . . . . . . . . 142
5.2.4 Forecasting of appointment demand . . . . . . . . . . . . . . . . . . 142
5.3 Integer linear programs for panel management . . . . . . . . . . . . . . 142
5.4 Defining model parameters using real-world data . . . . . . . . . . . . . 151
5.5 Numerical experiments . . . . . . . . . . . . . . . . . . . . . . . . . 154
5.5.1 Deterministic setting . . . . . . . . . . . . . . . . . . . . . . . . . 154
5.5.2 Simulation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 158

5.6 Conclusion and outlook . . . . . 165

6 Extensions to the panel management models 167

6.1 Further panel management integer linear programs . . . . . 167

6.2 A mixed integer linear program to define the physicians' capacities . . . . . 169

6.3 Combining the panel management models with the indirect queueing models . . 171

6.3.1 Constant appointment request rate . . . . . 172

6.3.2 Non-constant appointment request rate . . . . . 173

6.4 Appointment data from an urology group practice . . . . . 174

6.5 Selected numerical experiments . . . . . 178

6.5.1 Including the patient attribute gender . . . . . 179

6.5.2 Considering several resources . . . . . 182

6.5.3 Including patient physician assignment . . . . . 183

6.5.4 Minimizing variance . . . . . 185

6.5.5 Defining capacities . . . . . 187

6.6 Conclusion and outlook . . . . . 187

7 Conclusion and outlook 191

A Appendix 195

A.1 Further simulation results for the queueing model . . . . . 195

A.1.1 Realistic simulation for a constant individual arrival rate . . . . . 195

A.1.2 Realistic simulation for a non-constant individual arrival rate . . . . . 196

A.1.3 Results for rescheduling and rejected patients for a non-constant individual arrival rate . . . . . 196

A.1.4 Results for the same-day appointment probability and the physician utilization for a constant individual arrival rate . . . . . 197

A.1.5 Results for the same-day appointment probability and physician utilization for a non-constant individual arrival rate . . . . . 198

A.1.6 Sensitivity analysis for a non-constant individual arrival rate . . . . . 199

A.1.7 Results for using variable appointment offerings for a non-constant individual arrival time . . . . . 200

List of figures 201

List of tables 205

Bibliography 207

# 1 Introduction

Traditionally, health care is focused on the current medical problem of a patient. Depending on the problem, patients are referred to or directly visit a specialized medical facility. Further, treatment is typically divided into inpatient and outpatient care. Inpatients need to seek treatment in hospitals, having to stay at least one night. In contrast, outpatient care is typically administered by medical practices and outpatient clinics. This fragmented health delivery works for well-defined and short-term medical issues. However, the treatment, for example, of complex chronic diseases that involves several medical specialties challenges this paradigm. Therefore, for several years, we see a global trend towards integrated care (Leichsenring 2004). Integrated care signifies a closer collaboration between health care providers, shifting the perspective from a particular medical problem to the whole patient. Here, horizontal integration refers to integration on the same care level. An example would be a group medical practice employing physicians from several specialties. Vertical integration means strengthening the link between different care levels such as primary and secondary care (Gröne et al. 2001). An essential goal of health care integration is to achieve continuity of care, which means seeing the same provider several times for one or several related medical problems. In the best case, this provider knows of the patient's medical history, supervises the patient's treatment spanning care levels being in contact with other care providers, and gathering all relevant medical data (Gröne et al. 2001).

Let us consider the specific example of Germany. In 2003, MVZs (Medical Care Centers) were introduced as an alternative to the traditional single physician practice or group practice of physicians within the same specialty. In an MVZ, many practising doctors from different specialties can be employed and treat patients with statutory health insurance. Also, hospitals can fund MVZs, which allows them participation in ambulatory care (German Federal Ministry of Health 2021b). Further, so-called disease management programs were established. Those programs offer a structured treatment program for certain chronic diseases such as diabetes mellitus type 1 (National Association of Statutory Health Insurance Physicians 2021). Moreover, physicians and hospitals are urged to develop joined care concepts that can then be certified by health insurance providers. Finally, there are special general practitioner contracts where patients decide on one general practitioner who manages their entire care (German Federal Ministry of Health 2021a). In this case, we say that such a patient is part of the general practitioner's panel. More generally, the term physician panel describes the group of patients who regularly visit a specific physician and for which the physician feels responsible for. We conclude that the integrated care movement signifies that office-based physicians, especially general practitioners develop potentially bigger panels and stronger links with their panel patients.

## 1.1 Planning for medical practices with panels

We aim to investigate how the notion of a panel influences planning for medical practices with one or several physicians. On the one hand, having a panel means that we can collect and use data on an individual patient level, storing information on past visits, received services, assigned resources, service times, and more. On the other hand, it also means that we have to make sure that the physician or the physician group can manage the panel demand. To classify the opportunities and challenges for planning with a panel, we start by giving a short overview of planning for physicians and medical practices in general.

Physicians provide health services to patients. To do so, generally, the physician and the patient need to come together at the same place at the same time. Hence, planning for medical practices and physician means to match the supply of provider time with the demand of required service time of patients. More generally, we should include the additional physician and non-physician staff workload produced by patients besides their service time, e.g., administrative work. Planning is typically divided into hierarchical planning levels. On the strategic or long-term level, we decide on the total capacity, e.g., on the number of physicians and non-physician staff, including their working hours. On the tactical or mid-term level, we decide which type of future demand should be served by whom and when, e.g., through developing template schedules. This includes the decision on whether and how to use appointments. On the operational offline or short-term level, we assign the actual patient demand to providers and times based on the rules set at the tactical level. On the operational online level, we react to deviations from the plan (Hulshof et al. 2012).

We face different dimensions of demand uncertainty on the planning levels. On the operational level, we match actual patients and providers with remaining demand uncertainty on the attendance, punctuality, and service time. On the tactical level, we face uncertainty with respect to the actual number of future requests. However, we may match the predicted patient demand to build template schedules. On the strategic level, we may have none or only little information on the future patient demand that we can use to decide on the total provider capacity (Zonderland and Boucherie 2021). However, note that it is a simplification to assume that demand is independent of capacity. In general, capacity decisions also shape the demand and are not meant as a mere reaction to the demand stream.

How does the notion of a panel fit into this planning structure? First, we need to make more decisions, i.e., we need to decide on accepting or rejecting patient requests to enter the panel.

Here, we assume that patients once in the panel can not be thrown out by the physician. However, they can, of course, leave on their account. In Figure 1.1, we give an overview of the different possible patient request modes for a physician with panel patients.

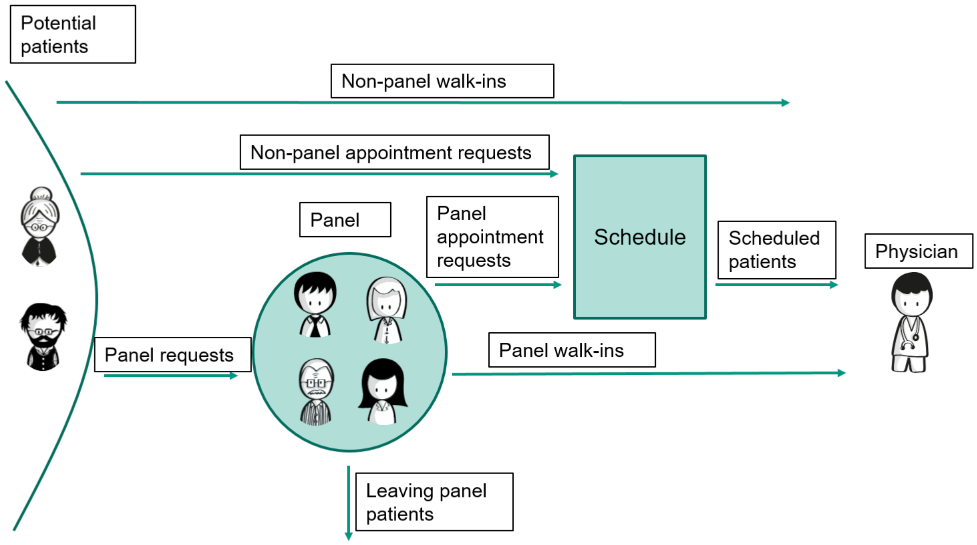

**Figure 1.1:** Patient request modes

The acceptance or rejection decision with respect to requests to enter the panel should be taken online, i.e., at the time of the request. However, the intake of a new panel patient means a long-term care commitment. Therefore, we can and should take potential future requests of panel patients into account for planning. We can do this by predicting the future panel demand based on historical panel data. This knowledge on future panel patient demand reduces uncertainty at the tactical level.

We shortly describe the different tactical planning approaches in the Operations Research literature for physicians and medical practices. Here, we comment on how the notion of a panel comes into play. We will then explain the research focus and contribution of this thesis.

There is a large body of literature on appointment scheduling; see for example Ahmadi-Javid et al. (2017a). There, appointment start times need to be assigned to a given set of patients, often intending to minimize physician idle time, physician overtime, and patient waiting time. Here, we can use individual patient information, e.g., gathered from historical appointment data, to reduce uncertainty in the model and obtain better results.

If we plan for a set of potential future patients, appointment schedules are called template schedules. Those template schedules reserve appointment slots for specific patient types. When a patient requests an appointment, the patient can be assigned to a corresponding open appointment slot. When we develop template schedules, we do not know the actual patients. However, we can use historical appointment data to define patient classes and reserve appointment slots for those classes. Note that using many patient classes reduces uncertainty for the template schedule. However, at the same time, it makes the later patient assignment more difficult. Therefore, some literature also focuses on the performance of template schedules when actual patients are assigned to and served by providers and time slots. Important performance measures that are typically taken into account are the indirect waiting time, also called lead time or access time, and the number of rejected patients. For example, Kuo et al. (2020) use a stochastic mixed-integer linear program together with a simulation to tradeoff scheduling efficiency and accessibility to care.

Another way of handling patient demand is to renounce appointments for future days and therefore also indirect waiting time and serve the patient demand on the day of request, possibly in overtime, see for example Cho and Cattani (2018). Here, it is vital to have some knowledge and control over the patient demand. Managing the panel can also influence the panel patient demand and reduce uncertainty, e.g., making it easier to switch from using appointments to same-day service.

Further literature focuses on matching supply and demand over more extended periods through deciding on provider capacity taking into account indirect waiting time. For example, Nguyen et al. (2018) determine the required number of physicians for an outpatient system with patient reentry to achieve indirect waiting time targets for new and re-visit patients. Here, historical appointment data can be used to estimate the probability of another visit for a given patient.

Note that until now, the (appointment) demand was always assumed as given and unmanageable. Finally, there is literature that again focuses on matching supply and demand over more extended periods supporting demand and capacity decision-making. For example, queueing models can be used to determine the distribution of indirect waiting times given the physician appointment capacity and the patient demand as in Green and Savin (2008). Here, the patient demand can be assumed to originate from a panel with a certain number of patients, i.e., the panel size. Hence, we can use those models to deduce results on appropriate panel sizes.

Instead of focusing on indirect waiting times, one can also assume that physicians work overtime to satisfy all demand. Ozen and Balasubramanian (2013), for example, redesign physician panels in a group practice through minimizing the maximal probability that the daily demand exceeds capacity, achieving a fair distribution of workload. Patients are not assumed to be indistinguishable as in the queueing models, but they can be classified to predict a single patient's demand better.

## 1.2 Focus of this thesis

In this thesis, our focus will be on models that match supply and demand over longer periods, e.g., several months or a year. We manage demand and supply considering performance indicators such as the distribution of indirect waiting times, the proportion of rejected patients, and the deviation between the physician's capacity and the workload produced by the panel in a period. In the following, we will shortly describe our models and contributions.

More precisely, we propose two modeling frameworks in this thesis. The motivation for the first modeling framework is to find the relationship between the physician's daily capacity, the panel size, and the stationary distribution of indirect waiting times of patients. The result and main contribution is a flexible analytical queueing model to represent the appointment backlog that can present many different settings by integrating queue length-dependent parameters such as the appointment request rate, the no-show probability, and the rescheduling probability. We further extend the model by considering queue length-dependent or even random service times and put a special focus on how the appointment request rate of a physician with panel patients can be modeled. A preliminary version of this model was published as:

> Anne Zander. Modeling Indirect Waiting Times with an M/D/1/K/N Queue. In *Proceedings of the Second KSS Research Workshop: Karlsruhe, Germany, February 2016. Ed.: P. Hottum*, volume 69 of *KIT Scientific Working Papers*, pages 110–119. Karlsruher Institut für Technologie (KIT), 2017.

The second modeling framework illustrates how to build integer linear programs (ILPs) to decide on the intake of new patients into panels period-wise over serval periods where we assume fixed physician capacities per period. The main objective is to minimize the sum of deviations between the expected panel workload and the physician's capacity over all considered periods. We also explain how those ILPs can be applied to decide on accepting or rejecting patient requests to enter the panel online while taking the future panel evolution into account. The corresponding chapter on this modeling framework was published as:

> Anne Zander, Stefan Nickel, and Peter Vanberkel. Managing the intake of new patients into a physician panel over time. *European Journal of Operational Research*, 294(1):391–403, 2021.

This framework is then extended by proposing further ILPs that consider capacities of several resources, e.g., physicians and non-physician staff simultaneously and capacities as decision variables. We also build an ILP that considers several overlapping physician panels to support group practices where a patient is assigned to one primary physician but can also be treated by the other physicians. The main contribution is the consideration of the temporal evolution of panel patients to decide on the intake of new patients into existing patient panels. Finally, we comment on how both model frameworks could be brought together. In Figure 1.2, we show the schematic interaction between the panel/demand management, capacity management and the appointment backlog queue.

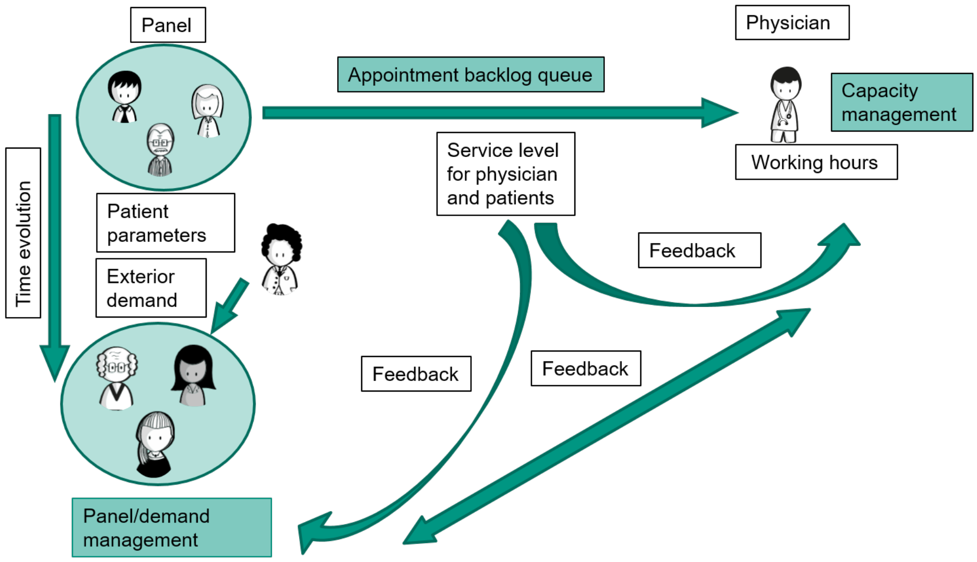

**Figure 1.2:** Schematic interaction between panel/demand management, capacity management and the appointment backlog queue

In this thesis, we further contribute by investigating which kind of data should be collected in the best case to support planning decisions in medical practices and compare it with examples of real-world data sets from several medical practices. We review planning and control decisions for medical practices indicating examples of corresponding recent Operations Research literature. We argue that to bridge the gap between theory and practice, modelers should focus on building models that use (potentially) available data and thoroughly explain the process of model parameters definition.

In fact, our second model framework is an example of such data-based modeling. It was built to be as simple as possible while still harvesting the majority of the optimization potential. Further, we explain how the model parameters can be defined based on real-world data with flexibility concerning the availability of data features. By contrast, the first model framework can be described as a very flexible and complex analytical queueing model that was motivated by a real-world setting. It is complex because it requires many parameters that are difficult to obtain in practice and the model accuracy is higher than it would be necessary for practice. Moreover, it is flexible because it allows representing a range of different settings, one of them being a physician with panel patients. Therefore, it is not easy to apply the queueing model in practice, but it can illustrate general relationships and has a mathematical value of its own.

The remainder of this thesis is structured as follows. Chapter 2 discusses the planning and control decisions that need to be taken by practice managers and examples of corresponding recent Operations Research literature. We illustrate the interactions between patients and medical practice and develop a best-case data set. In Chapter 3, we propose our queueing modeling approach to determine the stationary distribution of indirect waiting times. Chapter 4 showcases the relationships between parameters and performance measures, such as the distribution of indirect waiting times through conducting extensive numerical experiments. Chapter 5 presents a framework to build ILPs to decide on the intake of new patients into a panel over time considering the future panel development. Chapter 6 extends the previous chapter, presenting further ILPs, another real-world data set, and ideas of how to combine the two modeling frameworks. Last but not least, we draw a conclusion and give an outlook on future research in Chapter 7.

Note that, in the following, we will use the pronoun "we" because some of the presented results are joined work and because I like the reader to discover interesting findings together with me.

# 2 Ambulatory care logistics

In this chapter, we describe the planning and control decisions that have to be taken by a medical practice manager. We assign those decisions to planning levels and investigate which already known planning problems from the Operations Research literature can help guide those decisions. We name relevant patient attributes and interaction points between patients and the medical practice. From there, we derive a best-case data set that includes all the relevant attributes and interactions and compare it to several real-world data sets from medical practices. Besides the problem of data availability, we address the problem of defining model parameters from raw data. This chapter is joint work with Melanie Reuter-Oppermann. Her main contributions were the sections on location planning, layout planning, service design planning and workforce planning.

## 2.1 Introduction

In many countries, medical practices are responsible for delivering the majority of primary care and even some secondary and tertiary care. Hence, a considerable leverage to improve care is to support provider decisions to match supply and demand for their ambulatory care facilities. On the one hand, this benefits patients resulting in better access to care and less waiting time. On the other hand, it helps providers establish a reliable working environment avoiding financial loss and extensive overtime, making the medical profession more attractive.

Germany, like most western countries, faces a shortage of general practitioners, especially in rural areas. One cause is the aging population that needs more medical care (Schulz et al. 2016). Further, there are not enough medical students wanting to pursue a career in general medicine. Moreover, those who do, favor bigger cities and group practices that offer a better work-life balance (National Association of Statutory Health Insurance Physicians 2018).

In addition, in Germany, inhabitants can freely choose between medical practices. Unfortunately, this fact makes it more challenging to predict the demand streams, for example, when opening a new practice. However, especially in the case of general practitioners, patients generally see the same physician. In which case, the patient belongs to the panel of the physician. Patient behavior also influences the demand structure, e.g., walking in, not showing up, or being late.

Therefore, to match supply and demand, it is crucial to consider patient behavior. This can only be achieved through collecting the relevant data as a first step.

The main goal of this chapter is to provide decision support for medical practices to match supply and demand with a particular focus on the necessary data. To do so, we first define and characterize supply and demand. For the supply side, we focus on the planning and control decisions that have to be taken by the practice. We then assign those practice decisions to the three main planning levels: strategic (long-term), tactical (mid-term) and operational (short-term). We investigate which planning problems from the Operations Research literature can help guide those decisions and explain which data is necessary to solve the planning problems. For the demand side, we name relevant patient attributes and the interaction points between the patients and the practice. We further present factors that possibly influence patient behavior. From there, we derive the attributes of a potential best-case data set that includes all the relevant attributes and interactions.

Based on real-world appointment data sets from different practices, we show which data is generally available and missing compared to the best-case data set. Based on those findings, we explain which decisions can be supported with this generally available data. We further analyze which data is often missing to solve existing planning problems. Besides the problem of data availability, we address the problem of defining model parameters from raw data. Finally, we indicate future research proposing to investigate potentially relevant relationships between the environment and patient attributes.

This chapter is organized as follows. Section 2.2 lists the planning and control decisions that need to be taken by the practice sorted by planning level. In Section 2.3, a literature review on the relevant planning problems for medical practices is presented. Next, we describe how we define and document demand in Section 2.4. We elaborate on potential influence factors on patient demand behavior in Section 2.5 and present a best-case data set in Section 2.6. We analyze several real-world data sets in Section 2.7 and compare them to the best-case data set. Section 2.8 considers the problem of defining model parameters from raw data. Last but not least, in Section 2.9, we draw a conclusion and give an outlook on possible further research.

## 2.2 Planning and control decisions

As explained in the introduction, we want to provide decision support for matching supply and demand with a focus on data. We define supply as the available provider capacity and include the rules of how to assign those capacities, i.e., supply is defined as the result of the planning and control decisions to design and operate the health care delivery process. Similarly, we define demand as patients' time requirements and also include relevant patient attributes, interactions with the practice, and corresponding patient decisions and behavior. Note, that of course, demand and supply are interconnected. The demand is partly shaped by the supply decisions, and the supply should be designed to take the demand structure into account.

In this section, we mainly reference and apply the wording from the following review articles: Cayirli and Veral (2009), Hulshof et al. (2012) and Ahmadi-Javid et al. (2017a).

Let us shortly explain the main concrete objectives of matching supply and demand. On the operational level, the practice aims to fully utilize its planned capacity, i.e., working without breaks during planned working hours and not working outside of them. Hence, the practice's objectives are low idle time and low overtime. Another objective is the consideration of staff preferences when defining the capacity of resources. Here, the more in advance the capacity can be planed and the less the probability of changes in the plan, the better. Of course, the practice wants to deliver good patient care and therefore takes patients' objectives into account. Patients want access to care, short waiting times, and consideration of their preferences. Here, for patients with appointments, we distinguish between indirect waiting time, i.e., the time elapsed between the request and the actual appointment in days, and the direct waiting time, i.e., the time elapsed between planned appointment start time and actual start time. Further, patients may have preferences for specific physicians, days, and times of the day. Another aspect may be fairness concerning waiting times, i.e., aiming at similar waiting times for different patient groups.

In the following, we list the practice's plan and control decisions sorted by planning level. In general, decisions taken on higher planning levels guide the decisions on lower planning levels. We indicate related Operations Research planning problems in parentheses. We investigate those planning problems in more detail through literature reviews in Section 2.3.

We start with strategic decisions, i.e., long-term decisions. When opening or relocating a practice, one needs to decide on the:

- Location of the practice (location planning),
- Size of the practice (capacity planning, staffing, shift design),
- Design of the practice layout (layout planning), and,
- Target patient groups (case mix planning) and offering of medical treatments (service design planning).

Here, the size of the practice refers to the approximate number of physicians and non-physician staff, as well as rooms, equipment, and the approximate opening hours. If shifts are used, the shift design can already be defined on this level with possible adjustments in the tactical level. The layout of the practice comprises the size and the location of rooms and the assignment of functionalities to those rooms. Target patient groups, for example, may be publicly or privately insured patients. Practices also chose their main medical specialties and thereby also their medical treatments range.

Tactical decisions, i.e., mid-term, decisions are the:

- Fixing of an appointment system (capacity allocation, appointment scheduling),
- Assignment of staff to shifts (rostering), and,
- Capacity and panel management (capacity and panel management).

On the tactical level, the practice needs to decide if and how to handle appointments through defining an appointment system. The practice first decides on an appointment or access policy. In the traditional policy, every (non-urgent) patient needs to book an appointment ahead of time. In the advanced access or open access policy, patients can book appointments on the same (or maybe next-day) only and have to be seen by the physician on that day. A hybrid or carve-out policy lets patients schedule appointments ahead of time while also reserving for same-day demand. Further, the practice needs to decide if walk-ins, i.e., patients that do not give prior notice to their arrival, are allowed. Further, the appointment system defines to whom appointment start times are offered, for what, when, where, and with which physician or non-physician staff member. Also, it defines who is allowed to walk in, for what, when, and to which physician or non-physician staff. Here, we assigned the appointment system decision to the tactical level. However, note that significant changes in the appointment system should not be made regularly as they reduce system reliability and may confuse patients. Hence, some may argue that, for example, the decision on an appointment or access policy should be made on the strategical level.

Aiming at a reliable working environment, the actual working days and times of physicians and non-physician staff should be fixed beforehand. If shifts are used, the physicians and non-physician staff are assigned to the predefined shifts.

In general, patients will visit the same practice (and the same physician) repeatedly. In this case, we say the patient belongs to its panel. The panel size and composition significantly shape the demand. Therefore, it is vital to manage the panel, mainly by deciding who is allowed to enter. However, practices can and should also adjust their capacity to (non-avoidable) changes in demand through adjusting staffing, working hours, and shifts if applicable.

On the operational level, the practice needs to decide on:

- Accepting, rejecting or re-directing requests to walk in,
- Accepting, rejecting or re-directing requests for an appointment,
- Accepting or rejecting requests to join the panel,
- Assigning appointment requests to physicians, days and start times,
- Whom to serve next, and,
- Correcting actions to react to a deviation between plan and realization (also denoted as the operational online planning level).

After establishing an appointment system on the tactical planning level, the practice needs to apply those rules at the operational level deciding how to handle incoming requests. Here, we focus on the online case, meaning that requests have to be answered right away. For example, a walk-in can be admitted to the waiting room, send home, or maybe offered an appointment instead. For patients with appointment requests, the practice needs to assign an appointment start time compliant with their appointment system and fitting to the patient's preferences if possible. If several patients are waiting in the waiting room, the practice needs to decide whom to see next. Good plans are important. However, there are many reasons why the realization may deviate from the plan. In those cases, rules should be established on how to handle those deviations. This includes re-planning of appointments, staff, and resources.

## 2.3 Literature reviews

In this section, we aim to give an overview of Operations Research planning problems used to address the decisions listed in Section 2.2 and not to provide exhaustive literature reviews. The main planning problems are displayed in Figure 2.1. We put a focus on recent developments and shortly describe the necessary data of the mentioned problems. First, we list several review articles that consider relevant planning problems.

Hulshof et al. (2012) present a taxonomic classification of planning decisions in health care. The planning problems addressed in this section mainly coincide with those discussed in the ambulatory care services section. Rais and Viana (2011) present a survey on Operations Research applied to healthcare. Jack and Powers (2009) review articles on demand management, capacity management, and performance in health care. Zonderland and Boucherie (2021) review reviews on patient planning and scheduling, proposing a framework considering planning/hierarchical level, service, and the planning complexity. In this framework, we consider all planning levels in ambulatory care service, focusing on single appointment planning. Further, Zonderland (2021)

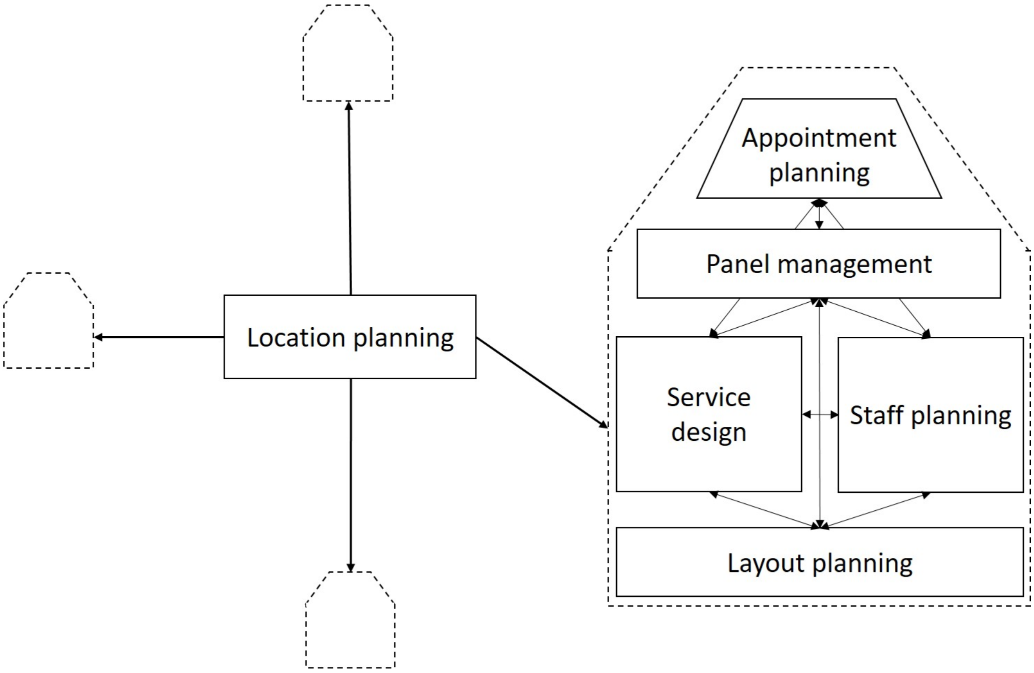

**Figure 2.1:** Illustration of planning problems for ambulatory care practices

reviews outpatient clinic optimization, where the outpatient clinic is part of a hospital. Still, there is much overlap with optimization of independent medical practices.

Note that models proposed in the literature can not always be assigned solely to a single planning level or planning problem since some of them take several decisions on different levels at once. We list such a model under one of the main planning problems addressed in the publication.

### 2.3.1 Location planning

Location planning has been widely studied within Operations Research in general (Drezner and Hamacher 2002) and also for many healthcare applications like locating ambulances (Reuter-Oppermann et al. 2017b), for example. Overviews on healthcare facility location planning can be found in Ahmadi-Javid et al. (2017b), Daskin and Dean (2005) and Güneş and Nickel (2015), for example. Still, fewer publications have addressed the location planning of (primary care) practices. One reason could be that general practitioner (GP) practices are managed individually in many countries, and their locations cannot be coordinated as easily as with other health services. Nevertheless, as GPs are a scarce resource in many countries worldwide, it becomes more and more crucial to place them at efficient locations and inform decision-makers about existing approaches and the possibility of receiving decision support.

Depending on the type of facility and the healthcare system, the importance of the different objectives for the individual problem can vary. Güneş and Nickel (2015) review healthcare facility location models and define the following objectives:

1. minimize the access cost for patients (e.g., travel cost, distance or travel time),
2. maximize covered demand and
3. maximize equity in access.

In the case of preventive care facilities, for example, visits are, in general, rare events, and often facilities only address certain patient types (Gu et al. 2010). On the contrary, many patients might see their GP regularly, and equity of access, as well as low travel costs, are both very important. Then, multicriteria approaches would be of high importance, as reviewed, for example, by Farahani et al. (2010).

The problem of locating primary care facilities is addressed in Abernathy and Hershey (1972), Parker and Srinivasan (1976), Hillsman (1980), Graber-Naidich et al. (2015), Reuter-Oppermann et al. (2019), Güneş et al. (2014), Panagiotis Mitropoulosa (2013), Ahmadi-Javid and Ramshe (2020), Tien and El-Tell (1984) and Hodgson et al. (1998) with varying objectives and constraints. Besides the locations of practices, some publications also address the size of the practices, i.e., the number of physicians working in each practice. In order to make the approaches useful for practice, they should be integrated into a decision support tool combined with a geographic information system (Reuter-Oppermann et al. 2017a). Reuter-Oppermann et al. (2019), for example, located GP practices in a German district using three model variations. These models were developed under two main basic requirements: (1) one practice is to be located that can be reached by as many inhabitants as possible, and (2) locate practices to cut down the driving time for all inhabitants to the next practice location to less than 15 minutes. The input data the authors used in their work included the demand (population), driving times, and the current GP locations.

The designs of primary care services and the underlying healthcare system significantly influence the location planning problem to be solved. The central aspect is whether patients can freely choose their practice or if they are assigned to one based on their home address. Suppose they can freely choose, then modeling their behavior when and where they see a general practitioner is very challenging. Panel information from existing practices would be a valuable input for location planning but is usually very difficult to obtain. Therefore, most publications either make assumptions, for example, based on distance. Others try to ensure spatial accessibility of practices for all inhabitants, as, for example, done by Schuurman et al. (2010). A review on concepts, methods, and challenges for spatial accessibility of primary care was published by Guagliardo (2004), for example.

If inhabitants are assigned to practices, the location problem is a districting problem of healthcare regions. Yanık and Bozkaya (2020) present a review on districting problems in healthcare with dedicated districting models for primary care. The idea is that all patients living within a district are served together. If primary care is centralized and districts are small enough, patients are served by one practice per region, and the districting problem is equivalent to a simple location-allocation problem.

Especially in rural areas or developing countries with a significant GP shortage, potentially low population density, but long distances between villages, mobile practices can be a good option to provide primary care to the inhabitants living in these regions. Then, the location problem is combined with a routing problem, as, for example, proposed by Hodgson et al. (1998). The strategic problem of defining which locations to visit when by mobile practices was studied by Büsing et al. (2021).

Important input parameters for location planning problems include the set of potential locations as well as the patient locations. In addition, a distance matrix between the two location sets must be computed. Assumptions on how and where patients attend a practice must be made. A standard assumption is that patients travel by car and choose the closest practice, while a maximum driving time of, e.g., $15$ minutes is targeted Reuter-Oppermann et al. (2019).

### 2.3.2 Layout planning

In Germany, for example, many single-handed practices are built into standard apartments within apartment buildings, as one reception, one waiting room, and one or two treatment rooms are sufficient to provide the desired level of care for their patients. This is potentially a reason why the literature on layout planning of practices is scarce. When installing group practices with several physicians and nurses that can treat more patients simultaneously, the layout of the practice becomes more important.

For the layout planning of ambulatory care practices, several input parameters are of importance. First of all, it is the number and the size of the rooms, sometimes together with a maximum floor plan size or outer walls. This includes if one doctor uses one treatment room to treat the patients one after another or two or more and moves between the rooms. In general, the aim is to minimize walking distances for physicians, but also nurses and maybe even patients. To model that, pathways are needed as input together with the frequencies. This can be challenging if it is an entirely new practice. Then, processes should be defined first, staff and patient numbers should be known or possible to estimate. If physicians use two rooms, many prefer to have a door between them to switch easily. This would then constraint the order of rooms. Many additional constraints might need to be considered. For example, in primary care practices, physicians might not want to be seen by patients who pick up a prescription to avoid delays in their schedules.

Once a layout is computed, a discrete-event simulation can be used to evaluate the expected walking distances and other indicators efficiently.

Within the healthcare literature, layout planning has been applied to the design of hospitals or hospital departments. Helber et al. (2016) specifically target large and complex hospitals within their research. From the operations management viewpoint, Vos et al. (2007) propose a framework to evaluate hospital designs. Several publications have addressed specific hospital departments, e.g., outpatient clinics (Vahdatzad and Griffin 2016) or emergency departments (Rismanchian and Lee 2017). Arnolds and Nickel (2013) targeted the layout planning of hospital wards considering multiple periods.

Literature on hospital layout planning has been summarised by Jamali et al. (2020). A review on layout planning in healthcare was presented by Benitez et al. (2019) and Arnolds and Nickel (2015), for example, and on facility layout planning in general by Pérez-Gosende et al. (2021).

### 2.3.3 Service design planning/case-mix planning/capacity planning

Before opening a new practice, the practice manager has to decide on the offered services and the corresponding target patient groups. We have only found one publication that addresses these problems for medical practices. One reason might be the difficulty of obtaining a helpful data basis to make those decisions. Another reason may be that practices are often either opened by a single physician or a group of physicians, and their specialties usually define the services that can be offered. The specialties and the fact that the physicians are registered with the statutory health insurance mostly define the target patient groups.

To the best of our knowledge, only Comis et al. (2021) have developed a simulation framework that allows us to model patient behavior and also to analyze the service design of primary care practices.

In some countries, health departments or similar legal structures decide what types of practices can be opened or taken over. In Germany, this is the Association of Statutory Health Insurance Physicians, for example. Still, practices can decide if they want to offer extra care services or dedicated office hours for certain patient or care types.

Service design has a strong interdependence with the practice layout, staff planning, appointment planning, and panel management. Based on the services and care types offered, matching patients can be accepted to the panel until the maximum panel size is reached. This directly relates to the number of physicians and nurses working in the practice as well as the opening hours and consultation times. The number of treatment rooms and the equipment, and the waiting room capacity should be planned accordingly.

Important input for service design planning is the (expected) demand for services, e.g., based on the inhabitants and existing practices in the catchment area. In addition, the interdependence with the other planning problems must be taken into account, e.g., regarding opening hours and the number of staff.

### 2.3.4 Personal/workforce planning

If a practice is not owned by the physicians but by health providers, physicians are usually employed. Then, the number of physicians to be hired must be determined. Then, shifts together with the number of physicians must be defined that match the consultation hours. The same holds for the medical assistants or nurses.

Overall, the following workforce planning problems have to be considered:

- staffing: number of physicians, number of medical assistants or nurses,
- shift design: number, length, and distribution of shifts within the week,
- rostering: assignment of shifts to staff,
- re-planning: re-assignment due to disruptions, e.g., staff sickness,

All planning problems have been addressed in the literature to a great extent for hospital nurses and physicians. In contrast, publications on ambulatory care workforce planning are scarce.

Stiglic and Kokol (2005) propose an approach for nurse scheduling in ambulatory care centers that provide primary care for patients in Slovenia. Many review papers have been published, including reviews on nurse rostering (Burke et al. 2004, Cheang et al. 2003), on, physician scheduling (Erhard et al. 2018), on general staff scheduling and rostering (Ernst et al. 2004) as well as workforce planning (De Bruecker et al. 2015).

Input parameters include the opening hours and the expected number of patients to be treated each day together with expected treatment durations.

### 2.3.5 Panel and capacity management

Many office-based physicians have a so-called panel, i.e., a group of patients who regularly visits the physician. If a big part of the demand comes from panel patients, it is essential to manage the panel to better control and forecast demand. How can we measure the workload produced by a panel? First, we need to define who belongs to the panel and who does not. Next, we need to measure the workload produced by single panel patients, i.e., number of visits, total time spent by the physician (and non-physician staff) to care, and organize the patient's care during a fixed period. Based on this historical data, we try to forecast the future panel demand and use this information to decide on accepting or rejecting patient requests to enter the panel.

The most straightforward approach to measure the panel workload is to determine the panel size, i.e., to count the number of panel patients. However, there is no general standard on which patients to count. Some consider the patients seen in the last two years (Margolius et al. 2018, Marx et al. 2011) others the number of patients seen in the last 18 months (Raffoul et al. 2016, Murray and Berwick 2003, Murray et al. 2007). Given a time frame, we determine the average workload per panel patient. Then, the question is, how to determine the maximal panel size given a physician's capacity. Murray et al. (2007) propose to divide the physician's capacity in a given period by the average time required for a single panel patient in that period. However, this approach completely ignores uncertainty in the problem structure.

To account for stochastic demand and sometimes also stochastic service times, several queuing models have been suggested. Green and Savin (2008) propose two queueing models to determine the relationship between panel size and the expected appointment backlog taking no-shows and rescheduling of no-shows into account. Zander (2017) extends their approach by including backlog-dependent request rate of panel patients. Liu and Ziya (2014) also present two queueing models and decide on the panel size and the service capacity to maximize the long-term average reward while constraining the expected access time. Izady (2015) uses several discrete-time queueing models with bulk service. In contrast, Zacharias and Armony (2017) consider the appointment backlog together with direct waiting time and also decide on the panel size and the service capacity to maximize the long-term average reward. Finally, Vanberkel et al. (2018) use a queueing network of multi-server queues to define the panel size of an oncology practice that balances demand from new patients and relapsed patients. The following information is necessary to apply the queueing models: panel size, appointment request rate, daily appointment capacity, service time (distribution), no-show probability (possibly dependent on the indirect waiting time), rescheduling probability of no-shows, and rewards and costs.

Patients have very different needs for medical attention. Therefore, it is reasonable to classify panel patients with similar attributes and needs instead of counting panel patients. Balasubramanian et al. (2010) propose a stochastic linear program to reassign patients to primary care physicians of a group practice with the objectives to minimize access time and to improve continuity of care. Here, they use a patient classification based on age and gender. Ozen and Balasubramanian (2013) minimize the maximal probability that the daily demand exceeds capacity in a group practice of primary care physicians to redesign physician panels. Finally, Zander et al. (2021) consider patient classification and take the future evolution of the panel into account. They propose deterministic integer linear programs that decide on the intake of new patients into panels with the primary objective to minimize the deviation between the expected panel workload and the physician's capacity over time.

For the models mentioned in this paragraph, the following information is necessary: panel composition, e.g., including patient information on age, gender, visit history, assigned physician; distribution for the number of visits in future periods per patient classes and expected future demand to join a panel.

Finally, we give an example of capacity planning/management without using the concept of a panel. Nguyen et al. (2018) present a capacity planning model to determine the required number of physicians for an outpatient system with patient reentry.

### 2.3.6 Appointment planning

In this section, we review planning problems that relate to the handling of appointments. Note that appointments are mainly considered for the primary resources, i.e., the physicians. This simplification assumes that the non-physician tasks corresponding to the appointment/patient-physician contact are performed along the way. For medical practices, we further focus on single appointments and predefined slots/appointment start times that can be assigned to patients. Further note that there is a significant overlap between the appointment planning and capacity planning/management literature since both problems are often addressed simultaneously.

Recently, Ahmadi-Javid et al. (2017a) reviewed optimization studies in outpatient appointment systems in healthcare. Further, reviews on appointment scheduling in healthcare are presented in Cayirli and Veral (2009) and Gupta and Denton (2008).

Ideally, before making decisions, we have some knowledge of the demand. Considering a practice with one or several panels, we may have some idea about the total patient demand over a longer period and are confronted with uncertainty about actual patient requests. Klute et al. (2019) predict outpatient appointment requests using machine learning and traditional models, concluding that one should test a variety of traditional, machine learning, and hybrid prediction methods to find the best one for a given data set.

As explained in Section 2.2, the practice manager has to decide if to work with appointments and, if yes, on an appointment system. This includes deciding on an access policy, on capacity allocation, i.e., which patient type is allowed to be seen by which physician (see also the section of panel and capacity management), and appointment scheduling. Scheduling appointments give the physician some control over the demand and help to reduce the uncertainty of incoming requests. However, if appointments are offered for future days, patients experience longer access times, and the practice is confronted with uncertainty due to no-shows, an issue that is well studied in the literature. Dantas et al. (2018) present an extensive literature review on no-shows in appointment scheduling, showing that the main determinants of no-shows are high lead time and prior no-show history. Other sources of uncertainty and possible attributes to classify patients often considered in the literature are: cancellations, walk-ins, unpunctual patients, and stochastic service times. Patients can be further distinguished by their time and physician preferences.

Considering access, the literature generally distinguishes between urgent and non-urgent requests. An urgent request should generally be treated on the day of the request, whereas non-urgent requests can also be treated on future days. Hence, the access policy and the walk-in rules should further indicate if urgent patients can walk in and if non-urgent patients can book a same-day appointment or even walk in. Cho and Cattani (2018) compare the traditional policy with the open access policy and conclude that physicians can serve more patients in the open access policy but maybe prefer the traditional policy because it allows them more control over the daily patient demand.

In appointment scheduling, we typically create a session template schedule for a physician by defining appointment slots. Later on, we assign actual patients to those predefined appointment slots upon their request. In the template schedule, we can differentiate between several patient classes with their individual parameters to reduce overtime, idle time, and patient waiting time. Note, however, that using many patient classes may lead to problems in assigning actual patients to slots resulting in higher access times or idle slots. In general, we use the same or similar template schedules every day. To counter the negative effects of no-shows, over-booking appointment slots and a maximal booking window may be used. For example, Leeftink et al. (2021) use an analytical queueing model with time-dependent no-show and cancellation rates to determine the optimal booking horizon to minimize the effects of no-shows and cancellations and the cost of rejecting patients.

In the following, we describe publications presenting models optimizing access time through deciding on a template schedule with and without consideration of capacity adjustments. Schacht (2018) uses a stochastic mixed-integer linear program to determine an appointment scheduling template that defines the allocation of walk-in and pre-scheduled appointment blocks in combination with the allocation of daily service times. They investigate when and how the weekly template schedule should be reconfigured to manage stochastic and seasonal patient load throughout the year to ensure access to care. Laan et al. (2018) develop a static and a dynamic appointment schedule with consideration of time-varying demand and capacity. Kuo et al. (2020) consider direct and indirect waiting time, i.e., access time, together, using a scenario-based stochastic mixed-integer linear program to define a template schedule and an inter-session simulation model to handle the actual appointment booking.

The majority of publications in appointment planning are on scheduling a set of given patients in one session, focusing on balancing overtime, idle time, and patient waiting time while possibly considering no-shows, walk-ins, unpunctual patients, stochastic service times, and lateness and interruption of the physician. Srinivas and Ravindran (2018) propose to use machine learning to classify patients with respect to their no-show probability. They then use this classification and a service duration classification to propose appointment scheduling rules. Using real-world test data in a simulation, their scheduling rules outperform the considered benchmark rules for all the clinic settings tested. Zacharias and Yunes (2020) design appointment schedules considering no-shows, non-punctuality, general stochastic service times, and unscheduled emergency walk-ins. In the case of punctual patients, they develop and implement an algorithm for globally minimizing a multimodular function over non-negative integer vectors in polynomial time. Kuiper et al. (2021b) calculate optimal appointment schedules considering no-shows and walk-ins and assuming phase-type distributed service times.

We now review some models that schedule actual patients. Feldman et al. (2014) present a static and a dynamic model to maximize the expected daily profit through deciding on a set of days to offer to patients for appointment booking. Patients have time preferences and may cancel or no-show. Zander and Mohring (2016) present a mixed-integer linear programming to determine a set of appointment start times to offer to an appointment requesting patient on a particular day. They take patient type-specific service times and time preferences into account. The aim is to achieve a full schedule by offering fitting appointments while taking future demand into account. Liu et al. (2019) consider non-sequential (one offer) and sequential (potentially several offers) appointment offerings to maximize the number of booked appointment slots where patients are assumed to have unknown time preferences. The authors show that sequential offering can significantly improve performance compared to non-sequential offering.

Finally, we give an example of a model supporting decisions on an operational online level. Samorani and Ganguly (2016) investigate the wait-preempt dilemma by building an analytical model that determines when a physician should wait to see a scheduled patient (who is late) or see an early patient right away.

The presented models in this section usually need some of the following parameter inputs: distribution of daily requests; the number of patients scheduled per day; service time distributions (possibly patient dependent); punctuality distributions (possibly patient dependent); cancellation and no-show probabilities (possibly patient, slot, or access time-dependent); walk-in probabilities (possibly slot dependent); costs for over time, idle time and patient waiting time; rewards for treating patients, and time and physician preferences of patients.

## 2.4 Demand definition and patient-practice interactions

In this section, we describe the patient-practice interactions that influence or define the demand. All those interactions should be documented such that the resulting data set can be used to guide future decisions.

Depending on the chosen appointment policy, there exist different contact points between patient and practice. In the most complex case, i.e., considering panel patients and appointments, we can differentiate between three main contact points. First, a new patient, who has not been seen by the practice before, contacts the practice to possibly join a panel. Next, a (panel) patient requests an appointment, and finally, the patient arrives for the appointment.

There may be further contact points if the patient announces to leave the panel, cancels the appointment, or leaves the practice before having been seen. In the following, we will sketch the patient behavior patterns along the process from the first contact with the practice until the realization of the appointment.

In general, a new patient looking for a first consultation or a (panel) patient with a request will do one of the following:

- Tries to book an appointment in person, via telephone or online (here booking of several appointments at once is also possible),
- Contacts the practice to inquire about the possibility to walk-in,
- Plans to walk into the practice without previous contact.

Next:

- No agreement is made on an appointment, or the possibility to walk-in (balking) and the patient seeks treatment elsewhere, tries again later, or stops seeking treatment,
- The patient books an appointment (this could also be a re-directed walk-in),
- The patient walks in.

If the patient decides to walk in:

- The patient is rejected as a walk-in and leaves,
- The patient waits for treatment and leaves before receiving treatment (reneging),
- The patient waits and is then treated.

If the patient decides to book an appointment:

- The patient cancels the appointment (reneging),
- The patient does not show up,
- The patient shows up.

If the patient shows up:

- The patient is early,
- The patient is late,
- The patient is on time.

Again, if the patient shows up:

- The patient leaves before receiving treatment (reneging),
- The patient waits and is then treated

After treatment, the patient might directly book a new appointment. Further, patients who started the process of seeking the treatment but dropped out along the way may contact again to start the process anew.

In Figure 2.2, we illustrate possible patient demand behavior. Note that patients dropping out during their quest for treatment/consultation can enter the system anew.

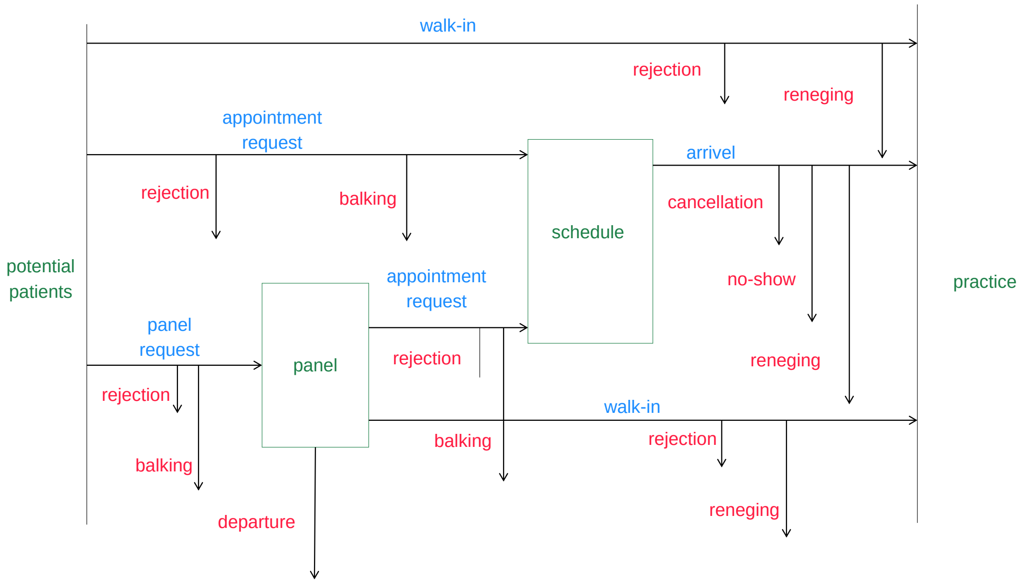

**Figure 2.2:** Possible patient demand behavior

## 2.5 Factors that influence patient demand

How patients behave along the interactions points described in Section 2.4 and how the practice reacts dependents on the appointment policy, patients attributes and possibly environmental factors. Here, we focus on aspects that the practice can know, observe and document.

A new patient will decide to seek treatments in a practice based on the publicly available information, the information received by contacting the practice, and personal recommendations. Therefore, to attract new patients, it is important to provide publicly available information, to be reachable by phone, and to build a generally good reputation. The following supply aspects may influence new patients to come for a first consultation:

- Location of the practice,
- Service offerings,
- Practice accepts new patients,
- Practice accepts publicly insured patients,
- Opening hours and consultation hours of the practice,
- Appointment booking modes (online or telephone),
- Appointment booking regulations,
- Access time/indirect waiting time,
- Availability of preferred time and physician for the first appointment,
- Waiting times in the practice for patients with appointments,
- Walk-in regulations,
- Waiting time in the practice for walk-in patients.

A new patient will decide to join the panel or not based on her overall service experience.

For a panel patient, the decision to walk in or to book an appointment may be based on the following supply aspects:

- Access time/indirect waiting time,
- Availability of preferred appointment time,
- Availability of preferred physician,

- Direct waiting time for patients with appointments,
- Waiting time for walk-ins.

Some of the mentioned factors that may influence patients' decisions can be controlled directly through the supply decisions, for example, the practice location or the opening hours. Other factors are semi-controllable because they depend not only on the supply decision but also on the demand and are therefore stochastic in nature, such as the (distribution of) indirect waiting time. Then, there are external factors that can not be controlled.

External factors that may influence the demand pattern may be:

- The season,
- The day of the week,
- The type of day (e.g., workday, holiday),
- Time of the day of the appointment,
- The weather.

Given the appointment policy, the practice may decide to accept or reject patients' requests to join the panel, to book an appointment or walk in based on the following patient attributes:

- Patient type (panel patient, new patient),
- Reason for the request (urgent problem, first consultation, other reasons),
- Attendance history of the patient,
- Insurance of the patient,
- Other patient attributes (health status, co-morbidities, age, etc.).

## 2.6 Best-case data set

Based on the previous sections, we give an extensive overview of potentially relevant data that should be documented if possible. Here, we structure the data by going over the different persons and objects separately.

We start with data that is related to a specific patient:

- Name/id,
- Birth date,

- Gender,
- Address,
- Insurance types,
- Assigned physician if any.

If some of those patient attributes change over time, the changes and the time of the changes together with the previous values should be documented.

If a new patient requests to join a panel, a panel patient requests to change panels or to leave the panel, we store the following data

- Patient,
- Day of request,
- Time of request,
- Corresponding appointment or walk-in event,
- Requested physician if any,
- If the patient does not join a panel, stays in the panel, or leaves the panel, what are the reasons?
- Assigned physician, if any.

Next, we list data that is related to an appointment request:

- Patient,
- Day of request,
- Time of request,
- Request mode (e.g., online, telephone, etc.),
- Times requested/offered (time of day, start time of appointment) if applicable,
- Services requested/offered (e.g., first consultation, specific treatment),
- Resources requested/offered (e.g., physician, non-physician staff),
- If no agreement is made, what are the reasons?
- Booked appointments, if there are any.

We now list data relevant for a booked appointment:

- Patient,
- Appointment request,
- Day of the booking,
- Time of the booking,
- Booking mode (e.g., online, telephone, etc.),
- Type (e.g., consultation, check-up, surgery, etc.),
- Planned day,
- Planned starting time,
- Planned duration,
- Planned services,
- Planned resources (e.g., physician, non-physician staff, equipment, room etc.),
- Cancellation (yes/no),
- No-show (yes/ no).

Instead of using a starting time and a duration, we can also note the booking of predefined time slots.

If the patient or practice cancels the appointment, we further consider:

- Cancellation day,
- Cancellation time,
- Cancellation reason,
- Cancellation unit (patient or practice),
- (New) booked appointment(s) if any.

Any change in a booked appointment can be documented as a cancellation together with the new appointment booked. This way, information on previous planning is saved.

If the patient shows up, we further consider:

- Patient arrival time,
- Reneging time in case of reneging,

- Start time,
- Duration,
- Services,
- Resources (e.g., physician, non-physician staff, equipment, room etc.).
- Notes.

If a patient walks in, we store the following data for this event:

- Patient,
- Patient arrival day,
- Patient arrival time,
- Booked appointment if any (in case of rejection),
- Reneging time (in case of reneging),
- Start time,
- Duration,
- Type (e.g., consultation, urgent problem, etc.),
- Services,
- Resources (e.g., physician, non-physician staff, equipment, room etc.).
- Notes.

To book appointments, we need to connect the patient with services, resources, and a booking time. Here we start with data related to a service:

- Service name,
- Resource requirements,
- Time requirements.

Appointments are booked ahead of time. Therefore, we need data on the future (planned) availability of resources and time slots. The following data is resource related:

- Resource name/id (e.g., physician, non-physician staff, equipment, room etc.),
- Time availability,

- Possible services,
- Possible non-patient-related work.

Any data changes before realization should be documented, storing the former and the new values together with the times those changes happened. After realization, the actual working time of resources should be documented. Note that the working time of a resource may include work that cannot be assigned to a specific patient. This type of work could be planned beforehand as well. However, here, we do not consider this explicitly.

For a combination of a resource and a time slot, we consider:

- Day of time slot,
- Start time of time slot,
- End time of time slot
- Resource,
- Possible booking modes (e.g., online appointment, walk-in slot, etc.),
- Possible patient types (e.g., panel patients, patients with urgent requests, etc.),
- Possible services,
- Start (and possibly end) time of booking availability if applicable.

Here, a time slot is considered to be the greatest common factor of all planned service durations. It can be part of an appointment or a walk-in event. We do not explicitly define it here, but of course, we can have further dependencies between time slots, booking modes, patient types, and services, e.g., a panel patient can only book service A via booking an online appointment. As explained before, resource and resource-time slot data should be expanded by a time component such that any changes can be retraced. This way, we know, for example, when a physician changed time availability due to a planned vacation. The realization of a resource and time slot usage can be reconstructed using the data on realized appointments or walk-in sessions. Note that a time slot might not be used entirely to serve one patient by one resource in the realization.

Of course, we can combine the different data sources to produce aggregated reports. For example, we can deduce the following (panel) patient-centered information:

- Distribution and type of requests,
- Distribution and type of visits,
- Administered services,

- Service duration (distributions),
- Distribution of cancellations and possibly following new requests,
- Distribution of missed appointments and possibly following new requests,
- Time preferences/availability,
- Physicians seen,
- Punctuality (distribution).

This patient-centered data could then be used to forecast future requests and visit behavior of that patient. Similarly, one could combine and aggregate the different data sources to come up with resource or resource-time-slot-centered data.

## 2.7 Real-world data sets

We were able to receive four (appointment) data sets. The data sets are from group practices of general practitioners (GP), urologists (U), cardiologists (C), and otolaryngologists (O) in Germany. The urology and cardiology practices use the same administration software. For three out of four data sets, a row describes an appointment. Sometimes those appointments can be walk-ins. However, this can only be identified by comparing the booking time and the planned start time of the appointment. However, a row in the data set from the general practitioner practice relates to a charged service for the patient. A subset of those rows relate to an actual patient-physician contact. In Table 2.1, we show the main appointment attributes and if they are available in the four data sets. Here, "X" stands for available, "(X)" stands for partly available, and "-" for not available.

We first notice that the most important information, i.e., the information of the day and planned start time of the appointment, the corresponding patient, services, and resources are available in all four data sets. This information is necessary for accounting, i.e., to receive payment from the insurance companies. However, the cardiology practice does not assign a patient ID to every patient, making it difficult to recognize patients that revisit. Three out of four practices document the day and time of the booking. No data is stored for any kind of rejected, balking, or reneging patients in all four cases. Another problem is that planned data entries are overwritten by the realized values, e.g., services and resources. Further, if an appointment is canceled, the corresponding data row is either deleted or overwritten such that it contains values for a new appointment. The urology and cardiology practices include a column that shows the day and time when the appointment was last changed. However, any previous values are lost. We also see that no data is stored about the patient arrival times and actual service duration.

| Data | GP | U | C | O |
|---|---|---|---|---|
| Patient id | X | X | (X) | X |
| Booking day and time | - | X | X | X |
| Booking mode | - | - | - | X |
| Day and planned start time | X | X | X | X |
| Planned duration | - | X | X | X |
| Planned services | - | - | - | - |
| Planned resources | - | - | - | - |
| Cancellation day and time | - | (X) | (X) | - |
| No-show | - | X | X | X |
| Arrival time | - | - | - | - |
| Start time | - | - | - | - |
| Duration | - | - | - | - |
| Services | X | X | X | X |
| Resources | X | X | X | X |

**Table 2.1:** Available information for booked appointments in the four data sets

The consequences of collecting the data the way those four practices do it are that the original demand (including rejection, balking, reneging, and preferences) can not be determined. It can not be determined if a slot stayed idle due to a late cancellation or due to low demand. Patients' unpunctuality cannot be included. Further, information on the actual service duration is not available and can therefore not be used to create better schedules in the future. For the general practitioner practice, no information of indirect waiting time can be deduced since the day and time of booking are not documented.

## 2.8 Defining model parameters from raw data

There is vast Operations Research literature on planning and control decisions for outpatient clinics. However, there are still few reports that show the actual application of the presented methods in practice. There are numerous reasons why methods are not applied. For example, Kuiper et al. (2021a) conduct interviews with ten outpatient clinics. They find that the appointment slots length used in those clinics is never based on data. The trade-off between idle time and waiting time widely used in theory is not recognized or applied. On the contrary, idle time is seen as a minor issue by clinicians because they can switch to other work. Still, in general, the clinics use tight schedules.

Besides numerous other requirements, one condition must be fulfilled to apply any method to solve a planning and control problem; namely, we need to define the model parameters. To this end, we first need (historical) data and second instructions on how to derive the model parameters from this data. We saw in Section 2.7 that data availability is already a big problem. Especially for strategic problems that need to be solved when opening a new practice, there is no self-collected data available. Then, decision-makers need to rely on publicly available data. Further, better decisions could probably be achieved if one had access to collected data from practices with similar features.

Modelers should also consider the second point. In case of available relevant real-world data of sufficient quality, the parameter definition process should be illustrated together with the model's description. However, when this kind of data is not available, we recommend that authors describe the necessary data and explain the derivation of the potential parameters to prove that their model can be applied in practice. This can be done using the potentially available data from Section 2.6. It would even be better if authors additionally explained how the parameters could be updated automatically over time given a stream of new data input.

Of course, not all models can be applied to every practice setting. In general, models can not capture every aspect of the real-world problem. Therefore, transforming data into parameters always entails aggregation and simplification. Sometimes this process would remove important features of the settings, probably delivering poor results. For example, it is not advisable to apply a scheduling model which does not consider no-shows to a setting with a high proportion of no-shows. However, as we have seen, it can also be the other way around. If the practice does not collect the relevant data, it can only apply models working with the provided data basis, which may also lead to poor results.

Let us consider an example to illustrate the difficulties of the parameter definition process. For example, queues are often used to model the appointment backlog (see Panel and Capacity Management in Section 2.3). There, we already make many assumptions that often do not hold in practice, e.g., that patients book the next available appointment, a time-independent request rate, and equal capacity per day. Next, we most likely are confronted with a data basis where (rejected) appointment requests are not collected. Hence, we have to define the appointment request rate based on realized appointments data ignoring any daily or weekly changes in the demand. In the end, without further research, it is not easy to assess if such a queueing model will deliver valuable results due to the simplifications in the model and the inaccuracy of the model parameters.

Therefore, to improve the impact of Operations Research models in practice, we need to educate practice managers and software companies about data collection, and we, as researchers, need to put more emphasis on building models tailored to specific practice settings with their potentially available data and explaining of how to derive the model parameters from this data.

## 2.9 Conclusion and outlook

In this chapter, we review the planning and control decisions that need to be taken by a practice manager when opening and running a medical practice. We assign those decisions to the planning levels and give some examples of Operations Research planning problems that address those decisions. We further define patient demand, including all relevant patient-practice interactions and possible influence factors on patient behavior. From there, we define a best-case data set and compare it to real-world data sets from medical practices. We comment on the consequences for model application if some data is not collected. We explain the importance of explaining how to define model parameters based on data.

On the one hand, we discovered that practices do not collect all relevant data for decision-making. For instance, data on appointment requests or data on the realized service times is often missing. The consequences are that many models from the literature are not or only partly applicable. On the other hand, many models from the Operations Research literature assume settings and parameters without making a connection to actual or possible data availability and without explaining the process of transforming raw data into model parameters. Consequently, we recommend that practice managers collect as much relevant data as possible. Here, they have to find a trade-off between data collection effort and the potential benefit. This trade-off decision can be supported by our overview of planning and control decisions (together with related planning problems and their needed data basis) and the presented best-case appointment data set. We further recommend that modelers aiming for implementation put a focus on the actual or potentially available data using our best-case data set and the process of transforming this data into model parameters.

We also want to use our knowledge of the typical and best case structure of appointment data sets to find new potentially relevant effects of patient behavior that should be integrated into models.

Access time can be deduced from typically available data. There has been done quite some research on the influence of access times on the no-show probability (Dantas et al. 2018). However, we should also investigate the influence of access times, for example, on the appointment request rate, on the cancellation probability, the probability to walk-in, the rate of requests to enter the panel, or the rate of patients that leave the panel, and many more.

In a best-case scenario, data on patient preferences would be collected. It would be interesting to investigate the effect of the fulfillment of those preferences on the cancellation and no-show probability. If data on actual service times were available, we could investigate the influence of the queue length of waiting patients in the practice on service times.

Further, we suggest future research to build a taxonomy of different practice settings, including a list of relevant data indicating their relative importance. In this context, an accessible meta collection of several data sets from real world-practices for every setting would be beneficial for decision-makers. This especially applies to decisions that are often taken before enough data to support the decision is available. Another idea would be to use performance improvement methods that do not rely on a vast data basis but make incremental changes based on the current performance.

# 3 An analytical queueing model to determine the distribution of indirect waiting times

This chapter presents a flexible analytical queuing model to investigate the relationship between the physician's daily capacity, the panel size, i.e., the number of panel patients, and the stationary distribution of indirect waiting times of patients. The queueing model can present many different settings by integrating queue length-dependent parameters such as the appointment request rate, the no-show probability, and the rescheduling probability. We further extend the model by considering queue length-dependent and random service times and put a particular focus on how the appointment request rate of a physician with panel patients can be modeled. A preliminary version of this model was published as:

> Anne Zander. Modeling Indirect Waiting Times with an M/D/1/K/N Queue. In *Proceedings of the Second KSS Research Workshop: Karlsruhe, Germany, February 2016. Ed.: P. Hottum*, volume 69 of *KIT Scientific Working Papers*, pages 110–119. Karlsruher Institut für Technologie (KIT), 2017.

## 3.1 Introduction

Besides providing professional medical services, physicians also need to consider other aspects such as waiting times and access to service to deliver patients an overall satisfactory service experience. In this work, we focus on access time, i.e., indirect waiting time, for patients, defined as the elapsed time between the day an appointment is made and the actual appointment day. Long indirect waiting times can lead to frustration and deterioration in patients' health. Some studies also show that long indirect waiting times increase the probability of a patient becoming a no-show (Gallucci et al. 2005), i.e., without or on short notice, the patient does not show up. If no other patient is available short term, the physician experiences idle time. Consequently, physicians need to control indirect waiting times to ensure access to service and to avoid no-shows.

We focus on physicians that operate under the traditional appointment policy, which means that every patient wanting to visit the physician has to book an appointment. For now, we assume that patients always book the next available appointment. Later, we will relax this assumption.

For this setting, we aim to build an analytical model on a tactical level that connects the number of appointment requests a physician receives, the physician's appointment offerings, and the resulting distribution of indirect waiting times.

We propose to model the appointment backlog as a queue. Hence patients join the queue when the appointment is made and exit the queue after service completion. Based on the number of patients in the queue at the arrival time of a new appointment request and the number of daily appointment offerings, we can determine the requesting patient's indirect waiting time. Therefore, we denote the queue as the indirect queue and the state of the queueing system, i.e., the number of patients waiting or getting treatment, as the (indirect) queue length.

To control indirect waiting times, the physician fixes an indirect waiting time service level, e.g., on average, patients should not wait more than two weeks for an appointment. Using our queueing model, the physician can deduce the amount of appointment demand the physician can manage given the appointment offerings or the other way around, the number of necessary appointment offerings given the appointment request pattern to achieve the indirect waiting time service level.

Important features of our basic model are the consideration of no-shows and rescheduling. Here, rescheduling refers to patients who make a new appointment on the day of an appointment. A physician operating under the traditional appointment policy may experience a range of potentially long indirect queue lengths which may in turn influence the queue parameters. Hence, into our basic model, we integrate an appointment request rate, a no-show probability, and a rescheduling probability dependent on the queue length.

We extend our basic model by considering a queue length-dependent and random service time. We further present different theoretical models for the appointment request rate of a physician with panel patients, i.e., patients who visit regularly. Here, we believe that the queue length impacts the appointment request rate since longer queue lengths mean that a significant portion of panel patients are already waiting for an appointment and probably will not requeste another one. Finally, we explain how we can relax the assumption that patients always book the next available appointment.

We run extensive numerical experiments and compare the queueing model results to the results of a simulation to validate our assumptions and approximations used in the queueing model.

We pursue several objectives with this very flexible, analytical queueing model. One is the modeling of medical practices to support tactical capacity decisions. We want to investigate the general indirect queue behavior in different settings to determine the different parameters' influence on the results. Even though the queueing model is built for a specific use case, we believe that the mathematical model has a value of its own. Therefore, we keep the queueing model as general as possible even if some features are not necessary for the original use case.

This chapter is organized as follows: First, we introduce relevant literature in Section 3.2 and explain our conceptual and mathematical modeling in Section 3.3. Then, we present the basic queueing model in Section 3.4. Section 3.5 comprises extensions to the basic queueing model. First, we explain how a queue length-dependent service time and even a general distributed queue length-dependent service time can be integrated. Then we present different models for the appointment request rate for a physician with panel patients. Lastly, we investigate how to relax the assumption on the next available appointment.

## 3.2 Literature review

First, we review the literature focusing on access time/indirect waiting time for appointments in health care.

The queueing model that we present is based on the $M/D/1/K$ queueing model of Green and Savin (2008). In their article, the authors develop two different queuing models ($M/D/1/K$ and $M/M/1/K$) to determine the relationship between the panel size, i.e., the number of panel patients of a physician, and the expected indirect queue length in steady state where they include the possibility of no-shows that directly reschedule. They assume that appointment requests are coming from the panel only and that this arrival rate is constant. For technical reasons, they use a dependency of the no-show probability from the indirect queue length at the time of the patient's service completion as a proxy for a no-show probability dependent on the indirect waiting time of the patient. Their goal is to determine the maximal panel size, which allows the physician to implement an open access policy where patients can only make appointments for the same day. They claim that an open access system can be installed if the expected probability of getting a same-day appointment is above a certain threshold, e.g., 80%. Hence, 80% of the time, the indirect waiting time is shorter than a day. With a known threshold, an upper bound for the panel size can be determined. They validate their queueing model through simulation.

The $M/D/1/K$ queueing model itself relies on the results of Garcia et al. (2002). Garcia et al. derive analytical expressions of the time-dependent probability distribution of an $M/D/1/K$ queue. They develop differential equations for the departure rates (dependent on the state of the queueing system and the time). These differential equations allow a numerical computation of the probability distribution of the queue length at each point in time. Numerical integration methods, such as Runge-Kutta, can be used. Furthermore, due to the particular structure of the differential equations Garcia et al. (2002) can also present an analytical solution even if the queue is initially not empty. Green and Savin (2008) adapt the differential equations of Garcia et al. (2002), including the possibility of rescheduling for no-shows, and use them to deduce the stationary probability distribution of the queue length.

Liu and Ziya (2014) present two single server queueing models ($M/M/1$) where they decide on the panel size and the service capacity to maximize the long-term average reward while constraining the expected access time. As in Green and Savin (2008), they assume that appointment requests are coming from the panel only and that this arrival rate is constant. They also increase the probability of being a no-show for longer access times using the same proxy as Green and Savin (2008). In contrast to Green and Savin (2008), the no-shows do not reschedule immediately. They assume a reward for treating a patient that showed up, a probability that a no-show slot can be filled by a walk-in patient and costs for overtime. In their first model, Liu and Ziya (2014) consider the physician's service capacity to be constant, and they keep the arrival rate (hence the panel size) variable. They show that there is a unique optimal arrival rate for increasing no-show probabilities (with respect to the indirect queue length at departure) that maximizes the long-term average reward. Liu and Ziya (2014) investigate the effect on the optimal value of the arrival rate dependent on the no-show probabilities. They show that the optimal arrival rate (and hence the panel size) increases if patients' show-up probabilities become less sensitive to additional appointment delays. In their second model, Liu and Ziya (2014) consider the arrival rate/panel size and the service rate/service capacity as variables, respectively. Again, optimal values for the arrival rate and the service rate can be determined. It is shown that for increasing show-up probabilities also the optimal service rate increases. The same effect is found again for the arrival rate: The optimal arrival rate (and hence the panel size) increases if patients' show-up probabilities become less sensitive to additional appointment delays.

Zacharias and Armony (2017) consider direct and indirect waiting time together. As in Liu and Ziya (2014), decisions on the panel size and the service capacities offered are made to maximize the long-term average daily reward. Again, there is a reward for every served patient, and there are costs for overtime. Also, they consider costs for direct and indirect waiting times. To model the appointment backlog (indirect waiting time), the authors use a G/D/1 queue with balking and batch service. Here, for a fixed panel size, the arrival rate is constant. For the waiting inside the doctor's office, they use a G/G/1 queue, including a fixed no-show probability and patients' unpunctuality. The two queueing models are approximated by heavy traffic diffusion limits. The authors find three cases for the optimal solution: Either it is not beneficial to have patients, or the clinic offers as many appointment slots as possible, or supply and demand are perfectly matched. In most of the considered parameter settings, the clinic should offer as many appointments as possible. The authors argue that this indicates that the open access policy is dominant in most cases, where clinics offer many appointments per day combined with a panel size such that patients get a same-day appointment with a high probability.

Izady (2015) presents three discrete-time queueing models (Model 1-3) with bulk service. Similar to Green and Savin (2008), a backlog-dependent no-show probability and rescheduling of no-shows is considered in Model 3. Again, the author uses the dependency of the no-show probability from the indirect queue length at the time of the patient's service completion as a proxy for the backlog-dependent no-show probability. In contrast to Green and Savin (2008), the author derives the waiting time distribution (Model 1), considers cancellations of appointments (Model 2), and considers a general arrival distribution. Izady tests his models with different arrival distributions. He shows that more capacity is needed for higher variability in the arrival distribution, and the impact of the no-show probability increases. He further shows that the impact of an increasing no-show probability is dependent on the arrival distribution. He uses Model 3 to analyze the influence of the panel size. He can reproduce the results of Green and Savin (2008), assuming a Poisson arrival process. However, using a discrete Weibull distribution with different standard deviations shows that a decrease in the arrival distribution variability leads to long backlogs for larger panel sizes.

In Zander (2017) the $M/D/1/K$ queueing model with rescheduling of no-shows of Green and Savin (2008) is further extended to an $M/D/1/K/N$ queueing model. The aim is to model a physician with panel patients with potentially larger panels than those suitable for open access. However, then a substantial part of the panel patients might be waiting in the indirect queue. Assuming that waiting patients do not book new appointments leads to an indirect queue length-dependent appointment request rate. Hence, the added feature here is a finite population, i.e., the panel, from where the appointment requests arrive which is symbolized by N in the Kendell notation of the queueing model. It is shown in Zander (2017) that the assumption on patients' booking behavior has a significant impact on the shape of the stationary probability distribution of the queue length.

Table 3.1 summarizes the main aspects of the queueing models from the literature and of our model that we will present in more detail in the next section. The second row of the table gives the Kendall notation of the queueing model if there is any. Here, the subscript $n$ stands for state-dependent distributions. We describe the arrival process in the third row. If no-shows and rescheduling are considered, the fourth and fifth row show the no-show probability function and the rescheduling probability function. Either a constant probability value is used, or there is a flexible dependency on the queue lengths. The last row denotes the method of calculating the stationary queue length distribution.

| Reference | Queueing model | Arrival process | No-show function | Rescheduling function | Solution method |
|---|---|---|---|---|---|
| Green and Savin (Green and Savin 2008) | $M/D/1/K$ | Poisson | flexible | constant for no-shows | analytical |
| Green and Savin (Green and Savin 2008) | $M/M/1/K$ | Poisson | flexible | constant for no-shows | analytical |
| Liu and Ziya (Liu and Ziya 2014) | $M/M/1$ | Poisson | flexible | - | analytical |
| Zacharias and Armony (Zacharias and Armony 2017) | $GI/D^x/1$ with balking | Poisson | constant | - | heavy traffic diffusion limit |
| Izady (2015) | discrete-time queueing model with bulk service, constant service times and one server | flexible arrival distribution | constant | constant for no-shows | analytical |
| Izady (2015) | discrete-time queueing model with bulk service, constant service times, one server and cancelations | flexible arrival distribution | constant | constant for no-shows | analytical |
| Izady (2015) | discrete-time queueing model with bulk service, constant service times and one server | flexible arrival distribution | flexible | constant for no-shows | analytical |
| Zander (2017) | $M/D/1/K/N$ | Poisson with specific arrival rate dependency on the queue length | flexible | constant for no-shows | analytical |
| This thesis | $M_n/D/1/K$<br>$M_n/D_n/1/K$<br>$M_n/G_n/1/K$ | Poisson with flexible arrival rate dependency on the queue length | flexible | flexible for shows and no-shows | analytical |

**Table 3.1:** Comparison of the queueing models from literature to model indirect waiting time

Outside the health care application area, Abouee-Mehrizi and Baron (2016) consider state-dependent queueing systems, namely $M_n/M_n/1$, $M_n/G_n/1$, $M_n/M_n/1/K$ and $M_n/G_n/1/K$. For all four queueing systems, they analytically derive the stationary queue length distribution at arbitrary times using the supplementary variable method and the stationary queue length distribution at arrival times using an embedded Markov chain.

## 3.3 Conceptual and mathematical modeling

As explained in the introduction, we focus on physicians who operate under the traditional appointment policy and potentially experience longer appointment backlogs. Furthermore, for now, we assume that patients always book the next available appointment. In this setting, our main objective is to control the appointment backlog or, more precisely, the indirect waiting time of patients on a tactical level. In the following, we will first explain our conceptual model and then our mathematical model.

To model the appointment backlog, we need to consider the arrival process of appointment requests and the appointment offerings to those requests. We are interested in the state of the system in the long run. Hence, we aim to determine the distribution of indirect waiting times given an arrival process and the appointment offerings.

Concerning the arrival process, appointment requests can arrive during working hours of the physician, e.g., via telephone or outside of the physician's working hours, e.g., via an online system. We do not consider cancelations or balking explicitly. Long-term cancelations are not included in the arrival process, and short-term cancelations are considered no-shows.

We want to include several effects for longer appointment backlogs or longer indirect queue lengths. First, the literature shows that longer indirect waiting times increase the probability of becoming a no-show (Gallucci et al. 2005). Both shows and no-shows, sometimes reschedule, i.e., book the next appointment on the day of an appointment, producing an additional appointment request stream. It is reasonable to assume that the probability of rescheduling also increases with increasing indirect waiting times both for shows and no-shows. Second, for physicians with panel patients, long appointment backlogs may contain a significant portion of appointments from panel patients. This fact will likely influence the arrival of general appointment requests, hence non-rescheduling requests. Third, for example, seasonal oscillations in the total arrival rate will become manifested in dependency of the total arrival rate on the appointment backlog for fixed appointment offerings because bigger arrival rates produce longer appointment backlogs. Therefore, we assume that the arrival process of appointment requests is dependent on the number of appointments in the appointment backlog (and on the appointment offerings) for general requests and on the current indirect waiting time for no-shows and rescheduling requests.

Concerning the appointment offerings, we assume that the physician decides on an average number of appointments slots to offer daily. The appointment slots are then assigned to requests on a first-come, first-serve basis. Here, we assume that there is a fixed booking horizon. Thus, when the appointment backlog reaches the booking horizon, appointment requests are rejected. We further assume that no-shows occupy capacity. Hence, an empty slot on short notice cannot be filled with another patient, and therefore, the physician stays idle during the slot. Also, we do not consider vacations. In our extended model, we consider a physician who changes the appointment offerings dependent on the queue length. Here, we assume the change is implemented from the next non-fully booked day. For example, in case of long queue lengths, the physician starts to plan for longer days to reduce the appointment backlog.

Next, we explain the basic mathematical modeling. We propose to model the appointment backlog as a queue: We extend the $M/D/1/K/N$ queueing model of Zander (2017) and hence also the $M/D/1/K$ queueing model of Green and Savin (2008) to an $M_n/D/1/K$ queueing model where we present analytical results for the stationary queue lengths distributions at different time points. The extended queueing models can be described as $M_n/D_n/1/K$ and $M_n/G_n/1/K$ in the Kendall notation.

Concerning the arrival process, we do not consider service vacations. In our mathematical model, appointment requests can only arrive during the working hours of the physician. General arrivals, i.e., non-rescheduling arrivals, occur according to a Poisson process with an arrival rate dependent on the indirect queue length, which corresponds to the size of the appointment backlog. The arrival process is represented by $M_n$ in the Kendall notation of the queueing model. The arrival rate function of non-rescheduling arrivals, or shortly the appointment request rate, is denoted by $\lambda$. If we are only interested in the queue length distribution at departure times and not in the general queue length distribution, we can relax the Poisson arrival assumption. We only need to define a general distribution for the number of arrivals during a service period without departures.

In the conceptual model, we assume that patients become no-shows or reschedule with a probability dependent on their indirect waiting time. The indirect queue length and the information on the appointment offerings can approximately represent the indirect waiting time at the time of arrival. However, in a queueing model, it is not tractable to store individual patient information, e.g., the queue length at arrival time, until the patient is served and that information is used. Here, Green and Savin (2008) argue that the length of the queue at departure represents the number of patients that arrived during the indirect waiting time and the service time of the considered patient. Therefore, the queue length at departure can also be seen as a proxy for indirect waiting time. Hence, we use a no-show function $\gamma$ that calculates the no-show probability dependent on the queue length directly after departure. The probability of rescheduling is also a function of the queue length directly after departure. For no-shows, the rescheduling function is denoted by $r^n$, and for shows, it is denoted by $r^s$.

Keep in mind that, in general, though we will not model this explicitly, the general arrivals the no-show probability and the rescheduling probability depend on the appointment offerings.

In our basic queueing model, we assume a single server queue representing one physician with deterministic service times of length $T$. Using the time unit days and ignoring the time of the day during which the practice does not offer appointments, this assumption reflects a physician who offers the same (average) number $\mu = \frac{1}{T}$ of appointments slots every day. However, note that we do not schedule patients to concrete days and time slots in the queueing model. Further, we assume a finite queue capacity of $K$ representing the finite booking horizon. We first assume a queue length-dependent service time in the model extensions and then a generally distributed service time with a distribution dependent on the queue length.

To avoid lengthy formulas, we introduce $\rho(k) = (1 - \gamma(k))r^s(k) + \gamma(k)r^n(k)$ and $\nu(k) = (1 - \gamma(k))(1 - r^s(k)) + \gamma(k)(1 - r^n(k))$ for $k = 0, \ldots, K$. Here, $\rho(k)$ and $\nu(k)$ denote the probabilities of rescheduling and not rescheduling of a patient that left $k$ patients in the system after his or her departure, respectively. Note, that in case of a rescheduling departure the patient itself who joins the queue immediately again is not counted. The basic model notation explained so far is summarised in Table 3.2.

| **Parameter** | **Description** |
|---|---|
| $T$ | service time |
| $\mu$ | (average) number of daily slots |
| $\lambda$ | arrival rate function |
| $K$ | queue capacity |
| $\gamma$ | no-show function |
| $r^n$ | rescheduling function for no-shows |
| $r^s$ | rescheduling function for shows |
| $\rho$ | rescheduling function |
| $\nu$ | non-rescheduling function |

**Table 3.2:** Basic model notation

Mathematically, the finite queue capacity ensures that the stationary queue length distribution exists. First, we determine the stationary queue length distribution at departure times. From there, we can deduce the indirect waiting time distribution in days, our primary interest. Besides, we also calculate the stationary queue length distribution (at arbitrary time points), the queue length distribution at departure times, differentiating rescheduling and non-rescheduling patients, and the queue length distribution at arrival times, differentiating rejected arrivals and non-rejected arrivals as well as rescheduling arrivals and non-rescheduling arrivals. Moreover, we calculate other measures such as the average no-show rate, the average rescheduling rate, the average rate of rejected patients, and more.

We present several model extensions. First, we explain how the appointment offerings per day can be made dependent on the indirect queue length to reflect a physician that adapts the appointments offerings on the current workload. Then, we propose different ways of modeling the appointment request rate for the case of a physician with panel patients. Finally, we explain how we can use several interconnected queueing models to represent a setting where patients do not always book the next available appointment.

We propose a second solution approach to determine the queue length distribution at departure times numerically via the transition matrix of the embedded Markov chain. Finally, even though it is not our main objective, we explain how our model can project the transient queue length development. The model features and extensions are summarized in Table 3.3.

| **Feature/Extension** | **Section** |
|---|---|
| Determination of the queue length distributions | Section 3.4.1 |
| Determination of the indirect waiting time distribution | Section 3.4.2 |
| Arrival process modeling | Section 3.4.3 |
| Transition matrix of the embedded Markov chain | Section 3.4.4 |
| Comment on the transient queue behavior | Section 3.4.5 |
| Description of other performance measures | Section 3.4.6 |
| Queue length-dependent service times | Section 3.5.1 |
| Appointment request rate models for a physician with panel patients | Section 3.5.2 |
| Capacity division between several demand streams | Section 3.5.3 |

**Table 3.3:** Model features and extensions

Our $M_n/D/1/K$ queueing model extends the $M/D/1/K$ model queuing model of Green and Savin (2008) through making the arrival rate and the rescheduling probability for show and no-shows dependent on the queue length. This extension allows considering longer appointment backlogs than appropriate for an open-access appointment policy. We further enhance the modeling of the departure rates compared to Green and Savin (2008) and Zander (2017), which we will later explain in more detail.

When we focus on the queue length distribution at departure times, which is the basis for determining the indirect waiting distribution, we are able to work with general arrival distribution just as Izady (2015). However, at the same time, we can use generally distributed service times.

Abouee-Mehrizi and Baron (2016) determine the stationary queue length distribution of an $M_n/G_n/1/K$ queue. We provide an alternative way of determining the same stationary queue length distribution where we can include the special feature of rescheduling. In contrast to Abouee-Mehrizi and Baron (2016), using our approach, we can also numerically calculate the transient queue length development.

## 3.4 Basic queueing model

In this section, we develop our basic queueing model. We start by determining the stationary queue length distribution at departures times and the stationary queue length distribution at arbitrary points in time in Section 3.4.1. We deduce the distribution of indirect waiting times measured in days in Section 3.4.2. Then, we go into detail about the arrival process of appointment requests in Section 3.4.3. We then deduce the transition matrix of the embedded Markov chain at departure times in Section 3.4.4 to be able to validate our analytical findings. We comment on using our results to determine the transient queue behavior via recursive numerical integration in Section 3.4.5 and describe further performance measures in Section 3.4.6.

### 3.4.1 Queue length distributions

To derive their queueing model Green and Savin (2008) use $D(k, t, t + \Delta t)$, the probability that a departure occurs within the time interval $(t, t + \Delta t)$ leaving behind exactly $k$ patients with $0 \leq k \leq K - 1$. Later, this approach leads to a good approximation of the queue length distribution in the case of a small proportion of rescheduling patients. However, we claim that in general a more detailed approach is necessary. This is because a departure leaving $k$ patients in the queueing system is the results of one of two situations. It could either be that immediately before departure the queueing system contained $k+1$ patients and the patient departing does not reschedule or immediately before departure the queueing system contained $k$ patients and the patient departing reschedules and therefore directly joins the queue again after departure. We will see later, for example, with regard to Lemma 1 that this distinction is needed. To represent these two situations, we define $D^{nr}(k, t, t + \Delta t)$ as the probability that a departure without rescheduling occurs within the time interval $(t, t + \Delta t)$ leaving behind exactly $k$ costumers with

$0 \leq k \leq K-1$ and $D^r(k,t,t+\Delta t)$ as the probability that a departure with rescheduling occurs within the time interval $(t,t+\Delta t)$ leaving behind exactly $k$ costumers with $1 \leq k \leq K$. Obviously, we have $D(k,t,t+\Delta t) = D^{nr}(k,t,t+\Delta t) + D^r(k,t,t+\Delta t)$ for $1 \leq k \leq K-1$. In the special case of $k=0$ and $k=K$ we have $D(0,t,t+\Delta t) = D^{nr}(0,t,t+\Delta t)$ and $D(K,t,t+\Delta t) = D^r(K,t,t+\Delta t)$. Based on the departure probabilities we can derive the corresponding departure rates:

$$\begin{aligned}
d(k,t) &= \lim_{\Delta t \to 0} \frac{D(k,t,t+\Delta t)}{\Delta t}, \qquad k = 0,\ldots,K,\\
d^{nr}(k,t) &= \lim_{\Delta t \to 0} \frac{D^{nr}(k,t,t+\Delta t)}{\Delta t}, \qquad k = 0,\ldots,K-1,\\
d^{r}(k,t) &= \lim_{\Delta t \to 0} \frac{D^{r}(k,t,t+\Delta t)}{\Delta t}, \qquad k = 1,\ldots,K.
\end{aligned}$$

We assume a Poisson arrival process where the arrival rate per time unit is dependent on the indirect queue length. Given $k$ patients in the system we denote the arrival rate by $\lambda(k)$. Hence, given $k$ patients in the system, the probability of an arrival in a short time interval $\Delta t$ is $\lambda(k)\Delta t$ while the probability of no arrival is $1-\lambda(k)\Delta t$. We will also need the probability distribution of the number of arrivals during a service period of a patient. We define $\alpha_k(i)$ as the probability that $i$ patients arrive during a service period of length $T$ given that there are currently $k = 1,\ldots,K-1$ patients in the system. Later on, we will explain in detail how to define this distribution including approximations and special cases.

The following lemma is based on Lemma 1 and on Proposition 1 in Green and Savin (2008) which are based on Lemma 1 and Proposition 2 in Garcia et al. (2002). The proof follows the proof of Proposition 2 in Garcia et al. (2002) with the difference that we use the departure rates $d^{nr}(k,t)$ instead of $d(k,t)$. Further, we use the Poisson arrival process that is dependent on the indirect queue length. By $p(k,t)$ we denote the probability that there are $k$ patients in the queueing system at time $t$.

**Lemma 1.** *Let* $p(0,0) = 1, p(k,0) = 0, \quad \forall k \in \{1,\ldots,K\}$ *and* $d(k,0) = 0, \quad \forall k \in \{0,\ldots,K-1\}$. *Then, at each time* $0 \leq t < T$, *all departure rates remain zero, and the probability distribution* $p(k,t)$ *obeys the following system of differential equations:*

$$\begin{aligned} \frac{dp(0,t)}{dt} &= -\lambda(0)p(0,t), && (3.1)\\ \frac{dp(k,t)}{dt} &= -\lambda(k)p(k,t) + \lambda(k-1)p(k-1,t), \quad k = 1,\ldots,K-1, && (3.2)\\ \frac{dp(K,t)}{dt} &= \lambda(K-1)p(K-1,t). && (3.3) \end{aligned}$$

*For* $t \geq T$ *the probability distribution* $p(k,t)$ *fulfills*

$$\begin{aligned} \frac{dp(0,t)}{dt} &= -\lambda(0)p(0,t) + d^{nr}(0,t), && (3.4)\\ \frac{dp(k,t)}{dt} &= -\lambda(k)p(k,t) - d^{nr}(k-1,t) + \lambda(k-1)p(k-1,t) \\ &\quad + d^{nr}(k,t), \quad k = 1,\ldots,K-1, && (3.5)\\ \frac{dp(K,t)}{dt} &= -d^{nr}(K-1,t) + \lambda(K-1)p(K-1,t). && (3.6) \end{aligned}$$

*Proof.* We start by proving the differential equations for $t \geq T$. We consider a time interval $(t, t+\Delta t)$ with $\Delta t \ll T$. We calculate the probability that the queueing system contains $k$ patients at time $t+\Delta t$ as the sum of probabilities of the possible events happening during the time interval $(t, t+\Delta t)$ for $k = 1,\ldots,K-1$. We assume that during a short time interval of length $\Delta t$ at most one event (arrival or departure) can occur.

1. At time $t$ the queueing system contained $k$ patients and no patient arrived until $t+\Delta t$ nor did a (non-rescheduling) patient depart. The probability for this is given by $p(k,t)(1 - \lambda(k)\Delta t) - D^{nr}(k-1,t,t+\Delta t)$.

2. At time $t$ the queueing system contained $k-1$ patients and a patient arrived between $t$ and $t+\Delta t$, increasing the probability by $p(k-1,t)\lambda(k-1)\Delta t$.

3. A time $t$ the queueing system contained $k+1$ patients and a non-rescheduling departure occurred during the time interval $(t, t+\Delta t)$ increasing the probability by $D^{nr}(k,t,t+\Delta t)$.

Note, that departures with rescheduling do not change the number of patients in the system and therefore should not be considered here. The probability $p(k, t + \Delta t), k = 1, \ldots, K - 1$, can then be written as:

$$\begin{aligned} p(k, t + \Delta t) =& p(k, t)(1 - \lambda(k)\Delta t) - D^{nr}(k - 1, t, t + \Delta t) \\ &+ p(k - 1, t)\lambda(k - 1)\Delta t \\ &+ D^{nr}(k, t, t + \Delta t), \quad k = 1, \ldots, K - 1. \end{aligned}$$

Subtracting $p(k, t)$ and dividing by $\Delta t$ yields:

$$\begin{aligned} \frac{p(k, t + \Delta t) - p(k, t)}{\Delta t} =& p(k, t)\left(\frac{1 - \lambda(k)\Delta t - 1}{\Delta t}\right) \\ &- \frac{D^{nr}(k - 1, t, t + \Delta t)}{\Delta t} \\ &+ p(k - 1, t)\lambda(k - 1) \\ &+ \frac{D^{nr}(k, t, t + \Delta t)}{\Delta t}, \quad k = 1, \ldots, K - 1. \end{aligned}$$

Now we let $\Delta t \to 0$ and get:

$$\begin{aligned} \frac{dp(k, t)}{dt} =& - \lambda(k)p(k, t) - d^{nr}(k - 1, t) \\ &+ \lambda(k - 1)p(k - 1, t) + d^{nr}(k, t), \quad k = 1, \ldots, K - 1. \end{aligned}$$

For $k = 0$, when the queueing system is empty a departure cannot happen. Therefore, we obtain:

$$\frac{dp(0, t)}{dt} = - \lambda(0)p(0, t) + d^{nr}(0, t).$$

For $k = K$, no non-rescheduling departure that leaves $K$ patients in the system can occur. Further, when there are $K$ patients in the queueing system, then an arrival will not change the state of the system since this arrival will be blocked because of the finite queue capacity. Hence, we have:

$$\frac{dp(K, t)}{dt} = - d^{nr}(K - 1, t) + \lambda(k - 1)p(k - 1, t).$$

In the case of $t < T$ departures cannot occur. Thus, the differential equations for $t < T$ result from the equations for $t \geq T$ omitting the departure terms. □

The following proposition is similar to Proposition 1 in Green and Savin (2008). The proof follows the proof given in their online supplement with the differences that we integrate our assumption that the appointment request rate is dependent on the queue length, that we integrate our rescheduling functions, that we find a different formula for $k = K - 1$ and an additional formula for $k = K$.

**Proposition 1.** *For any time $t \geq T$, we have:*

$$\begin{aligned} d(0,t) =& p(0,t-T)\lambda(0)\nu(0)\alpha_1(0) \\ &+ d(1,t-T)\nu(0)\alpha_1(0), \end{aligned} \tag{3.7}$$

$$\begin{aligned} d(k,t) =& p(0,t-T)\lambda(0)\nu(k)\alpha_1(k) \\ &+ p(0,t-T)\lambda(0)\rho(k)\alpha_1(k-1) \\ &+ \sum_{i=1}^{k+1} d(i,t-T)\nu(k)\alpha_i(k+1-i) \\ &+ \sum_{i=1}^{k} d(i,t-T)\rho(k)\alpha_i(k-i), \quad k = 1, \ldots K-2, \end{aligned} \tag{3.8}$$

$$\begin{aligned} d(K-1,t) =& p(0,t-T)\lambda(0)\nu(K-1)(1-\sum_{i=0}^{K-2}\alpha_1(i)) \\ &+ p(0,t-T)\lambda(0)\rho(K-2)\alpha_1(K-2) \\ &+ \sum_{i=1}^{K} d(i,t-T)\nu(K-1)(1-\sum_{j=0}^{K-i-1}\alpha_i(j)) \\ &+ \sum_{i=1}^{K-1} d(i,t-T)\rho(K-2)\alpha_i(K-1-i), \end{aligned} \tag{3.9}$$

$$\begin{aligned} d(K,t) =& + p(0,t-T)\lambda(0)\rho(K-1)(1-\sum_{i=0}^{K-2}\alpha_1(i)) \\ &+ \sum_{i=1}^{K} d(i,t-T)\rho(K-1)(1-\sum_{j=0}^{K-i-1}\alpha_i(j)). \end{aligned} \tag{3.10}$$

*Proof.* We will first derive the equations for $k = 1, \dots, K-2$. There are four different possibilities that result in the queueing system having $k$ patients immediately after (a rescheduling or non-rescheduling) departure at time $t + \Delta t$ based on the state of the queueing system at time $t - T$ where again $\Delta t \ll T$. In the following, we describe each of these possible four scenarios and calculate the corresponding probabilities:

1. The queue was empty at $t-T$, an appointment request arrived between $t-T$ and $t-T+\Delta t$, the patient showed up or not and did not reschedule after service completion, and during the service time of the patient, $k$ new appointment requests arrived. The probability of this scenario is:
   $p(0, t-T)\lambda(0)\Delta t e^{-\lambda(0)\Delta t}\nu(k)\alpha_1(k)$.
2. The queue was empty at $t-T$, an appointment request arrived between $t-T$ and $t-T+\Delta t$, the patient showed up or not and rescheduled after service completion, and during the service time of the patient, $(k-1)$ new appointment requests arrived. The probability of this scenario is:
   $p(0, t-T)\lambda(0)\Delta t e^{-\lambda(0)\Delta t}\rho(k-1)\alpha_1(k-1)$.
3. A (rescheduling or non-rescheduling) departure occurred during $(t-T, t-T+\Delta t)$ and left $i = 1, \dots, k+1$ patients in the queueing system, the first patient in the queue showed up or not and did not reschedule after service completion, and during the service time of the patient $(k+1-i)$ new appointment requests arrived. The probability of this scenario is:
   $D(i, t-T, t-T+\Delta t)\nu(k)\alpha_i(k+1-i)$.
4. A (rescheduling or non-rescheduling) departure occurred during $(t-T, t-T+\Delta t)$ and left $i = 1, \dots, k$ patients in the queueing system, the first patient in the queue showed up or not and rescheduled after service completion, and during the service time of the patient $(k-i)$ new appointment requests arrived. The probability of this scenario is:
   $D(i, t-T, t-T+\Delta t)\rho(k-1)\alpha_i(k-i)$.

Summing up the probabilities of each scenario yields:

$$\begin{aligned}
D(k, t, t+\Delta) = & p(0, t-T)\lambda(0)\Delta t e^{-\lambda(0)\Delta t}\nu(k)\alpha_1(k) \\
& + p(0, t-T)\lambda(0)\Delta t e^{-\lambda(0)\Delta t}\rho(k-1)\alpha_1(k-1) \\
& + \sum_{i=1}^{k+1} D(i, t-T, t-T+\Delta t)\nu(k)\alpha_i(k+1-i) \\
& + \sum_{i=1}^{k} D(i, t-T, t-T+\Delta t)\rho(k-1)\alpha_i(k-i), \\
& k = 1, \dots, K-2.
\end{aligned}$$

In the case of $k = 0$, a rescheduling patient is not possible and hence scenarios 2 and 4 are not possible. Hence, we obtain:

$$\begin{aligned} D(0,t,t+\Delta) =& p(0,t-T)\lambda(0)\Delta t e^{-\lambda(0)\Delta t}\nu(0)\alpha_1(0) \\ &+ D(1,t-T,t-T+\Delta t)\nu(0)\alpha_1(0). \end{aligned}$$

In the case $k = K-1$, the terms corresponding to scenarios 1 and 3 change due to the limited queue capacity. The arrival of $K-1$ ($K-i$, respectively) patients has the same effect as more than $K-1$ patient arrivals ($K-i$, respectively). Also in contrary to Green and Savin (2008) the sum of the third term runs until $K$ as a departure leaving $K$ patients behind is indeed possible if this departure is a rescheduling departure:

$$\begin{aligned} D(K-1,t,t+\Delta) =& p(0,t-T)\lambda(0)\Delta t e^{-\lambda(0)\Delta t}\nu(K-1)(1-\sum_{i=0}^{K-2}\alpha_1(i)) \\ &+ p(0,t-T)\lambda(0)\Delta t e^{-\lambda(0)\Delta t}\rho(K-2)\alpha_1(K-2) \\ &+ \sum_{i=1}^{K} D(i,t-T,t-T+\Delta t)\nu(K-1)(1-\sum_{j=0}^{K-i-1}\alpha_i(j)) \\ &+ \sum_{i=1}^{K-1} D(i,t-T,t-T+\Delta t)\rho(K-2)\alpha_i(K-1-i). \end{aligned}$$

In contrary to Green and Savin (2008) we also consider the case $k = K$. In this case only the terms corresponding to the scenarios with rescheduling, 2 and 4, remaind valid. Again as for $k = K-1$ the arrival of $K-1$ ($K-i$, respectively) patients has the same effect as more than $K-1$ patient arrivals ($K-i$, respectively). Therefore, we have:

$$\begin{aligned} D(K,t,t+\Delta) =& p(0,t-T)\lambda(0)\Delta t e^{-\lambda(0)\Delta t}\rho(K-1)(1-\sum_{i=0}^{K-2}\alpha_1(i)) \\ &+ \sum_{i=1}^{K} D(i,t-T,t-T+\Delta t)\rho(K-1)(1-\sum_{j=0}^{K-i-1}\alpha_i(j)). \end{aligned}$$

Dividing by $\Delta t$ and taking the limits with $\Delta t \to 0$ results in the departure rates and thus concludes the proof. □

Because we are working with a queueing system with a finite capacity a stationary queue length probability distribution exists. Thus, also stationary departure rates exist. Let the stationary probabilities and the stationary departure rates be defined as

$$\begin{aligned}
\pi(k) &= \lim_{t\to\infty} p(k,t), \quad k = 0,\ldots,K,\\
d^*(k) &= \lim_{t\to\infty} d(k,t), \quad k = 0,\ldots,K,\\
d^{nr*}(k) &= \lim_{t\to\infty} d^{nr}(k,t), \quad k = 0,\ldots,K-1,\\
d^{r*}(k) &= \lim_{t\to\infty} d^{r}(k,t), \quad k = 1,\ldots,K.
\end{aligned}$$

The following lemma is based on a part of the proof of Proposition 2 in the online appendix of Green and Savin (2008) where the differences are that we again integrate our assumption that the appointment request rate is dependent on the queue length and that we use $d^{nr*}$ instead of $d^*$.

**Lemma 2.** *For the stationary probability distribution and the stationary departure rate the following relationship holds:*

$$d^{nr*}(k) = \pi(k)\lambda(k), \qquad k = 0,\ldots,K-1. \tag{3.11}$$

*Proof.* We let $t \to \infty$ in the second set of equations in Lemma 1. In steady state we have $\frac{dp(k,t)}{dt} = 0$ for all $k = 0,\ldots K-1$. Hence, equation (3.4) yields $d^{nr*}(0) = \pi(0)\lambda(0)$. The equations for $k = 1,\ldots,K-2$ can be deduced via induction. Assume that $d^{nr*}(k) = \pi(k)\lambda(k)$ is true for a fixed $k$. Equation (3.5) for $k+1$ gives: $\lambda(k+1)\pi(k+1) + d^{nr*}(k) = \lambda(k)\pi(k) + d^{nr*}(k+1)$. Using the equation for $k$ we obtain: $\lambda(k+1)\pi(k+1) = d^{nr*}(k+1)$. The equation for $k = K-1$ can be deduced directly letting $t \to \infty$ in equation (3.6). □

The proof of the following proposition mimics part of the proof of Proposition 2 in the online appendix of Green and Savin (2008) where the differences are that we assume a queue length-dependent appointment request rate, integrate our rescheduling functions and obtain a slightly different recursion due to a small calculation error in the online appendix of Green and Savin (2008).

**Proposition 2.** *The values of the stationary departure rates* $d^*(k)$ *for* $k = 1, \dots, K$ *can be determined knowing the value of* $d^*(0)$ *using the recursion* $f(k) := \frac{d^*(k)}{d^*(0)}$ *with*

$$f(0) =1, \tag{3.12}$$

$$f(1) =\frac{1}{\nu(0)\alpha_1(0)} - 1, \tag{3.13}$$

$$\begin{aligned} f(k+1) =& \frac{1}{\nu(k)\alpha_{k+1}(0)} \left(f(k) - \nu(k)\alpha_1(k) - \rho(k-1)\alpha_1(k-1)\right) \\ & - \frac{1}{\nu(k)\alpha_{k+1}(0)} \sum_{i=1}^{k} f(i) \left(\nu(k)\alpha_i(k+1-i) + \rho(k-1)\alpha_i(k-i)\right), \\ & k = 1, \dots, K-2, \end{aligned} \tag{3.14}$$

$$\begin{aligned} f(K) =& \frac{1}{\nu(K-1)} \left( f(K-1) - \nu(K-1)(1 - \sum_{i=0}^{K-2} \alpha_1(i)) - \rho(K-2)\alpha_1(K-2) \right) - \\ & - \frac{1}{\nu(K-1)} \sum_{i=1}^{K-1} f(i) \left( \nu(K-1)(1 - \sum_{j=0}^{K-i-1} \alpha_i(j)) + \rho(K-2)\alpha_i(K-1-i) \right). \end{aligned} \tag{3.15}$$

*Proof.* We start by letting $t \to \infty$ in the equations in Proposition 1 and substituting $\pi(0)\lambda(0)$ by $d^{nr*}(0) = d^*(0)$ which is valid due to Lemma 2. We obtain:

$$d^*(0) =d^*(0)\nu(0)\alpha_1(0) + d^*(1)\nu(0)\alpha_1(0), \tag{3.16}$$

$$\begin{aligned} d^*(k) =& d^*(0)\nu(k)\alpha_1(k) \\ & + d^*(0)\rho(k-1)\alpha_1(k-1) \\ & + \sum_{i=1}^{k+1} d^*(i)\nu(k)\alpha_i(k+1-i) \\ & + \sum_{i=1}^{k} d^*(i)\rho(k-1)\alpha_i(k-i), \quad k = 1, \dots K-2, \end{aligned} \tag{3.17}$$

$$\begin{aligned} d^*(K-1) =& d^*(0)\nu(K-1)(1 - \sum_{i=0}^{K-2} \alpha_1(i)) \\ & + d^*(0)\rho(K-2)\alpha_1(K-2) \\ & + \sum_{i=1}^{K} d^*(i)\nu(K-1)(1 - \sum_{j=0}^{K-i-1} \alpha_i(j)) \\ & + \sum_{i=1}^{K-1} d^*(i)\rho(K-2)\alpha_i(K-1-i), \end{aligned} \tag{3.18}$$

$$\begin{aligned} d^*(K) =& d^*(0)\rho(K-1)(1 - \sum_{i=0}^{K-2} \alpha_1(i)) \\ & + \sum_{i=1}^{K} d^*(i)\rho(K-1)(1 - \sum_{j=0}^{K-i-1} \alpha_i(j)). \end{aligned} \tag{3.19}$$

Together, we have $K$ linearly independent equations ((3.16),(3.17),(3.18)) for $K+1$ unknown variables. Equation (3.19) does not yield any additional information.
It is trivial to determine $f(0)$, while $f(1)$ is obtained by rearranging (3.16):

$$\begin{aligned} f(0) =& 1, \\ f(1) =& \frac{1}{\nu(0)\alpha_1(0)} - 1. \end{aligned}$$

The remaining expressions $f(k+1)$ for $k = 1, \ldots, K-2$ can be derived from (3.17) by first dividing the equation by $d^*(0)$:

$$\begin{aligned} f(k) =& \nu(k)\alpha_1(k) + \rho(k-1)\alpha_1(k-1) \\ & + \sum_{i=1}^{k} f(i)\nu(k)\alpha_i(k+1-i) + \sum_{i=1}^{k} f(i)\rho(k-1)\alpha_i(k-i) \\ & + f(k+1)\nu(k)\alpha_{k+1}(0), \end{aligned}$$

and then by isolating $f(k+1)$:

$$\begin{aligned} f(k+1) =& \frac{1}{\nu(k)\alpha_{k+1}(0)} \left(f(k) - \nu(k)\alpha_1(k) - \rho(k-1)\alpha_1(k-1)\right) \\ & - \frac{1}{\nu(k)\alpha_{k+1}(0)} \sum_{i=1}^{k} f(i) \left(\nu(k)\alpha_i(k+1-i) + \rho(k-1)\alpha_i(k-i)\right), \\ & k = 1, \ldots, K-2. \end{aligned}$$

Similarly, we derive the expression for $f(K)$ by dividing equation (3.18) by $d^*(0)$:

$$\begin{aligned} f(K-1) =& \nu(K-1)(1 - \sum_{i=0}^{K-2} \alpha_1(i)) + \rho(K-2)\alpha_1(K-2) \\ & + \sum_{i=1}^{K-1} f(i)\nu(K-1)(1 - \sum_{j=0}^{K-i-1} \alpha_i(j)) + \sum_{i=1}^{K-1} f(i)\rho(K-2)\alpha_i(K-1-i) \\ & + f(K)\nu(K-1), \end{aligned}$$

and then by isolating $f(K)$:

$$
\begin{aligned}
f(K) =& \frac{1}{\nu(K-1)} \left( f(K-1) - \nu(K-1)(1 - \sum_{i=0}^{K-2} \alpha_1(i)) - \rho(K-2)\alpha_1(K-2) \right) \\
& - \frac{1}{\nu(K-1)} \sum_{i=1}^{K-1} f(i) \left( \nu(K-1)(1 - \sum_{j=0}^{K-i-1} \alpha_i(j)) + \rho(K-2)\alpha_i(K-1-i) \right).
\end{aligned}
$$

□

Later we want to calculate values for $d^*$. But using the recursion $f$ we might run into numerical problems especially for large values of $K$ and a very small or large $d^*(0)$. Then, the values $f(k)$ will become very large or very small, respectively. To avoid this problem we propose another way of calculation. Instead of $f(k) = \frac{d^*(k)}{d^*(0)}$ we use $g(k) = \frac{d^*(k)}{d^*(k-1)}$. Because $d^*(k)$ and $d^*(k-1)$ will not differ from each other as much as $d^*(k)$ and $d^*(0)$ the recursion $g$ will be computable for lager values of $K$ too.

**Corollary 3.** *Let* $g(k) := \frac{d^*(k)}{d^*(k-1)}$*, than:*

$$
\begin{aligned}
g(1) =& f(1), & (3.20) \\
g(k+1) =& \frac{1}{\nu(k)\alpha_{k+1}(0)} \left( 1 - \frac{1}{\prod_{j=1}^{k} g(j)} \left( \nu(k)\alpha_1(k) + \rho(k-1)\alpha_1(k-1) \right) \right) \\
& - \frac{1}{\nu(k)\alpha_{k+1}(0)} \left( \sum_{i=1}^{k} \frac{1}{\prod_{j=i+1}^{k} g(j)} \left( \nu(k)\alpha_i(k+1-i) + \rho(k-1)\alpha_i(k-i) \right) \right), \\
& k = 1, \dots, K-2, & (3.21) \\
g(K) =& \frac{1}{\nu(K-1)} \left( 1 - \frac{1}{\prod_{j=1}^{K-1} g(j)} \left( \nu(K-1)(1 - \sum_{j=0}^{K-2} \alpha_1(j)) + \rho(K-2)\alpha_1(K-2) \right) \right) \\
& - \frac{1}{\nu(K-1)} \left( \sum_{i=1}^{K-1} \frac{1}{\prod_{j=i+1}^{K-1} g(j)} \left( \nu(K-1)(1 - \sum_{j=0}^{K-i-1} \alpha_i(j)) + \rho(K-2)\alpha_i(K-1-i) \right) \right). \\
& & (3.22)
\end{aligned}
$$

*Proof.* We use the recursion for $f$ to deduce a recursion for $g$. Obviously, we have $g(1) = f(1)$. Of course, we have $f(k) = \frac{d^*(k)}{d^*(0)} = \frac{d^*(k)}{d^*(k-1)} \frac{d^*(k-1)}{d^*(k-2)} \cdots \frac{d^*(1)}{d^*(0)} = g_k \cdot g_{k-1} \cdots g_1$. Inserting this relation into (3.14) and (3.15) concludes the proof. □

With help of the recursion $g$, we do not need to start with determining $d^*(0)$ first and then all other values $d^*(k)$ for $k = 1, \dots, K$. Instead, we start by determining $d^*(l)$ for a certain $l \in \{1, \dots K\}$ and then calculate $d^*(k)$ for $k = 0, \dots, l-1$ and for $k = l+1, \dots, K$.

**Corollary 4.** *Let* $l \in \{1, \dots K\}$. *For* $k = 0, \dots, l-1$, *we have:*

$$d^*(k) = \frac{1}{\prod_{j=k+1}^{l} g(j)} d^*(l), \tag{3.23}$$

*whereas for* $k = l+1, \dots, K$, *we have:*

$$d^*(k) = \prod_{j=l+1}^{k} g(j) d^*(l). \tag{3.24}$$

*Proof.* The formulas are evident by using the definition of $g$. □

Based on the recursions for $d^*$, we are now able to calculate the distribution $q$ of the number of patients in the queueing system immediately after departures.

**Proposition 5.** *Using the recursion f, the distribution q of the number of patients in the queueing system immediately after departures is given by:*

$$q(k) = \frac{f(k)}{\sum_{i=0}^{K} f(i)}, \qquad k = 0, \dots, K, \tag{3.25}$$

*or using the recursion g for* $k = 0, \dots, l-1$*:*

$$q(k) = \frac{\frac{1}{\prod_{j=k+1}^{l} g(j)}}{\sum_{i=0}^{l-1} \frac{1}{\prod_{j=i+1}^{l} g(j)} + 1 + \sum_{i=l+1}^{K} \prod_{j=l+1}^{i} g(j)}, \tag{3.26}$$

*for* $k = l$*:*

$$q(k) = \frac{1}{\sum_{i=0}^{l-1} \frac{1}{\prod_{j=i+1}^{l} g(j)} + 1 + \sum_{i=l+1}^{K} \prod_{j=l+1}^{i} g(j)}, \tag{3.27}$$

*and for* $k = l+1, \dots, K$*:*

$$q(k) = \frac{\prod_{j=l+1}^{k} g(j)}{\sum_{i=0}^{l-1} \frac{1}{\prod_{j=i+1}^{l} g(j)} + 1 + \sum_{i=l+1}^{K} \prod_{j=l+1}^{i} g(j)}. \tag{3.28}$$

*Proof.* The stationary departure rate that leaves $k$ patients in the queueing system after departure is given by $d^*(k)$. Hence, the total stationary departure rate is given by $\sum_{i=0}^K d^*(i)$. Thus, the probability that a departure leaves exactly $k$ patients in the queueing system after departure is given by $q(k) = \frac{d^*(k)}{\sum_{i=0}^K d^*(i)}$. Using the definition of the recursion $f$, we obtain: $q(k) = \frac{f(k)d^*(0)}{\sum_{i=0}^K f(i)d^*(0)} = \frac{f(k)}{\sum_{i=0}^K f(i)}$. Similarly, we can substitute $d^*(k)$ using (3.23) and (3.24) and obtain the other formulas. □

Unfortunately, until now we can only determine the stationary departure rate $d^*$ correct up to a scaling parameter. Thus, we need one more equation. One such possible equation is given by the next proposition. To understand the proof of Proposition 6 we need to introduce three definitions.

**Definition 1.** An *M/G/1//N queue* is a queueing system with one server, a generally distributed service time, a finite population set with $N$ elements, a finite system capacity of $N$ and a Poisson arrival process, where the mean number of arrivals is the mean individual arrival rate of one element of the finite populations times the number of elements of the populations currently not waiting or getting served.

**Definition 2.** The term *busy period* denotes a period that starts with the arrival of a patient to an empty queue and ends with the first departure of a patient that leaves an empty queue behind.

**Definition 3.** The term *idle period* denotes a period that starts with the departure of a patient that leaves an empty queue and ends with the first arrival of a patient to the empty queue.

**Proposition 6.** *It holds:*

$$d^*(0) = \frac{1}{\frac{T}{q(0)} + \frac{1}{\lambda(0)}} = \frac{\lambda(0)q(0)}{T\lambda(0) + q(0)}. \tag{3.29}$$

*Proof.* In Chapter 4 of Takagi (1993) Takagi explains Takác's method (Takács 1962) to determine the queue length at an arbitrary point in time based on the queue length distribution at departure times for M/G/1//N queues. The start of this method also works in our case. We calculate the rate of transitions from system state 0 to system state 1 due to an arrival. The probability that the queue is empty after a service completion is $q(0)$. Therefore, $\frac{1}{q(0)}$ is the number of patients treated during a busy period. Takagi explains that this can be understood by considering a long time period during which a large number $M$ of patients is served. This period includes $Mq(0)$ busy periods on average. Hence, dividing the number of served patients by the average number of busy periods yields the average number of patients served per busy period: $\frac{M}{Mq(0)} = \frac{1}{q(0)}$. Hence, the mean duration of a busy period is $\frac{T}{q(0)}$. The mean duration of an idle period is $\frac{1}{\lambda(0)}$.

This is true because the duration of an idle period is exponentially distributed with parameter $\lambda(0)$ and mean $\frac{1}{\lambda(0)}$. Together, the mean length of a cycle of a busy and idle periods is $\frac{T}{q(0)} + \frac{1}{\lambda(0)}$. The number of cycles per unit time is hence given by $\frac{1}{\frac{T}{q(0)}+\frac{1}{\lambda(0)}}$. This number must be equal to the rate of transitions from system state $1$ to system state $0$ which is $d^*(0)$. Hence, we have:

$$d^*(0) = \frac{1}{\frac{T}{q(0)} + \frac{1}{\lambda(0)}} = \frac{\lambda(0)q(0)}{T\lambda(0) + q(0)}.$$

□

**Corollary 7.** *We can determine the stationary departure rates $d^*(k)$ for $k = 0, \dots, K$ via:*

$$d^*(k) = q(k)\frac{\lambda(0)}{T\lambda(0) + q(0)}, \qquad k = 0, \dots, K. \tag{3.30}$$

*Proof.* By definition, as explained in the proof of Proposition 5, we have $q(k) = \frac{d^*(k)}{\sum_{i=0}^{K} d^*(i)}$ for $k = 0, \dots, K$. Hence, $\frac{d^*(0)}{q(0)} = \sum_{i=0}^{K} d^*(i)$. At the same time, due to Proposition 6, we have $\frac{d^*(0)}{q(0)} = \frac{\lambda(0)}{T\lambda(0)+q(0)}$. Therefore, we obtain $d^*(k) = q(k)\sum_{i=0}^{K} d^*(i) = q(k)\frac{\lambda(0)}{T\lambda(0)+q(0)}$ for $k = 0, \dots, K$. □

**Lemma 3.** *Assuming the stationary departure rates $d^*(k)$ for $k = 0, \dots, K$ as given, we can determine the stationary departure rates for rescheduling departures and non-rescheduling departures via:*

$$d^{nr*}(k) = \sum_{i=0}^{k} d^*(i) \prod_{j=i}^{k-1} \left(-\frac{\rho(j)}{\nu(j)}\right), \qquad k = 0, \dots, K-1, \tag{3.31}$$

$$d^{r*}(k) = \sum_{i=0}^{k-1} -d^*(i) \prod_{j=i}^{k-1} \left(-\frac{\rho(j)}{\nu(j)}\right), \qquad k = 1, \dots, K, \tag{3.32}$$

*or via:*

$$d^{nr*}(K-k) = \sum_{i=0}^{k-1} -d^*(K-i) \prod_{j=i+1}^{k} \left(-\frac{v(K-j)}{\rho(K-j)}\right), \qquad k = 1, \dots, K, \tag{3.33}$$

$$d^{r*}(K-k) = \sum_{i=0}^{k} d^*(K-i) \prod_{j=i+1}^{k} \left(-\frac{v(K-j)}{\rho(K-j)}\right), \qquad k = 0, \dots, K-1. \tag{3.34}$$

*Proof.* At first, we show:

$$\begin{aligned}
d^{nr*}(0) =& d^*(0),\\
d^{r*}(K) =& d^*(K),\\
d^{r*}(k) =& \frac{\rho(k-1)}{\nu(k-1)} d^{nr*}(k-1), \qquad k = 1,\dots,K,\\
d^{nr*}(k) =& d^*(k) - d^{r*}(k), \qquad k = 1,\dots,K-1.
\end{aligned}$$

The first, the second, and the fourth equation are evident from the definition of $D(k,t,t+\Delta t)$, $D^{nr}(k,t,t+\Delta t)$ and $D^r(k,t,t+\Delta t)$ and by letting $\Delta t \to 0$ and then letting $t \to \infty$. Further, the total stationary departure rate with a number of $k = 1,\dots,K$ patients in the queueing system immediately before departure is given by: $d^{r*}(k) + d^{nr*}(k-1)$. We also know that if the queueing system contains $k$ patients immediately before a departure then the departure will be a non-rescheduling departure with probability $\nu(k-1)$ because the patient leaves behind $k-1$ patients in the system (the rescheduling patient is not counted). Hence, we obtain: $\frac{d^{nr*}(k-1)}{d^{r*}(k)+d^{nr*}(k-1)} = \nu(k-1)$. Rewriting this equation and inserting $1-\nu(k-1) = \rho(k-1)$ shows the third equation. To determine the non-rescheduling and rescheduling departure rates we can either start with $d^{nr*}(0) = d^*(0)$ or with $d^{r*}(K) = d^*(K)$ and use the third and fourth equation iteratively. □

Note that we should chose between Equations (3.31) and (3.32) and Equations (3.33) and (3.34) depending on the ratio between $\rho(k)$ and $\nu(k)$. Numerically, it is easier to work with products of numbers smaller than one. Hence, if we have in general $\frac{\rho(k)}{\nu(k)} < 1$, we should opt for Equations (3.31) and (3.32) and otherwise for Equations(3.33) and (3.34).

We are also able to calculate the distributions $q^{nr}$ and $q^r$ of the number of patients in the queueing system immediately after non-rescheduling and rescheduling departures.

**Lemma 4.** *We have:*

$$q^{nr}(k) = \frac{d^{nr*}(k)}{\sum_{i=0}^{K-1} d^{nr*}(i)}, \qquad k = 0,\dots,K-1, \tag{3.35}$$

*and*

$$q^{r}(k) = \frac{d^{r*}(k)}{\sum_{i=1}^{K} d^{r*}(i)}, \qquad k = 1,\dots,K. \tag{3.36}$$

*Proof.* The formulas are evident by the definition of $d^{nr*}$ and $d^{r*}$. □

Note that $q^r$ and $q^{nr}$ are quotients of sums of $d^{nr*}$ and $d^{r*}$ values. Therefore, it suffices to insert scaled $d^{nr*}$ and $d^{r*}$ values calculated based on scaled $d^*$ values via Lemma 3. Hence, to determine $q^r$ and $q^{nr}$, we do not need Proposition 6 and Corollary 7.

Now we are in position to compute the distribution $\pi$ of the number of patients in the queueing system at arbitrary points in time.

**Proposition 8.** *For the stationary queue length distribution, we have:*

$$\pi(k) = \frac{\lambda(0)}{\lambda(k)(T\lambda(0) + q(0))} \sum_{i=0}^{k} q(k) \prod_{j=i}^{k-1} \left( -\frac{\rho(j)}{\nu(j)} \right), \qquad k = 0, \dots, K-1, \tag{3.37}$$

$$\pi(K-k) = \frac{\lambda(0)}{\lambda(k)(T\lambda(0) + q(0))} \sum_{i=0}^{k-1} -q(K-i) \prod_{j=i+1}^{k} \left( -\frac{v(K-j)}{\rho(K-j)} \right), \qquad k = 1 \dots, K, \tag{3.38}$$

$$\pi(K) = 1 - \sum_{k=0}^{K-1} \pi(k). \tag{3.39}$$

*Proof.* From Lemma 2 we have $\pi(k) = \frac{d^{nr*}(k)}{\lambda(k)}$ for $k = 0, \dots, K-1$. Here, we use Lemma 3 to substitute $d^{nr*}$ and finally apply Corollary 7. To fix $\pi(K)$ we use $\pi(K) = 1 - \sum_{k=0}^{K-1} \pi(k)$. □

Note that in Proposition 8 we can further insert the relations of Proposition 5 to eliminate $q$ and work with the recursions $f$ and $g$ instead.

In summary, to determine the queue length distribution $\pi$, we start by defining the recursions $f$ and $g$ as ratios of the stationary departure rates $d^*(k), \forall k \in \{0, \dots, K\}$ using Proposition 1. With $f$ or $g$ we determine the queue length distribution immediately after departures $q$ applying Proposition 5. The queue length distribution immediately after departures is then used to determine the stationary departure rates $d^*(k), \forall k \in \{0, \dots, K\}$ via Proposition 6 and Corollary 7. Applying Lemma 3 yields the stationary departure rates for rescheduling patients $d^{*r}(k), \forall k \in \{1, \dots, K\}$ and non-rescheduling patients $d^{*nr}(k), \forall k \in \{0, \dots, K-1\}$. Finally, invoking Lemma 2, we obtain the queue length distribution $\pi$ based on $d^{nr*}$.

Comparing our approach to that of Green and Savin (2008), we also define the recursion $f$ for the stationary departure rates $d^*$. But, in addition, we use the recursion $g$ to avoid numerical problems in the implementation. Instead of using an equilibrium equation, we use Proposition 6 (Takác's method) to come up with the missing equation to define the stationary departure rates $d^*$. The equilibrium equation used by Green and Savin (2008) adapted to a non-constant arrival rate is given by $\sum_{k=0}^{K-1} \lambda(k)\pi(k) = \frac{1}{T} \sum_{k=1}^{K} (1 - \rho(k-1))\pi(k)$. The idea is to equate the overall (non-rejected) arrival rate with the overall departure rate, where the term $\frac{1}{T} \sum_{k=1}^{K} \rho(k-1)\pi(k)$

should describe the rescheduling patients. Though it is true that the overall departure rate is $1/T$ when the queue is not empty, the average departure rate given a queue of length $k$ in steady state is not necessarily $1/T$. This is because - in contrast to the arrival rate - the departure rate is dependent on the progression of the service. A departure only happens after a period of $T$ after the start of a service. Therefore, if we saw every service progression with the same probability given a certain queue length $k$, we would have an average departure rate of $1/T$. However, this is not true. Taking the example of $k = 1$, we notice that in steady state a queue length of $k = 1$ will mainly be seen at the beginning of a service period (after an arrival to an empty queue, or after a departure) hence resulting in an average departure rate less than $1/T$. Thus, the equilibrium equation as given here is not correct. This observation also explains why we need to introduce the departure rates in the first place. Another contribution is Lemma 3 to determine $d^{r*}$ and $d^{nr*}$ based on $d^*$. Using Lemma 2, we calculate the queue length distribution $\pi$ based on $d^{nr*}$ whereas Green and Savin (2008) use the same lemma but with $d^*$ instead of $d^{nr*}$. If we had $\nu(k) = 1$ (no one reschedules) in Lemma 3 then we would have $d^{nr*}(k) = d^*(k)$ for all $k = 0, \ldots, K-1$. Hence, it would not make a difference. Still, if $1 - \nu(k)$ is small for all $k = 0, \ldots, K-1$, we have $d^{nr*}(k) \approx d^*(k)$ for all $k = 0, \ldots, K-1$ and the resulting stationary queue length distributions are similar.

### 3.4.2 Indirect waiting time distribution

Besides the queue length distribution, we are also interested in the distribution of indirect waiting times measured in days. To compute the indirect waiting time distribution we first determine the stationary queue length distribution as seen by arrivals. In our case the well known PASTA (Poisson arrivals see time averages) property does not apply due to the relationship between the arrival rate and the number of patients in the queueing system. We use Bayes' theorem to determine the queue length distribution as seen by arrivals. Here, we differentiate between arrivals including rejected requests (due to the queue capacity) and not including rejected requests. We further include or exclude rescheduling patients or focus on rescheduling patients entirely.

**Proposition 9.** *In the following we define $d^{r*}(0) = 0$. The stationary queue length distribution $p_a^{rej,res}$ at arrival times including rejected requests and including rescheduling patients is given by:*

$$p_a^{rej,res}(k) = \frac{\lambda(k)\pi(k) + d^{r*}(k)}{\sum_{l=0}^{K} \left(\lambda(l)\pi(l) + d^{r*}(l)\right)}, \qquad k = 0, \ldots, K. \tag{3.40}$$

*The stationary queue length distribution $p_a^{rej}$ at arrival times including rejected requests and excluding rescheduling patients is given by:*

$$p_a^{rej}(k) = \frac{\lambda(k)\pi(k)}{\sum_{l=0}^{K}\lambda(l)\pi(l)}, \qquad k = 0, \dots, K. \tag{3.41}$$

*The stationary queue length distribution $p_a^{onlyres}$ at arrival times only focusing on rescheduling patients is given by:*

$$p_a^{onlyres}(k) = \frac{d^{r*}(k)}{\sum_{l=1}^{K} d^{r*}(l)} = q^r(k), \qquad k = 1, \dots, K. \tag{3.42}$$

*The stationary queue length distribution $p_a^{res}$ at arrival times excluding rejected requests and including rescheduling patients is given by:*

$$p_a^{res}(k) = \frac{d^{*}(k)}{\sum_{l=0}^{K} d^{*}(l)} = q(k), \qquad k = 1, \dots, K. \tag{3.43}$$

*The stationary queue length distribution $p_a$ at arrival times excluding rejected requests and excluding rescheduling patients is given by:*

$$p_a(k) = \frac{d^{nr*}(k)}{\sum_{l=0}^{K-1} d^{nr*}(l)} = q^{nr}(k), \qquad k = 0, \dots, K-1. \tag{3.44}$$

*Proof.* We start with the stationary queue length distribution at arrival times including rejected requests and including rescheduling patients. Let $X_t$ be the number of patients in the system at time $t$ where we assume steady state has been reached. Then the probability that there are $k = 0, \dots, K$ patients in the system at the time of an arrival (that may be rejected or rescheduling) is given by:

$$\begin{aligned}
&P(X_t = k|\text{arrival at } t)\\
=&\frac{P(\text{arrival at } t|X_t = k)P(X_t = k)}{P(\text{arrival at } t)}\\
=&\frac{P(\text{non-res. arrival at } t|X_t = k)P(X_t = k) + P(\text{res. arrival at } t|X_t = k)P(X_t = k)}{P(\text{non-res. arrival at } t) + P(\text{res. arrival at } t)}\\
=&\frac{P(\text{non-res. arrival at } t|X_t = k)P(X_t = k) + \frac{P(X_t=k|\text{res. arrival at } t)P(\text{res. arrival at } t)}{P(X_t=k)}P(X_t = k)}{\sum_{l=0}^{K} P(\text{non-res. arrival at } t|X_t = l)P(X_t = l) + P(\text{res. arrival at } t)}\\
=&\frac{P(\text{non-res. arrival at } t|X_t = k)P(X_t = k) + P(X_t = k|\text{res. arrival at } t)P(\text{res. arrival at } t)}{\sum_{l=0}^{K} P(\text{non-res. arrival at } t|X_t = l)P(X_t = l) + P(\text{res. arrival at } t)}.
\end{aligned}$$

Of course we know that $P(X_t = l) = \pi(l)$ in steady state. Further, the probability of a non-rescheduling arrival in a short time interval $\Delta t$ is $\lambda(l)\Delta t$ if $l$ patients are in the system. The probability of a rescheduling arrival in a short time interval $\Delta t$ is $(\sum_{l=1}^{K} d^{r*}(l))\Delta t$. Because rescheduling arrivals are patients that just left service and rescheduled the distribution of the queue length at rescheduling arrival times is equivalent to the queue length distribution at rescheduling departure times. Therefore, we have $P(X_t = k|\text{res. arrival at } t) = q^r(k) = \frac{d^{r*}(k)}{\sum_{l=1}^{K} d^{r*}(l)}$. Together, we obtain:

$$P(X_t = k|\text{arrival at } t) = \frac{\lambda(k)\pi(k) + d^{r*}(k)}{\sum_{l=0}^{K} (\lambda(l)\pi(l) + d^{r*}(l))}, \qquad k = 0, \dots, K.$$

The formulas for the other queue length distributions are similarly deduced. □

Note that for the Equations (3.42), (3.43) and (3.44) we do not need to know the exact values of $\pi$ as $\pi(K)$ does not appear in the formulas. Even more, as for those equations the probability values are quotients of sums of $d^{r*}$ and $d^{nr*}$ values, it suffices to insert scaled $d^{nr*}$ and $d^{r*}$ values calculated based on scaled $d^*$ values via Lemma 3. Hence, Proposition 6 and Corollary 7 are not needed to determine Equations (3.42), (3.43) and (3.44).

When calculating indirect waiting times we do not want to include rejected patients. A non-rescheduling arrival to the queue is rejected if and only if the queue is at full capacity. A rescheduling arrival is never rejected. Hence, we just have to exclude the non-rescheduling arrivals to a full queue. Therefore, in the following, $p = p_a^{res} = q$ denotes the stationary distribution of the queue length at arrival times excluding rejected requests and including rescheduling patients which is equivalent to the queue length distribution immediately after departures. Based on $p$ we can deduce the actual distribution of indirect waiting times. We are interested in the indirect waiting time measured in days. That is why we do not go into detail considering the remaining service time of a patient being served at the time of an arrival. First, let us assume that the physician has a working time per day corresponding to $\mu \in \mathbb{Z}_+$ time slots of length $T$ each. Consistent with the queueing model, we assume that patients arriving to an empty queue are served immediately (not necessarily at the beginning of a time slot). However, if the queue is non-empty when one day ends and another one starts, patients start being served at the beginning of the first slot of the day. Therefore, experiencing longer queue lengths means that patients generally start being serviced at the beginning of a time slot. Hence, in the following analysis we assume that a new patient arriving to an non-empty queue will start being serviced at the beginning of a time slot. In case of a patient arriving to an empty queue, we assume that he or she is being served on the day of the request regardless of his or her arrival time. Note that this may induce a maximal overtime of $T$ if the patient arrives during the last time slot of the day.

**Proposition 10.** *Let $\mu \in \mathbb{Z}_+$ and $K_l = \left\{k \in \{0, \dots, K\} : l = \left\lfloor \frac{k}{\mu} \right\rfloor \right\}$. Then, the distribution of indirect waiting times $\iota$ in days is given by:*

$$\iota(l) = \sum_{k \in K_l} \frac{(l+1)\mu - k}{\mu} p(k) + \sum_{k \in K_{l-1}} \frac{k - (l-1)\mu}{\mu} p(k), \qquad l = 0, \dots, \left\lfloor \frac{K}{\mu} \right\rfloor + 1. \quad (3.45)$$

*Proof.* If the queue is empty the patient will be treated immediately on the day of the request. If the queue length is 1 and the patient arrives during the first $\mu - 1$ slots of the day, the patient will get treatment on the day of request. If the request arrives in the last slot of the day then one patient in the system is treated at that time and hence the requesting patient has to wait until the first slot of the next day. Therefore, the patient will experience an indirect waiting time of zero days with probability $\frac{\mu-1}{\mu}$ and an indirect waiting time of one day with probability $\frac{1}{\mu}$. In general, a patient arriving to a queue with length $k = l\mu + i, i = 0, \dots, \mu - 1$ has a probability of $\frac{\mu-i}{\mu} = \frac{(l+1)\mu-k}{\mu}$ to wait $l$ days and a probability of $\frac{i}{\mu} = \frac{k-l\mu}{\mu}$ to wait $l + 1$ days. □

In the following, we extend Proposition 10 for non-integer values of $\mu$. We interpret a non-integer $\mu$ as the average value of the number of appointment slots offered per day. Of course there are numerous constellations that lead to a certain non-integer value of $\mu$. We do not want to make complicated calculations for each constellation but instead give an approximation for the distributions of indirect waiting times based on the average value of appointment slots offered.

**Corollary 11.** *Let $\mu > 0$ and $K_l = \left\{k \in \{0, \dots, K\} : l = \left\lfloor \frac{k}{\mu} \right\rfloor \right\}$. Then, an approximate distribution of indirect waiting times $\iota$ in days is given by:*

$$\iota(l) = \sum_{k \in K_l} \frac{(l+1)\mu - k}{\mu} p(k) + \sum_{k \in K_{l-1}} \frac{k - (l-1)\mu}{\mu} p(k), \qquad l = 0, \dots, \left\lfloor \frac{K}{\mu} \right\rfloor + 1. \quad (3.46)$$

### 3.4.3 Arrival process

As explained before, we assume a Poisson arrival process for the non-rescheduling appointment requests where the arrival rate $\lambda(k)$ per time unit is dependent on the indirect queue length $k$. Therefore, ignoring departures, the time between two arrivals is exponentially distributed with rate $\lambda(k)$, where $k$ is the number of patients in the system after the first arrival. The exponential distribution is the only memoryless continuous-time distribution. This memoryless feature ensures that the arrival time is not dependent on the last arrival time, which allows us to come up with the differential equations in Lemma 1. Remember that given $k$ patients in the system, the probability of an arrival in a short time interval $\Delta t$ is $\lambda(k)\Delta t$ while the probability of no arrival is $1 - \lambda(k)\Delta t$.

To complete the basic model, we need to determine $\alpha_k(i)$, i.e., the probability that $i$ patients arrive during a service period of length $T$ given that there are $k = 1, \ldots, K-1$ patients in the system when the service starts. In general, for a time period of length $t$ with no departure, we define the probability of $i$ arrivals with $k$ patients present at the start of the interval as $\alpha_k(i,t)$. For $i = 0$ the arrival rate is equal to $\lambda(k)$ during the whole time period of length $t$. Hence, we have $\alpha_k(0,t) = e^{-\lambda(k)t}$.

**Lemma 5.** *For $i > 0$ and $k \in \{1, \ldots, K-1\}$, $\alpha_k(i, \cdot)$ is defined by the differential equation:*

$$\frac{d\alpha_k(i,t)}{dt} = -\lambda(k+i)\alpha_k(i,t) + \lambda(k+i-1)\alpha_k(i-1,t) \tag{3.47}$$

*with boundary condition $\alpha_k(i,0) = 0$.*

*Proof.* The probability that $i$ patients arrive during an time interval of length $t + \Delta t$ is equal to the probability that $i$ patients arrived during $[0,t]$ and no one arrives during $[t, t+\Delta t]$ plus the probability that $i-1$ patients arrived during $[0,t]$ and one patient arrives during $[t, t+\Delta t]$. Thus, we have:

$$\begin{aligned}\alpha_k(i, t+\Delta t) &= \alpha_k(i,t)\alpha_k(0,\Delta t) + \alpha_k(i-1,t)\alpha_k(1,\Delta t)\\ &= \alpha_k(i,t)(1-\lambda(k+i)\Delta t) + \alpha_k(i-1,t)\lambda(k+i-1)\Delta t\\ &= \alpha_k(i,t) - \lambda(k+i)\alpha_k(i,t)\Delta t + \lambda(k+i-1)\alpha_k(i-1,t)\Delta t.\end{aligned}$$

Subtracting $\alpha_k(i,t)$ and dividing by $\Delta t$ leads to the differential equation. The boundary condition means that the probability of any number of arrivals during a time interval of length zero is zero. □

In the following, we derive a formula for $\alpha_k(1,t)$ in case of $\lambda(k) \neq \lambda(k+1)$. The formulas for $i > 1$ can be determined iteratively. However, note that they become increasingly complicated. We apply separation of variables and variations of constants to solve the first order linear differential Equation (3.47). The corresponding formula yields:

$$\begin{aligned}\alpha_k(1,t) &= e^{-\lambda(k+1)t}\int_0^t \lambda(k)e^{-\lambda(k)s}e^{\lambda(k+1)s}ds\\ &= \frac{\lambda(k)}{\lambda(k+1)-\lambda(k)}\left(e^{-\lambda(k)t} - e^{-\lambda(k+1)t}\right).\end{aligned} \tag{3.48}$$

In special settings, we can provide simple analytical formulas for $\alpha_k(i,t)$. Let us assume that $\lambda$ is a linear function of the form $\lambda(k) = mk + n \geq 0$ for $k = 0, \ldots, K$ with $m \neq 0$. Then $\alpha_k(\cdot, t)$ follows a negative binomial distribution $NB(\frac{n}{m} + k, 1 - e^{-mt})$, as explained, for example, in Shin (2000):

$$\alpha_k(i,t) = \binom{\frac{n}{m} + k + i - 1}{i} e^{-(mk+n)t} \left(1 - e^{-mt}\right)^i, \qquad i \geq 0. \tag{3.49}$$

Here, $\binom{\cdot}{\cdot}$ stands for the extended binomial coefficient where non-integer values are allowed in the first entry. In Section 3.5.2, we present a set of formulas for $\alpha_k(i,t)$ for the case of a physician who treats panel patients only. These formulas are a special case of Equation (3.49).

Going through Section 3.4, we notice that only Lemma 1 (and hence also Lemma 2) relies on the assumption of a Poisson arrival process dependent on the queue length. All other proportions, corollaries, and lemmas also work with an arrival rate for an empty queue and a general arrival distribution during a service period dependent on the number of patients in the system at the beginning of the service. This is true because in the formulas $\lambda(k)$ only appears in the context of $\alpha_k$ or as $\lambda(0)$, where we have to set $\pi(0)\lambda(0)$ equal to $d^*(0)$ in Proposition 2. We further notice that Lemma 1 and Lemma 2 are only relevant to deduce the queue length distribution but not to deduce the queue length distribution at departure times and hence the indirect waiting time distribution. Thus, if we are not interested in the exact queue length distribution or dependent performance measures, we can use arrival distributions $\alpha_k$ for $k = 1, \ldots, K - 1$ different to the distributions defined above.

A first simple possibility is to assume a constant arrival rate $\lambda(k)$ during a service period where $k = 1, \ldots, K - 1$ is the number of patients present at the beginning of the service period. More generally, for a time interval of length $t$ without a departure, we then obtain a Poisson distribution with parameter $\lambda(k)t$. Hence, for $i \geq 0$, we have

$$\alpha_k(i,t) = \frac{(\lambda(k)t)^i}{i!} e^{-\lambda(k)t}. \tag{3.50}$$

The differentiation between Equation (3.48) and Equation (3.50) is mainly of mathematical interest. In practice, the values for $\lambda(k)$ will change only slightly with $k$ (or not at all for close values of $k$). Hence, even if we wish to deduce the queue length distribution as well, we believe using Equation (3.50) is a reasonable approximation. Note that the basic queueing model itself already contains other approximations of relationships observed in practice.

Another example for the arrival distributions $\alpha_k$ is given in Section 3.5.1.

### 3.4.4 Embedded Markov chain immediately after departures

Instead of using our analytical approach to determine the queue length distribution immediately after departures, we can use the embedded Markov chain (immediately after departure) and numerically calculate the corresponding distribution. Therefore, we need to construct the transition matrix $\Phi$ with dimensions $(K+1) \times (K+1)$ where $\Phi_{ij}$ is the probability to transition from system state $i-1$, i.e., $i-1$ patients in the system after a departure to system state $j-1$, i.e., $j-1$ patients in the system after the next departure. We have to solve $q\Phi = q$ for $q = (q(0), \dots, q(K))$ with $\sum_{k=0}^{K} q(k) = 1$. Then, $q$ is the stationary distribution of the number of patients in the system immediately after a departure.

**Proposition 12.** *Setting* $\beta_k(i) = 1 - \sum_{i=0}^{K-k-1} \alpha_k(i)$*, the transition matrix looks as follows:*

$$\begin{pmatrix}
\nu(0)\alpha_1(0) & \nu(1)\alpha_1(1) + \rho(0)\alpha_1(0) & \dots & \nu(K-1)\beta_1(i) + \rho(K-2)\alpha_1(K-2) & \rho(K-1)\beta_1(i) \\
\nu(0)\alpha_1(0) & \nu(1)\alpha_1(1) + \rho(0)\alpha_1(0) & \dots & \nu(K-1)\beta_1(i) + \rho(K-2)\alpha_1(K-2) & \rho(K-1)\beta_1(i) \\
0 & \nu(1)\alpha_2(0) & \dots & \nu(K-1)\beta_2(i) + \rho(K-2)\alpha_2(K-3) & \rho(K-1)\beta_2(i) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & \nu(K-1)\beta_{K-1}(i) + \rho(K-2)\alpha_{K-1}(0) & \rho(K-1)\beta_{K-1}(i) \\
0 & 0 & \dots & \nu(K-1) & \rho(K-1)
\end{pmatrix}$$

*Proof.* First note that the probability to transition form state $0$ to any state $k \in \{0, \dots, K\}$ is equal to the probability to transition from state $1$ to $k \in \{0, \dots, K\}$. This is true because per definition in the time interval between the last departure that left an empty system and the first arrival to the empty system no other arrival can occur. Hence, the length of the time interval is irrelevant for the calculation of the transition probability. Further, only state transitions from state $k \in \{1, \dots, K\}$ to states $\{k-1, \dots, K\}$ are possible, since at most one patient can leave the system from one departure to the next one.

A transition from state $k \in \{1, \dots, K-1\}$ to state $k-1$ happens if the patient who leaves the system does not reschedule and no new patients arrive. Hence, this happens with probability $\nu(k-1)\alpha_k(0)$. For $k = K$, every newly arriving patient is rejected, hence the transition probability form state $K$ to state $K-1$ is $\nu(K-1)$.

A transition from state $k \in \{1, \dots, K-2\}$ to state $l \in \{k, \dots, K-2\}$ happens if the patient leaving does not reschedule and $l-k+1$ new patients arrive or if the patient reschedules and $l-k$ patients arrive. The corresponding transition probability is $\nu(l)\alpha_k(l-k+1) + \rho(l-1)\alpha_k(l-k)$.

A transition from state $k \in \{1, \dots, K-1\}$ to state $K-1$ happens if the leaving patient does not reschedule and more than $K-k-1$ new patients arrive or if the leaving patient reschedules and $K-k-1$ new patients arrive. Therefore, the transition probability is given as $\nu(K-1)(1 - \sum_{i=0}^{K-k-1} \alpha_k(i)) + \rho(K-2)\alpha_k(K-k-1)$.

A transition from state $k \in \{1, \ldots, K-1\}$ to state $K$ only happens if the leaving patient reschedules and more than $K-k-1$ new patients arrive. Hence, the transition probability is given as $\rho(K-1)(1 - \sum_{i=0}^{K-k-1} \alpha_k(i))$.

A transition from state $K$ to $K$ happens if the leaving patient reschedules. Every newly arriving patient is rejected. The corresponding transition probability is $\rho(K-1)$. □

Solving the system of linear equations $q\Phi = q$ and normalizing $q$ yields the queue length distribution at departure times in steady state.

### 3.4.5 Transient queueing model behavior

The paper of Green and Savin (2008) builds on the paper of Garcia et al. (2002). Garcia et al. (2002) derive analytical expressions of the time-dependent probability distribution of $M/D/1/K$ queues initialized in an arbitrary deterministic. Hence, Garcia et al. (2002) focus on transient results, whereas Green and Savin (2008) use this approach to deduce stationary results. Like Green and Savin (2008), we are mainly interested in stationary results. Still, we want to note here that similar to Garcia et al. (2002), the equations in Lemma 1 and Proposition 1 can be used to compute the probability distribution of the queue for any point in time via a recursive numerical integration. However, we cannot develop an analytical solution as Garcia et al. (2002) due to the more complex problem structure, mainly due to the queue length-dependent appointment request rate.

### 3.4.6 Other performance measures

Besides the different stationary distributions we have calculated so far other performance measure are of interest. We investigate the following rates:

- Average arrival rate,
- (Average) hidden demand rate,
- (Average) no-show rate,
- (Average) rescheduling rate,
- (Average) non-rescheduling rate,
- (Average) effective arrival rate and
- Average rate of rejected patients.

Here, we define a rate as a number of events per time unit given a certain queue length. The average rate is a number of events per unit time in steady state. Note that we need to have the stationary queue length distribution for some rates.

We define the average arrival rate as the average rate of arriving appointment requests in steady state excluding rescheduling requests. The average arrival rate is easy to calculate as $\sum_{k=0}^{K} \pi(k)\lambda(k)$. Here, we include $k = K$ in the summation as this reflects the arriving requests when the system capacity is reached. Note that this demand will be rejected.

Due to long indirect waiting times during which patients do not make new appointments because they already have an appointment it may happen that patients book less appointments than they normally would (if they experienced fewer or no indirect waiting times). Therefore, we define the hidden demand rate as the rate of patient requests that is not realized due to the fact that the patients are already waiting which is $\lambda(0) - \lambda(k)$ for a queue length of $k \in \{0, \ldots, K\}$. The average hidden demand rate can then be determined as $\sum_{k=0}^{K} \pi(k)(\lambda(0) - \lambda(k))$.

A system having $k$ patients immediately after the departure happens with rate $d^{nr*}(k-1)+d^{r*}(k)$. A proportion of $\gamma(k-1)$ of those departures will be no-shows. Hence, the no-show rate given a queue length of $k \in \{1, \ldots, K\}$ is $\gamma(k-1)(d^{nr*}(k-1) + d^{r*}(k))$. The average no-show rate is given by $\sum_{k=1}^{K} \gamma(k-1)(d^{nr*}(k-1) + d^{r*}(k))$.

The average rate of rescheduling patients is given by $\sum_{k=1}^{K} d^{r*}(k)$. Note that $d^{r*}(k)$ is the rate of rescheduling departures leaving $k$ patients behind independently of the queue length. Hence, in steady state, the rescheduling rate given a queue length of $k \in \{1, \ldots, K\}$ is $\frac{d^{r*}(k)}{\pi(k)}$.

The average rate of non-rescheduling patients is given by $\sum_{k=0}^{K-1} d^{nr*}(k)$. The queue length before the departure of a non-rescheduling patient leaving behind $k$ patients is $k+1$. Hence, in steady state, given a queue length $k \in \{1, \ldots, K\}$ the non-rescheduling rate is $\frac{d^{nr*}(k-1)}{\pi(k)}$.

The complete rate of patient requests also includes rescheduling patients. Therefore, we define the effective arrival rate as the total arrival rate plus the rate of patients rescheduling appointments given a certain queue length. Then, we have an effective arrival rate $\varepsilon(k)$ given by $\lambda(k) + \frac{d^{r*}(k)}{\pi(k)}$ for $k \in \{0, \ldots, K\}$ where we define $d^{r*}(0) = 0$. The average effective arrival rate is given by $\sum_{k=0}^{K} \pi(k)\lambda(k) + \sum_{k=1}^{K} d^{r*}(k)$.

A patient is rejected if the capacity of the queue is reached. Hence, the average rate of rejected patients is given by $\pi(K)\lambda(K)$. Note that a rescheduling patient is never rejected.

Instead of working with average rates we can also work with proportions with respect to served patients (including no-shows). We will investigate the steady state proportions:

- Proportion of no-shows,
- Proportion of rescheduling patients,
- Proportion of non-rescheduling patients,
- Proportion of rejected patients,
- Proportion of time that the queue is empty and
- Proportion of idle time.

In the following lemma we determine the relationship between proportion and average rate. This way the proportions of no-shows, rescheduling patients, non-rescheduling patients, and rejected patients are easily calculated.

**Lemma 6.** *A patient type with an average rate $r$, i.e., the average number of patients of that type served in a time unit, has a proportion of $p = \frac{rT}{1-\pi(0)}$ with respect to all served patients including no-shows.*

*Proof.* Let us assume a patient type with proportion $p$. In a large time interval $S$, the queue is non-empty during a time of $S(1 - \pi(0))$. If the queue is non-empty there is one patient served every $T$ units of time. Hence, during the time interval $S$ a number of $\frac{S(1-\pi(0))p}{T}$ patients of the considered type is served. Hence, the rate of patients of a certain type being served is $\frac{(1-\pi(0))p}{T}$. Rearranging the equation gives the result. □

The proportion of time that the queue is empty is given as $\pi(0)$.

Idle time as we understand it here can be due to an empty queue or due to a no-show. In a large time interval of length $S$ the queue will be empty during a total time of $\pi(0)S$. Further, we will experience $S(\sum_{k=0}^{K-1} \gamma(k)(d^{nr*}(k) + d^{r*}(k+1)))$ no-shows with a total time of idleness due to no-shows being $ST(\sum_{k=0}^{K-1} \gamma(k)(d^{nr*}(k) + d^{r*}(k+1)))$. Hence, the proportion of time the physician is idle is given by $\pi(0) + T(\sum_{k=0}^{K-1} \gamma(k)(d^{nr*}(k) + d^{r*}(k+1)))$.

From the proportion of idle time we can deduce the physician's utilization which is one minus the proportion of idle time, and hence $1 - \pi(0) - T(\sum_{k=0}^{K-1} \gamma(k)(d^{nr*}(k) + d^{r*}(k+1)))$.

We can also calculate the probability of getting a same-day appointment assuming that the physician works $\mu$ time slots per day of length $T$ using Proposition 10 and Corollary 11 with $k = 0$. Note, that this definition is different form the definition of the same-day appointment probability given in Green and Savin (2008). They define the same-day appointment probability as the probability that an arrival sees a queue length of less than $\mu$.

## 3.5 Model extensions

In this section, we present several model extensions to our basic queueing model. We start with integrating a queue length-dependent service time and even a generally distributed service time dependent on the queue length in Section 3.5.1. In Section 3.5.2, we consider a physician with panel patients and propose theoretical models for the appointment request rate. Finally, in Section 3.5.3, we explain how to relax the assumption that patients always take the next available appointment.

### 3.5.1 Variable appointment offerings

Until now we assumed a fixed service time $T$ for every patient. In this section we want to investigate how a queue length-dependent service time or even a random service time can be integrated into our queueing model.

Let us start with a physician adapting the service time to the workload hence to the length of the indirect queue. Note that in practice this does not mean that the physician actually offers shorter time slots when the appointment backlog is long. Instead, we measure the service time in days and assume that the physician offers an average number $\mu(k) = \frac{1}{T(k)}$ of time slots per day and that this number is dependent on the queue length $k$. Hence, for longer queue lengths, the physician switches to working longer hours to reduce the backlog faster. If the queue length shortens, the physician can switch back to working shorter hours. In a realistic setting, of course, patients book time slots on concrete days, which generally cannot be shifted. However, the physician can plan for short days after the last fully booked day in the schedule. For our model, we assume that the physician can switch to longer or shorter days immediately.

Going through Section 3.4, we observe that $T$ only appears explicitly in Propositions 6 and 8 and in Corollary 7. These are the propositions necessary to deduce the queue length distribution. Otherwise, for the computation of the queue length distribution immediately after departures and the indirect waiting time distribution, we have to consider the dependency on $T$ for $\alpha_k(i)$, i.e., the probability that $i$ patients arrive during a service period given that there are currently $k = 1, \ldots, K-1$ patients in the system and the possibly implicit dependency on $T$ for the appointment request function $\lambda$, the no-show function $\gamma$ and the rescheduling functions $r^n$ and $r^s$. Let us assume a queue length-dependent service time $T = T(k)$ where the service time of a patient is dependent on the queue length $k$ at the beginning of the service. Here, the patient to be served is included. Hence, considering the arrival probabilities during a service time, we just have to set $\alpha_k(i) = \alpha_k(i, T(k))$ where $\alpha_k(\cdot, t)$ can, for example, be an arrival distribution as defined in Section 3.4.3. Here, we have to keep in mind that the Equations (3.48), (3.49) and (3.53) are compatible with Lemmas 1 and 2 whereas Equation (3.50) is not. Hence, if we are only interested in the queue length distribution immediately after departures and the indirect waiting time distribution, we do not need Lemmas 1 and 2, Propositions 6 and 8 and Corollary 7. We simply insert an arrival distribution of our choice.

Suppose we only want to determine the indirect waiting time distribution and assume that the number of patients arriving during a time interval of length $t$ follows a Poisson distribution as in Equation (3.50). Note that both $\lambda$ and $T$ only appear together as the product $\lambda T$. Hence, if we assume no implicit dependencies on $T$, we see that changing the service time $T$ is equivalent to keeping $T$ constant and changing $\lambda$ instead.

If we want to determine the queue length distribution as well, we have to stick with an appropriate arrival distribution and we need to adjust Propositions 6 and 8 and Corollary 7. In the deduction of Equation (3.29) of Proposition 6, we determine the mean duration of a busy period by multiplying the mean number of patients served during a busy period with the service duration $T$. Now, instead of $T$, we need to insert the expected value of $T$. The probability that there are $k \geq 2$ patients present at the beginning of a service is equivalent to the probability that there are $k \geq 2$ left in the queue after a departure which is $q(k)$. One patient present at the beginning of a service can be the result of one of two situations: Either a departure occurred leaving behind one patient or a patient arrived to an empty queue. The probability of the first situation is $q(1)$. The probability of the second situation is then given by $1 - \sum_{k=1}^{K} q(k) = q(0)$. Hence, the expected value of $T$ is $ET = q(0)T(1) + \sum_{k=1}^{K} q(k)T(k)$. Hence, Propositions 6 and 8 and Corollary 7 remain valid if we insert $ET$ instead of $T$ into the formulas. This model extension results into an $M_n/D_n/1/K$ queueing model.

We can even go further and assume a general distributed service time dependent on the queue length. For $k = 1, \dots, K$, we denote the density function of the general distributed service time by $b_k$ with mean $\overline{b_k}$. We assume that $b_k$ is absolutely continuous and that $\overline{b_k}$ is finite and positive for all $k = 1, \dots, K$. Then, we can define $\alpha_k(i) = \int b_k(t)\alpha_k(i, t)dt$ for $k = 1, \dots, K - 1$. This way, we can approximately model a physician who randomly changes the appointment offerings dependent on the queue length. Note again that the change in the service time is immediate in the mathematical model while in reality the physician will probability change the appointment offerings for the next not fully booked day. Again, we have to adjust Propositions 6 and 8 and Corollary 7 if we want to determine the queue length distribution. Here, the expected value of the service time is given as $ET = q(0)\overline{b_1} + \sum_{k=1}^{K} q(k)\overline{b_k}$. This model extension results into an $M_n/G_n/1/K$ queueing model.

We also have to adjust the calculation of the indirect waiting time in days in Proposition 10 and in Corollary 11. In this case, we define $\mu(k)$ as the expected appointment offerings for a queue of length $k$, i.e., $\mu(k) = \int b_k(t)\frac{1}{t}dt$. Further, we set $\mu(0) = 1$ per default.

**Corollary 13.** *Let $\mu(k) > 0, \forall k = 0, \dots, K-1$ and $K_l = \left\{k \in \{0, \dots, K\} : l = \left\lfloor \frac{k}{\mu(k)} \right\rfloor \right\}$. Then, an approximate distribution of indirect waiting times $\iota$ in days is given by:*

$$\begin{aligned}\iota(l) = &\sum_{k\in K_l} \frac{(l+1)\mu(k) - k}{\mu(k)} p(k) \\ &+ \sum_{k\in K_{l-1}} \frac{k - (l-1)\mu(k)}{\mu(k)} p(k), \qquad l = 0, \dots, \max_{k=1,\dots,K} \left\lfloor \frac{k}{\mu(k)} \right\rfloor + 1. \end{aligned} \tag{3.51}$$

Let us investigate the appointment request function $\lambda$, the no-show function $\gamma$ and the rescheduling functions $r^n$ and $r^s$ in more detail. All four functions depend on the queue length. However, as explained in the introduction, part of this dependency may reflect a dependency on the indirect waiting time. Hence, for example, we may have $\lambda(k) = \lambda(k, l(k))$ with $l(k)$ being the average number of waiting days in case of a queue of length $k$. We know from Proposition 10 and Corollaries 11 and 13 that a patient arriving to a queue with length $k = l\mu(k) + i, l \in \mathbb{N}_0, 0 \leq i < \mu(k)$ has a probability of $\frac{\mu(k)-i}{\mu(k)}$ to wait $l$ days and a probability of $\frac{i}{\mu(k)}$ to wait $l+1$ days. Therefore, the average number of waiting days is:

$$l(k) = \frac{\mu(k) - i}{\mu(k)} l + \frac{i}{\mu(k)}(l+1) = l + \frac{i}{\mu(k)} = \frac{k}{\mu(k)}. \tag{3.52}$$

In the following, if we will make use of Equation (3.52) whenever we assume a dependency on the indirect waiting time.

When working with a non-constant service time, we need to change the formula to deduce a proportion given an average rate in Section 3.4.6. Here again, we substitute the service time $T$ by the expected value of $T$. Further, we can add two other interesting performance measures: The expected service time and the expected number of daily appointment offerings. As already explained above, the expected service time is given as $q(0)T(1) + \sum_{k=1}^{K} q(k)T(k)$ in case of a queue length-dependent service time and as $ET = q(0)\overline{b_1} + \sum_{k=1}^{K} q(k)\overline{b_k}$ in case of a random service time. Similarly, the expected number of daily appointment offering is given as $q(0)\mu(1) + \sum_{k=1}^{K} q(k)\mu(k)$.

### 3.5.2 Physicians with panel patients

In this section, we consider the case of a physician treating only panel patients, i.e., patients that visit regularly. Let the panel size, i.e., the number of panel patients, be $N$. In the simplest case, we can assume that every patient has the same individual appointment request rate $\eta$ independent of the current queue length as in Green and Savin (2008). Then, in the following, we will speak of a constant total arrival rate $\lambda = N\eta$.

Next, we assume that panel patients who have already booked an appointment and are waiting for their treatment will not book another appointment. We assume an individual appointment request rate $\eta$ for patients not waiting for treatment or getting treatment. If there are $k$ patients waiting in the queue, only $N - k$ patients will potentially book appointments. Hence, the appointment request rate is given by $\lambda(k) = \eta(N - k)$ if the queue length is $k$. In the following, we present three different theoretical models for the appointment request rate $\lambda$ in the case of a physician that only treats panel patients.

For the first part of this section, we assume that $\eta$ is constant and, therefore, not dependent on the queue length. In the following, we will shortly refer to this modeling as the case with a constant individual arrival rate (and a queue length-dependent total arrival rate). For this special case, we can deduce simple formulas for $\alpha_k(i, t)$, i.e., the probability that $i$ patients arrive during a time interval of length $t$ with no departures given that there are $k = 1, \dots, K - 1$ patients in the system at the beginning of the time interval. The number of appointment requests follows a Binomial distribution $B_{(N-k),p}$ with $(N - k)$ trials, and $p$ being the probability that one specific patient requests an appointment during a time interval of length $t$. We know that the time until an individual patient's request is exponentially distributed with parameter $\eta$. Therefore, the probability of the next request happening in a time interval of length $t$ is $p = 1 - e^{-\eta t}$. We then obtain the probability that $i$ patients arrive during $t$, given that $k$ patients are in the queue:

$$\alpha_k(i, t) = \binom{N-k}{i} \left(1 - e^{-\eta t}\right)^i \left(e^{-\eta t}\right)^{(N-k-i)}. \tag{3.53}$$

Note, that Equation 3.53 is a special case of Equation 3.49 with $m = -\eta$ and $n = N\eta$.

Let us go back to the beginning and assume that $\eta$ may, in fact, depend on the queue length. Hence, we have $\lambda(k) = \eta(k)(N - k)$. In the following, we will shortly refer to this modeling as the case with a non-constant individual arrival rate. We assume an average number of appointments (rescheduling appointments included) $\delta$ per time unit that a panel patient would book and attend if there was no indirect waiting time. In reality, patients do experience indirect waiting times. Let us assume that in this situation, they try to stick to their attended appointment rate $\delta$ if possible through adapting $\eta, r^n$, and $r^s$. Therefore, in general, the functions $\eta, r^n$, and $r^s$ should increase with increasing indirect waiting times. Of course, if indirect waiting times go up too high, patients will not be able to keep their appointment rate $\delta$ no matter what the choice of $\eta, r^s$ and $r^n$ is. Considering all three functions $\eta, r^s$ and $r^n$ simultaneously is quite complex and leaves room for much flexibility. Therefore, we will assume for a moment that both $r^s$ and $r^n$ are zero.

**Lemma 7.** *Assuming that the rescheduling probabilities $r^s$ and $r^n$ are zero and that a patients sticks to the attended appointment rate $\delta$ if possible, the individual arrival rate of a patient is given by:*

$$\eta(k) = \frac{1}{\frac{1-\gamma(k)}{\delta} - l(k)}, \qquad k \textit{ such that } \frac{1-\gamma(k)}{\delta} \geq l(k) + 1, \tag{3.54}$$

$$\eta(k) = 1, \qquad k \textit{ such that } \frac{1-\gamma(k)}{\delta} < l(k) + 1. \tag{3.55}$$

*Proof.* We consider a patient that experiences a queue length of $k$ and an indirect waiting time of $l(k)$ days as given by Equation (3.52). After the day of an attended appointment, there are two different outcomes. The patient requests a new appointment with rate $\eta(k)$ and waits $l(k)$ days and then either shows up or not. If patients do not show up, they will again request an appointment with rate $\eta(k)$ and wait $l(k)$ days and so on. In summary, the time difference in days between two consecutive attended appointments is given by:

$$\begin{aligned}
&(1-\gamma(k)) \sum_{m=0}^{\infty} \gamma(k)^m (m+1) \left( l(k) + \frac{1}{\eta(k)} \right) \\
=& (1-\gamma(k)) \left( l(k) + \frac{1}{\eta(k)} \right) \sum_{m=0}^{\infty} \gamma(k)^m (m+1) \\
=& (1-\gamma(k)) \left( l(k) + \frac{1}{\eta(k)} \right) \sum_{m=0}^{\infty} \frac{\partial}{\partial \gamma} \gamma(k)^{m+1} \\
=& (1-\gamma(k)) \left( l(k) + \frac{1}{\eta(k)} \right) \frac{\partial}{\partial \gamma} \sum_{m=0}^{\infty} \gamma(k)^{m+1} \\
=& (1-\gamma(k)) \left( l(k) + \frac{1}{\eta(k)} \right) \frac{\partial}{\partial \gamma} \gamma(k) \sum_{m=0}^{\infty} \gamma(k)^{m} \\
=& (1-\gamma(k)) \left( l(k) + \frac{1}{\eta(k)} \right) \frac{\partial}{\partial \gamma} \frac{\gamma(k)}{1-\gamma(k)} \\
=& (1-\gamma(k)) \left( l(k) + \frac{1}{\eta(k)} \right) \left( \frac{1}{1-\gamma(k)} + \frac{\gamma(k)}{(1-\gamma(k))^2} \right)
\end{aligned}$$

$$\begin{aligned}
=& (1-\gamma(k)) \left( l(k) + \frac{1}{\eta(k)} \right) \left( \frac{1}{(1-\gamma(k))^2} \right) \\
=& \left( l(k) + \frac{1}{\eta(k)} \right) \left( \frac{1}{1-\gamma(k)} \right).
\end{aligned} \tag{3.56}$$

Due to our assumption (3.56) should be equal to $\frac{1}{\delta}$ if possible:

$$\frac{1}{\delta} = \left(l(k) + \frac{1}{\eta(k)}\right)\left(\frac{1}{1-\gamma(k)}\right).$$

Rearranging yields

$$\eta(k) = \frac{1}{\frac{1-\gamma(k)}{\delta} - l(k)}.$$

Of course this definition for $\eta(k)$ only makes sense for queue lengths $k$ such that $\frac{1-\gamma(k)}{\delta} > l(k)$. Remember that we use days as the unit of time. It is then reasonable to limit the individual appointment request rate $\eta$ to a maximal value of $1$ corresponding to one appointment request per day. □

Finally, to be able to make a comparison with the results of Green and Savin (2008), let us assume $r^s = 0$ and $r^n = 1$.

**Lemma 8.** *Assuming that $r^s = 0$ and $r^n = 1$ and that a patients sticks to the attended appointment rate $\delta$ if possible, the individual arrival rate of a patient is given by:*

$$\eta(k) = \frac{1}{\frac{1}{\delta} - \frac{l(k)}{1-\gamma(k)}}, \qquad k \text{ such that } \frac{1}{\delta} \geq \frac{l(k)}{1-\gamma(k)} + 1, \tag{3.57}$$

$$\eta(k) = 1, \qquad k \text{ such that } \frac{1}{\delta} < \frac{l(k)}{1-\gamma(k)} + 1. \tag{3.58}$$

*Proof.* We consider a patient that experiences a queue length $k$ and an indirect waiting time of $l(k)$ days. After the day of an attended appointment, there are again two different outcomes. The patient requests a new appointment with rate $\eta(k)$ and waits $l(k)$ days and then either shows up for his or her requested appointment or not. If patients do not show up, they will reschedule with probability $1$ and wait for $l(k)$ days and so on.

In summary, the time difference in days between two consecutive attended appointments is given by:

$$\begin{aligned}
&(1-\gamma(k))\sum_{m=0}^{\infty}\gamma(k)^m\left((m+1)\,l(k)+\frac{1}{\eta(k)}\right)\\
=&(1-\gamma(k))\sum_{m=0}^{\infty}\gamma(k)^m\,(m+1)\,l(k)+(1-\gamma(k))\sum_{m=0}^{\infty}\gamma(k)^m\frac{1}{\eta(k)}\\
=&(1-\gamma(k))\,l(k)\frac{1}{(1-\gamma(k))^2}+(1-\gamma(k))\,\frac{1}{1-\gamma(k)}\frac{1}{\eta(k)}\\
=&\frac{l(k)}{1-\gamma(k)}+\frac{1}{\eta(k)}. \qquad (3.59)
\end{aligned}$$

Again, (3.59) should be equal to $\frac{1}{\delta}$ if possible:

$$\frac{1}{\delta}=\frac{l(k)}{1-\gamma(k)}+\frac{1}{\eta(k)}.$$

Rearranging yields

$$\eta(k)=\frac{1}{\frac{1}{\delta}-\frac{l(k)}{1-\gamma(k)}}.$$

Again, this definition for $\eta(k)$ only makes sense for $k$ such that $\frac{1}{\delta}>\frac{l(k)}{1-\gamma(k)}$. Limiting $\eta$ to a maximal value of $1$ corresponding to one appointment per day gives the result. □

#### Patient-dependent individual arrival rates

So far, we assumed that every patient has the same individual arrival rate function $\eta$. In reality, given a fixed queue length, patients have different individual arrival rates. We will include this into the model via dividing the panel size $N$ into subgroups of sizes $N_1, N_2, \dots, N_c$ with $\sum_{i=1}^{c} N_i = N$ where each subgroup $i$ of panel patients has the same individual arrival rate function $\eta_i$. In the following, we illustrate the model extension for patient dependent individual arrival rates using two subgroups with sizes $N_1$ and $N_2$ with $N_1+N_2=N$. Note that the same procedure can be used for more than two subgroups with some more calculation effort.

Having two panel-subgroups with sizes $N_1$ and $N_2$ with individual arrival rate functions $\eta_1$ and $\eta_2$, we have to adjust the calculation of the total arrival rate function $\lambda$. For a single panel patient group of size $N$ with individual arrival rate function $\eta$, we have $\lambda(k)=\eta(k)(N-k)$. To model two patients groups exactly, we need $\lambda$ to be dependent on two variables: the number of patients of subgroup 1 and the number of patients of subgroup 2. To avoid the model becoming too complicated, we approximate this two-dimensional problem using a total arrival rate function that is still only dependent on the queue length. We do this via fixing a partitioning into the two subgroups for every queue length.

**Lemma 9.** *The approximate total arrival rate for a queue length $k = 0, \dots, K$ for two panel subgroups with sizes $N_1$ and $N_2$ and individual arrival rates $\eta_1(k) < \eta_2(k)$ is given as $\lambda(k) = (N_1 - a(k)k)\eta_1(k) + (N_2 - (1 - a(k))k)\eta_2(k)$ with*

$$a(k) = \frac{-B(k) + \sqrt{B(k)^2 - 4A(k)C(k)}}{2A(k)} \tag{3.60}$$

*with $A(k) = k\eta_2(k) - k\eta_1(k)$, $B(k) = N_1\eta_1(k) + N_2\eta_2(k) + k\eta_1(k) - k\eta_2(k)$ and $C(k) = -N_1\eta_1(k)$.*

*Proof.* Given a queue length $k$, we define $a(k)$ as the proportion of subgroup 1 patients that are part of $k$. Then, we experience an arrival rate of subgroup 1 patients of $(N_1 - a(k)k)\eta_1(k)$ and an arrival rate of $(N_2 - (1 - a(k))k)\eta_2(k)$ of subgroup 2 patients. The proportion of subgroup 1 patients $a(k)$ should then correspond to the proportion of subgroup 1 patient arrivals in relation to the total arrivals. Hence, we have:

$$a(k) = \frac{(N_1 - a(k)k) \cdot \eta_1(k)}{(N_1 - a(k)k) \cdot \eta_1(k) + (N_2 - (1 - a(k))k) \cdot \eta_2(k)}.$$

This quadratic equation can be solved for $a(k)$ with solutions:

$$a(k)_{1,2} = \frac{-B(k) \pm \sqrt{B(k)^2 - 4A(k)C(k)}}{2A(k)}.$$

Assuming $\eta_1(k) < \eta_2(k)$, we determine that $a(k)_1 \geq 0$ and $a(k)_2 \leq 0$. Since $a(k)$ has to be non-negative, we conclude that $a(k) = a(k)_1$. Having determined $a(k)$, we obtain the total arrival rate $\lambda(k) = (N_1 - a(k)k)\eta_1(k) + (N_2 - (1 - a(k))k)\eta_2(k)$ that can be plugged into the queueing model. Note, that we still assume the same no-show and rescheduling behavior for all patients. □

#### Mix of panel and non-panel patients

So far, we have assumed a physician that only treats panel patients. In reality, physicians often experience the largest share of their patient demand from panel patients and treat non-panel patients, i.e., patients who only show up once. To model non-panel patients, we assume a non-panel arrival rate $r^{np}$ that may or may not dependent on the queue length. As in Section 3.5.2, we would need a total arrival rate function $\lambda$ dependent on two variables: the number of panel patients and the number of non-panel patients. Again, we approximate the two-dimensional problem using a total arrival rate function that is still only dependent on the queue length. We do this via fixing a partitioning in panel and non-panel patients for every queue length.

**Lemma 10.** *The approximate total arrival rate for a queue length $k = 0, \ldots, K$ for a mix of panel and non-panel patients with panel size $N$ and a non-panel patients arrival rate $r^{np}(k)$ is given as $\lambda(k) = (N - a(k)k)\eta(k) + r^{np}(k)$ with*

$$a(k) = \frac{(N\eta(k) + k\eta(k) + r^{np}) - \sqrt{(N\eta(k) + k\eta(k) + r^{np})^2 - 4(k\eta(k))(N\eta(k))}}{2(k\eta(k))}. \tag{3.61}$$

*Proof.* Given a queue length $k$, we define $a(k)$ as the proportion of panel patients that are part of the patients in the queue $k$. Then, we experience an arrival rate of panel patients of $(N - a(k)k)\eta(k)$ and an arrival rate of $r^{np}(k)$ of non-panel patients. The proportion of panel patients $a(k)$ should then correspond to the proportion of panel patient arrivals in relation to the total arrivals. Hence, we have:

$$a(k) = \frac{(N - a(k)k)\eta(k)}{(N - a(k)k)\eta(k) + r^{np}(k)}.$$

Again, this equation can be solved for $a(k)$ with solutions:

$$a(k)_{1,2} = \frac{(N\eta(k) + k\eta(k) + r^{np}) \pm \sqrt{(N\eta(k) + k\eta(k) + r^{np})^2 - 4(k\eta(k))(N\eta(k))}}{2(k\eta(k))}.$$

Both, $a(k)_1$ and $a(k)_2$ are non-negative, but $a(k)$ should also fulfill $a(k) \leq 1$ to reflect the fact that $a(k)$ is a proportion. For $a(k)_1 = \frac{(N\eta(k)+k\eta(k)+r^{np})+\sqrt{(N\eta(k)+k\eta(k)+r^{np})^2-4(k\eta(k))(N\eta(k))}}{2(k\eta(k))}$ we have:

$$a(k)_1 > \frac{N\eta(k) + k\eta(k)}{2k\eta(k)} = \frac{N + k}{2k} \geq \frac{2k}{2k} = 1.$$

Hence, we have $a(k) = a(k)_2$ and we obtain the total arrival rate $\lambda(k) = (N - a(k)k)\eta(k) + r^{np}(k)$ that can be used in the queueing model. □

Of course, it is also possible to combine a mix of panel and non-panel patients with panel patient dependent individual arrival rates via considering several panel subgroups as in Section 3.5.2

and a non-panel patient arrival rate. Again, note that we still assume the same no-show and rescheduling behavior for all patients.

### 3.5.3 Capacity division

So far, we have considered single queues. Here, we want to consider a physician or a practice that experiences several demand streams with designated capacities, i.e., appointment offerings. For example, a physician may block slots for specific patient types, e.g., privately insured patients are always treated in the afternoon. Here is another example. Talking to different practices, we realized that those offering online appointment booking to their patients often have to deal with two systems simultaneously: the administrative practice software and the online appointment booking system. Often those two systems are not interconnected. Consequently, physicians block slots in their appointment schedule managed in their administrative software and offer those slots for online booking. In certain time intervals, they check the received online bookings and transfer them to their general appointment schedule in the administrative system. In general, we distinguish the following cases:

- independent demand streams with dedicated distinct capacities,
- independent demand streams with overlapping capacities,
- interconnected demand streams with dedicated distinct capacities,
- interconnected demand streams with overlapping capacities.

The example of the physician blocking slots for privately insured patients can be represented by the case of independent demand streams with dedicated distinct capacities because a patient can either be privately insured or not. If, for example, non-used capacity by a demand stream is unblocked to be used by another demand stream after a certain time frame, we are in the case of independent demand streams with overlapping capacities.

The example of the two non-connected appointment booking systems can be represented by interconnected demand streams with dedicated distinct capacities. The demand streams are interconnected because a patient can decide to book online or, for example, via telephone. Having booked an appointment, independent of the booking mode used, the patient will probably not book another appointment. If, for example, unused online booking capacity is unblocked for general booking after a certain time frame, we are in the case of interconnected demand streams with overlapping capacities.

In the case of independent demand streams with dedicated distinct capacities, we can model each demand stream together with its dedicated capacity as a single queue. In the case of independent demand streams with overlapping capacities, we will present an example where we build an approximate model where every demand stream has its own queue, but the queue parameters are dependent on the stationary distributions of the other queues. In the cases of interconnected demand streams with dedicated distinct or overlapping capacities, we would need to build more-dimensional queueing models. We will not consider those cases in further detail here.

In the following, we explore independent demand streams with dedicated distinct or overlapping capacities in mode detail. We explain how to decide on a fair capacity division between two independent demand streams and how we can relax the assumption that patients always book the next available appointment using several interconnected queues.

#### Fair capacity division between two independent demand streams

We assume a physician with two independent demand streams with designated distinct capacities. We want to answer the following question: How should the total available capacity be divided between the two streams such that patients of both streams experience the same expected indirect waiting time? Here, we restrict ourselves to the case of constant total appointment offerings $\mu$.

To answer the question, we model the two demand streams' appointment schedules as indirect queues using the subscripts $1$ and $2$ to distinguish the parameters of the two streams. Let us assume that both queues have the same booking horizon $B$ in days which leads to the queueing capacities $K_1 = \lfloor \mu_1 B \rfloor$ and $K_2 = \lfloor \mu_2 B \rfloor$. Hence, we are searching for $\mu_1$ and $\mu_2$ with $\mu_1 + \mu_2 = \mu$ such that the expected indirect waiting times in days are the same for the two patient groups. In general, we cannot provide an analytical solution but have to approximate by determining the indirect waiting times for different capacity divisions. Last but not least, comparing the two queueing models to one single queueing model with appointment offerings $\mu$ and booking horizon $B$ in days (and therefore a queue capacity of $K = \lfloor \mu B \rfloor$), we can quantify the benefit of having one schedule instead of two schedules.

Let us investigate the example of a physician with panel patients in more detail. Let us assume that the total panel size $N$ can be divided into to subgroups $N_1$ and $N_2$ with their own capacities $\mu_1$ and $\mu_2$. Further, we assume the same arrival behavior for both patient groups. Let the expected indirect waiting times in days be $m_1$ and $m_2$ for the two patient groups, respectively.

For example, if we have $\lambda_{1,2} = N_{1,2}\eta$, we could formulate the problem as:

$$\min |m_1 - m_2| \tag{3.62}$$

s.t.

$$\mu_1 + \mu_2 = \mu \tag{3.63}$$

$$K_{1,2} = \lfloor \mu_{1,2} B \rfloor \tag{3.64}$$

$$\lambda_{1,2} = N_{1,2}\eta \tag{3.65}$$

$$\mu_1, \mu_2 > 0. \tag{3.66}$$

#### Interconnected queues representing patients' time preferences

Until now, we assumed that patients always book the next available appointment. Of course, this is not true in general. For example, patients might book check-up appointments further into the future. Also, patients with non-urgent matters may have preferences concerning the day and the time of the day of their appointment. Here, we will not investigate the general problem but a simplified version where we make particular assumptions on patients' booking behavior.

We assume several appointment types where one type is defined by a number of slots that a patient tries to book into the future. Suppose the intended slot is already booked, then the patient books the next available slot after that. For example, we could have one appointment type, e.g., urgent requests, where patients want to be treated right away and, therefore, always book the next available appointment. If patients only book appointments of that type, we can apply our queueing model directly. Besides, we could have check-up appointments that are always booked three weeks into the future. If patients only book appointments of that type, we can again apply our queueing model. We can virtually shift the booking time to three weeks later, resulting in a queue where every patient books the next available appointment. Note that defining appointment types this way allows us to reuse the queueing model. Further, for every appointment type considered, the queue parameters can be set individually. For example, an indirect waiting period of three weeks will be different for a patient that booked the next available appointment but would have booked an earlier appointment if possible and a patient that wanted to book an appointment further in the future. In the following, we will first investigate the setting of one appointment type where patients try to book $x_1 > 0$ slots into the future. Then, we consider two appointment types together.

Let us assume that every patient tries to book $x_1$ slots into the future and, if that is not possible, books the next available appointment after that. We further assume a fixed number of average appointment offerings per day $\mu$. Then, the queue forms at $x_1$ slots into the future. Therefore, we need to adjust dependencies on the indirect waiting time and the number of patients waiting or getting treated. Let $l_1(k)$ be the average number of indirect waiting days and $m_1(k)$ be the average number of patients waiting for treatment given a queue length of $k$. For $l_1$ we can use Equation (3.52) adding the $x_1$ slots, hence we have $l_1(k) = \frac{x_1+k}{\mu}$. The more complicated aspect is that patients whose appointments are scheduled less than $x_1$ slots into the future are not part of the queue anymore and, therefore, not represented in the queue length. To determine the number of patients whose appointments are scheduled less than $x_1$ slots into the future, we would have to keep track of the queue's history. However, this is not possible without building a much more complicated model. Therefore, we simplify by working with a constant number of patients whose appointments are scheduled less than $x_1$ slots into the future. In steady state, the proportion of time that the queue is non-empty is $(1 - \pi_1(0))$. Hence, we assume an average number of patients whose appointments are scheduled less than $x_1$ slots into the future of $z_1 = (1-\pi_1(0))x_1$. Then, the total number of patients waiting for treatment or getting treatment is approximately given by the current queue length plus $z_1$. Hence, we have $m_1(k) = k + z_1$.

Let us investigate the setting of a physician with only panel patients further. We assume a panel size of $N$. The function $l_1$ should be used to define the no-show function and the rescheduling functions, if necessary. We have an appointment request rate $\lambda_1(k) = \eta_1(k)(N - m_1(k)) = \eta_1(k)(N_1 - k - z_1)$ with $(1 - \pi_1(0))x_1 = z_1$. We need to solve this problem numerically testing different values of $z_1$ until $(1 - \pi_1(0))x_1 = z_1$ is approximately fulfilled. Here, bigger values of $z_1$ result in a lower appointment request rate, which, in turn, leads to a higher probability of an empty queue and hence to smaller values of $(1 - \pi_1(0))x_1$.

Now, let us assume that we have two appointment types, one where patients book at least $x_1$ slots into the future (type 1) and one where patients book at least $x_2 < x_1$ slots into the future (type 2). We further assume a fixed number of average appointment offerings per day $\mu$ for both appointment types. In case of high utilization, both patient types behave equally and book the next available appointment, which can, in principle, be represented by one single queueing model. However, then, we can only define one set of queue parameters for both appointment types.

In general, we can approximately consider two interconnected queueing models for the two appointment types. Type 1 can be handled as explained above via virtually shifting the booking time to $x_1$ slots later. Type 2 forms its queue in the empty phases of the first queue starting form $x_2$ slots in the future, where we ignore that those empty phases are not always of lengths that are multiples of the service time. We interconnect the two queues such that the parameter input for one queue is dependent on the queue length distribution of the other queue.

**Lemma 11.** *Given two appointment types that are booked at least $x_1$ and $x_2$ slots into the future with $x_2 < x_1$ with a common queue capacity $K$ and a common number of appointments per day $\mu = \frac{1}{T}$, the approximate average number of patients waiting for treatment $m_{1,2}(k)$ and the approximate average number of waiting days $l_{1,2}(k)$ given a queue length of $k \in \{0, \dots, K_1\}$ or $k \in \{0, \dots, K_2\}$ of appointment type 1 or 2, respectively, are:*

$$m_1(k) = k + (1 - \pi_1(0))x_1, \tag{3.67}$$

$$m_2(k) = k + (1 - \pi_2(0))\pi_1(0)x_2, \tag{3.68}$$

$$l_1(k) = \frac{k + \sum_{i=\lceil (x_1 - x_2)\pi_1(0) \rceil}^{K_2} i\pi_2(i)}{\mu}, \tag{3.69}$$

$$l_2(k) = \frac{x_2 + \frac{k}{\pi_1(0)}}{\mu}, \qquad \forall 0 < k \leq (x_1 - x_2)\pi_1(0), \tag{3.70}$$

$$l_2(k) = \frac{x_2 + k + (x_1 - x_2)(1 - \pi_1(0)) + EQL_1}{\mu}, \qquad \forall (x_1 - x_2)\pi_1(0) < k \leq K_2. \tag{3.71}$$

*The adjusted queue capacities are:*

$$K_1 = K - x_1 - \sum_{i=\lceil (x_1 - x_2)\pi_1(0) \rceil}^{K_2} i\pi_2(i), \tag{3.72}$$

$$K_2 = K - x_2 - (x_1 - x_2)(1 - \pi_1(0)) - EQL_1. \tag{3.73}$$

*The adapted service times are:*

$$T_1(k) = \frac{\overline{\varepsilon}_2(k) + \varepsilon_1(k)}{\varepsilon_1(k)} T, \tag{3.74}$$

$$T_2(k) = \frac{\frac{1 - \pi_1(0)}{T} + \varepsilon_2(k)}{\varepsilon_2(k)} T, \qquad \forall 0 < k \leq (x_1 - x_2)\pi_1(0), \tag{3.75}$$

$$T_2(k) = \frac{\overline{\varepsilon}_1(k) + \varepsilon_2(k)}{\varepsilon_2(k)} T, \qquad \forall (x_1 - x_2)\pi_1(0) < k \leq K_2. \tag{3.76}$$

*Here, $\pi_1$ and $\pi_2$ are the queue length distributions in steady state, $EQL_1$ is the expected queue length of the first queue, $\varepsilon_1$ and $\varepsilon_2$ are the effective arrival rates and $\overline{\varepsilon}_1$ and $\overline{\varepsilon}_2$ are average effective arrival rates with*

$$\overline{\varepsilon}_1 = \sum_{k=0}^{K_1} \pi_1(k)\lambda_1(k) + \sum_{k=1}^{K_1} d_1^{r*}(k), \tag{3.77}$$

$$\overline{\varepsilon}_2 = \sum_{k=\lceil (x_1 - x_2)\pi_1(0) \rceil}^{K_2} \left(\pi_2(k)\lambda_2(k) + d_2^{r*}(k)\right). \tag{3.78}$$

*Proof.* We first define $m_{1,2}(k)$ given a queue length $k$ for the two appointment types, respectively. Just as in the single queue case, the number of patients waiting for a type 1 or type 2 appointments can be defined as $m_{1,2}(k) = k + z_{1,2}$ with $z_1 = (1 - \pi_1(0))x_1$ and $z_2 = (1 - \pi_2(0))\pi_1(0)x_2$.

Next, we define $l_2$. For slots booked between $x_2$ and $x_1$ slots into the future, type 2 appointments are made in the formerly empty phases of the first queue. Hence, for a queue length $k \leq (x_1 - x_2)\pi_1(0)$ the average number of booked slots by both appointments types is $\frac{k}{\pi_1(0)}$. Hence, we obtain Equation (3.70). For $k > (x_1 - x_2)\pi_1(0)$, on average, there are type 2 appointment booked which are further than $x_1$ slots into future. Thus, we add the expected queue length of the first queue to the average number of appointments booked by both appointment types. The result is Equation (3.71).

Now, we define $l_1$. Here, we need to add the average number of type 2 appointments booked more than $x_1$ slots into the future which is $\sum_{i=\lceil (x_1-x_2)\pi_1(0)\rceil}^{K_2} i\pi_2(i)$ where $K_2$ is the queue capacity of the second queue. Hence, we obtain (3.69).

We assume a fixed booking horizon of $K$ slots for both appointment types together. Then for type 1, the booking horizon $K_1$ should be set to the original booking horizon $K$ minus $x_1$ and minus the expected number of occupied slots by type 2 appointments that lie more than $x_1$ slots into the future. The booking horizon $K_2$ for type 2 appointments is given as the original booking horizon minus $x_2$ minus the expected number of slots occupied by type 1 appointments that lie more than $x_2$ and less than $x_1$ appointments into the future and minus the expected queue length of type 1 appointments.

It remains to show the adjustments to the service times. We integrate the fact that the queue capacity is split between both appointment types via virtually extending the service time dependent on the queue lengths. Type 1 appointments arrive with an effective arrival rate of $\varepsilon_1(k) = \lambda_1(k) + \frac{d_1^{r*}(k)}{\pi_1(k)}$ given a queue of length $k$. The average effective arrival rate of type 2 appointments that are booked at least $x_1$ slots into the future is approximately given by $\bar{\varepsilon}_2 = \sum_{k=\lceil (x_1-x_2)\pi_1(0)\rceil}^{K_2} (\pi_2(k)\lambda_2(k) + d_2^{r*}(k))$. Considering a number $S$ of served patients, we experience a share of $\frac{\varepsilon_1(k)}{\varepsilon_1(k)+\bar{\varepsilon}_2}$ of type 1 appointments. Hence, the (virtual) service times for a type 1 appointment is given as $ST : \frac{\varepsilon_1(k)}{\varepsilon_1(k)+\bar{\varepsilon}_2}S = \frac{\varepsilon_1(k)+\bar{\varepsilon}_2}{\varepsilon_1(k)}T$.

For type 2 appointments given a queue length of $k \leq (x_1 - x_2)\pi_1(0)$ the effective arrival rate is given as $\varepsilon_2(k) = \lambda_2(k) + \frac{d_2^{r*}(k)}{\pi_2(k)}$. The average arrival rate of type 1 patients is $\frac{1-\pi_1(0)}{T}$ which shows the result.

The deduction for the (virtual) service time in case of $k > (x_1 - x_2)\pi_1(0)$ mimics the above deduction for type 1 appointments. □

Let us assume that both demand streams, i.e., appointment types 1 and 2, stem from the same patient panel of size $N$, such that an appointment of one type does not influence the appointment request rate for the other type. Then, the functions $l_1$ and $l_2$ should be used to define the no-show functions and the rescheduling functions, if necessary. For the appointment request rates, we have $\lambda_{1,2}(k) = \eta_{1,2}(k)(N - m_{1,2}(k))$. Again, the solution can only be found approximately through an iterative process. We start with the first queue and set any influence of the second queue to zero. Then, as described for the single queue, additionally integrating the adapted service time, we find the queue length distribution for the first queue. With those results we iteratively determine the queue length distribution of the second queue. Then, we again determine the distribution for the first queue and go on until a predefined accuracy is achieved. The calculations for a single queue become easier, if we do not consider rescheduling and hence $d^{r*}$. A rough approximation would be to set $\varepsilon(k) = \lambda(k) + \frac{1}{T}\rho(k-1)$. Note that we are not able to proof rigorously that this iterative process will always produce a solution.

The same theoretical approach can be used in the case of three or more different patient types. Then, the second type books appointments in the gaps left by the first type and the third type books appointments in the gaps left by the second type.

# 4 Queuing model implementation and numerical experiments

This chapter extends the previous theoretical chapter on a flexible analytical queuing model to investigate the relationship between the physician's daily capacity, the panel size, and the stationary distribution of indirect waiting times of patients. We shortly present a real-world data analysis and then implement the model and build a simulation to verify the queueing model's correctness and validate used approximations. We further investigate the general indirect queue behavior in extensive numerical experiments.

## 4.1 Introduction

In this chapter, we first analyze a real-world appointment data from a medical practice to show that our model assumptions of queue length-dependent parameters are reasonable. Next, we verify and validate the queueing model from Chapter 3. Moreover, we conduct numerical experiments to study further mathematical approximations and investigate the parameters' influence on the performance measures. To do so, we implement the queueing model and a simulation model.

Using the simulation model, we first verify the correctness of our mathematical model and its implementation by mimicking the mathematical queueing system completely. Second, we validate the simplifications and approximations used in modeling. Starting with the simplifications used in the mathematical model compared to the conceptual model, we investigate the effects of not scheduling patients to concrete days and time slots in the queueing model. Moreover, we investigate the assumption that the no-show probability and the rescheduling probability depend on the queue length instead of the indirect waiting time. Furthermore, we look into assuming that changes in the number of daily appointment offerings can be implemented immediately instead of starting with the next non-fully book day. For the extended model versions, we review the consequence of using one-dimensional modeling in the case of panel patients with different individual arrival rates and in the case of a mix of panel and non-panel patients.

In addition, using our queueing model only, we investigate other model approximations aiming for a model that is easier to apply in practice but still delivers valuable results. First, we check the effects of modeling the probability $\alpha_k(i, t)$ of $i$ arrivals with $k$ patients present at the start of the interval $t$ as a Poisson distribution instead of assuming a negative binomial distribution (in case of a linear total arrival rate function) or the accurate general distribution. Further, we investigate the effect of using arrival distributions $\alpha_k(\cdot, t)$ with different coefficients of variation. Additionally, we look into merging the general arrivals and the rescheduling arrivals in the total arrival rate function and set the rescheduling probability to zero. We finally compare the different queue length distributions to find out if it is necessary to compute the queue length distribution at arbitrary time points in practice.

Our numerical experiments study the following primary parameters: total arrival rate function, panel size, individual arrival rate, the (average) number of daily appointment offerings and the queue capacity. Remember that for determining the queue length distributions, the formulas are dependent on $\rho(k)$ and $\nu(k) = 1 - \rho(k)$ which are the probabilities of rescheduling and not rescheduling of a patient that left $k$ patients in the system after departure, respectively. $\rho(\cdot)$ and $\nu(\cdot)$ depend on the no-show function and the rescheduling functions for shows and no-shows. Hence, instead of testing for different no-show functions and rescheduling functions of shows and no-shows, it is enough to consider different functions for $\rho$. We will further see later that we can integrate the rescheduling function $\rho$ into the total arrival rate, meaning that it is enough to consider different total arrival rate functions to cover effects on the queue length distributions. Concerning the total arrival rate function, we consider the following cases: constant total arrival rate; the same constant individual arrival rate for all panel patients; the same queue length-dependent individual arrival rate for panel patients; different constant individual arrival rats for panel patients; and a mix of panel and non-panel patients, where all panel patients have the same constant individual arrival rate.

We mainly study the influence of the parameters mentioned above on the primary performance measures: queue length distribution, expected queue length, indirect waiting time distribution and expected indirect waiting time. We further consider same-day appointment probability, physician utilization, rescheduling rate, and rejection rate.

The remainder of this chapter is organized as follows: First, we shortly describe a real-world data set in Section 4.2 and then briefly address the model and simulation implementation in Section 4.3 including the simulation output analysis. In Section 4.4, we present the numerical experiments. Section 4.5 comments on the model application in practice, and in Section 4.6 we draw a conclusion and present an outlook on future research.

## 4.2 Real-world data analysis

In this section, we present some descriptive statistics for an appointment data set from a cardiology group practice. We received data from the years 2016 and 2017 for 3 physicians. Every row in the data set represents a booked appointment with possible information on the patient ID, the day and time, the booking day and time, and the assigned physician. Further, it is indicated if the patient showed up or not and whether the patient is a panel patient. We define that a patient reschedules if the patient books a new appointment on the day of an appointment. We find that out of $4732$ appointments corresponding to patients who had a patient id and showed up more than once in the considered time frame, almost $323$ appointments, i.e., $0.07\,\%$, correspond to patients who rescheduled. Only $179$ appointments correspond to no-shows, from which $27$ rescheduled. This means that the probability for no-shows to reschedule was $0.15$, whereas it was $0.07$ for shows. From $9160$ appointments that could be assigned to a physician, only $12\,\%$ booked the next available appointment. We study if the number of patients a physician sees on average per day is dependent on the fullness of the schedule. Here, we use the number of already booked appointments in the schedule as a proxy for the fullness. In Figure 4.1, we see that the physicians tend to serve more patients when they experience a fuller schedule.

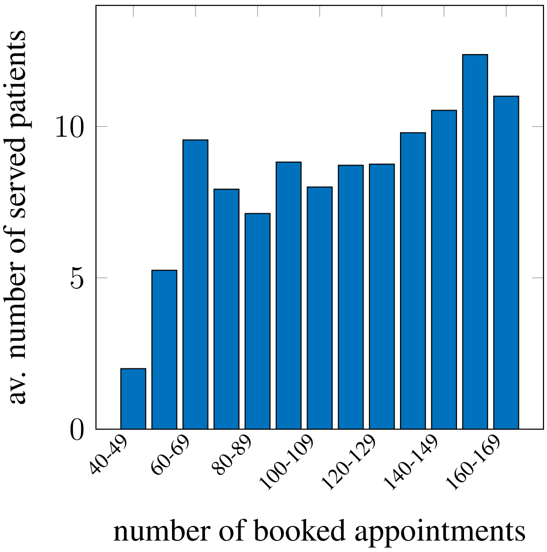

**Figure 4.1:** Average number of patients seen per day dependent on the number of appointments booked in the physician schedule

Next, we investigate if there is a relationship between the fullness of the schedule and the number of booked appointments per day. Indeed, we find in Figure 4.2 that the average number of booked appointments per day increases with the number of already scheduled appointments. This is true considering all appointment requests or appointment requests excluding rescheduling requests. On the right side of Figure 4.2, we illustrate the distribution of indirect waiting times.

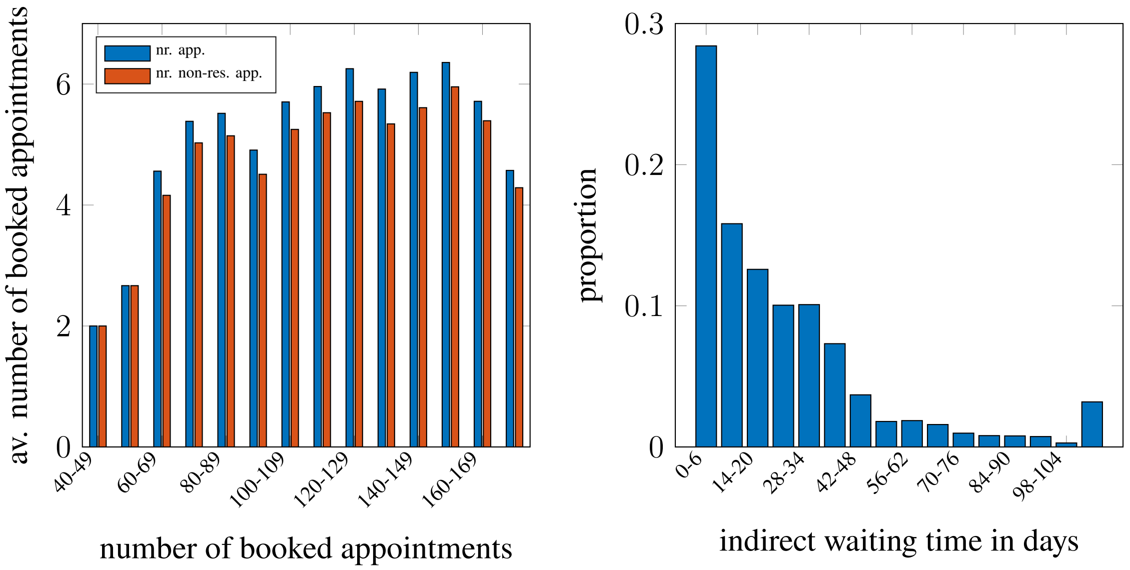

**Figure 4.2:** Number of appointment bookings per day including and excluding rescheduling patients dependent on the number of appointments booked in the physician schedule and the distribution of indirect waiting time in days

In Figure 4.3, we show the relationship between the indirect waiting times and the no-show and rescheduling probability, respectively. We see no apparent connection between indirect waiting time and no-show probability in this data set. However, we observe that the rescheduling probability increases with indirect waiting time. This effect is even more substantial for panel patients.

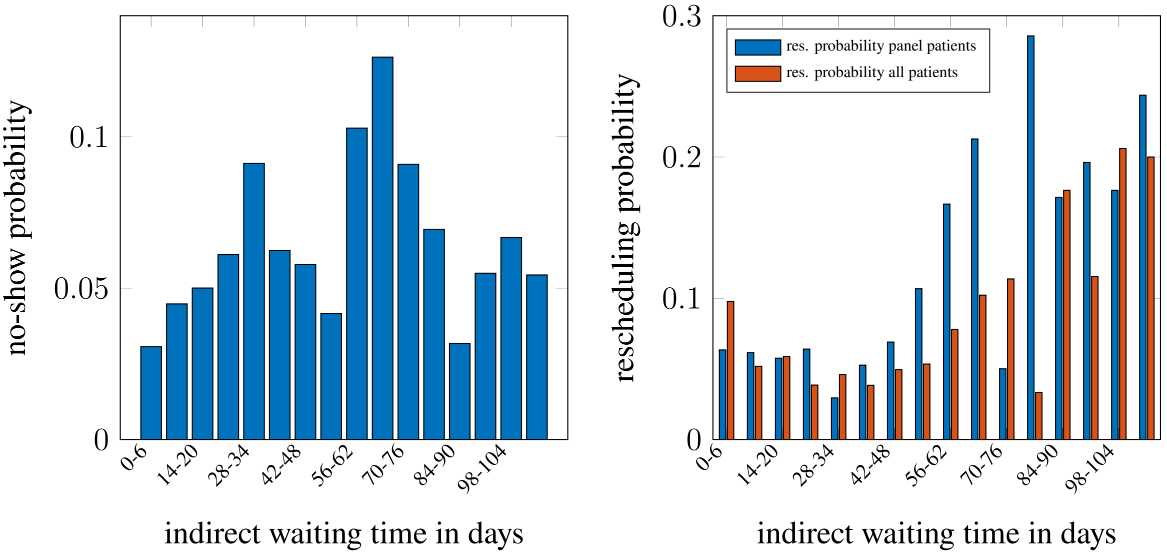

**Figure 4.3:** Probability of being a no-show or a rescheduling patient given the indirect waiting time in days

The analysis of this one exemplary data set from a medical practice shows that our model assumptions of queue length (or indirect wilting time) dependent parameters are reasonable. Unfortunately, the data set is not suitable to define model parameters because the assumption that patients always book the next available appointment is not fulfilled.

## 4.3 Model and simulation implementation

We implemented the mathematical queueing model, including all extensions in Java. The program extends the program written by David Koza to implement the approach of Green and Savin (2008).

Program inputs are

- the queue capacity $K$,
- the service time $T$,
- the appointment request rate function $\lambda$,
- the no-show function $\gamma$,
- the rescheduling function for no-shows $r^n$,
- the rescheduling function for shows $r^d$,
- and the probabilities that $i$ patients arrive during a service period of length $T$ given that there are $k = 1, \dots, K-1$ patients in the system when the service starts, $\alpha_k(i)$.

In the case of a physician with panel patients only, we can choose between the modeling examples explained in Section 3.5.2. For the case of a constant individual arrival rate, $\eta$ is an input variable. For a non-constant individual arrival rate, the average number of attended appointments per time unit $\delta$ is the input. Both input variables are then, of course, used to determine the total arrival rate $\lambda$. In addition, we can include patient-dependent individual arrival rates and choose to work with a mix of panel and non-panel patients. For patient individual arrival rates, the size of the subgroups and their individual arrival rates are needed as inputs. In the case of mixing panel and non-panel patients, the non-panel arrival rate is another input. Also, we can opt to use queue length-dependent (expected) appointment offerings per day, hence define a function $\mu$, according to Section 3.5.1.

Program outputs are the general queue length distribution and the queue length distributions

- after departures,
- after departures of rescheduling patients,

- after departures of non-rescheduling patients,
- at arrival times,
- at arrival times of non-rejected patients,
- at arrival times of non-rescheduling patients,
- at arrival times of non-rejected and non-rescheduling arrivals,
- and at arrival times of rescheduling arrivals.

Further outputs are

- the stationary departure rate,
- the stationary departure rate for rescheduling patients,
- the stationary departure rate for non-rescheduling patients,
- the indirect waiting time distribution in days,
- the distribution of treated patients per day,
- the average total arrival rate,
- the (average) hidden demand rate
- the (average) no-show rate
- the (average) rescheduling rate,
- the (average) effective arrival rate,
- the average rate of rejected patients,
- the proportion of no-shows,
- the proportion of rescheduling patients,
- the proportion of rejected patients,
- the proportion of time the queue is empty,
- the proportion of time the physician is idle,
- the same day appointment probability,
- and the average physician utilization.

Besides the Java implementation, MatLab has been used to determine the queue length distribution immediately after departures using the transition matrix of the embedded Markov chain as explained in Section 3.4.4.

### 4.3.1 Simulation

We further built a discrete-event simulation of the queueing system in AnyLogic $8$. In the following, we comment on how to address the simplifications and assumptions mentioned in Section 4.1 in the simulation.

To measure indirect waiting time (in days) in the simulation model, we have to schedule patients to actual days. To do so, we assume that every day consists of a number of appointment slots. Like in the queueing model, an arriving patient to an empty queue will always be treated immediately if the patient does not arrive in the last time slot of the day. A patient who arrives during the last time slot of the day will be treated on the day of request, possibly causing overtime. The same logic is applied for a patient that arrives to a non-empty queue. If the last scheduled patient is planned to finish before the last time slot of the day, the patient is scheduled directly afterward on the same day, possibly causing overtime. Otherwise, the patient is scheduled for the next day, possibly causing idle time. This approach is slightly different from the simulation approach of Green and Savin (2008). In their simulation, they also work with fixed service periods. However, in the situation of an empty queue, an arriving patient is not served immediately but has to wait until the start of the following service period. Green and Savin (2008) argue that this assumption makes the simulation more realistic because it reflects the time delay between appointment request and actual treatment, for example, if the patient is calling and cannot show up immediately. Building the simulation like that has the advantage that we can clearly match the different appointment slots to days. No overtime or idle time will be incurred. On the other hand, we argue that simulating a direct treatment of an arriving patient reflects the situation of patients who walk in.

In the simulation, we can keep track of the arrival times of all patients. Hence, we can easily make the no-show probability and the rescheduling probability dependent on the actual experienced indirect waiting time.

We are also able to represent variable appointment offerings per day realistically in our simulation. By assigning patients to actual days and not altering this decision later, we cannot immediately implement a change in the number of offered appointment slots per day. Instead, having planned a day, we check the length $k$ of the queue and assign $\mu(k)$ time slots to the following day.

Further, the simulation allows us to analyze the situation of patient individual arrival rates and the situation of a physician with panel and non-panel patients realistically. This is because we can easily count patients of a specific type in the queue in the simulation.

Now, we describe the additional inputs for the simulation besides those already defined for the mathematical model. One input is the initial queue length, so we do not always have to start our simulation with an empty queue. Here, at the beginning of the simulation, the number of patients as indicated by the initial queue length is fed into the queueing system. A further input parameter sets the time when to start collecting data and statistics on various performance measures. Using boolean parameters, we can decide to simulate different variations:

- decide to measure indirect waiting time (remember that this also alters the way of servicing),
- make the no-show function (and the rescheduling functions) dependent on the queue length at departure times or at arrival times,
- select on one of the variants to model the total arrival rate of panel patients,
- decide on simulating panel patients with different individual arrival rates (in this case, binomial panel arrivals are used),
- decide on simulating a mix of panel and non-panel patients (in this case, binomial panel arrivals are used).

The simulation model has the following additional outputs compared to those from the mathematical model:

- worked overtime if indirect waiting time is measured,
- idle time if indirect waiting time is measured,
- number of panel patients in the queue per arrival rate group in the case of different individual arrival rates for panel patients,
- number of panel patients in the queue in the case of a mix of panel and non-panel patients,
- development of the queue length over time.

### 4.3.2 Simulation output analysis

As explained in Section 4.3.1 our simulation model outputs a wide range of performance measures. However, our primary interests are the queue length distribution, the indirect waiting time distribution, and the corresponding means. In order to obtain reliable simulation outputs, we need to decide on an appropriate output analysis approach.

We denote the queue length in our simulation at time $t$ as $L(t)$. The queue length is always a discrete value, but the time interval between two different queue length values can be of any positive length. Hence, the queue length produces continuous-time data in our simulation which means that we have to store the queue length values and the lengths of the time interval between queue length changes. The mean queue length corresponding to a time period $[t_1, t_2]$ can be computed as $\int_{t_1}^{t_2} L(t)dt$. Contrary, the indirect waiting time $W_i$ of the $i$-th non-rejected patient in the simulation yields discrete-time data, which means that we only have to store the $W_i$-values and no additional time intervals. The mean indirect waiting time for patients $i_1$ to $i_2$ can be computed as $\sum_{i=i_1}^{i_2} \frac{W_i}{i_2-i_1+1}$.

In our case, we are interested in the stationary analysis of our queueing system. We are faced with two challenges. First, we know that we will experience an initial transient phase at the beginning of the simulation until the steady state of the queue is approximately reached. Therefore, it is advised to find the initial transient phase's length and start data collection afterward. Second, we need to determine the accuracy of our estimated expected values $\widehat{L}$ and $\widehat{W}$ with $L$ and $W$ being the actual expected values of the queue length and the indirect waiting time in steady state, respectively.

We use standard methods to address the two challenges as explained by Law (2007). We first apply Welch's method to handle the problem of the initial transient. This method was first mentioned in Welch (1981) and Welch (1983) and is tailored to discrete-time data. Hence, we transform the continuous-time data to discrete-time data via batching to apply this method to the queue length. We use time intervals of length $d$ and set $L_i = \int_{(i-1)d}^{id} L(t)dt$. Now, we illustrate Welch's method for the indirect waiting time. The same procedure can be used for transformed queue length data. The first step is to run $R$ replications of length $M$ of the simulation with $W_{rm}$ being the indirect waiting time of the $m$-th patient in the $r$-th replication. The data produced during one simulation replication will be highly correlated (e.g., the indirect waiting time of a patient is correlated to the indirect waiting time of the subsequent patient). However, data across replications is independent. Therefore, we define $\overline{W}_{\bullet m} = \frac{1}{R}\sum_{r=1}^{R} W_{rm}$ for $m = 1, \ldots, M$. The resulting process $\overline{W}_{\bullet m}, m = 1, \ldots, M$ has the same mean as the single replications: $E(\overline{W}_{\bullet m}) = E(W_{rm}), r = 1, \ldots, R$ and because of the independence of the $W_{rm}$'s for a fixed $m$ and $r = 1, \ldots, R$ we have: $Var(\overline{W}_{\bullet m}) = \frac{Var(W_{rm})}{R}, r = 1, \ldots, R$. The next step is to plot $\overline{W}_{\bullet m}, m = 1, \ldots, M$ and to increase the number of replications and to use moving averages to smooth out high-frequency oscillations. We then choose $l$ such that $\overline{W}_{\bullet l}$ seems to have converged. All data points $\overline{W}_{\bullet m}, m = 1, \ldots, l$ are then considered to be part of the initial transient and are omitted in the following analysis.

To approach the second problem, we use the replication/deletion approach to determine the standard error of the mean estimators and to construct a confidence interval. We again use $R$ replications of length $M$ of our simulation. We define $\overline{W'}_{r\bullet} = \frac{\sum_{m=l+1}^{M} W_{rm}}{M-l}, r = 1, \ldots, R$ where $l$ is the warm up period as defined by Welch' s method. Then the $\overline{W'}_{r\bullet}, r = 1, \ldots R$ are independent and identically distributed random variables with $E(\overline{W'}_{r\bullet}) \approx W$. Further, $\widehat{W} := \overline{W'}_{\bullet\bullet} = \frac{\sum_{r=1}^{R} \overline{W'}_{r\bullet}}{R}$ is an approximately unbiased point estimator for $W$ and $\widehat{Var}(\overline{W'}_{\bullet\bullet}) = \frac{S^2}{R}$ is an unbiased point estimator of $Var(\overline{W'}_{\bullet\bullet})$ where $S^2 = \frac{\sum_{r=1}^{R} (\overline{W'}_{r\bullet} - \overline{W'}_{\bullet\bullet})^2}{R-1}$ is the sample variance of $\overline{W'}_{r\bullet}, r = 1, \ldots R$. Finally, $\widehat{W} \pm t_{R-1,1-\alpha/2}\sqrt{\frac{S^2}{R}}$ is an approximate $100(1-\alpha)$ confidence interval for $W$ where $t_{R-1,1-\alpha/2}$ is the upper $1-\alpha/2$ critical point for the $t$ distribution with $R-1$ degrees of freedom.

## 4.4 Numerical experiments

This section presents our extensive numerical experiments using the implemented mathematical queueing model and the simulation. We start with defining our basic parameter settings in Section 4.4.1. Next, we comment on the runtime of our model implementation in Section 4.4.2. In Sections 4.4.3 and 4.4.4, we verify the mathematical queueing model and its implementation and then validate used simplifications in the basic queueing model. Additionally, we take a closer look at the arrival process modeling, showing that we can approximate the accurate distribution of the number of patients arriving given a certain queue length in a time interval by a Poisson distribution in Section 4.4.5.

Further, in Section 4.4.6, we present numerical experiments with our basic queueing model with a focus on the queue length distributions and the indirect waiting time distribution. However, we also show results for rescheduling and rejected patients and compare our results with those of Green and Savin (2008). Moreover, we investigate the influence of the individual arrival rate and the queue capacity on the results.

Section 4.4.7 examines the consequences of integrating rescheduling arrivals into the total arrival rate, and Section 4.4.8 studies the queue length distribution for linear total arrival rates. Section 4.4.9 compares the expected indirect waiting times resulting from using arrival distributions with different coefficients of variation.

The following Section 4.4.10 shows the effects of assuming queue length-dependent appointment offerings on the indirect waiting time distribution. Next, in Sections 4.4.11 and 4.4.12, we present the resulting distributions assuming several different individual appointment request rates for panel patients and a mix of panel and non-panel patients. Finally, we show an example for a fair capacity division for two independent demand streams in Section 4.4.13. An overview of our numerical experiments can be found in Table 4.1.

| **Experiments** | **Section** |
|---|---|
| Basic parameter settings | Section 4.4.1 |
| Runtime of the model implementation | Section 4.4.2 |
| Model and implementation verification | Section 4.4.3 |
| Model validation | Section 4.4.4 |
| Approximation of the arrival process | Section 4.4.5 |
| Results for the basic queueing model | Section 4.4.6 |
| Integration of rescheduling arrivals into the total arrival rate | Section 4.4.7 |
| Results for linear total arrival rates | Section 4.4.8 |
| Results for arrival distributions with different coefficients of variation | Section 4.4.9 |
| Results for variable appointment offerings | Section 4.4.10 |
| Results for different constant individual arrival rate of panel patients | Section 4.4.11 |
| Results for a mix of panel and non-panel patients | Section 4.4.12 |
| Results for a fair capacity division between demand streams | Section 4.4.13 |

**Table 4.1:** Model features and extensions

Throughout this section, we use the different theoretical models from Section 3.5.2 to define the total arrival rate function $\lambda$. Either we use a constant total arrival rate with $\lambda = N\eta$, $N$ being the panel size and $\eta$ being the individual arrival rate. Or we use a non-constant total arrival rate. There, we differentiate between a constant individual arrival rate, i.e., $\lambda(k) = \eta(N-k)$ and a non-constant individual arrival rate, i.e., $\lambda(k) = \eta(k)(N-k)$. Remember that, in the second case, we are restricted to either setting $r^n = r^s = 0$ or $r^n = 1$ and $r^s = 0$.

### 4.4.1 Basic parameter settings

For our numerical experiments, we use the same basic parameter settings as in Green and Savin (2008). These should be realistic for a system representing a physician schedule. They should therefore be appropriate to study the general model behavior in this use case. Further, we can compare our results to the results of Green and Savin (2008). Unfortunately, as stated in Section 4.2, we were not able to define model parameters from real-world data due to the assumption that patients always book the next available appointment. We will elaborate on the possibilities and challenges of adopting the model in practice in Sections 4.5 and 4.6.

In their setting, Green and Savin (2008) assume a constant arrival rate $\lambda = \eta N$ with $N$ being the panel size and $\eta$ being the individual appointment request rate of panel patients. They further assume a service time $T$ of $0.05$ days which corresponds to $20$ appointment slots per day. Finally, say set the queue capacity $K$ to $400$, which corresponds to a booking horizon of $20$ days or $4$ weeks.

Concerning the no-show function Green and Savin (2008) use two different data sets to define it. Here, we stick to the no-show function that fits the Columbia MRI data. They show that an exponential function of the form $\gamma^d(l) = \gamma_{max} - (\gamma_{max} - \gamma_{min})e^{-l/C}$ fits the considered data set well. Here, $\gamma^d(l)$ denotes the no-show probability of a patient with an indirect waiting time of $l$ days. $\gamma_{max}$ and $\gamma_{min}$ are parameters denoting the maximal and minimal no-show probabilities, respectively, and $C$ is a sensitivity parameter. To use this no-show function in our model, we need to convert the dependency of the indirect waiting time in days to the indirect queue length at the time of departure. We use our considerations from Equation (3.52). With $\mu(k) = \mu = \frac{1}{T}$ a patient experiencing queue length $k$ waits on average $l(k) = \frac{k}{\mu(k)}$ days. Therefore, we define our no-show function as $\gamma(k) = \gamma^d(l(k))$ depicted in Figure 4.4. Note that here we do not consider weekend days as waiting days.

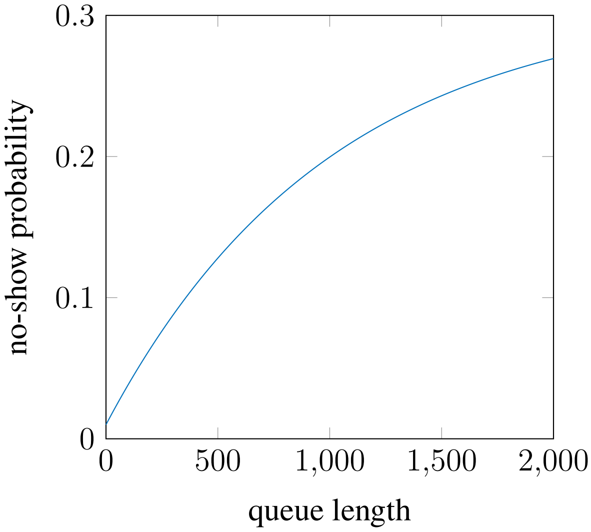

**Figure 4.4:** No-show probability dependent on the queue length

Further, Green and Savin (2008) consider a constant rescheduling probability $r^n$ of no-shows, which they set to $1$ in their experiments and do not consider a non-zero rescheduling probability $r^s$ of shows. We summarize the basic parameter settings in Table 4.2.

| Parameter | Value | Description |
|---|---|---|
| $\eta$ | $0.008/\text{day}$ | individual arrival rate |
| $T$ | $0.05$ days | service time |
| $\gamma_{min}$ | $0.01$ | minimal no-show probability |
| $\gamma_{max}$ | $0.31$ | maximal no-show probability |
| $C$ | $50$ | no-show function sensitivity parameter |
| $K$ | $400$ | queue capacity |
| $r^n$ | $1$ | rescheduling probability of no-shows |
| $r^s$ | $0$ | rescheduling probability of shows |

**Table 4.2:** Basic parameter settings from Green and Savin (2008)

In the case of a physician with panel patients, we can either assume a constant total arrival rate of $\lambda = \eta N$ as in Green and Savin (2008), where $\eta$ is the individual request rate and $N$ the panel size or we use a queue length-dependent total arrival rate $\lambda(k) = \eta(k)(N-k)$ as explained in Section 3.5.2. In the case of a constant $\eta$, principally, we can consider different arbitrary rescheduling functions for shows and no-shows. Working with a non-constant $\eta$ means we either have to stick to $r^n = 0$ or $r^n = 1$ and $r^s = 0$. In both cases, we have to fix the parameter $\delta$, the average number of attended appointments per time unit (in case of no indirect waiting time). We use formulas (3.54), (3.55) and (3.57), (3.58) inserting $i = 0$ and fixing $\eta(0) = \eta = 0.008$ as given by the parameter settings from Green and Savin (2008). We then rearrange for $\delta$. In the case of $r^s = r^n = 0$, we have:

$$\eta = \frac{1}{\frac{1-\gamma(0)}{\delta} - l(0)} \Rightarrow \delta = \frac{1-\gamma(0)}{\frac{1}{\eta} + l(0)} = (1-\gamma^d(0))\eta = (1-\gamma_{min})\eta = 0.00792.$$

In the case of $r^s = 0$ and $r^n = 1$, we have:

$$\eta = \frac{1}{\frac{1}{\delta} - \frac{l(0)}{1-\gamma(0)}} \Rightarrow \delta = \frac{1}{\frac{1}{\eta} + \frac{l(0)}{1-\gamma(0)}} = \eta = 0.008.$$

All other parameter settings will be explained together with the corresponding numerical experiment.

### 4.4.2 Runtime of the model implementation

This section investigates the runtime of our model implementation for increasing values of the queueing capacity $K$ using a Microsoft Surface Intel(R) Core(TM) i7-8650U with 16 GB RAM. When the queue capacity increases by one, we also have to calculate one more value for the queue length distributions. For this analysis, we use a constant total arrival rate $\lambda = \eta N$ with $N = 2344$. All other parameters are set as explained in Section 4.4.1. We plot the runtime of our implementation (the calculation of all queue length distributions and other performance measures are included) dependent on the queue capacity $K$ in Figure 4.5. We see an increase in runtime with increasing queue capacity. However, the calculation of all relevant performance measures can still be done in a few seconds also for larger queue capacities.

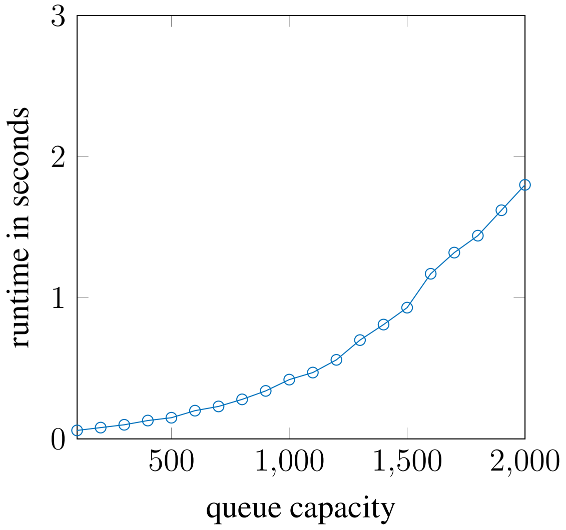

**Figure 4.5:** Runtime of the Java implementation dependent on the queue capacity for N=2344 with a constant total arrival rate

For larger panel sizes and queue capacities using the recursion $g$ in the implementation becomes relevant. Setting the queue capacity to $K = 2400$ and the panel size to $N = 2500$ Java can no longer calculate results using the recursion $f$ due to $f$ values that become too large. Using $g$ instead of $f$, we do not run into numerical problems and are still able to solve instances for larger panel sizes and larger queue capacities that are of a realistic size, i.e., correspond to several months.

### 4.4.3 Verification of the mathematical model and its implementation

We verify the mathematical model and its implementation in two ways. First, we compare the calculated queue length distribution at departure times with the one determined using a Matlab program based on the transition matrix of the embedded Markov chain as explained in Section 3.4.4. Second, we use our simulation with parameter settings such that the queueing model is complectly mimicked and compare the queue length distributions.

To compare the results of the Java with the Matlab implementation, we use a constant total arrival rate $\lambda = \eta N$ with $N = 2344$. All other parameters are set as explained in Section 4.4.1. The values of the two queue length distributions at departure times are the same up to the fourth decimal place. This indicates that both approaches have been implemented correctly. The resulting queue length distributions at departure times can be seen in Figure 4.6.

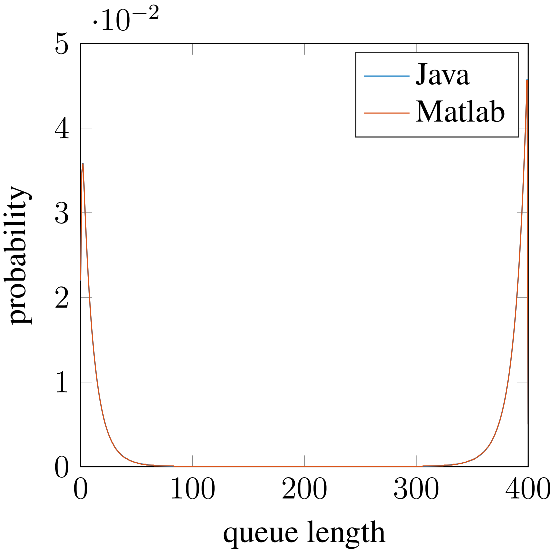

**Figure 4.6:** Queue length distributions at departure times calculated using the Java and the Matlab implementation for a constant total arrival rate

To verify the queueing model with the simulation, we again use the same parameters settings as in Section 4.4.1 together with three different panel sizes $N = 2300, 2344$ and $2400$. In the simulation, we mimic the queueing model and therefore make the no-show probability dependent on the queue length at departure times and do not assign patients to concrete days and time slots. To analyze the simulation output, we apply the methods explained in Section 4.3.2 where we focus on the queue length distribution and the expected queue length in steady state.

We will explain the simulation output analysis in detail for $N = 2300$. We simulate the queueing system for $5000$ days which corresponds to approximately $20$ years assuming $5$ working days per week. We use $15$ different replications. We collect data on the queue length and the times when the queue length changes during the simulation. One simulation replication takes a few seconds to run. We use a Matlab program to perform our simulation output analysis. As explained in Section 4.3.2 we start by batching the queue length data to transform the continuous-time data into discrete-time data. Here, we choose batches of length one day. Then we apply Welch's method to determine the length of the initial transient phase. The results of plotting the moving average of $\overline{L}_{\bullet m}$ for different numbers of replications and different moving average windows can be seen in Figure 4.7. Here, the moving average window describes the number of neighboring values used to calculate the moving average of $\overline{L}_{\bullet m}$. $15$ replications together with a moving average window of $1000$ seem to be sufficient. From Figure 4.7 we conclude that it is enough to exclude the first $1000$ days and consider them as part of the initial transient phase. Please note that this warm-up period corresponds to approximately $4$ years.

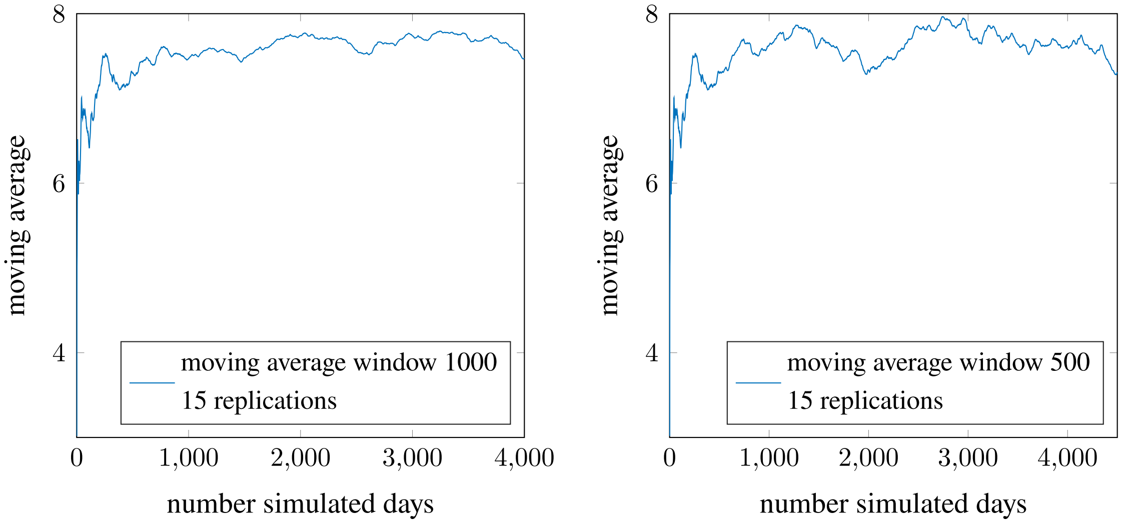

**Figure 4.7:** Moving average of $\overline{L}_{\bullet m}$

We apply the replication/deletion method to estimate the expected queue length and find a $95\%$ confidence interval for the estimator. Again using $15$ replications and data from $5000$ simulated days, we obtain $\widehat{L} = 7.61$ and the following $95\%$ confidence interval: $[7.46, 7.76]$. The results fit quite well, considering that we determined an expected queue length of $7.58$ using our queueing model. We further use the collected data (after the initial transient phase) to reconstruct the queue length distribution, as can be seen in the left part of Figure 4.8. Comparing the distribution for $N = 2300$ to the corresponding simulated distribution in the right part of Figure 4.8, we see that the simulated results match the analytical results quite well. In the following, when we state simulation results, we will not go into detail about the application of Welch's method and the replication/deletion method anymore.

For $N = 2400$, we already expect a high expected queue length value. Therefore, to avoid a very long initial transient phase, we start our simulation with an initial queue length of $390$ instead of $0$. Applying the methods from Section 4.3.2 to the simulation output for the panel size of $N = 2400$ using $15$ replications, $5000$ simulated days and an initial transient phase of $1000$ days, we obtain $\widehat{L} = 392.11$ with a $95\%$ confidence interval of $[391.95, 392.27]$. Again, those results match the analytically calculated expected queue length of $392.34$. The corresponding queue length and simulated queue length distribution can once again be seen in the left and right parts of Figure 4.8.

The case of a panel size $N = 2344$ is challenging to simulate. Looking at the analytical queue length distribution in the left part of Figure 4.8, we see, for example, that the probability of experiencing a queue length of 200 is $2.31 \cdot 10^{-6}$. Hence, even for a long simulation run, it is unlikely to experience such a queue length at least once. Therefore, we run two simulation experiments with initial queue lengths $0$ and $400$. We expect both simulations to reconstruct the left and right part of the queue length distribution, respectively. Simulating with an initial queue length of $0$ shows that none of the $15$ replications reaches queue length values above $60$. Therefore, the assumption that this simulation reconstructs the left part of the queue length distribution seems reasonable. In fact, for $5000$ simulated days with a warm-up phase of $1000$ days, we experience an estimated value of $\widehat{L} = 11.4$ with a $95\%$ confidence interval of $[11.06, 11.74]$ which is close to the analytically calculated expected queue length of $10.98$ for the left part of the queue length distribution. Starting with a queue length of $400$ we obtain $\widehat{L} = 388.53$ with a $95\%$ confidence interval $[388.24, 388.82]$ which is not too far away from the analytically calculated expected queue length of $387.76$ for the right part of the queue length distribution. One problem, however, remains. Simulating the two destitution parts separately means that we do not have any information on how to scale the two parts to create a joint distribution. Here, we use the cumulative probabilities of the queue length distribution from our mathematical model to scale the two parts to create a joint distribution. In the right part of Figure 4.8, you can then see the scaled simulated left and the right part of the queue length distribution.

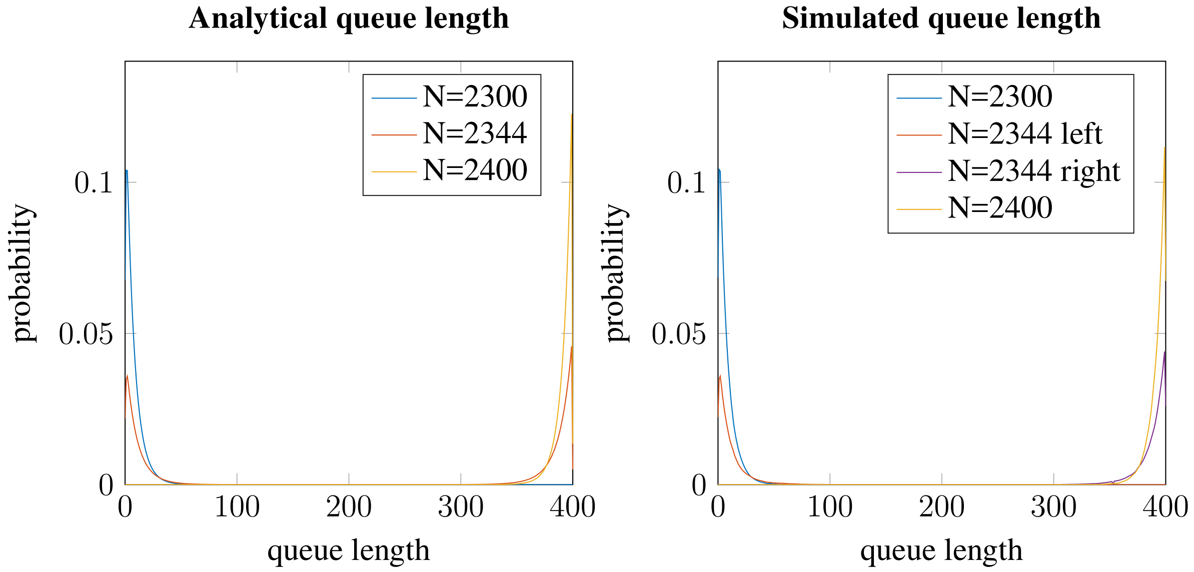

**Figure 4.8:** Calculated and simulated queue length distributions for $N = 2300$, $N = 2344$ and $N = 2400$ for a constant total arrival rate

The results of our simulation studies where we mimicked the queueing system show that we can be confident about the analytical results from the queueing system and the results from the simulation.

### 4.4.4 Validation of simplifications in the basic model

In this section, we validate two simplifications used in the mathematical model. One is the simplification to not schedule patients to concrete days and time slots in the queueing model. The other is the assumption that the no-show probability and the rescheduling probability depend on the queue length instead of the indirect waiting time. Hence, in the simulation, we schedule patients to the next available slot on a concrete day. This way, we can measure indirect waiting time and make the no-show function dependent on the patients' indirect waiting time in days. Note that this also alters the handling of the workload in the queue as explained in Section 4.3.1. We consider the scenario of a constant total arrival rate, again using the basic parameter setting from Section 4.4.1 and two different panel sizes $N = 2300$ and $N = 2400$. As simulation output, we focus on the indirect waiting time distribution and the expected indirect waiting time.

We start with a panel size of $N = 2300$. Using 15 replications, $100,000$ simulated patients with an initial transient phase of $20,000$ patients, we obtain $\widehat{W} = 0.35$ and a $95\%$ confidence interval of: $[0.34, 0.36]$. Even though the simulation does not mimic the queue length but instead pictures a more realistic scenario, the estimated expected indirect waiting time does not differ much from the analytically determined value of $0.38$. Further, also the indirect waiting time distribution determined by the simulation is very similar to the analytically determined distribution as can be seen in Figure 4.9.

For $N = 2400$ using 15 replications, an initial queue length of $400$ and $100,000$ simulated patients with an initial transient phase of $20,000$ patients, we obtain $\widehat{W} = 19.59$ and a $95\%$ confidence interval of: $[19.58, 19.6]$. The estimated expected indirect waiting time is again close to the analytically determined value of $19.62$. Again, we see in Figure 4.9 that the calculated and simulated indirect waiting time distributions are very similar.

The results show that the simplified assumptions in the queueing model, such as the dependency of the no-show function on the queue length at departure times and the inobservance of concrete days and slots when scheduling appointments, do not have a significant impact on the results which backs up the value of the queueing model. More results for a realistic simulation can be found in Sections A.1.1 and A.1.2.

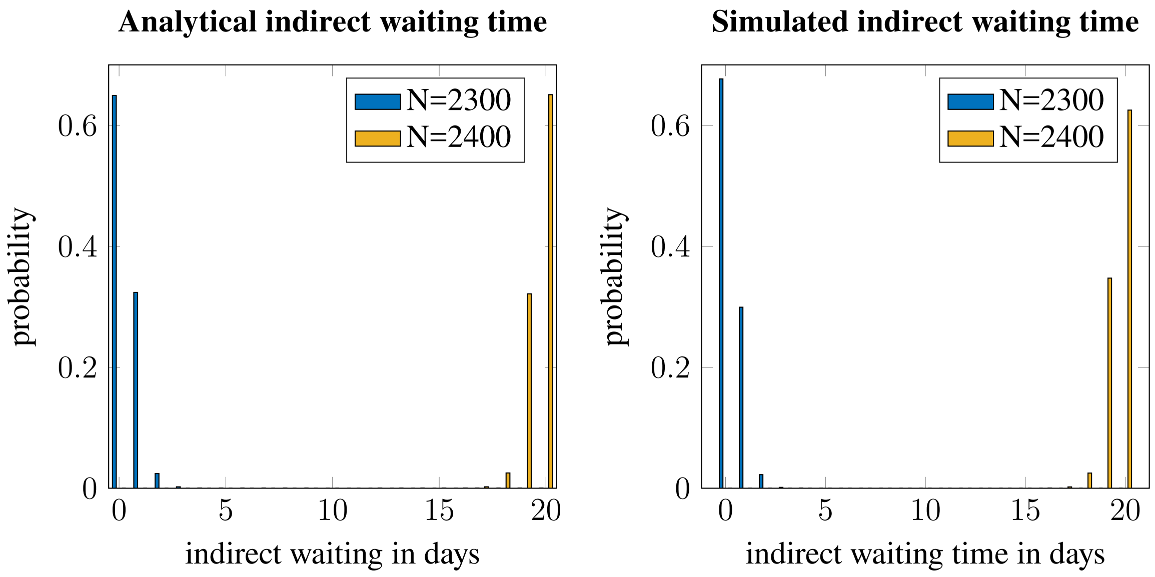

**Figure 4.9:** Calculated and simulated indirect waiting time distributions for panel sizes $N = 2300$ and $N = 2400$

### 4.4.5 Approximation of the arrival process

In Section 3.4.3, we discussed the different possibilities to define $\alpha_k(i,t)$, the probability of $i$ arrivals with $k$ patients present at the start of the interval $t$. If we aim to find the exact values of the queue length distribution, we should use Equation (3.48) and its derivatives for the total arrival rate function. In the case of a linear total arrival rate function, $\alpha_k(\cdot,t)$ follows a negative binomial distribution, shown in Equation (3.49). This can further be simplified to a binomial distribution in the case of a physician with panel patients and a constant individual arrival rate, as can be seen in Equation (3.53).

Suppose we are not interested in the queue length distribution or only in approximated values of the queue length distribution. In that case, we can model $\alpha_k(\cdot,t)$ as a Poisson distribution as can be seen in Equation (3.50). Here, we compare the (approximated) queue length distributions when using either the correct binomial distribution or the approximated Poisson distribution for $\alpha_k(\cdot,t)$ in the case of a physician with panel patients and a constant individual arrival rate. We use a panel size of $N = 2500$ and otherwise the basic parameter settings of Green and Savin (2008).

In Figure 4.10, we plot the sections of the two queue length distributions where they differ the most. As expected, using the Poisson distribution for $\alpha_k(\cdot,t)$, we slightly overestimate the expected number of patients arriving during a time interval. Hence, we experience higher probabilities for larger queue lengths. Still, the curves lie close to each other. Therefore, in the following, we will use the Poisson distribution in our numerical experiments if not stated otherwise.

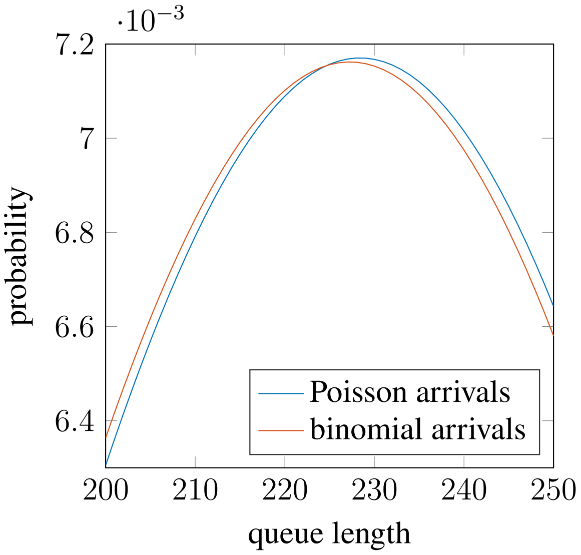

**Figure 4.10:** Section of the queue length distributions for panel size $N = 2550$ using the binomial and the Poisson distribution for arrivals during a time interval for a constant individual arrival rate

### 4.4.6 Basic model results

In this Section, we run numerical experiments with the basic queueing model. First, we investigate the queue length distribution and the indirect waiting time distribution. We also study queue lengths distributions at arrival and departure times. We then show results for the proportion of rescheduling and rejected patients. Then, we reproduce figures from Green and Savin (2008) and compare our results to theirs. Finally, we consider sensitivity of the results with respect to the individual arrival rate and the queue capacity.

#### Results for the queue length and the indirect waiting time distributions

Here, we present the resulting queue length distributions and indirect waiting distributions for the basic parameter settings from Section 4.4.1 combined with our three models for the appointment request rate $\lambda$ and panel sizes that correspond to a small, medium, and large expected queue length/indirect waiting time.

For the constant total arrival rate $\lambda = \eta N$, we illustrate the queue length distributions and indirect waiting time distributions for the panel sizes $N = 2300$, $2344$ and $2400$ in Figure 4.11. For the queue length, we obtain the expected values $7.58$, $214.53$ and $392.34$ for the panel sizes $2300$, $2344$ and $2400$. For the indirect waiting time in days, we obtain the expected values $0.38$, $11.17$, and $19.62$. Interestingly, the queue length and the indirect waiting time distribution for the panel size $N = 2344$ show two spikes which means that the queue is either very small or very large. Therefore, in this case, one must be careful with the expected queue length or expected indirect waiting time value that will rarely be reached. Let us consider the queue length distribution from the panel size $N = 2344$ to be made of two separated distributions for the two spikes. Then, we can compute the expected queue length for queue lengths equal to or smaller than $200$ as $10.98$ and as $387.76$ for queue lengths equal to or larger than $201$.

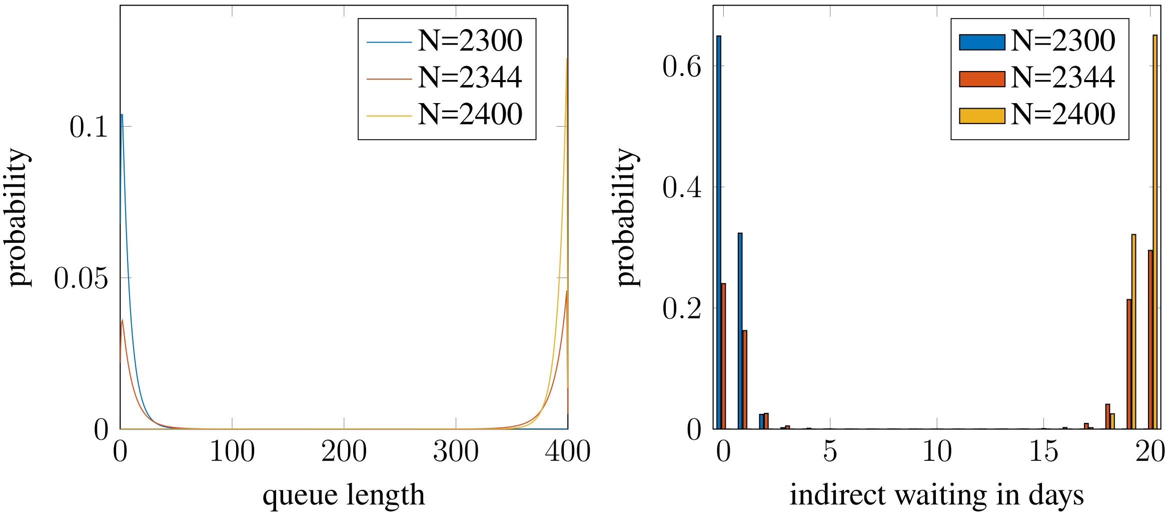

**Figure 4.11:** Queue length and indirect waiting time distributions for $N = 2300$, $N = 2344$ and $N = 2400$ for a constant total arrival rate

For a constant individual arrival rate with a total arrival rate $\lambda(k) = \eta(N - k)$, we compare the indirect waiting time distributions for the three panel sizes $N = 2300,\ 2540$ and $2800$ in Figure 4.12.

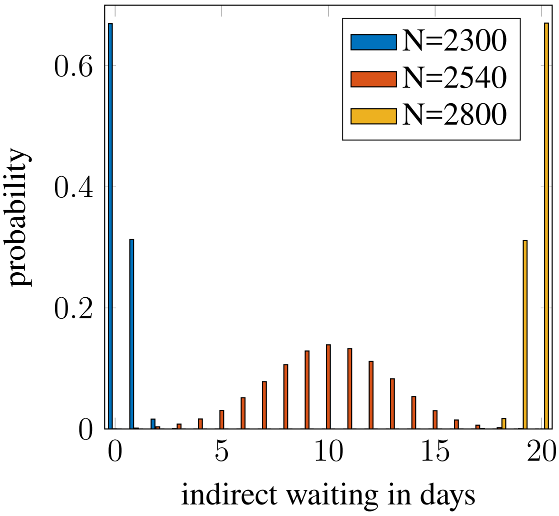

**Figure 4.12:** Indirect waiting time distributions for $N = 2300$, $N = 2540$ and $N = 2800$ for a constant individual arrival rate

In comparison to Figure 4.11, we observe a different behavior for the medium panel size. Instead of having two spikes (one for small and one for large indirect waiting times values), the distribution only shows one maximum and smaller variance. The analytically determined expected indirect waiting times in days are $0.35$, $9.99$ and $19.65$ for the three panel sizes $N = 2300$, $2540$ and $2800$.

For a non-constant individual arrival rate with a total arrival rate $\lambda(k) = \eta(k)(N - k)$, in Figure 4.13, we compare the indirect waiting time distributions for the three different panel sizes $N = 2300$, $2340$ and $2360$. The curves show similar behavior to the curves in the case of a constant total arrival rate. Especially, the distribution for a medium panel size of $N = 2340$ shows two spikes, one for small and one for large indirect waiting times. The expected indirect waiting times in days are $0.38$, $10.67$ and $19.56$ for the panel sizes $N = 2300$, $2340$ and $2360$, respectively.

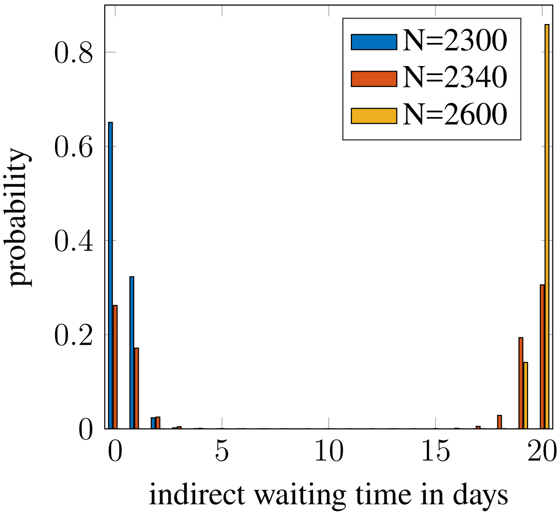

**Figure 4.13:** Indirect waiting time distributions for $N = 2300$, $N = 2340$ and $N = 2360$ for a non-constant individual arrival rate

See Sections A.1.1 and A.1.2 of the Appendix to learn more about the resulting indirect waiting time distributions for a constant and non-constant individual arrival rate in a realistic simulation.

#### Results for the queue length distributions at departure and arrival times

Here, we compare the queue length distribution with the queue length distributions immediately after departure times or at arrival times for a constant total arrival rate and a constant individual arrival rate with one panel size each. Again, we use the parameter settings from Section 4.4.1.

In the case of a constant total arrival rate, the PASTA (Poisson arrivals see time averages) property applies. The queue length distribution is identical to the queue length distribution at arrival times excluding rescheduling arrivals. In Figure 4.14, we compare the queue length distribution $\pi$ with the queue length distribution at arrival times $p$ and the different departure queue length distributions $q$, $q^{nr}$ and $q^{r}$ for the panel size $N = 2345$. We focus on a small range of queue lengths, such that the differences between the distributions can be seen. The queue length distribution immediately after rescheduling departure times differs the most from the other distributions. This is because rescheduling happens mainly for longer queue lengths and hence the probability of a short queue given a rescheduling departure is lower than for a non-rescheduling departure. However, in general, the differences are very small, indicating that for simplicity, the queue length distribution immediately after departure times can be used as a proxy for the queue length distribution. The queue length immediately after departure times distribution can easily be calculated numerically as explained in Section 3.4.4.

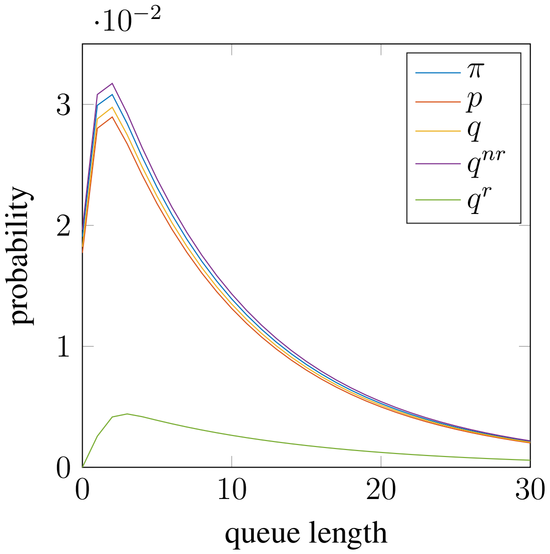

**Figure 4.14:** Comparison of the different queue length distributions for panel size $N = 2345$ for a constant total arrival rate

In Figure 4.15, we compare the queue length distribution with the queue length distributions at departure times for the case of a constant individual arrival rate for a panel size of $N = 2550$. For a non-constant total arrival rate $\lambda(k) = \eta(N - k)$, the PASTA property does not apply anymore. We see a difference between the queue length distribution and the queue length distribution at arrival times excluding rescheduling, which is almost identical to the queue length distribution immediately after non-rescheduling departure times due to a small value of $\pi(K)$ for $N = 2500$. Because the total arrival rate decreases with increasing panel sizes, non-rescheduling arrivals see shorter queues compared to the queue length distribution. Again, we focus on a small range of queue lengths, such that the differences between the distributions can be seen. Again, the queue length distribution immediately after rescheduling departure times differs the most from the other distributions. However, in general, the differences are very small, indicating again that for simplicity, the queue length distribution immediately after departure times can be used as a proxy for the queue length distribution.

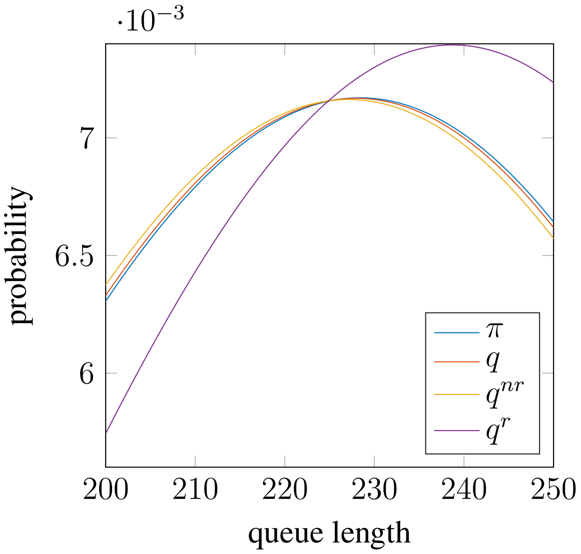

**Figure 4.15:** Comparison of the different queue length distributions for panel size $N = 2550$ for a constant individual arrival rate

### Results for rescheduling and rejected patients

In Figure 4.16, we show the proportion of rescheduling and rejected patients for a constant total arrival rate and a constant individual arrival rate for different panel sizes. Note that here, for $r^n = 1$, the proportion of rescheduling patients is identical with the proportion of no-shows. For a constant total arrival rate, we see that both the proportion of rescheduling and the proportion of rejected patients steeply increase when a certain panel size is reached until the proportion of rescheduling patients reaches its maximum, and the proportion of rejected patients increases further linearly. We see a similar behavior of the curves for a constant individual arrival rate but with a smoother transition.

See Section A.1.3 to learn more about the proportions of rescheduling and rejected patients in the case of a non-constant individual arrival rate.

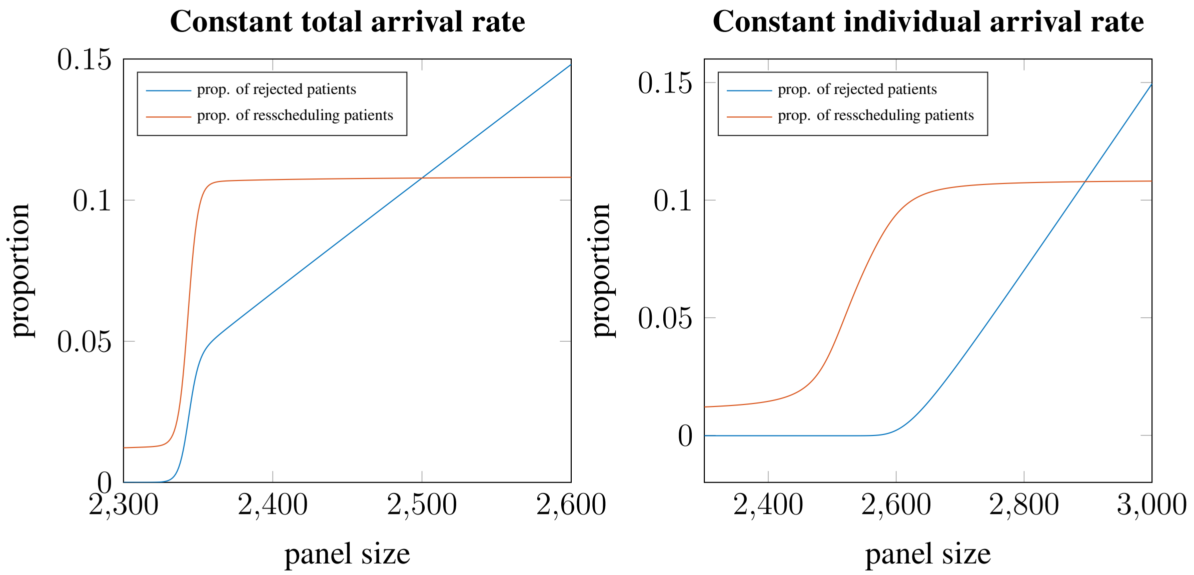

**Figure 4.16:** Proportion of rescheduling and rejected patients for a constant total and individual arrival rate

### Comparison with the results from Green and Savin (2008)

In this section, we reproduce Figures 2, 3 and 4 from Green and Savin (2008) for their $M/D/1/K$ queueing model using our implementation. Figure 2 shows the expected appointment backlog in days dependent on the panel size with and without no-shows. Figure 3 shows the same-day appointment probability dependent on the panel size with and without no-shows. Finally, Figure 4 shows the physician's utilization dependent on the expected appointment backlog in days with no-shows.

We start with using a constant total arrival rate as in Green and Savin (2008). Note that the authors transform a queue length $i$ into an indirect waiting time of $\frac{i}{\mu}$ days which is the expected indirect waiting times in days according to Equation (3.52). We calculate the expected indirect waiting time in days based on the indirect waiting times in days distribution. The two approaches yield slightly different results because the relationship between the expected waiting time in days calculated based on the distribution and the expected queue length is non-linear.

Instead of changing the panel size, the physician can also change the (average) daily number of appointment slots offered $\mu$ to achieve a certain service level. Hence, Figure 4.17 shows the expected indirect waiting time in days dependent on the panel size and the daily capacity of appointments $\mu$ with a fixed panel size of $N = 2344$. Here, we differentiate between models without no-shows, with no-shows and a no-show rescheduling probability of $0.5$, and with no-shows and a no-show rescheduling probability of $1$. Note that most results do not differ whether we have a model with no-shows and a rescheduling probability of $r^n = 0$ or a model without no-shows. This is because a slot with a no-show patient is still served.

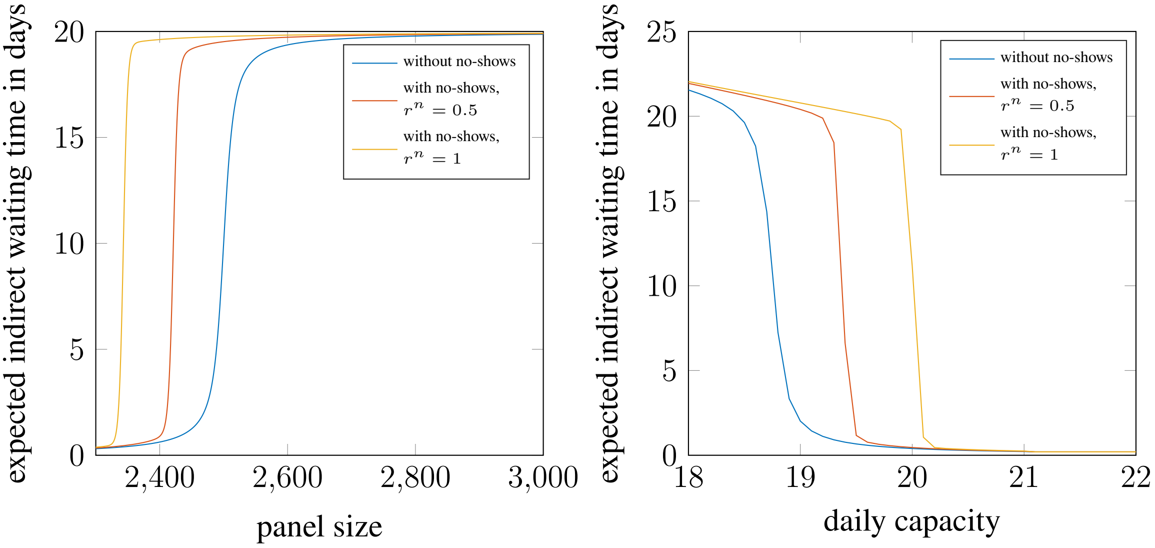

**Figure 4.17:** Expected indirect waiting time in days dependent on the panel size and the daily capacity (with $N = 2344$) for a constant total arrival rate

Comparing our results with those in Figure 2 of Green and Savin (2008), we observe a similar behavior of the curves. We see a steep increase in the expected indirect waiting time once a certain panel size is exceeded. This increase is a little smoother and starts for larger panel sizes if no-shows reschedule less often or are not considered at all. However, we see a difference in the results. The increase in the expected indirect waiting time starts for smaller panel sizes in our case. Interestingly, our results for $r^n = 1$ are very similar to those presented in Figure 1 in Zander (2017). Therefore, we conclude that the differences to the results of Green and Savin (2008) are not mainly due to the usage of the departure rates $d^{r*}$ and $d^{nr*}$ instead of $d^*$ alone since both Green and Savin (2008) and Zander (2017) only use $d^*$. Hence, there must be different reasons, e.g., small calculation errors in the formulas for the recursion $f$ in Green and Savin (2008). However, based on the comparison of the results in Zander (2017) and our results, we observe that the difference due to using $d^*$ alone instead of $d^{r*}$ and $d^{nr*}$ seems to be small, at least for this special instance. As remarked before in Section 3.4, this is due to a generally small difference between $d^{nr*}$ and $d^*$ which in turn is due to small values of $1 - \nu = 1 - (1 - \gamma(k))(1 - r^s(k)) + \gamma(k)(1 - r^n(k)) = \gamma(k)$.

We now recreate Figure 3 of Green and Savin (2008), showing the same-day appointment probability dependent on the panel size with and without no-shows. As explained before in Section 3.4.6, we use a different definition of the same-day appointment probability. Figure 4.18 shows our results for the same-day appointment probability dependent on the panel size and the daily capacity for a panel size $N = 2344$. In contrast to Green and Savin (2008), a patient seeing a queue length of less than $\mu = 20$ patients will not automatically get a same-day appointment, e.g., if the patient arrives during the $12^{th}$ time slot of the day and $10$ patients are already waiting. Therefore, our curves lie underneath the curves of those from Green and Savin (2008).

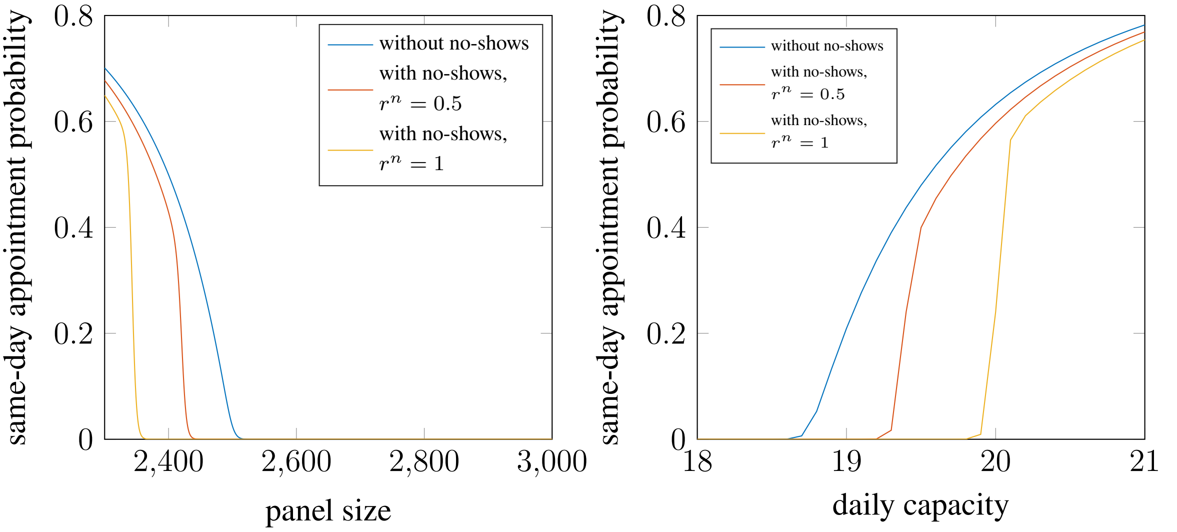

**Figure 4.18:** Same-day appointment probability dependent on the panel size and on the daily capacity (with $N = 2344$) for a constant total arrival rate

Next, we reproduce Figure $4$ of Green and Savin (2008) that shows the physician's utilization dependent on the expected appointment backlog in days with no-shows. In Section 3.4.6, we defined the proportion of idle time of the physician, which is, in fact, one minus the physician's utilization. Figure 4.19 shows our results for the physician utilization. Here, we have to differentiate between a model without no-shows and a model with no-shows and a rescheduling probability of $r^n = 0$. Comparing our results to those of Green and Savin (2008), we observe similar behavior of the curves in the case of $r^n = 1$. However, the maximum and minimum physician utilization values reached in our case are approximately $0.93$ and $0.89$, respectively. Whereas, for Green and Savin (2008) those values are approximately $0.94$ and $0.92$. We believe that our results are more accurate because for a queue length of $20$ days, we obtain an approximate no-show probability of $\gamma^d(20) = 0.11$. At the same time, the probability of an empty queue is almost zero. Therefore, an approximate value of $0.89 = 1 - 0.11$ for the physician utilization in the extreme case of $20$ days indirect waiting makes sense. Comparing the results for the different models with and without no-shows, we observe, as expected, that the physician utilization tends to $1$ in the case with no no-shows. In contrast, physician utilization first rises and reaches a maximal value before decreasing for the models with no-shows. The increase at first is because the probability of an empty queue decreases with longer indirect waiting times. However, at the same time also the proportion of no-shows increases with the indirect waiting time. Therefore, physician utilization begins to decline at some point. Hence, even though the system is overloaded and patients experience long indirect waiting times, the physician is under-utilized due to no-shows. For $r^n = 0$, the maximal utilization value of $\approx 0.9663$ is reached for a panel size of $2471$ and an indirect waiting time of $2.15$ days. For $r^n = 0.5$ the maximal utilization of $\approx 0.9542$ is reached for a panel size of $2408$ and an indirect waiting of $1.41$ days. Moreover, for $r^n = 1$, the maximal utilization of $\approx 0.9316$ is reached for a panel size of $2332$ and an expected indirect waiting of $1.11$ days.

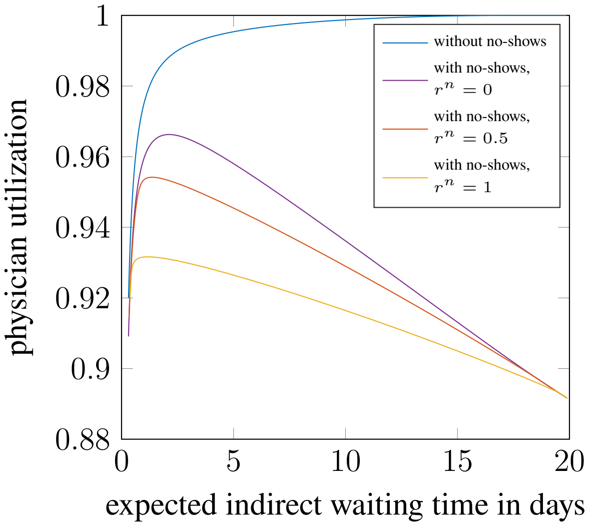

**Figure 4.19:** Physician utilization dependent on the indirect waiting time in days for a constant total arrival rate

Next, for a constant individual arrival rate $\eta$ with a total arrival rate function $\lambda(k) = \eta(N - k)$, we show the curves for the expected indirect waiting time dependent on the panel size and the daily capacity for panel size $N = 2540$. The results can be seen in Figure 4.20.

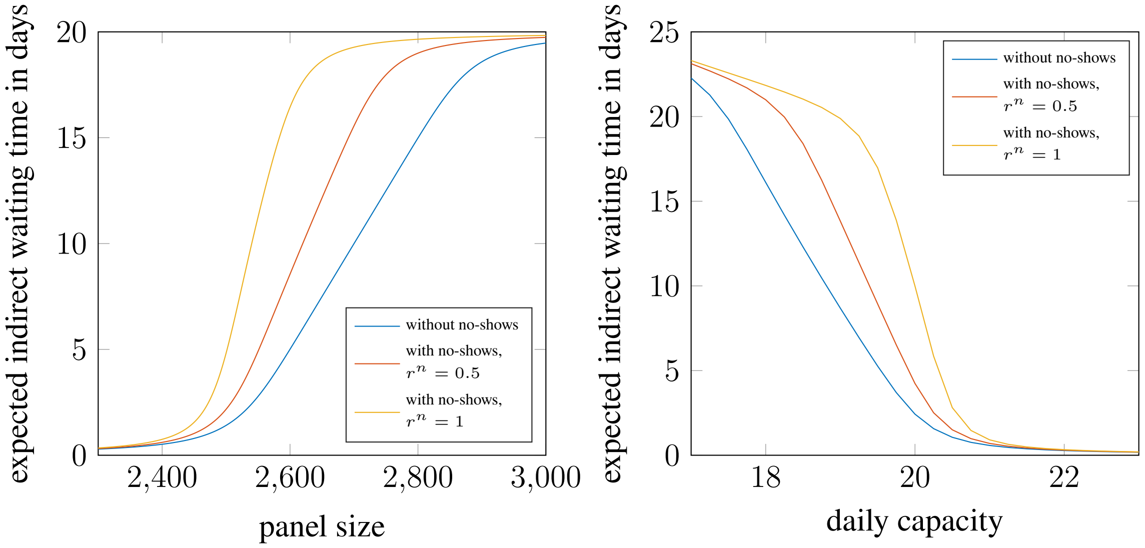

**Figure 4.20:** Expected indirect waiting time in days dependent on the panel size and on the daily capacity (with $N = 2540$) for a constant individual arrival rate

In comparison to Figure 4.17, we notice a much smoother transition from an empty queue to a full queue for all three curves. This is because the appointment request rate decreases with increasing panel size.

Finally, for a non-constant individual arrival rate $\eta$ with a total arrival rate function $\lambda(k) = \eta(N-k)$, we show the curves for the expected indirect waiting time dependent on the panel size and the daily capacity for panel size $N = 2340$. The results can be seen in Figure 4.21. Similar to the case of a constant total arrival rate, we see a very steep increase in the expected indirect waiting time once a particular panel size is exceeded. As expected, this increase happens for smaller panel sizes compared to Figure 4.17.

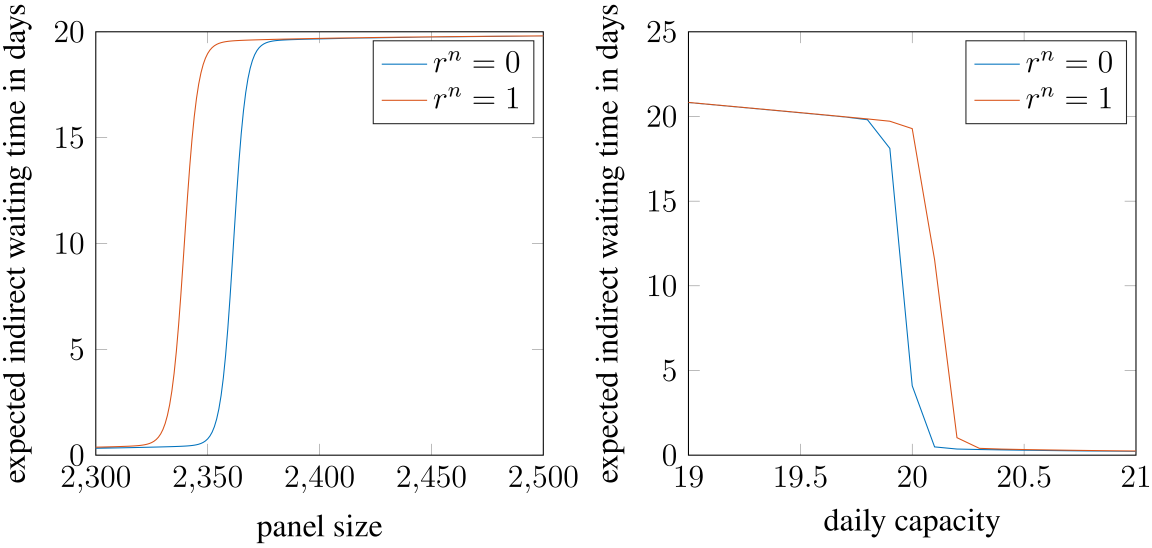

**Figure 4.21:** Expected queue length dependent on the panel size and on the daily capacity (with $N = 2340$) for a non-constant individual arrival rate

See Sections A.1.4 and A.1.5 to learn more about the same-day appointment probability and the physician utilization for a constant and non-constant individual arrival rate.

### Sensitivity analyses for the individual arrival rate and the queue capacity

First, we investigate the sensitivity with respect to the individual arrival rate $\eta$. We plot the expected indirect waiting time in days dependent on the panel size for different values of $\eta$ for a constant total and individual arrival rate in Figure 4.22. We see that the results are quite sensitive with respect to $\eta$. Interestingly, we see that a difference of $0.005$ in the individual arrival rate produces an approximate difference of $150$ in the panel size for both graphs.

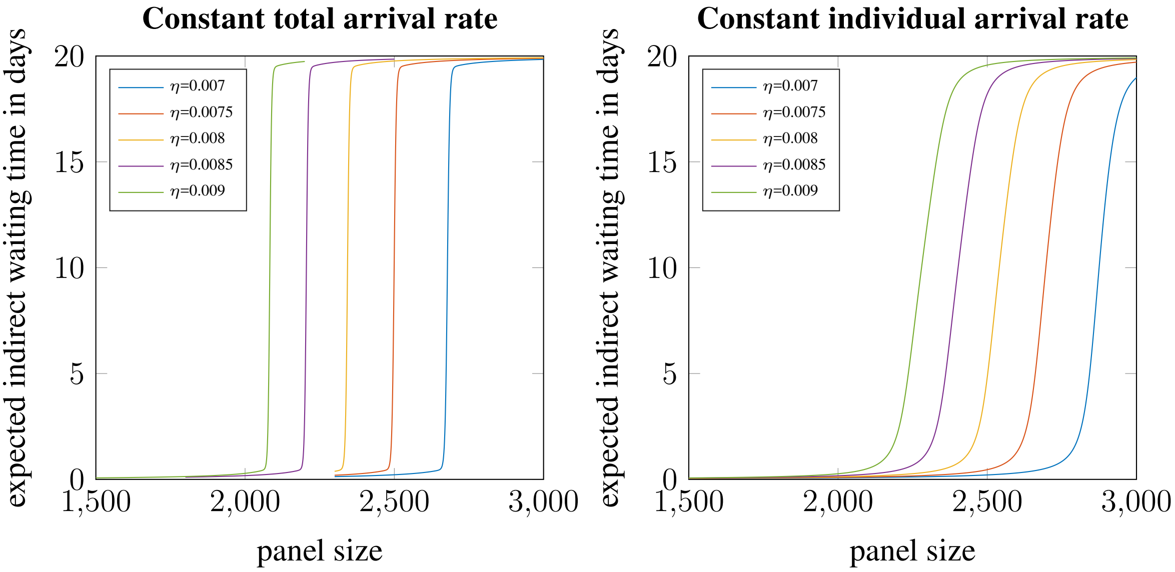

**Figure 4.22:** Expected indirect waiting time in days dependent on the panel size for different values of the individual arrival rate for a constant total and individual arrival rate

See Sections A.1.6 to learn more about the sensitivity with respect to the individual arrival rate for a non-constant individual arrival rate.

Next, we investigate the influence of the queue capacity on the panel size and the rate of rejected patients aiming to keep the same expected indirect waiting time of 11.17 days.

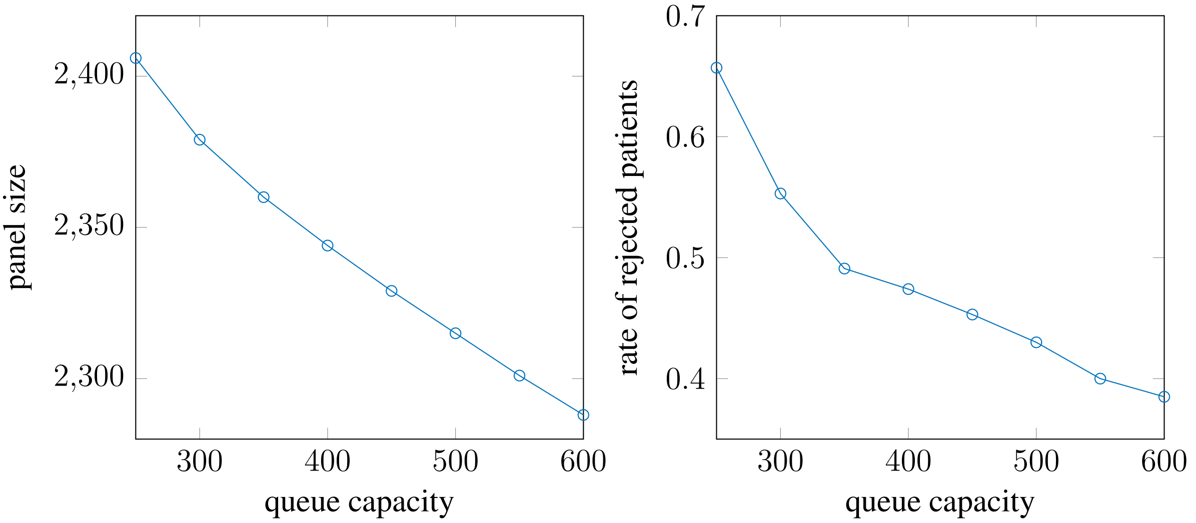

**Figure 4.23:** Panel size and rate of rejected patients dependent on the queue capacity such that the expected indirect waiting time remains approximately constant for a constant total arrival rate

For a constant total arrival rate, in Figure 4.23, we see that choosing the queue capacity functions as a trade-off between the panel size and the rate of rejected patients. By choosing a small queue capacity, one can manage a bigger panel size. However, patients also experience higher rejection probabilities, whereas a large queue capacity stands for a smaller panel size and a lower rejection rate.

In the case of a constant individual arrival rate, we see the same tradeoff between the panel size and the rate of rejected patients aiming to keep the same expected indirect waiting time of $10.67$ days. However, the decrease in the panel size and the rate of rejected patients happens faster in Figure 4.24 compared to the case of a constant total arrival rate.

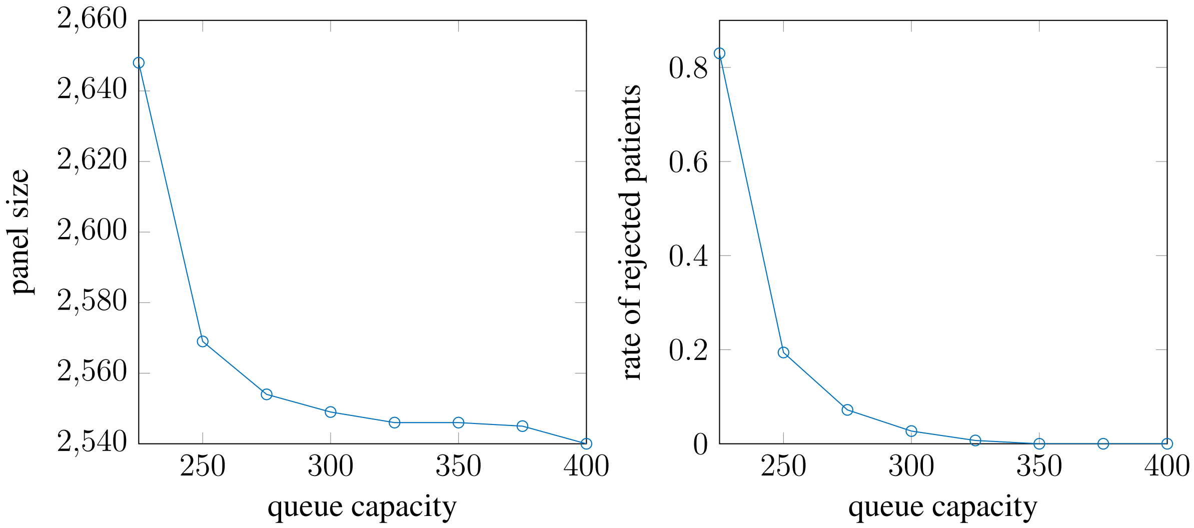

**Figure 4.24:** Panel size and rate of rejected patients dependent on the queue capacity such that the expected indirect waiting time remains approximately constant for a constant individual arrival rate

### 4.4.7 Integrating rescheduling arrivals into the total arrival rate

In this section, we research the effect of integrating the rescheduling arrivals into the total arrival rate. In this case, we set the rescheduling probabilities to zero and instead add the rate of rescheduling patients to the total arrival function $\lambda$. We do this via setting $\lambda$ equal to the effective arrival rate $\varepsilon(k) = \lambda(k) + \frac{d^{r*}(k)}{\pi(k)}$ (see Section 3.4.6). An approximation to the effective arrival rate would be $\lambda + \frac{1}{T}\rho(\cdot - 1)$. Remember that $\rho(k-1)$ is the rescheduling probability of a patient who leaves $k-1$ patients in the queue after the patient departs and directly reschedules (the patient itself is not counted). So approximately assuming a rate of departing patients of $\frac{1}{T}$ during the time the queue is of length $k$, we can approximate the rate of rescheduling patients as $\frac{1}{T}\rho(k-1)$. As we have argued before, the average departure rate given a queue of length $k$ in steady state is generally not equal to $1/T$. This is because we generally do not see an equal distribution of service progressions.

In the left part of the Figures 4.25 and 4.26, we compare the effective arrival rate $\varepsilon$ with the function $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ where we define $\rho(-1) = 0$ with panel sizes $N = 2344$ and $N = 2540$

for a constant total arrival rate and for a constant individual arrival rate, respectively. As expected we see a difference between the two curves especially in the case of a constant total arrival rate confirming that given a queue length $k \in \{1, \ldots, K\}$ the rescheduling departure rate $\frac{dr^*(k)}{\pi(k)}$ is not identical to $\frac{1}{T}\rho(k-1)$. However, in both cases the two curves still lie close to each other.

In the right part of the Figure 4.25, we compare the queue length distribution for $N = 2344$ with the queue length distributions resulting from a model with the total arrival rate equal to the effective arrival rate, or equal to $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ for a constant total arrival rate. The three curves are similar; however, the models using the effective arrival rate as the total arrival rate or $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ as the total arrival rate slightly overestimate the probability of small and underestimate the probability of long queue lengths. The model using the effective arrival rate deviates even more from the original model than the model using $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$.

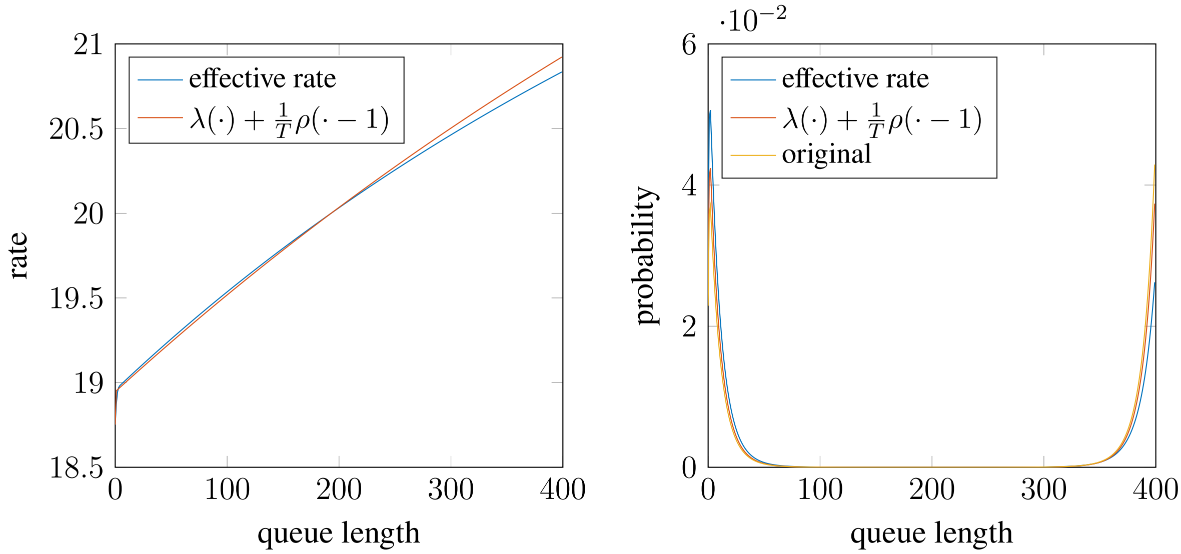

**Figure 4.25:** Comparison of the effective arrival rate for $N = 2344$ with the function $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ and the resulting queue lengths distributions using the original model, the effective arrival rate as the total arrival rate, and $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ as the total arrival rate for a constant total arrival rate

In the right part of the Figure 4.26, we compare the queue length distribution for $N = 2540$ with the queue length distributions resulting from a model with the total arrival rate equal to the effective arrival rate, equal to the effective arrival rate for the queue length range $[100, 300]$ and with continuous constant extension for $[0, 99]$ and $[301, 400]$ or equal to $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ for a constant individual arrival rate.

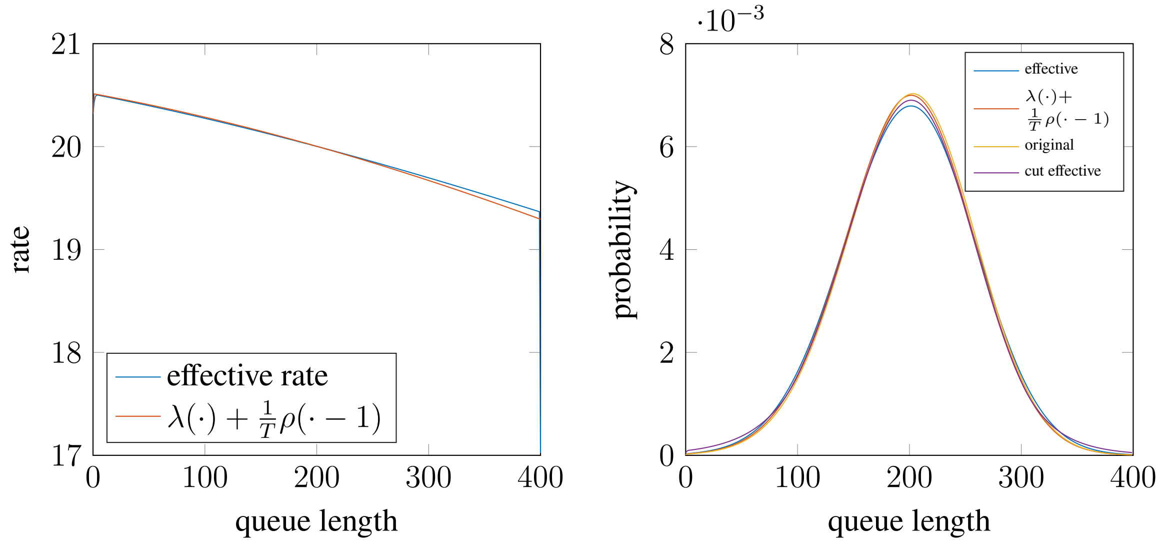

**Figure 4.26:** Comparison of the effective arrival rate for $N = 2540$ with the function $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ and the resulting queue lengths distributions using the original model, the effective arrival rate as the total arrival rate, a cut version of the effective arrival rate as the total arrival rate and $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ as the total arrival rate for a constant individual arrival rate

Although the four curves are similar, we see significant differences with greater differences between the original model and the usage of the effective arrival rate as the total arrival rate compared to the original model and the usage of $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ as the total arrival rate.

In practice, we may not know the underlining model of our data and be able to differentiate between general arrivals and rescheduling arrivals. Instead, we could use the observed values of the effective arrival rate as an input. Further, we may not be able to observe every queue length and hence extrapolate the effective arrival rate. Our test shows that even using a not very sophisticated extrapolation for the less probable queue lengths approximates the queue length distribution of the original model quite well.

Here is another interesting observation. The effective rate for the constant total arrival rate increases with increasing queue lengths, which means the longer the queue, the longer it gets, and the shorter the queue, the shorter it gets. This explains the resulting queue length distribution with two spikes. On the other hand, the effective rate for the constant individual arrival rate decreases with increasing queue length, meaning that long queue lengths tend to get shorter and short queue lengths tend to get longer. This results in the queue length distribution with one maximum close to the expected queue length.

### 4.4.8 Results for linear total arrival rates

In this section, we study linear total arrival rates where we assume the rescheduling probabilities to be zero. First, we vary the slope keeping the point $(200, 20)$ on the linear function. The results can be seen in Figure 4.27. We see that the expected queue length remains approximately the same, namely $200$. The shape of the queue length distribution, however, changes with the slope $m$. The distribution variance increases with $m$ such that we observe two spikes when $m > 0$, a uniform distribution for $m = 0$ and a distribution with one maximum for $m < 0$.

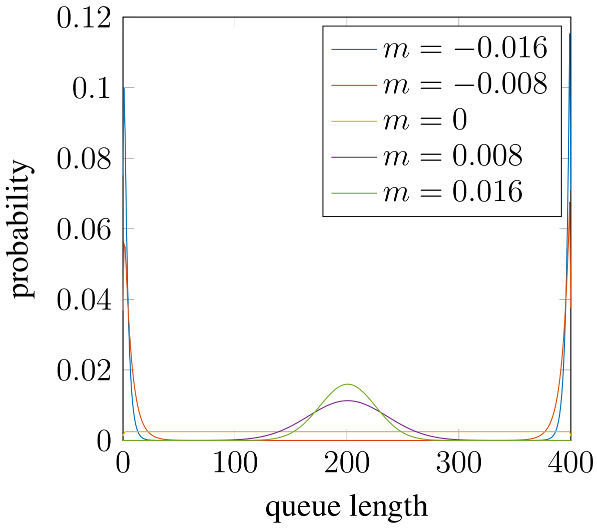

**Figure 4.27:** Comparison of the queue length distributions for different linear total arrival rates such that $200m + n = 20$

Next, we keep the same slope $m = -0.008$ but vary the intercept $n$. We see in Figure 4.28 that the shape of the queue length distribution stays the same, but the maximum, which is reached for the expected queue length, shifts. It shifts to the queue length $q$ with $20 = mq + n$.

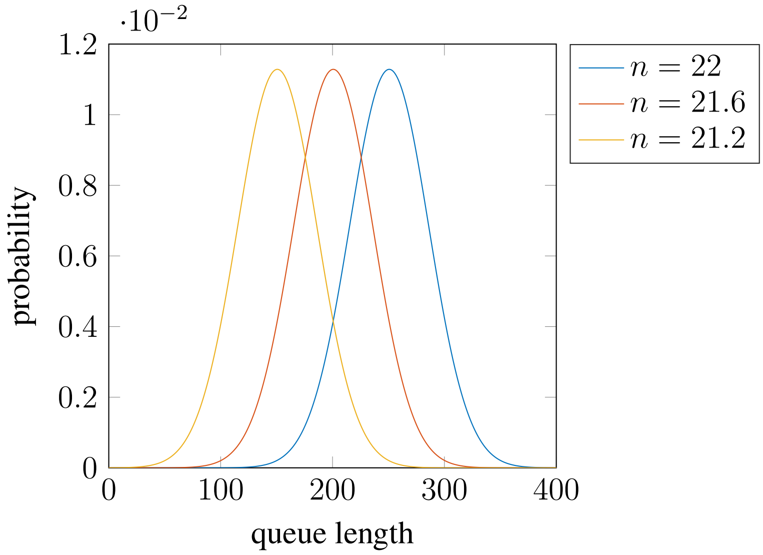

**Figure 4.28:** Comparison of the queue length distributions for different linear total arrival rates such that $m = -0.008$

### 4.4.9 Results for different arrival distributions

This section compares the expected indirect waiting times resulting from using distributions for the number of arriving patients during a service period with different coefficients of variation. We consider the case of a constant total arrival rate and compare using Poisson arrival as in Equation (3.50) with using a binomial distribution $B(n,p)$ with $n$ trials and success probability $p$. The expected value of the binomial distribution $np$ is set to the expected value of the Poisson distribution but with different coefficients of variation $1-p$. The smaller we choose $n$, the bigger is $p$, and the smaller is the coefficient of variation. In Figure 4.29, we compare Poisson arrivals with binomial arrivals for $n = 2,3$ and $4$. We observe that the increase in the expected waiting time happens later and is steeper for smaller coefficients of variation. These findings confirm the results of Izady (2015) who used a discrete Weibull distribution with a varying coefficient of variation.

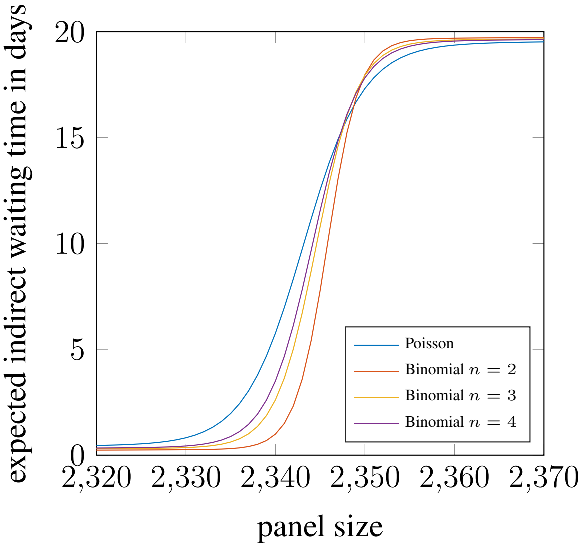

**Figure 4.29:** Comparison of the expected indirect waiting time dependent on the panel size for different arrival distributions for a constant total arrival rate

### 4.4.10 Results for variable appointment offerings

Now, we investigate if we can alter the form of the indirect waiting time distribution using a queue length-dependent number of offered appointment slots per day as explained in Section 3.5.1.

We start with a constant total arrival rate and use a panel size of $N = 2344$. If we assume a constant number of appointments slots offered per day, namely $20$, we find an expected indirect waiting time of $11.17$ days. Let us assume that the physician offers $19$ time slots per day as long as the queue length is less than $50$. If the queue length lies between $50$ and $150$ the physician offers $20$ slots, and for a queue length of $150$ or more, the physician offers $21$ slots. Putting this into the model, we find an expected indirect waiting time of $2$ days. The expected number of appointment slots offered per day is $19.3$, which means that the physician works less on average, and at the same time, patients experience less indirect waiting time. Even better, looking at the distribution of the indirect waiting time in Figure 4.30, we see that there is only one spike instead of two. This means that the expected indirect waiting time is a far better indicator for the actual indirect waiting time than in the case of a constant number of offered appointment slots per day. In conclusion, a little bit of flexibility concerning working hours can lead to a much better performance of the whole system. Please note that even though the results of a variable number of appointment slots offered per day are very promising, we assume in our theoretical model that the physician can switch from a workday with, for example, $21$ slots to a workday with $20$ slots immediately. In reality, patients have concrete appointment days and times, and therefore we assume that the new working hours can only be applied for new incoming patients and not for those already waiting.

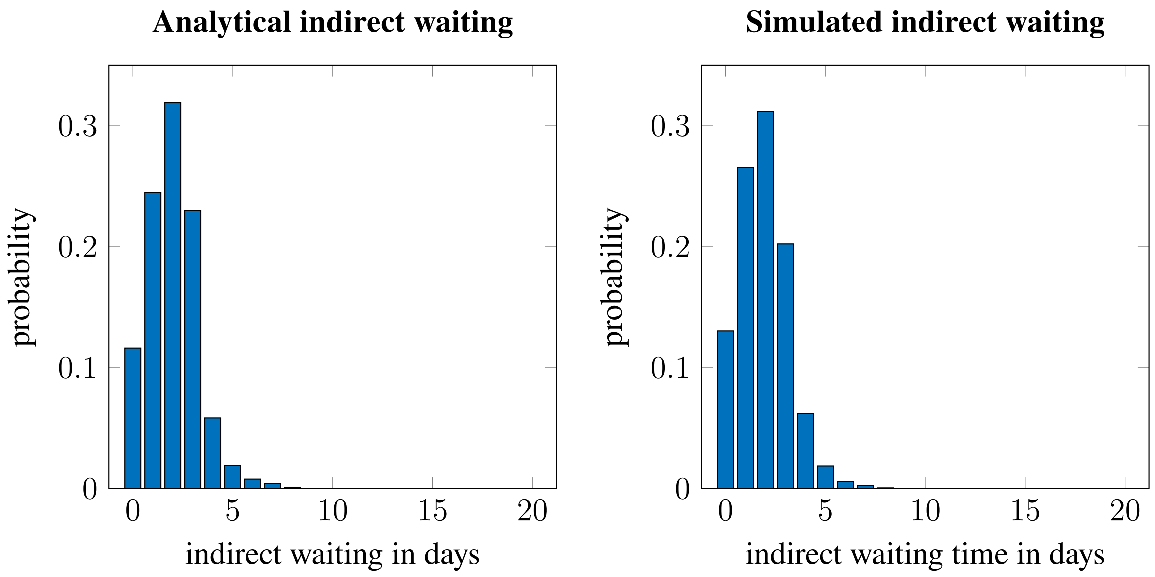

**Figure 4.30:** Analytical and simulated indirect waiting time distribution for N=2344 with a variable number of offered slots per day for a constant total arrival rate

We use simulation to include this effect. Using $15$ replications, an initial queue length of $400$, a total number of $100,000$ simulated patients with an initial transient phase of $20,000$ patients we obtain $\widehat{W} = 1.9$ and a $95\%$ confidence interval of: $[1.83, 1.97]$. The estimated expected indirect waiting time is again close to the analytically determined value of $2$. This shows a good performance of our queueing model; even so, it relies on simplified assumptions. Also, the indirect waiting time distribution produced by the simulation is very similar to the analytically calculated one, as can be seen in Figure 4.30.

We also investigate the case of a constant individual arrival rate with panel size $N = 2540$ using a variable number of offered appointment slots per day. If we assume a constant number of appointment slots offered per day, namely $20$, we find an expected indirect waiting time of $10$ days. Let us assume that the physician offers $19$ time slots per day as long as the queue length is less than $50$. If the queue length is between $50$ and $150$, $20$ slots are offered, and for a queue length of $150$ or more, the physician offers $21$ slots. Putting this into the model, we find an expected indirect waiting time of $6.3$ days. The expected number of appointment slots offered per day is $20.23$, which means that the physician works a little bit more on average, but at the same time, patients experience less indirect waiting time. Figure 4.31 shows the indirect waiting time distribution. In comparison to Figure 4.12, we obtain an indirect waiting time distribution with a much smaller variance.

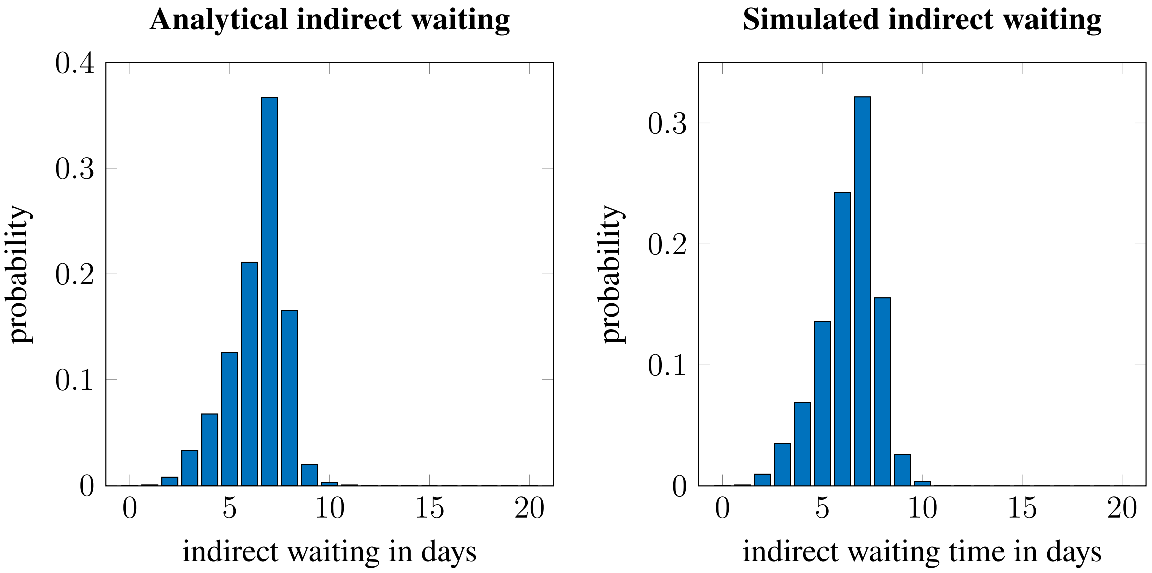

**Figure 4.31:** Analytical and simulated indirect waiting time distribution for $N = 2540$ and a variable number of appointment slots offered per day for a constant individual arrival rate

We validate the analytical results via simulation. Using $15$ replications, an initial queue length of $130$, a total number of $100,000$ simulated patients with an initial transient phase of $20,000$ patients we obtain $\widehat{W} = 6.31$ and a $95\%$ confidence interval of: $[6.27, 6.34]$. Again, there is only a slight difference between the simulated and the analytically calculated expected indirect waiting time. Further, also the queue length distributions are very similar, as can be seen in Figure 4.31.

Next, we investigate the effects of random service times in the case of a constant individual arrival rate with panel size $N = 2540$. Here, we assume that the number of appointment slots offered per day follows a Poisson distribution (where we cut the option to offer zero appointments) with an expected value equal to the number of appointment slots offered for a given queue length as in the previous experiments. In the mathematical model, the service time can change from one patient to the next. In reality, we plan whole days with the same service time. To take this into account in the model implementation, we weigh the probability of a service time with the corresponding number of appointments per day. Then, we find an expected indirect waiting time of $6.4$ days. The expected number of appointment slots offered per day is $20.24$. The values are very similar to those with a deterministic number of appointment slots offered. Also, the indirect waiting time distribution in the left part of Figure 4.32 is very similar to the distribution in the deterministic case.

Again, we validate the analytical results via simulation. Using $15$ replications, an initial queue length of $130$, a total number of $100,000$ simulated patients with an initial transient phase of $20,000$ patients we obtain $\widehat{W} = 6.19$ and a $95\%$ confidence interval of: $[6.14, 6.24]$. Again, only a small difference exists between the simulated and the analytically calculated expected indirect waiting time. Further, comparing the indirect waiting time distribution, we observe a higher variance in the simulated distribution compared to the analytical one in Figure 4.32.

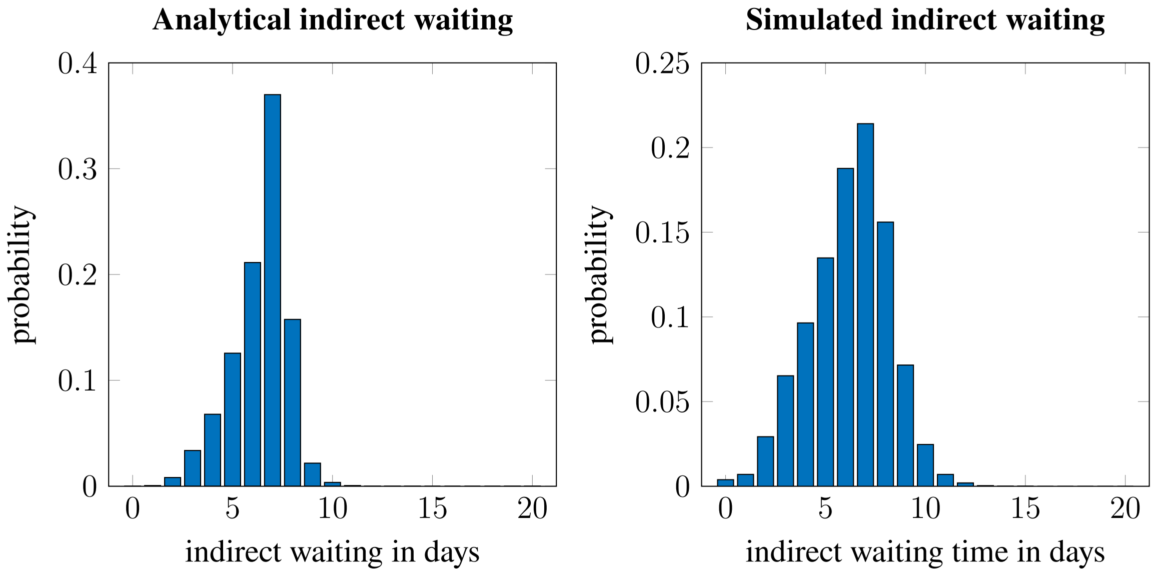

**Figure 4.32:** Analytical and simulated indirect waiting time distribution for $N = 2540$ and a random number of appointment slots offered per day for a constant individual arrival rate

Remember form Section 3.5.1 that for the determination of the indirect waiting time with the number of arrivals following a Poisson distribution, changing $T(k)$ is equivalent to chancing $\lambda(K)$ if we assume no implicit dependencies on $T$. Therefore, if we integrate the rescheduling arrivals into the total arrival rate as explained in Section 4.4.7, we see that changing $T(k)$ approximately reduces to changing $\lambda$. Therefore, for practical applications it is enough to study the influence of different total arrival rate functions $\lambda$ without rescheduling patients or queue length-dependent service times.

See Section A.1.7 in the Appendix to learn more about the effect of using variable appointment offerings in case of a constant and non-constant individual arrival rate.

### 4.4.11 Results for different constant individual arrival rates for panel patients

In this section, we relax the assumption of a single constant individual appointment request rate that is the same for all patients. Using our queueing model with a panel size of $N = 2540$, we can compare the already determined expected indirect waiting time of $9.99$ days in the case of one patient group (with an individual request rate of $0.008$) with the expected indirect waiting times assuming two patient groups with sizes $N_1 = N_2 = 1270$ and individual arrival rates $\eta_1$ and $\eta_2$ with $\eta_1 + \eta_2 = 0.008$. We use the total arrival rate $\lambda$ calculated as explained in Section 3.5.2 in our queueing model. In Figure 4.33, we show the queue length distributions for $\eta_1 = 0.008, 0.006, 0.004$ and $\eta_1 = 0.002$. We observe that the higher the variation in the individual arrival rates, the smaller the expected queue length. In a model with two different individual arrival rates, there is a higher probability of finding patients with the biggest individual arrival rate in the queue. In turn, this leads to a smaller total arrival rate given a certain queue length compared to the original model with one individual arrival rate $\eta_1 = \eta_2 = 0.008$. Smaller total arrival rates then lead to shorter queue lengths.

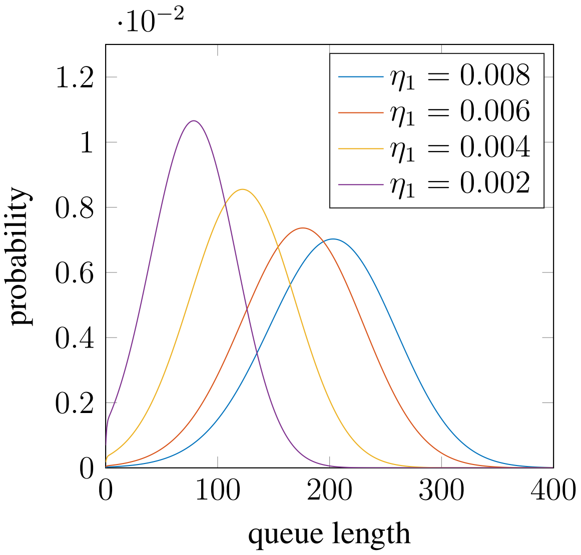

**Figure 4.33:** Comparison of the queue length distributions for two subgroups $N_1 = N_2 = 1270$ with individual arrival rates $\eta_1 + \eta_2 = 0.008$

In Figure 4.34, we compare the analytical determined indirect waiting time distribution for two subgroups $N_1 = N_2 = 1270$ with individual arrival rates $\eta_1 = 0.006$ and $\eta_2 = 0.01$ with the realistically simulated indirect waiting time distribution. We make the no-show function dependent on the queue length at arrival times and assign patients to concrete days and timeslots.

As we model patients individually in our simulation model, we can assign individual appointment request rates that are different for each patient. Using $15$ replications, an initial queue length of $200$, a total number of $100,000$ simulated patients with an initial transient phase of $40,000$ patients we obtain $\widehat{W} = 8.69$ and a $95\%$ confidence interval of: $[8.54, 8.83]$. The estimated expected indirect waiting time is close to the analytically determined value of $8.67$. This shows the validity of the approximation used in the analytical model, namely that we still use a total arrival rate dependent on the queue length instead of using a two-dimensional model.

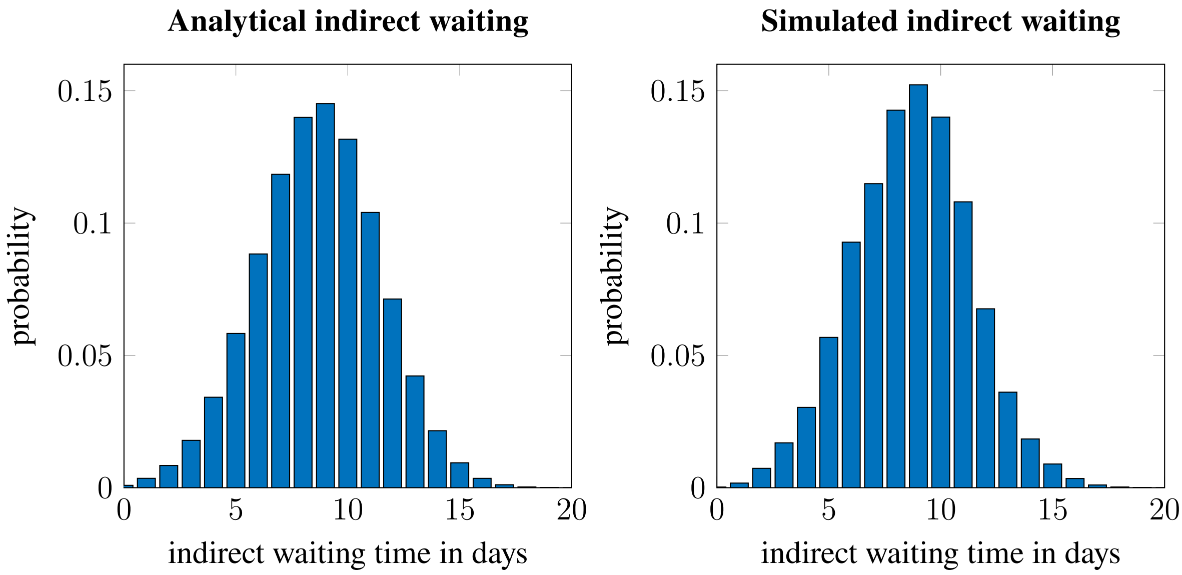

**Figure 4.34:** Analytical and simulated indirect waiting time distribution for two subgroups $N_1 = N_2 = 1270$ with individual arrival rates $\eta_1 = 0.006$ and $\eta_2 = 0.01$

### 4.4.12 Results for a mix of panel and non-panel patients

Here, we investigate the queue length distribution for a physician with a mix of panel and non-panel patients. In Figure 4.35, we plot the queue length distribution for panel sizes $N$ and constant non-panel arrival rates $r^{np}$: $N = 2500, r^{np} = 0$; $N = 2300, r^{np} = 1.6$; $N = 1500, r^{np} = 8$ and $N = 700, r^{np} = 14.4$ with constant individual arrival rate $\eta = 0.008$ for panel patients. For increasing values of $r^{np}$ the queue length distribution shifts to the right and changes shape from flat to peak. We see a transition between a model with a constant individual arrival rate to a model with a constant total arrival rate.

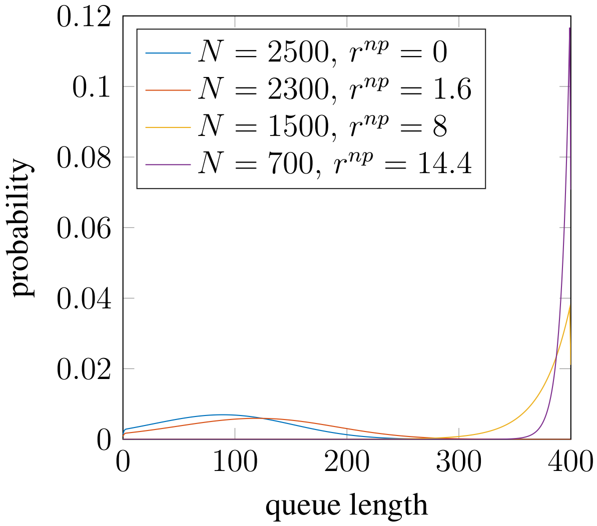

**Figure 4.35:** Comparison of the queue length distributions for different panel sizes $N$ and arrival rates of non-panel patients $r^{np}$ for a constant individual arrival rate for panel patients

In Figure 4.36, we compare the analytical determined indirect waiting time distribution for $N = 2300$, $r^{np} = 1.6$ with the realistically simulated indirect waiting time distribution.

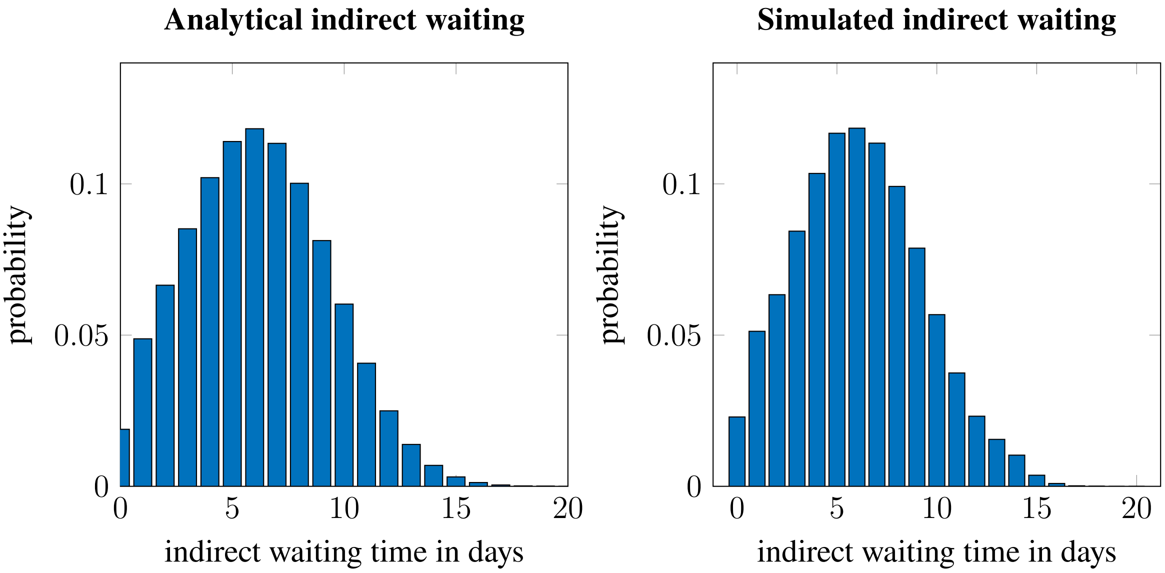

**Figure 4.36:** Analytical and simulated indirect waiting time distribution for $N = 2300$ and $r^{np} = 1.6$ for a constant individual arrival rate

We make the no-show function dependent on the queue length at arrival times and assign patients to concrete days and timeslots. Additionally, we model panel patients individually. The non-panel patients are represented via a constant Poisson arrival rate. Using $15$ replications, an initial queue length of $200$, a total number of $100,000$ simulated patients with an initial transient phase of $20,000$ patients we obtain $\widehat{W} = 6.11$ and a $95\%$ confidence interval of: $[5.96, 6.27]$. The estimated expected indirect waiting time is close to the analytically determined value of $6.15$. Similar to the experiment with different individual arrival rates, this shows the validity of the approximation used in the analytical model, namely that we still use a total arrival rate dependent on the queue length instead of using a two-dimensional model.

### 4.4.13 Results for fair capacity division between two independent demand streams

In this section, we run experiments to investigate two separated schedules as explained in Section 3.5.3. We assume a physician with a total number $N = 2600$ of panel patients, all having the same constant individual arrival rate, and with a capacity of $\mu = 20$ appointment slots per day. These panel patients are split up into two patient groups $N_1$ and $N_2$ with capacities $\mu_1$ and $\mu_2$ per day. Assuming a booking horizon of 20 days each patient group has a queue capacity of $K_{1,2} = 20\mu_{1,2}$. For $\mu_1 = 1, \dots 10$ and $\mu_2 = \mu - \mu_1$, we aim to find $N_1$ and $N_2 = N - N_1$ such that the difference in the expected indirect waiting time for both groups is minimal. This is done via enumeration. The results can be seen in the left side of Figure 4.37.

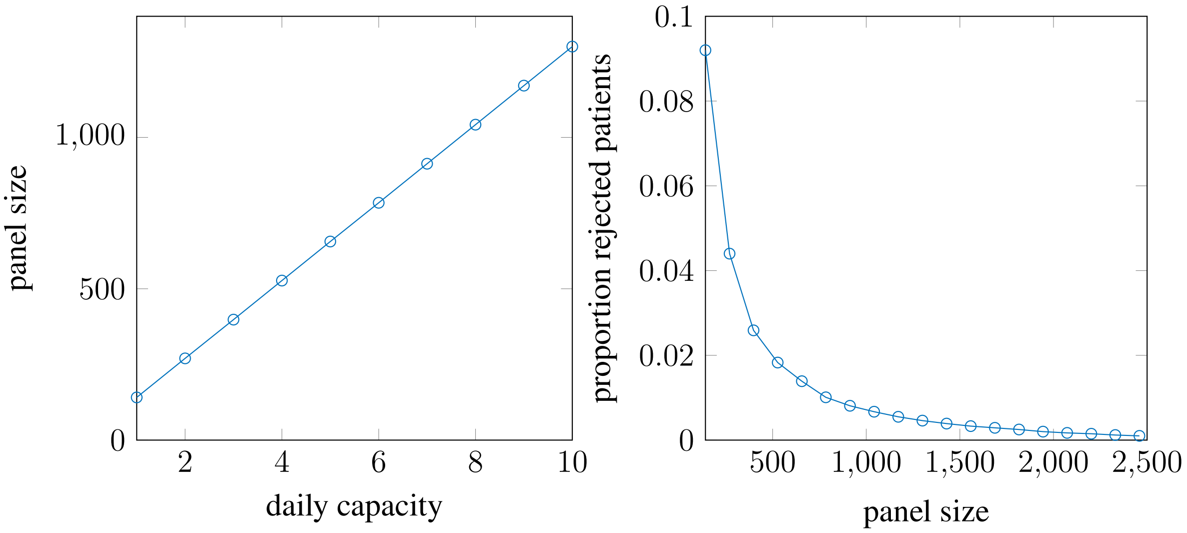

**Figure 4.37:** Panel size $N_1$ dependent on the daily capacity $\mu_1$ and proportion of rejected patients dependent on the panel size $N_1$ for a constant individual arrival rate

Interestingly, a very similar indirect waiting time of approximately $15.4$ to $15.7$ days can be found for every value of $N_1$. Further, the curve shows an almost linear relationship. In fact, as one might have expected, we approximately have: $\frac{N_1}{N_2} = \frac{\mu_1}{\mu_2}$. However, note that for $N = 2600$ and $\mu = 20$, the indirect waiting time in days is higher with a value of $16.5$ days. A single queue with a panel size of $N = 2589$ would reach an indirect waiting time of $15.6$ days. At first sight, the results might seem counterintuitive, two queues having a smaller indirect waiting time than one single queue. This is because we did not consider rejected patients so far. On the right side of Figure 4.37, we see that queues with bigger panel sizes have a smaller proportion of rejected patients.

## 4.5 Application in practice

In this section, we shortly answer the question how our queueing model, possibly with approximations, can be applied in practice. Our model is designed for physicians operating under the traditional appointment policy, where every patient has to book an appointment. However, in practice, many physicians experience walk-ins and reserve capacity for same-day demand. Our model can still be applied, focusing on appointment demand and the capacity reserved for this demand. Next, we need to know if the physician reserves capacity for different patient types. In this case, we have to work with several separate queueing models, as in Section 3.5.3. Furthermore, to apply the model, the majority of patients need to book the next available appointment or the next available appointment after a some fixed time interval to use interconnected queuing models, as explained in Section 3.5.3.

In the following, we will address the model features and explain their relevance and application in practice. In our mathematical model, appointment requests can only arrive during the working hours of the physician. Therefore, in practice, all appointment requests coming in outside the physician's working hours, e.g., appointment requests from an online booking system, need to be distributed such that an arrival distribution solely during the working hours can be constructed.

The mathematical model further differentiates between non-rescheduling arrivals and rescheduling arrivals. This differentiation is the reason for the complicated model structure. However, in Section 4.4.7 of the numerical experiments, we saw that integrating the rescheduling requests into the general requests, i.e., using the effective arrival rate as the total arrival rate, only leads to small differences in the results. Therefore, in practice, it should be enough to define the arrival function $\lambda$, including all requests and setting the rescheduling probability functions $r^s$ and $r^n$ to zero.

When defining the dependency of $\lambda$ on the queue length, we have to keep in mind that this dependency, in turn, may depend on the appointment offerings. This means that when we change, for example, the model parameter of daily appointment offerings $\mu$, we also need to adjust the rescheduling functions, and the appointment request function accordingly.

In the queueing model, the total arrival rate is dependent on the queue length. However, from historical data, we will, in general, not be able to determine the total arrival rate accurately for all possible queue lengths. As we saw in the numerical experiments in Section 4.4.7, it seems to be enough to use a function $\lambda$ which is accurate for queue lengths around the expected queue length, hence exactly for the range of queue lengths for which we should have enough data. It is theoretically possible to work with general arrival distributions. However, we would not advise doing so in practice. When we simultaneously investigate the dependency on the queue length, it will be challenging to find the correct arrival distribution based on data.

If we have very sparse historical data but still want to build a queueing model, we can use a theoretical appointment request function. Also, in the case of an excellent data basis, we may fit a theoretical appointment request function. For example, if we distinguish panel and non-panel patients. If panel and non-panel patients share the same capacities, we can use a appointment request rate $\lambda$, as explained in Section 3.5.2. If they do not share capacities, we have to use two separate queueing models, one for panel and one for non-panel patients. For non-panel patients, we could assume a constant $\lambda$ independent of the queue length as in Green and Savin (2008). In case of only panel patients, we can refer to Section 3.5.2 and use $\lambda(k) = \eta(k)(N - k)$ where $\eta(k)$ is the individual patient request rate. As we saw in the experiments, having patients with different individual arrival rates leads to significant different results compared to the case of a single individual arrival rate.

In our queueing model, we assume a single server queue with a physician who offers the same number $\mu$ of appointments every day in the basic model with a possible extension of queue length-dependent random appointment offerings. In practice, physicians probably have a repeating weekly schedule where they, for example, offer the same number of appointments on Monday, Tuesday, and Thursday but work fewer hours on Wednesday and Friday. In this case, we need to approximate and have to use the average appointment offerings per day in our queueing model.

Further, in the queueing model, we assume a finite queue capacity representing the finite booking horizon. We saw in Section 4.4.6 that the queue capacity could be chosen to find a tradeoff between the maximal panel size and the rejection rate given a fixed service level. If no booking horizon is used in practice, we have to choose a queue capacity $K$ big enough such that the influence on the results is negligible.

The distribution of indirect waiting times is calculated based on the queue length distribution at arrival times, excluding rejected requests and including rescheduling patients, which is equivalent to the queue length distribution immediately after departures. In practice, if only the distribution of indirect waiting times is of interest, it is enough to determine the queue length distribution immediately after departures through using the transition matrix of the embedded Markov chain. From there, we apply Proposition 10 or Corollaries 11 and 13 to define the distribution of indirect waiting times. If the transient queue development or other performance measures, e.g., the departure rates, are of interest, our queueing model should be used. Here, we can still decide if we want to integrate the rescheduling requests into the general requests or not.

## 4.6 Conclusion and outlook

The motivation for the presented research in Chapters 3 and 4, was to investigate the relationship between the physician's daily capacity, the panel size, and the stationary distribution of indirect waiting times of patients. The result was a flexible analytical queueing model that goes beyond the needed accuracy of the use case and can potentially be used in other application areas.

Our queueing model delivers exact results for the queue length distribution in the case of a linear queue length-dependent total arrival rate, a queue length-dependent rescheduling probability, and queue length-dependent random service times. For small queue capacities, we can do the same for non-linear total arrival rate functions. However, the determination of the corresponding exact arrival distributions becomes increasingly challenging to compute. The model also yields exact results for the queue length distributions at departure times for non-Poisson arrival processes if we have exact formulas for the number of patients to arrive during a time interval given the queue length. We obtain approximate results for the queue length distribution for larger queue capacities in the case of a non-linear total arrival rate function or a general arrival distribution.

Further contributions are the theoretical total arrival rate models motivated by a physician with panel patients. Here, for example, we can compute the exact queue length distribution for the case of a finite population because this case reduces to a linear total arrival rate function. We further showed how to model the total arrival rate to derive an approximate queue distribution in the case of different individual arrival rates for this population and in the case of a mix of a population arrival stream and a constant arrival stream.

Moreover, we propose how to approximately determine the indirect waiting time distribution in days under the assumption that the queueing model represents the booking of future appointments. Finally, for this situation, we develop a solution approach to relax the assumption that new arrivals are always assigned to the next available appointment using interconnected queueing models.

We implemented the mathematical queueing model and a simulation that can be adjusted to either mimic the queueing model or produce realistic results. Using the model implementation and simulation together, we verified the queueing model and validated the simplifications necessary to model the use case of a physician schedule. In particular, we validated the assumption that an indirect waiting time dependency can be approximately modeled as a queue length-dependent dependency. Note that the model implementation yields fast results with respect to the queue lengths distributions. Further, using the queueing model we can compute, for example, queue length distribution with two spikes which is not possible using the simulation model.

For practical purposes, we showed that for approximate results, we could integrate the rescheduling arrivals into the total arrivals, use a Poisson arrival distribution and change the arrival function

$\lambda$ instead of the service time $T$ to include queue length-dependent service times. This means we can cover most of the modeling effects by studying different total arrival rate functions without considering rescheduling. We found out that the effective arrival rate (or the total arrival rate if rescheduling patients are integrated) shapes the form of the queue length distribution. An increasing effective arrival rate produces a queue length distribution with two spikes, whereas a decreasing effective arrival rate yields a distribution with one maximum. We also saw that the queue length distribution at departure times is a good enough approximation to the queue length distribution.

We discovered that the no-show function (as a part of the rescheduling function) only influences the provider utilization. Here, if the probability of being a no-show increases with the queue length, we saw that the provider utilization increases with increasing indirect waiting time, reaches a maximum, and then decreases again.

Finally, we found that choosing the queue capacity corresponds to a tradeoff between a maximal panel size and the rate of rejected patients for a fixed expected indirect waiting time.

There are many possibilities for future research considering the use case of a physician schedule. First, we would like to conduct experiments for interconnected queues and build a corresponding simulation. Further, it would be interesting to integrate the possibility of different daily capacities for every weekday. This way, we could validate the assumption of using an average appointment offerings per day in the queueing model. Similarly, it is reasonable to assume that the total arrival rate varies per weekday. Therefore, the use of a total arrival rate independent of the weekday should be validated in a simulation.

We could try to decide on the queue length-dependent service time to minimize the variance of the indirect waiting time distribution while constraining the expected number of served patients per day. It would further be interesting to calculate and simulate examples of interconnected queues.

The most exciting continuation of this work would be to consider real-world data to deduce model parameters and then compare the model results with the historical distributions. How this can be done when patients actually schedule the next available appointment was discussed in the previous section. However, in the case of appointment data from physicians, we will rarely see that most patients book the next available appointment or behave in the way needed to apply interconnected queues. Here, an idea would be to work with the number of already booked appointments in the schedule when booking to approximate the queue length to collect data on the number of arrivals per day given a certain queue length. Here, we have to take into account that the number of appointments booked on a day is dependent on the opening hours of the practice that day if the appointments are mainly made by telephone. The opening hours of the practice are not necessarily the same as the working time of the considered physician. Therefore, some data transformation is necessary.

Having this data, we should check if there is an actual dependency of the total arrival rate on the queue length and if there is enough data available per queue length to define an actual distribution for the number of arrivals per day. Another problem arising here is that periods with more arrivals (for example, when the influenza is spreading) also produce longer queue length and a dependency of the total arrival rate on the queue length. However, the question is if a queue length-dependent total arrival rate can reproduce the queue length distribution in this case.

Another problem is the definition of the daily capacity in retrospect. We will probably only have estimates on the working hours per day and historical data on the number of patients treated per day. Sometimes days will be overbooked or will include idle times of the physician. In connection with patients that do not always book the next available appointment, it is then difficult to decide, for example, if a short day was due to low demand or due to a low fixed capacity. Another option would be to assume a random capacity.

Moreover, the dependency of the no-show probability on the indirect waiting time or directly on the queue length should be investigated. Here, we should think about how to take into account waiting days that fall on weekends. Further, we should investigate the effects of assuming that appointment requests only arrive during the working hours of the physician.

A first step before using real-world data would be to use a simulation to produce data. The data could then be used to define a model and calculate results which can then be compared to the simulation results. This environment could, for example, be used to simulate seasonal oscillations of the total arrival rate.

We should further investigate other possible application areas, especially where the next available appointment assumption is better kept than in the physician schedule context. One idea would be to move away from systems with virtual queues representing appointment bookings to actual queues. Imagine, for example, the waiting line for a popular food truck. Then the total arrival rate will probably depend on the queue length such that long queue lengths scare potential customers away. Also, there may be some very enthusiastic customers who rejoin the queue after being serviced.

# 5 Managing the intake of new patients into a physician panel over time

In this chapter, we propose deterministic integer linear programs that decide on the intake of new patients into panels over time while considering the future panel development. The main objective is to minimize the deviation between the expected panel workload and the physician's capacity over time. We conduct experiments using parameters based on real-world data. We show that, even in an uncertain environment, the expected differences between workload and capacity over time can be significantly reduced, considering several future periods instead of one. Using a detailed classification of new patients decreases the expected differences further. This chapter is joined work with Stefan Nickel and Peter Vanberkel. The following text matches the accepted manuscript of the following publication:

> Anne Zander, Stefan Nickel, and Peter Vanberkel. Managing the intake of new patients into a physician panel over time. *European Journal of Operational Research*, 294(1):391–403, 2021.

## 5.1 Introduction

Rural regions in Germany face a shortage of medical care provided by office-based physicians. One reason for this is demographic change. Even though the population is decreasing in total, the increasing number of older patients who need more medical attention implies an increasing demand for medical care for most medical specialties of office-based physicians (Schulz et al. 2016). This situation is further aggravated due to the physicians themselves getting older and retiring (Association of Statutory Health Insurance Physicians of Rhineland-Palatinate 2016).

Every year the German National Association of Statutory Health Insurance Physicians surveys medical students on a number of work-related topics (National Association of Statutory Health Insurance Physicians 2018). The results of the survey show that there are not enough medical students interested in specializing in the field of general medicine. Further, a third of the medical students do not want to work in locations with less than $10,000$ inhabitants. In addition, the preference for working as an employee in a group practice increases. For more than $90$ percent of medical students, work-life balance is of high importance. More than 80 percent of medical students value flexible working hours as very important.

Therefore, a way for rural areas to attract potential physicians is to offer them flexible working hours with less administrative work in group practices where physicians can be employed. To ensure an attractive working environment and good access to care, it is essential to balance supply and demand within the practice and across physicians taking continuity of care into account. Note that in Germany, patients can choose their physicians freely. However, regarding office-based physicians and especially general practitioners, patients usually stay with one physician who manages their (primary) health care. We say that the patient belongs to the panel of the physician. However, we observe that practices reject new patients due to the physician's already very high workload.

A similar situation exists in many developed countries where the population is aging. In Canada, for example, there is a chronic shortage of family physicians. This shortage disproportionately impacts rural areas, as many physicians prefer to work in urban settings (Pong and Pitblado 2005). A number of initiatives, including adding Nurse Practitioners, forming Collaborative Family Practice Teams (Nova Scotia Health Authority 2020), and increasing Telehealth options, aim to supplement a family physician's capacity. Despite this, access to primary care and determining the appropriate panel for family physicians remains a challenge (Urban 2019).

Besides predictable working hours, the remuneration of office-based physicians in Germany who mainly treat patients insured with the statutory health insurance is another reason to match supply and demand over time. A big part of the remuneration of office-based physicians is budgeted, meaning that physicians receive less payment per case when the budget is exhausted. As a consequence, working more hours may not result in more pay.

The demand for health care of a panel changes over time due to patients leaving and entering the panel but also due to changes in the health status of patients, i.e., in general, older patients need more medical attention than younger patients. Therefore, to balance supply and demand over time, it is crucial to manage the panel. Here, the main adjustable parameter is the decision about accepting exterior demand to enter the panel.

In this work, we focus on matching supply and demand for physicians and patients on a tactical level. We assume that the most significant share of appointment demand comes from patients that have visited before and from whom we know the visiting history. We consider equal-sized periods and use age and the number of visits in the last period to classify patients and build a distribution for the number of visits in the following period.

We build integer linear programs to decide whether or not to accept external demand per period to balance the physician's working hours with the expected workload produced by panel patients for all considered periods. Further programs integrate constraints including the variance of the panel's workload, and consider group practices with several physicians where we decide on the patient-physician assignment additionally to the acceptance decision.

To evaluate our programs in a realistic setting, where we take immediate acceptance or rejection decisions for every emerging patient request to enter the panel, we simulate the exterior demand as well as the panel evolution based on real-world data while our integer linear programs manage the decision on accepting or rejecting exterior demand. We can lower the deviation between workload and capacity significantly. However, there remains variance in the workload to be managed. For that reason, our approach should be used together with suitable operational models for appointment planning.

To the best of our knowledge, this article is the first that takes the temporal evolution of a patient panel into consideration to decide on the intake of new patients into existing patient panels. We further classify patients by their number of visits to the physician. We will show that this classification allows to predict the number of visits in the future far better than other patient attributes as, for example, age.

The remainder of this paper is organized as follows. Section 5.2 reviews the relevant literature. In Section 5.3, we present our integer linear programs (ILPs). In Section 5.4, we describe a real-world data set from a general practitioner group practice and define model parameters. We report on our numerical experiments in Section 5.5. Here, we present deterministic results as well as results from a simulation where the ILPs are solved to guide decisions in a stochastic environment. Section 5.6 provides concluding remarks and defines lines of future research.

## 5.2 Literature review

We investigate the literature on matching supply and demand for practices and physicians on a tactical level for physicians with panel patients. The essential characteristics of a panel are its size and composition. When considering several physicians in a group practice with individual panels, the patient-physician assignment is another important aspect. Therefore, we start by researching the literature streams panel sizing, case mix, workload distribution, and continuity of care. Since we want to use predictions of the future workload produced by a patient, we also investigate the field of appointment demand forecasting.

### 5.2.1 Panel sizing

In the last $15$ years, quite some literature on panel sizing in health care was published. The general idea is to determine the maximal size of a physician panel to ensure a manageable workload and access to care for the physician's patients. The most straightforward approach to the problem is to divide the physician's time capacity in a given period by the expected time required for medical attention of a single patient in that period (Murray et al. 2007). However, this approach does not consider the sources of variability that make the panel size problem more complex. The demand for health care is variable because we do not know when an individual patient is going to ask for medical attention and how much time is required to treat the patient.

The demand is further dependent on the appointment policy the practice is operating under. Under the traditional appointment policy, a practice only sees patients that have booked appointments. On the other side of the spectrum, there is the open or advanced access policy where practices see patients on the same day of their request. There are also appointment policies that reserve some capacity for same-day demand and use the remaining capacity for appointment booking. Especially if appointments are booked days and sometimes weeks or months into the future, some patients cancel their appointments or do not show up, leaving the physician idle. Consequences are that physicians sometimes have to work overtime or are idle even if there is enough demand. Further, appointment backlogs are forming. Therefore, it is necessary to define to what extent those consequences are tolerated, for example, via setting service levels.

In 2008, for the case of an open access policy, Green and Savin (2008) propose queueing models to determine the relationship between the panel size and the expected appointment backlog taking no-shows and rescheduling of no-shows into consideration. Based on an accepted probability of not getting a same-day appointment (assuming the physician does not work overtime), they calculate a maximal panel size. More recently, Liu and Ziya (2014) present two single server queueing models. They decide on the panel size and the service capacity to maximize the long-term average reward/profit while constraining the expected access time. Izady (2015) presents several discrete-time queueing models with bulk service. Zacharias and Armony (2017) consider the appointment backlog together with direct waiting time, i.e., the waiting time of patients in the practice. As in Liu and Ziya (2014), they decide on the panel size and the service capacity to maximize the long-term average daily reward. A different approach was taken by Vanberkel et al. (2018). They use a queueing network of multi-server queues to define the panel size of an oncology practice that balances demand from new patients and relapsed patients.

In our approach, we do not consider a static panel. Instead, we take into account the evolution of the panel over time. Therefore, we have to match supply and demand not only now but also in the future. However, to use our models, the physician needs to compute a target capacity for every considered future period, i.e., a maximal workload he or she can manage in a given period to comply with predefined service levels. To this end, one of the so far presented panel sizing models can be applied.

In practice, determining a manageable panel size is not enough. The physician also has to define which patients belong to her panel to determine the current panel size. That is easier said than done. On the one hand, patients leaving the panel do not always say that they will not visit again. So, the physician has to wait some time before she can remove those patients from the panel. On the other hand, some patients remain with the same physician but visit rarely. Therefore, the physician might wrongfully be inclined to remove those, too. There is no standard definition of how to determine the panel size. Some use the number of patients seen by the physician in the last two years (Margolius et al. 2018, Marx et al. 2011) others the number of patients in the last 18 months (Raffoul et al. 2016, Murray and Berwick 2003, Murray et al. 2007). Of course, the panel sizes change dependent on the considered time frame and are not comparable with each other. We will see in the data section that the considered time frame indeed has a significant impact on the panel size when experiencing a high proportion of patients that visit the physician rarely. We, therefore, argue that panel size as the only measure of workload is not enough. At least the corresponding average appointment request rate should be indicated. Therefore, to estimate the workload produced by a panel, we count patients and categorize them with respect to the number of visits per period.

### 5.2.2 Case-mix and workload distribution

Balasubramanian et al. (2010) use a stochastic linear program to optimally reassign patients to primary care physicians of a group practice with the objectives to minimize access time and to improve continuity of care. To do so, they first use a patient classification based on age and gender. Then they evaluate the resulting panels via simulation of the practice appointment scheduling system. The authors further use a more sophisticated patient classification based on more factors besides age and gender, such as specific medical conditions. The simulation shows that using the optimal panel design compared to the original panel design reduces the waiting time and the number of redirections of patients to other physicians. The improvements were similar using the new classification.

Ozen and Balasubramanian (2013) propose to minimize the maximal overflow frequency in a group practice of primary care physicians to redesign physician panels. The overflow frequency measure was first defined by Green et al. (2007). It represents the probability that the daily demand exceeds capacity. This measure is tailored for practices operating under advanced access. A practice should aim at a relatively low overflow frequency to ensure timely access to care. Our models can be used for different appointment policies. Hence, we refrain from using the measure overflow frequency. However, we can include constraints on the standard deviation of the workload of the panel.

### 5.2.3 Continuity of care

Our models can be used to manage the panel of a whole practice or a single physician. Even if treating a patient panel together in a group practice allows for more flexibility, every physician should have her own panel to ensure continuity of care. Continuity of care is associated with a fewer number, and fewer costs of emergency department visits (Dreiher et al. 2012, Marshall et al. 2016). Further, lower continuity of care in primary care is associated with a higher mortality rate of older patients (Maarsingh et al. 2016, Wolinsky et al. 2010). The literature review on continuity of care and quality care outcomes from Van Servellen et al. (2006) further lists positive relationships between continuity of care and patient satisfaction, early diagnosis of patients' conditions, improved compliance to treatment as well as reduced resource consumption.

### 5.2.4 Forecasting of appointment demand

In general, to classify patients to predict the number of visits, different factors are considered in the literature. Those factors include age, gender, number of morbidities, specific chronic diseases, region, and socioeconomic status (Balasubramanian et al. 2010, Ozen and Balasubramanian 2013, Riens et al. 2012). We will show later that, for example, age has a considerable influence on the average number of visits per year. However, the visit history of a panel patient allows for a more accurate prediction of future visits for this individual patient. Therefore, we will use the non-stationary attributes age and the number of visits in the last period to classify patients. Besides, we show that we can easily add more stationary attributes, such as gender, to improve the patient classification.

## 5.3 Integer linear programs for panel management

Our work aims to match supply and demand on a tactical level for physicians and their panel patients over time through managing the intake of new patients. Specifically, we want to decide whether to accept or reject a requesting patient to enter a panel. Particularly, we assume that once patients are admitted to the panel, they can not be removed by the physician. However, patients can, of course, decide to drop out of the panel or, if they do not give notice, are assumed to have left the panel after a predefined time frame not seen by the physician.

We discretize time and aim at matching supply and demand for the resulting periods. Here, supply is the physician's time capacity measured in the number of visits the physician can serve, and the demand is the generated workload by the panel also measured in number of visits. We categorize panel patients based on different attributes, the two basic ones being age and the number of visits per period. Note that age and the number of visits are non-stationary patient attributes. Therefore, these two attributes are in the core of our modeling and are treated differently than stationary attributes such as gender that can easily be integrated later. New patients who want to enter the panel are categorized based on the same attributes or a subset of those attributes. We assume that the number of visits per period of a panel patient may change over time. Thus, we define a distribution for the number of visits in the next period for a patient class (defined as patients with the same attribute values).

The decisions we need to make are whether to accept or to reject requests to enter the panel. We aim to find the least complex model that can exploit most of the potential of the decisions in minimizing the workload-capacity-deviation. Therefore, we start with a deterministic approach with known exterior demand, where we work with expected workload values. We will see that panel patients' expected workload in future periods can be represented by linear functions. The same is true for the variance of the workload, another objective to be used later on. Hence, we will build ILPs, where we decide on the number of patients per patient class to be added in each considered period. At the end of this section and in Section 5.5.2, we will comment on how these ILPs can be applied to take individual patient acceptance or rejection decisions in a realistic setting. We will further show that the benefit of using more complex models, including uncertainty such as Markov decision processes or stochastic linear integer programs would probably not yield significantly better results.

We classify panel patients with respect to age and the number of visits in a considered period. Therefore, we define $N$ as the set of age categories $N = \{0, 1, \ldots, n-1\}$ and $M$ as the set of visit categories $M = \{0, 1, \ldots, m-1\}$. A patient belonging to visit category $j \in M$ visits the physician an expected number of $f_j$ times in one period. We consider the evolution of the panel over time, taking into account the expected panel workload for several future periods. Here, a length of a period corresponds to the difference between successive age categories meaning that patients transition from on age category to the next from one period to the next. We assume that patients of age category $n-1$ leave the panel in the next period. To cover every period until a starting panel does not contribute to the future workload anymore, we look at a maximum of $n$ periods into the future. Hence, we define $T$ as the set of periods $T = \{0, 1, \ldots, n\}$.

Let $q_{ijl}$ be the probability that a patient who belonged to age category $i \in N \backslash \{n-1\}$ and visit category $j \in M$ last period belongs to the visit category $l \in M$ this period. To model patients that leave the panel after not having seen the physician for a number of periods, we can define several visit categories with an expected number of zero visits per period. Those visit categories stand for patients not having visited the physician in one, two and more periods. The last zero visits category can then be defined as the category for patients having left the panel via setting every transition probability to other visit categories to zero.

We define $p_{kijl}$ as the probability that a patient of age category $i \in N$ and visit category $j \in M$ will be in visit category $l \in M$ in $k \in \{0, \dots, n-i-1\}$ periods. By definition, we have:

$$p_{0ijl} = \begin{cases} 1 & \forall i \in N, j \in M, l = j, \\ 0 & \forall i \in N, j \in M, l \in M \backslash \{j\}, \end{cases} \tag{5.1}$$

$$p_{kijl} = \sum_{h=0}^{m-1} q_{(i+k-1)hl} p_{(k-1)ijh} \quad \forall i \in N \backslash \{n-1\}, j \in M, k \in \{1, \dots, n-i-1\}, l \in M. \tag{5.2}$$

We define $o_{kij}$ as the expected workload of a patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods. Then, we have:

$$o_{kij} = \begin{cases} \sum_{l=0}^{m-1} f_l p_{kijl} & \forall i \in N, j \in M, k \in \{0, \dots, n-i-1\}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

We denote the variance of the distribution $(q_{ijl})_l$ as $\sigma_{ij}^2$. Then, by $u_{kij}$ we define the variance of the workload of a patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods. Hence, we have:

$$u_{0ij} = 0 \quad \forall i \in N, j \in M, \tag{5.4}$$

$$u_{1ij} = \sigma_{ij}^2 \quad \forall i \in N \backslash \{n-1\}, j \in M, \tag{5.5}$$

$$u_{kij} = \begin{cases} \sum_{l=0}^{m-1} p_{(k-1)ijl} \sigma_{(i+k-1)l}^2 & \forall i \in N \backslash \{n-1\}, j \in M, k \in \{1, \dots, n-i-1\}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

Now, let us consider a single panel. We define the number of patients in the starting panel (period $k = 0$) as $v_{0ij}$ for age category $i \in N$ and visit category $j \in M$. At the end of a period $k \in T \backslash \{n\}$ a decision is taken to impanel a number of patients $w_{kij}$ belonging to age category $i \in N$ and visit category $j \in M$. Last but not least, $v_{kij}$ is the expected number of patients belonging to age category $i \in N$ and to visit category $j \in M$ in period $k \in T$. Now, we are able to determine the expected workload $\sum_{j=0}^{m-1} \sum_{i=0}^{n-1} f_j v_{kij}$ of the panel in a period $k \in T$. In period $k \in T$ the panel consists of aged patients that belonged to the original panel $(v_{0ij})_{ij}$ and of added patients from preceding periods $(w_{hij})_{ij}, h \in \{0, \dots, k-1\}$. The original panel patients of age category $i \in N$ and $j \in M$ contribute $o_{kij}$ required time each whereas added patients of age category $i \in N$ and visit category $j \in M$ from period $h \in \{0, \dots, k-1\}$ contribute $o_{(k-h-1)ij}$ required time each. Hence, we have:

$$\sum_{j=0}^{m-1} \sum_{i=0}^{n-1} f_j v_{kij} = \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( o_{kij} v_{0ij} + \sum_{h=0}^{k-1} o_{(k-h-1)ij} w_{hij} \right). \tag{5.7}$$

Similarly, we determine the total variance of the workload of the panel in a period $k \in T$:

$$\sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( u_{kij} v_{0ij} + \sum_{h=0}^{k-1} u_{(k-h-1)ij} w_{hij} \right). \tag{5.8}$$

In the following, we will present some examples of ILPs using the linear terms (5.7) and (5.8). We differentiate our examples with respect to the main component in the objective, the patient attributes used to categorize panel patients besides age and number of visits per period, the number of physicians considered, and the patient attributes used to categorize new patients.

| Model | Main component in the objective | Other patient attributes | Several physicians | Classification of new patients |
|---|---|---|---|---|
| (E AN) | Expected (E) workload | – | No | age (A),<br>number of visits (N) |
| (E AA) | Expected (E) workload | – | No | aggregated age (AA) |
| (E G ANG) | Expected (E) workload | Gender (G) | No | age (A),<br>number of visits (N),<br>gender (G) |
| (E SP AN) | Expected (E) workload | – | Yes (SP) | age (A),<br>number of visits (N) |
| (SD SP AN) | Standard deviation (SD) workload | – | Yes (SP) | age (A),<br>number of visits (N) |

**Table 5.1:** Overview of the different models

In the objective function, we minimize the deviation of the expected workload from the capacity, or we minimize the standard deviation of the workload using a summation approach or a min-max approach over the considered periods and physicians. The expected workload of a panel might match the capacity in all the considered periods. Still, when the actual workload is realized, the difference between workload and capacity might be substantial due to a high variance in the workload. Hence, it is reasonable to consider the standard deviation of the workload.

As an example, we consider gender as an additional patient attribute besides age and number of visits. New patients can be categorized with respect to a subset of the attributes of panel patients (age, number of visits per period, and gender) or even to aggregated attributes. For example, it might not be possible or desirable to classify new patients according to age and number of visits. Maybe, new patients should only be categorized by age but measured in larger time units than defined by the length of a period. Hence, we present an example were we categorize patients with respect to age groups, where one age group contains several age categories. Table 5.1 gives an overview of the different models we are going to present. Based on this classification of models other models can easily be generated.

Let us start with our first model. We assume that we want to balance supply and demand for one physician over the next $t \in T$ periods. We define $c_k$ as the capacity of the physician in period $k \in \{1, \ldots, t\}$ measured in number of visits per period. We further assume that during a period $k \in \{0, \ldots, t-1\}$ the physician experiences exterior demand (in number of patients) to enter the panel $d_{kij}$ according to age and visit category $i \in N$ and $j \in M$. We define our first ILP as:

$$(E\ AN) \qquad \min \sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( o_{kij} v_{0ij} + \sum_{h=0}^{k-1} o_{(k-h-1)ij} w_{hij} \right) - c_k \right| \tag{5.9}$$

s.t.

$$w_{kij} \leq d_{kij} \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M \tag{5.10}$$

$$w_{kij} \in \mathbb{Z}_0^+ \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M \tag{5.11}$$

The objective function (5.9) minimizes the sum over all periods of the absolutes values of the differences between the workload of the panel and the capacity of the physician. The first set of constraints (5.10) assures that we can not add more patients of a certain age and a certain visit category in a period than there is demand of such patients in this period. The third set of constraints (5.11) forces the decision variables to be non-negative integers. We do not explicitly linearize the absolute values in the objective function (5.9) here but note that this can be done using standard methods. Table 5.2 summarizes the parameters and variables used so far.

| **Symbol** | **Description** |
|---|---|
| **Sets** | |
| $N$ | Set of age categories $N = \{0, 1, \ldots, n-1\}$ |
| $M$ | Set of visit categories $M = \{0, 1, \ldots, m-1\}$ |
| $T$ | Set of periods $T = \{0, 1, \ldots, n\}$ |
| **Parameters** | |
| $t$ | Number of considered periods in the optimization $t \in \{1, \ldots, n\}$ |
| $q_{ijl}$ | Probability that a patient who belonged to age category $i \in N \backslash \{n-1\}$ and visit category $j \in M$ last period belongs to the visit category $l \in M$ |
| $p_{kijl}$ | Probability that a patient of age category $i \in N$ and visit category $j \in M$ will be in visit category $l \in M$ in $k \in \{0, \ldots, n-i-1\}$ periods |
| $o_{kij}$ | Expected workload of a patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods |
| $c_k$ | Capacity of the doctor in period $k \in \{1, \ldots, t\}$ measured in number of visits per period |
| $f_j$ | Expected number of visits per period for a patient belonging to visit category $j \in M$ |
| $d_{kij}$ | Exterior demand in period $k \in \{0, \ldots, t-1\}$ according to age and visit category $i \in N$ and $j \in M$ |
| $v_{0ij}$ | Number of patients in the starting panel (period 0) that belong to age category $i \in N$ and to visit category $j \in M$ |
| **Decision variables** | |
| $w_{kij}$ | Number of patients belonging to age category $i \in N$ and visit category $j \in M$ to be added to the panel in period $k \in \{0, \ldots, t-1\}$ |
| $v_{kij}$ | Expected number of patients belonging to age category $i \in N$ and to visit category $j \in M$ in period $k \in \{1, \ldots, t\}$ |

**Table 5.2:** Basic model notation

For our next ILP we group several age categories together resulting in a set $E = \{0, \ldots, e-1\}$ of age groups where each age group consists of several age categories. We define the decision variables $x_{kg}$ as the number of added patients of age group $g \in E$ in period $k \in \{0, \ldots, t-1\}$. We assume that we know the probability that a patient of a certain age group is a patient of a certain age category. Hence, we further define $d_{kg}$ as the exterior demand in period $k \in \{0, \dots, t-1\}$ according to age group $g \in E$ and $b_{kig}$ as the probability that a random patient of age group $g \in E$ belongs to age category $i \in N$ in period $k \in \{0, \dots, t-1\}$. For example, we can assume for the demand $d_{kg} = \sum_{i \in g} \sum_{j \in M} d_{kij}$ using the demand parameters from the basic ILP (E AN). Then, $b_{kig}$ can be defined as $b_{kig} = \frac{\sum_{j \in M} d_{kij}}{d_{kg}}$. We further assume that we know the probability $r_{kij}$ that a patient of a age category $i \in N$ belongs to the visit category $j \in M$ in period $k \in \{0, \dots, t-1\}$. Then, $r_{kij}$ can be defined as $r_{kij} = \frac{d_{kij}}{\sum_{j \in M} d_{kij}}$. To obtain our new model formulation, we substitute the first set of constraints (5.10) with $x_{kg} \leq d_{kg}, \forall k \in \{0, \dots, t-1\}, g \in E$ and eliminate $w_{hij}$ in ILP (E AN) via setting $w_{hij} = r_{hij} b_{hig} x_{hg}$ where $g$ is the age group such that $i \in g$. We obtain the following ILP:

$$(E\ AA) \qquad \min \sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( o_{kij} v_{0ij} + \sum_{h=0}^{k-1} o_{(k-h-1)ij} r_{hij} b_{hig} x_{hg} \right) - c_k \right| \tag{5.12}$$

s.t.

$$x_{kg} \leq d_{kg} \qquad \forall k \in \{0, \dots, t-1\}, g \in E \tag{5.13}$$

$$x_{kg} \in \mathbb{Z}_0^+ \qquad \forall k \in \{0, \dots, t-1\}, g \in E \tag{5.14}$$

We summarize the model notation in addition to the basic notation in Table 5.3.

| Symbol | Description |
|---|---|
| $E$ | Set of age groups $E = \{0, \dots, e-1\}$ |
| $x_{kg}$ | Number of added patients of age group $g \in E$ in period $k \in \{0, \dots, t-1\}$ |
| $d_{kg}$ | Expected exterior demand in period $k \in \{0, \dots, t-1\}$ according to age group $g \in E$ |
| $r_{kij}$ | Probability of a random patient of age category $i \in N$<br>belonging to the visit category $j \in M$ in period $k \in \{0, \dots, t-1\}$ |
| $b_{kig}$ | Probability of a random patient of age group $g \in E$<br>belonging to age category $i \in N$ in period $k \in \{0, \dots, t-1\}$ |

**Table 5.3:** Additional model notation

We can easily integrate stationary patient attributes into the model. For example, we might realize that gender has a significant influence on the number of visits per timer period. Then, we introduce gender dependent parameters $d^b_{kij}$, $o^b_{kij}$ and $v^b_{0ij}$ as well as gender dependent decision variables $w^b_{kij}$ with $b \in \{f, m\}$. We obtain another ILP:

$$(E\,G\,ANG) \qquad \min \sum_{k=1}^{t} \left| \sum_{b\in\{f,m\}} \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( o^b_{kij} v^b_{0ij} + \sum_{h=0}^{k-1} o^b_{(k-h-1)ij} w^b_{hij} \right) - c_k \right| \tag{5.15}$$

s.t.

$$w^b_{kij} \le d^b_{kij} \qquad \forall k \in \{0,\dots,t-1\}, i \in N, j \in M, b \in \{f,m\} \tag{5.16}$$

$$w^b_{kij} \in \mathbb{Z}^+_0 \qquad \forall k \in \{0,\dots,t-1\}, i \in N, j \in M, b \in \{f,m\} \tag{5.17}$$

Now, imagine a group practice with several physicians who all have their own patient panel. In this case, when accepting a new patient we also have to decide to which physician the patient should be assigned. Hence, we use physician dependent parameters and variables. In order to balance the workload between the physicians we can use a min-max optimization approach. Let $A$ be the set of physicians. Then, we obtain yet another ILP:

$$(E\,SP\,AN) \qquad \min z \tag{5.18}$$

s.t.

$$\sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( o^a_{kij} v^a_{0ij} + \sum_{h=0}^{k-1} o^a_{(k-h-1)ij} w^a_{hij} \right) - c^a_k \right| \le z \qquad \forall a \in A \tag{5.19}$$

$$\sum_{a\in A} w^a_{kij} \le d_{kij} \qquad \forall k \in \{0,\dots,t-1\}, i \in N, j \in M \tag{5.20}$$

$$w^a_{kij} \in \mathbb{Z}^+_0 \qquad \forall k \in \{0,\dots,t-1\}, i \in N, j \in M, a \in A \tag{5.21}$$

So far, all presented ILPs minimize the difference between workload and capacity. Now, we give an example where we minimize variance. As before, imagine a group practice with several physicians who all have their own patient panel. As a first step, using an ILP for the whole practice without differentiation based on the physicians, we decide on patients to be accepted on the practice level, the result being $w_{kij}$ for $i \in N$, $j \in M$ and $k \in \{0,\dots,t-1\}$. In a second step, we minimize the maximal variance of the physicians' expected workloads while constraining for a small difference between workload and capacity for every physician. This approach yields a fair allocation of patients to physicians in the sense that the variability of the panel demand is similar for each physician. For example, the physicians then experience similar distributions of overtime and idle time. Let $s^a$ be a small constant dependent on $a \in A$. We obtain:

$$(SD\,SP\,AN) \quad \min z \tag{5.22}$$

s.t.

$$\sum_{k=1}^{t}\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(u^a_{kij}v^a_{0ij} + \sum_{h=0}^{k-1} u^a_{(k-h-1)ij}w^a_{hij}\right) \leq z \qquad \forall a \in A \tag{5.23}$$

$$\sum_{k=1}^{t}\left|\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(o^a_{kij}v^a_{0ij} + \sum_{h=0}^{k-1} o^a_{(k-h-1)ij}w^a_{hij}\right) - c^a_k\right| \leq s^a \qquad \forall a \in A \tag{5.24}$$

$$\sum_{a \in A} w^a_{ijk} = w_{kij} \tag{5.25}$$

$$w^a_{kij} \in \mathbb{Z}^+_0 \qquad \forall k \in \{0, \dots, t-1\}, i \in N, j \in M, a \in A \tag{5.26}$$

Of course, many more versions and extensions are possible. For example, our model can also be used to redesign panels in a group practice. To this end, we model patients eligible for reassignment as new patients in an ILP such as ILP (E SP AN) or ILP (SD SP AN). Instead of minimizing variance, we could also opt to minimize the maximal overflow frequency (on the period level) as in Ozen and Balasubramanian (2013). But, note that this modeling leads to a non-linear objective function. Further, we could easily include weights for the considered periods in the optimization to discount periods that lie further into the future.

In our ILPs, we decide on the number of new patients differentiated by categories and groups to impanel for all considered $t$ periods. However, the ILPs presented should be solved at least once every period. Therefore, we will only act on decisions for the first period. The decisions for future periods are merely taken to account for the future panel development. Further, every time a program is solved, the starting panel can be adjusted. In this way, we can, for example, account for patients that have left the panel.

Until now, we have assumed that we know the exact demand for this period and the upcoming periods. In reality, we might have an estimation for the expected demand during a period based on historical data or based on demographic data on the region. To better adapt to the uncertainty in the current period, we can quickly adjust the programs such that we can solve them several times during a period or even every time a request to enter the panel occurs. To this end, we constrain that the already accepted demand during the current period is added to the panel. Moreover, we adjust the remaining demand until the end of the current period. However, the optimization still relies on the demand forecast for the remaining part of the current period and future periods. We will show in Section 5.5 that using our ILPs to guide the acceptance decisions of new patients is still meaningful even if only the expected future demand with our without patient classification is known.

Note that we could also use the actual time a patient spends with his physician, e.g., measured in minutes, to define visit categories. The issue here is that such data is often not available or reliable. Hence, we work with the number of visits.

## 5.4 Defining model parameters using real-world data

To test our models, we use a real-world data set from a group practice for general medicine that is run by $3$ physicians. One of them is working full time and the other two practice approximately $50$ percent part-time. They care for their patients together meaning that they have one panel for the whole practice. We have data from the years $2010$ to $2014$. Every row in the data represents a chargeable service for the patient (this does not necessarily mean a patient-physician contact) on a specific day. Examples for such service charges are general treatment charges, laboratory charges, charges for chronically ill patients, emergency charges, and blood pressure measurement charges. Every day where at least one service is documented for a patient is counted as a visit of that patient in our model. We assume that every visit produces workload for the practice (for the physicians or the other staff of the practice). We cannot determine the time, personal, and equipment requirements for the different services from the data. Therefore, we simplify by assuming the same time required for every visit partitioned proportionally between the physicians.

The critical columns of the data for us are the practice-specific patient-ID, the birthdate of the patient, and the date of the service. Unfortunately, there is no column indicating gender or any other stationary patient attribute. We count $7,472$ patients who visited the practice $220,710$ times in those $5$ years.

We decide on using a period length of one year. We believe this is a good choice since a year is long enough to level out seasonal effects between periods. Hence, it is reasonable to look at the number of visits last period to predict the number of visits next period instead of taking into account several preceding periods. As we are interested in the time evolution of the panel over the years, we need to define who belongs to the panel. Having a data set of $5$ years, we define that a patient who does not show up for $4$ years in a row left the panel and that a patient who did not visit the practice during the last $4$ years is new. In reality, we may have panel patients that do not show up for $4$ years in a row without leaving the panel. In our model, those patients are then counted as new patients when they show up again. The practice assigns ascending patient-IDs to new patients. Therefore, we know that patients with patient-IDs smaller than the current highest patient-ID must have visited the practice before. Hence, we can determine how often a new patient, in our definition, is, in fact, a panel patient that did not show for 4 years. Indeed, in 2014, 179 of the 756 new patients, according to our definition, have visited the practice before.

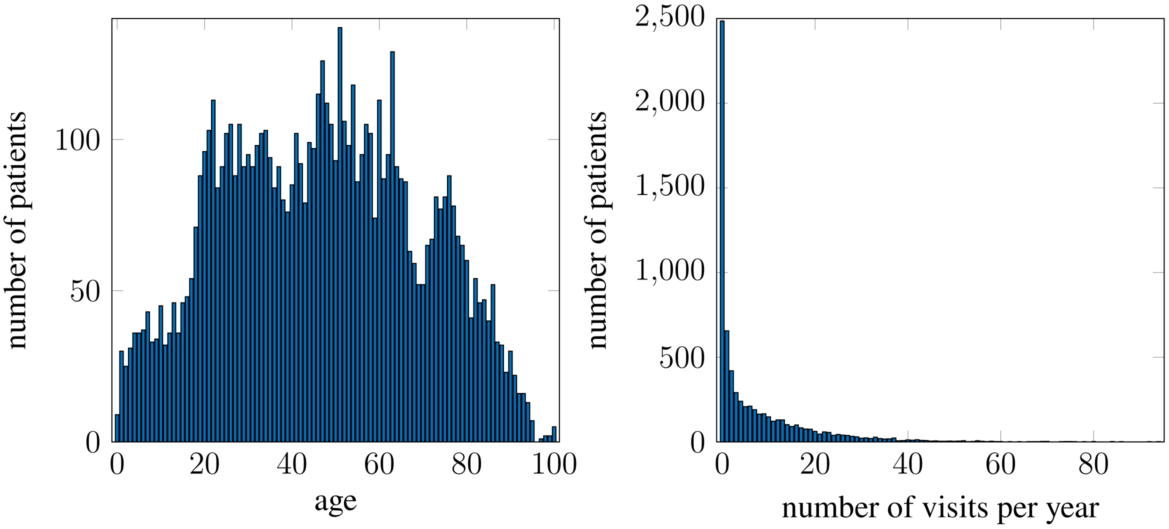

**Figure 5.1:** Number of panel patients by age and by number of visits per year in 2014

At the end of 2014, we find 6864 patients in the panel with an average number of visits of 7.11. Figure 5.1 shows the number of panel patients by age and by the number of visits for 2014. Note that many panel patients, i.e., 2485 out of 6864, do not show in 2014. To come back to our argument that panel size as an only measure of workload is not enough, we investigate the resulting panel sizes if we define a different time frame for belonging to the panel. Considering only patients that visited in 2014 yields 4379 patients. Looking back 2 or 3 years yields 5423 or 6273, respectively. Note that the average number of visits for panel patients in 2014 differs for the different time frames. It is 7.11 for our definition of looking back 4 years, and it is 11.15 when we only consider patients that visited this year.

To reflect the definition of new patients and leaving patients in our model, we include 4 zero visits categories. Visit category 0 contains all patients that did not visit this year and not in the last three years. By definition, patients in this visit category will stay in this visit category for all future periods and are not considered part of the panel anymore. Visit category 1 contains all patients that did not visit this year and the last two years, but visited the practice three years ago. Visit categories 2 and 3 are defined similarly. There are few patients with a very high number of visits per year. Hence, we group several numbers of visits such that a visit category consists of at least 5% of all made observations in the 5 considered years. This grouping results in a total of 17 visit categories.

Because the length of a period is a year, the age categories reflect the age in years. We decide to work with $100$ age categories reflecting the ages $0$ to $99$. A patient of age $99$ automatically leaves the panel in the following period.

For the numerical experiments we assume a constant exterior demand $d_{kij}$ in period $k \in \{0, \ldots, t-1\}$ according to age and visit category $i \in N$ and $j \in M$ equivalent to the composition of accepted new patients in $2014$. Note that the real exterior demand in $2014$ may have been higher. Unfortunately, we do not have records of rejected exterior demand.

To show that age and number of visits last period are predictors for the number of visits this period, we perform regression analyses. Using $17,648$ observations and including ages between $20$ and $80$, the coefficient of determination for an age-dependent average number of visits per year is $0.937$ with an additional number of $0.21$ visits per year. However, looking at individual patients, using $21,811$ observations of all age categories, a regression model including age and number of visits last year to predict the number of visits next year yields an adjusted coefficient of determination value of $0.631$. A regression model considering the number of visits last period as an only explanatory variable yields an adjusted coefficient of determination value of $0.627$. Hence, even if there is a strong correlation between age and number of visits on average the influence of age at an individual patient level is small compared to the influence of the number of visits last period.

To compute the transition probabilities $q_{ijl}$ that a patient who belonged to age category $i \in N$ and visit category $j \in M$ last period belongs to the visit category $l \in M$ this period, we group age categories together such that there are at least $1500$ observations in every considered age group. Even with the data aggregation used, for some age groups and visit categories, there are few observations ($< 100$) to define the transition probabilities. This problem appears mainly for the bigger visit categories and very young or old patients. Therefore, and also to smooth the tails of the transition distributions, we perform small data changes to keep the data plausible. For example, if there is a positive probability to transition from a visit category $j$ to $l$ and $l + 2$, we add a virtual observations such that there is also a positive probability to transition from $j$ to $l + 1$. We add (or remove) $286$ virtual observations based on $21,918$ real observations.

Based on the transition probabilities and the average numbers of visits per period, we can determine $o_{kij}$, the expected workload of a single patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods as well as $u_{kij}$, the variance of the workload of a single patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods. Using the panel of $2014$ as the starting panel, we find that the variance of the workload after one period is $247,710$, which translates into a standard deviation of $497.7$. Hence, the standard deviation corresponds to approximately $1$ percent of the starting panel's expected workload after one period.

We do not fix specific capacities $c_k,\ k \in \{1, \dots, t\}$ here. We will still orientate ourselves on the number of visits in the year $2014$ for the numerical experiments assuming a good match between demand and supply for the year $2014$.

## 5.5 Numerical experiments

In this section, we first present numerical experiments in a deterministic setting, i.e., with deterministic exterior demand and expected future panels and later on experiments using a simulation to include uncertainty. In the following, when solving an ILP, we stop calculations whenever we reach an absolute MIP gap of $t$, i.e., the number of considered periods in the optimization, to speed up computation time. The MIP gap corresponds to a maximal average deviation of $1$ visit from the optimal solution for every considered period. To solve our models, we use CPLEX 12.8 on an Intel 1.9 GHz PC with 16 GB. For $t < 10$, the computation time is maximal $0.2$ seconds setting all parameters to those determined in Section 5.4. Particularly, the starting panel is set to the panel in $2014$. The capacity, as well as the exterior demand to enter the panel, are assumed to be the same over all considered periods. Higher values of $t$ are neither practical (in terms of demand forecasting) nor necessary. In fact, as we will see later in the numerical experiments, considering two to four future periods already yields a significant improvement compared to considering only the next period. Solving the model without accepting MIP gaps can take a very long time, even for small values of $t$. However, due to the inherent variance of the workload, it is unnecessary to solve the models to optimality.

### 5.5.1 Deterministic setting

In our analyses, we consider three ILPs for a single panel, the first one being ILP (E AN), where we consider the exterior demand on the level of age and number of visits. In ILP (E A), we categorize new patients according to age only (similar to the presented ILP (E AA) but without age groups). In ILP (E), we do not categorize new patients. We just count them. We further consider three scenarios with a length of $10$ periods. In a scenario, we solve an ILP once per period. The panel that results from applying the decisions of period $0$ of the ILP solution is used as the starting panel for the next period. Note that those resulting panels are expected panels and therefore contain non-integer values. For every scenario, we then compare the three ILPs with a varying number of considered periods $t$ in the optimization. In the following, we consider using $t = 1$ periods in the optimization as the base case. For example, using ILP (E) with $t = 1$ reduces to determine the expected workload of the current panel in the next period and to divide the difference between capacity and this expected workload by the expected workload of a new patient to determine the number of new patients that should be taken in.

The first scenario considers a physician who starts a new practice working part-time with $25$ percent. In the second scenario, we consider a physician working part-time with $25$ percent, who increases her working hours to $50$ percent. In the third scenario, we consider a group practice of three full time working physicians that reduces its capacity to two full time working physicians. Note that we choose those three scenarios in order to illustrate the benefits of considering more than one period at a time.

We define the parameters for all scenarios as determined in Section 5.4, apart from the starting panel and the capacity. In particular, we assume the same exterior demand for all considered periods. Note that we were only able to define the exterior demand for the last year of the appointment data in Section 5.4 due to the definition of new patients and the small number of years covered in the data. Further, to show the benefit of our models, it is sufficient to use a big enough exterior demand that covers a wide range of different patient classes to be generally able to reach the set capacity and to allow decision flexibility. Depending on the ILP, we aggregate the demand data accordingly, as described in Section 5.3. The panel in $2014$ from Section 5.4 corresponds to two full time working physicians. For Scenario $1$, we start with an empty panel and use a capacity of $6,074$ visits per period, which corresponds to $1/8$ of the total visits in $2014$. Scenario $2$ uses a capacity of $6,074$ for the first $4$ periods and from there on a capacity of $12,148$. Scenario $3$ uses a capacity of $48,593$. To simulate the three scenarios later, we need integer-valued starting panels for Scenarios $2$ and $3$. Therefore, we randomly remove patients from or add patients to the $2014$ panel. We continue until the panel exhibits the necessary workload of $1/8$ or $3/2$ of the workload of the $2014$ panel. For the third scenario, we further let the panel with $3$ physicians age for $5$ periods with no intake of new patients. The resulting panel is the starting panel for Scenario $3$. This approach shortens an otherwise long time where no patient is added.

The objectives of our ILPs minimize the sum over all considered periods $t$ of the absolutes values of the differences between the workload of the panel and the capacity of the physician, in short, the sum of differences. To compare the ILPs and to analyze the influence of $t$, we use the sum over all $10$ considered periods of the differences as an output.

In Figure 5.2, we compare the sum of differences for the three ILPs with a varying number of considered periods $t$ for Scenario $1$. We see that the sum of differences decreases with the number of considered periods $t$ for all three ILPs. The sum of differences decreases substantially from $t = 1$ to $t = 2$ and from $t = 2$ to $t = 3$ for ILP (E AN) and ILP (E A). The sum of differences of ILP (E) decreases from $t = 1$ to $t = 2$ and then remains on that level. We see no consistency in the relationship between the sum of differences of the 3 ILPs for $t = 1$ and $t = 2$. However, for $t \geq 3$, we observe that the sum of differences of ILP (E AN) is lower than the sum of differences of ILP (E A), which in turn is lower than the sum of differences of ILP (E). Considering values of $t > 5$ does not lead to significant improvements in the sum of differences. In total, we see that the lowest sum of differences $4,591$ is reached for ILP (E AN) using $t = 5$ periods. This value is $10$ percent lower than the sum of differences of ILP (E AN) using $t = 1$ periods and $17$ percent lower compared to the sum of differences of ILP (E) using $t = 1$ periods. Further, comparing the sum of differences for $t = 5$, we see a decrease of $1$ percent between ILP (E) and ILP (E A) and a decrease of $2$ percent between ILP (E A) and ILP (E AN).

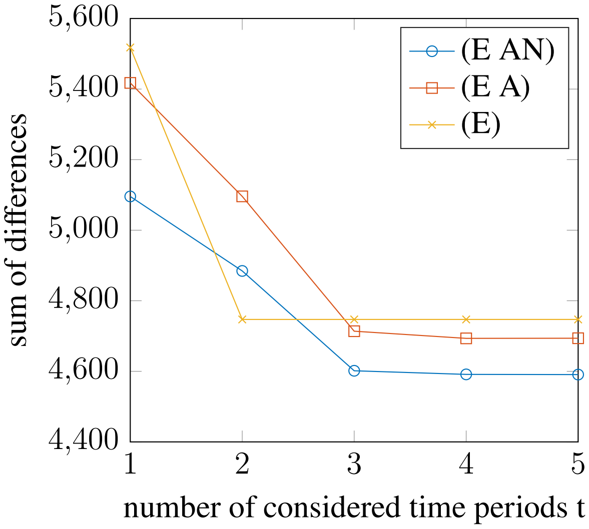

**Figure 5.2:** Sum of differences for the three ILPs and different values of $t$ in Scenario 1

How can we interpret those findings? Analyzing $o_{kij}$, i.e., the expected workload of a single patient of age category $i \in N$ and visit category $j \in M$ in $k$ periods, defined by the data presented in Section 5.4, we realize that the expected number of visits of a random newly impaneled patient increases over the first periods. Therefore, adding a lot of new patients to the panel simultaneously without considering future effects leads to an overload in future periods. In our scenario, looking $t = 3$ periods into the future is already enough to foresee those effects and to lower the sum of differences significantly. We further see that being able to classify new patients by age or even by age and number of visits is beneficial to decrease the sum of differences further than without the classification of new patients. However, of course, there is a lower bound for the sum of differences. The demand per period, which corresponds to $2,515$ visits per period, is not enough to reach the capacity of $6,075$ visits in the first two periods even though the patients added in the first period require more visits in the second period.

In Figure 5.3 for Scenario 2 we see a similar behavior to Scenario 1. Again, the sum of differences decreases with the number of considered periods $t$. The slope of the decrease is flattening faster for ILP (E) than for ILP (E AN) and ILP (E A). The lowest sum of differences $3180$ is reached by ILP (E AN), considering $t = 6$ periods. This value is $45$ percent lower compared to the sum of differences of ILP (E AN) using $t = 1$ periods. Further, comparing the sum of differences for

$t = 6$, we see a decrease of $7$ percent between ILP (E) and ILP (E A) and a decrease of $8$ percent between ILP (E A) and ILP (E AN).

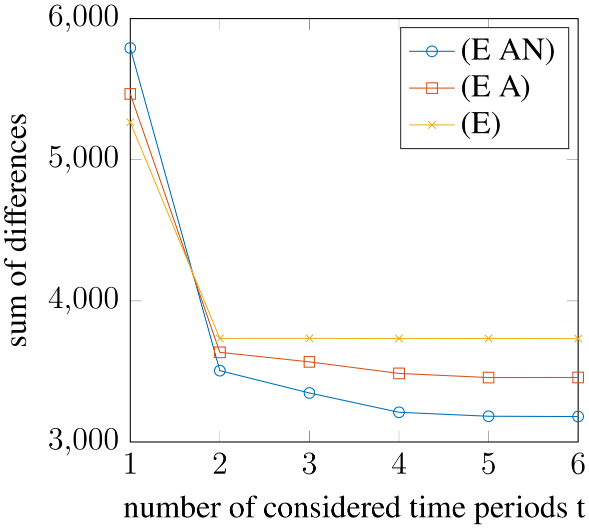

**Figure 5.3:** Sum of differences for the three ILPs and different values of $t$ in Scenario 2

The behavior of the curves in this scenario can again be explained by a high intake of new patients that produce overload in future periods for small values of $t$. Again, classifying new patients helps to lower the sum of differences further. Depending on the classification it is enough to consider two to four future periods to lower the sum of differences significantly. The sum of differences can not decrease more because the change from the first capacity to the next one takes two to three periods due to the limited demand.

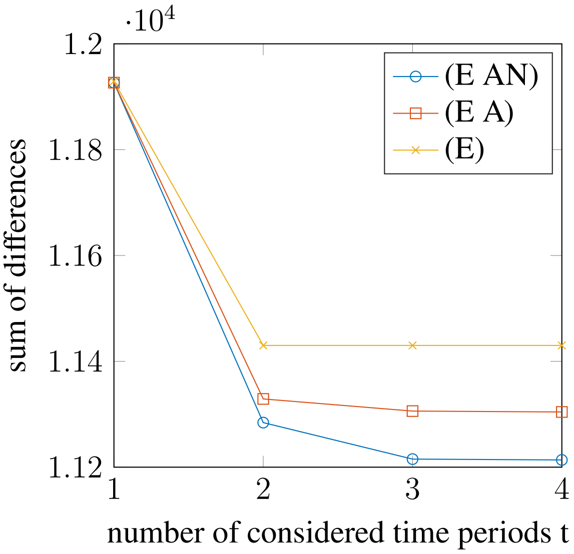

**Figure 5.4:** Sum of differences for the three ILPs and different values of $t$ in Scenario 3

In Figure 5.4, we again see a similar curve behavior for Scenario $3$ when compared to Scenario $1$ and $2$. Nevertheless, the interpretation is different. To lower the workload of the panel, new patients are not accepted in the first periods. However, once the workload reaches the new capacity, the demand is not enough to compensate for the decrease in workload over the following periods if we use $t = 1$. Looking $t = 2$ periods into the future is enough to foresee that effect for ILP (E A) and ILP (E). For ILP (E AN), looking $3$ periods into the future decreases the sum of difference again slightly compared to $t = 2$. Again, classifying new patients helps to lower the sum of differences further.

Because it takes the first two periods to decrease the panel's workload close to the desired capacity, we experience a high sum of difference. However, to compare the results with our simulation results later on, we omit the first two periods in the resulting sum of differences. This reduction is possible because the differences for period $1$ and $2$ are the same for all considered values of $t$. For $t = 4$, this leads to sums of differences of $549$, $640$, and $765$ for the ILPs (E AN), (E A), and (E), respectively. For $t = 1$, this leads to sums of differences of $1,262$, $1,262$, and $1,264$ for the ILPs (E AN), (E A), and (E), respectively. Here, comparing those results for ILP (E AN) between $t = 1$ and $t = 4$, we experience a reduction of $44$ percent. Further, comparing the sum of differences for $t = 4$, we see a decrease of $16$ percent between ILP (E) and ILP (E A) and a decrease of $14$ percent between ILP (E A) and ILP (E AN).

For all three scenarios, in a deterministic setting, we see that considering the future panel evolution decreases the deviation between workload and capacity over time. Here, we find that it not necessary to consider numerous future periods, but instead, two to four periods suffice. Classifying new patients helps decrease the deviation further but does not have as much impact as considering future periods. Here, the decision on the detail of new patient classification can be understood as a trade-off between fairness, i.e., little to no classification, and effectiveness, i.e., a significant further decrease of the sum of differences.

Unfortunately, we are not able to test the other ILPs presented in Section 5.3 due to data unavailability.

### 5.5.2 Simulation

To show that our deterministic ILPs are still useful in a stochastic setting, we built a discrete event simulation in AnyLogic 8. We simulate the requests to enter the panel, the acceptance decisions concerning those requests, and the resulting panel evolution from period to period.

We simulate the three scenarios combined with the three ILPs and a varying number of considered periods in the optimization, as presented before in Section 5.5. Assuming the starting panel with its transition probabilities and the expected number of visits per patient class as given, the essential input parameters are the number of considered periods $t$, the capacities $c_k, k \in \{1, \ldots, t\}$, the expected exterior demands $\mu_{kij}, k \in \{0, \ldots, 2t-1\}, i \in N, j \in M$, and the 99-percentile exterior demands $p^{99}_{kij}, k \in \{0, \ldots, 2t-1\}, i \in N, j \in M$.

At the beginning of a period, we decide on upper bounds for the number of patients to be added this period for each patient class using Algorithm 1 which we explain in more detail later. When a patient arrives requesting to join the panel, she is accepted if the current number of new patients that have been accepted in her patient class is below or equal to the defined upper bound; otherwise, the patient is rejected.

At the end of the period the category transitions are simulated and the workload-capacity-deviation is calculated for the period. Again, the main output is the sum of the absolute differences between the workload of the panel and the capacity for the considered periods, here $10$, which is now a random variable. To obtain statistically relevant results, we run numerous replications for each considered setting, i.e., a combination of scenario and ILP. Figure 5.5 shows a flowchart of the simulation.

In a first step, we consider the stochastic transition of patients from one visit category to the next; the demand remains deterministic, i.e., the exterior demand to enter the panel is known for all considered periods. In a second step, we include stochastic demand. Hence, the rates of Poisson distributions for the patient demand of every patient class is known for all considered periods.

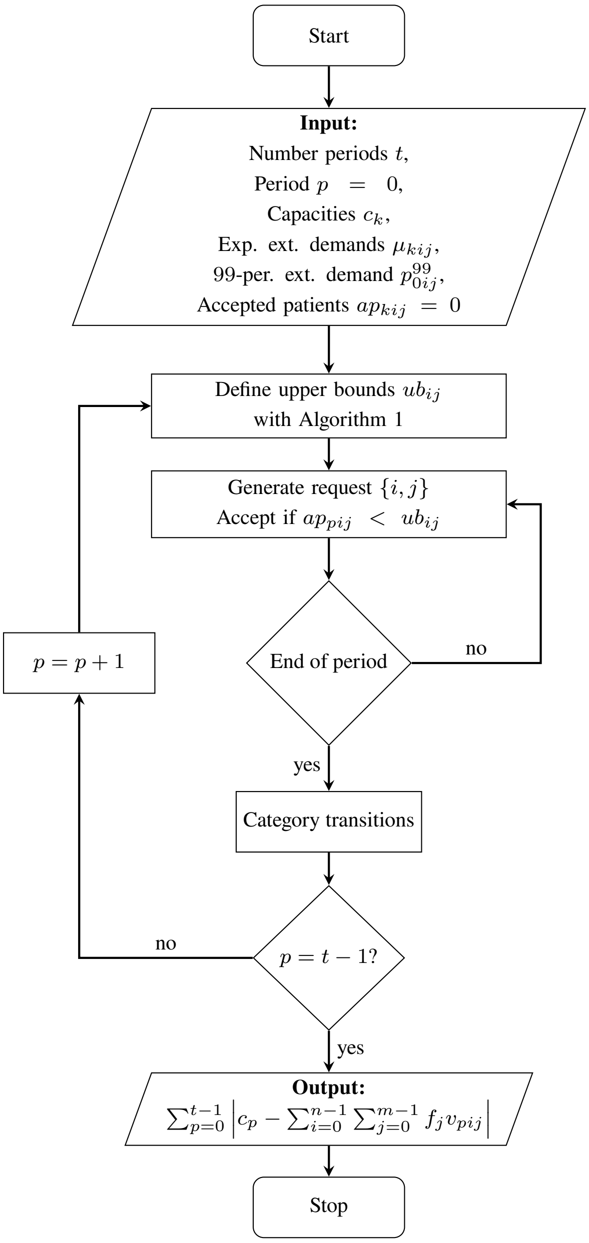
Start
Input:
Number periods $t$,
Period $p = 0$,
Capacities $c_k$,
Exp. ext. demands $\mu_{kij}$,
99-per. ext. demand $p^{99}_{0ij}$,
Accepted patients $ap_{kij} = 0$
Define upper bounds $ub_{ij}$
with Algorithm 1
Generate request $\{i, j\}$
Accept if $ap_{pij} < ub_{ij}$
$p = p + 1$
End of period
no
yes
Category transitions
no
$p = t - 1$?
yes
Output:
$\sum_{p=0}^{t-1} \left| c_p - \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} f_j v_{pij} \right|$
Stop

**Figure 5.5:** Simulation flowchart

Only considering the panel evolution, i.e., the stochastic transition of patients from one visit category to another, we solve the ILP once at the beginning of every period. We use the solution of our ILP for the first period to define the number of patients to be added per patient class in the considered period, just as we did in our first numerical experiments.

Including demand variability in the simulation, we could solve the ILP using the expected number of patients per class as the demand input. Then, we could use the solution of the first period to define upper bounds on the number of patients to be added for each patient class for the current period. The problem with this approach is that if more patients arrive than expected, they will never be accepted even if it would be beneficial. One possibility to overcome this problem would be to resolve the ILP whenever a patient arrives who would be rejected. This approach leads to a high number of ILPs that we need to solve during a period. In practice, this is feasible because one ILP can be solved in less than a second. However, in the simulation, the high number of ILPs that need to be solved leads to very long, not feasible run times. We decided to take a different approach and solve the ILP several times only at the beginning of a period with different demand inputs to obtain upper bounds on the number of patients to be accepted for each patient class for the considered period. Algorithm 1 shows this process exemplary for Period $p = 0$ with a classification of new patients with respect to age and the number of visits.

**Algorithm 1:** Definition upper bounds

```
Data: Expectec exterior demand $\mu_{kij}, k \in \{0, \dots, t-1\}, i \in N, j \in M$,
99-percentile of the exterior demand $p^{99}_{0ij}, i \in N, j \in M$
Result: Upper bounds $ub_{ij}, i \in N, j \in M$

$NUB = \{\{i,j\} : i \in N, j \in M\}$;
$UB = \{\}$;
$d_{0ij} = p^{99}_{0ij}, i \in N, j \in M$;
$d_{kij} = \mu_{kij}, k \in \{1, \dots, t-1\}, i \in N, j \in M$;
Solve ILP;
while $\exists i \in N, j \in M : w_{0ij} \geq \mu_{0ij}$ do
    for $i \in N, j \in M : w_{0ij} \geq \mu_{0ij}$ do
        $ub_{ij} = w_{0ij}$;
        $UB = UB \cup \{i,j\}$;
        $NUB = NUB \backslash \{i,j\}$;
    end
    $d_{0ij} = \mu_{0ij}, \{i,j\} \in UB$;
    $d_{0ij} = p^{99}_{0ij}, \{i,j\} \in NUB$;
    Solve ILP;
end
$ub_{ij} = w_{0ij}, \{i,j\} \in NUB$;
```

For the ILP version with only one patient class, i.e., where we only count patients but do not classify them, we run the ILP with no exterior demand intake restriction in the first period.

However, we constrain the intake of new patients for the following periods by the expected demand. This way, the solution of the ILP for the first period yields an upper bound on the number of patients to be added this period, taking the expected demand in future periods into account. For the two ILPs with patient classes, the approach is similar. We always constraint the intake of new patients by the expected demand for all periods except the first one. For the first period, we determine the 99 percent quantile of every Poisson distribution for every patient class. We use this quantile as the exterior demand input in the first period. The solution for the first period then yields the number of patients to be added if the demand is high.

In general, we will see some patient classes where the solution value is bigger than the expected demand. In this case, we will fix the upper bounds for those patient classes to the solution value. Of course, in general, the upper bound will not be reached when the demand is realized, which means that the ILP solution contains more workload from those patient classes than what will be available. In turn, the other patient classes are under-represented. Therefore, we solve the model again, constraining the exterior demand intake for those patient classes where we already set an upper bound by the expected demand. We set the demand input for the other patient classes to the 99 percent quantile. Again, we will probably have some patient classes where the solution value exceeds the expected demand. In those cases, we fix the upper bound to the solution value, and solve the model again, constraining the exterior demand intake by the expected demand for those patient classes. We continue until the solution value is below or equal to the expected demand for all patient classes. For those classes where we have not yet set the upper bound, we set the upper bound to the solution value of this last ILP.

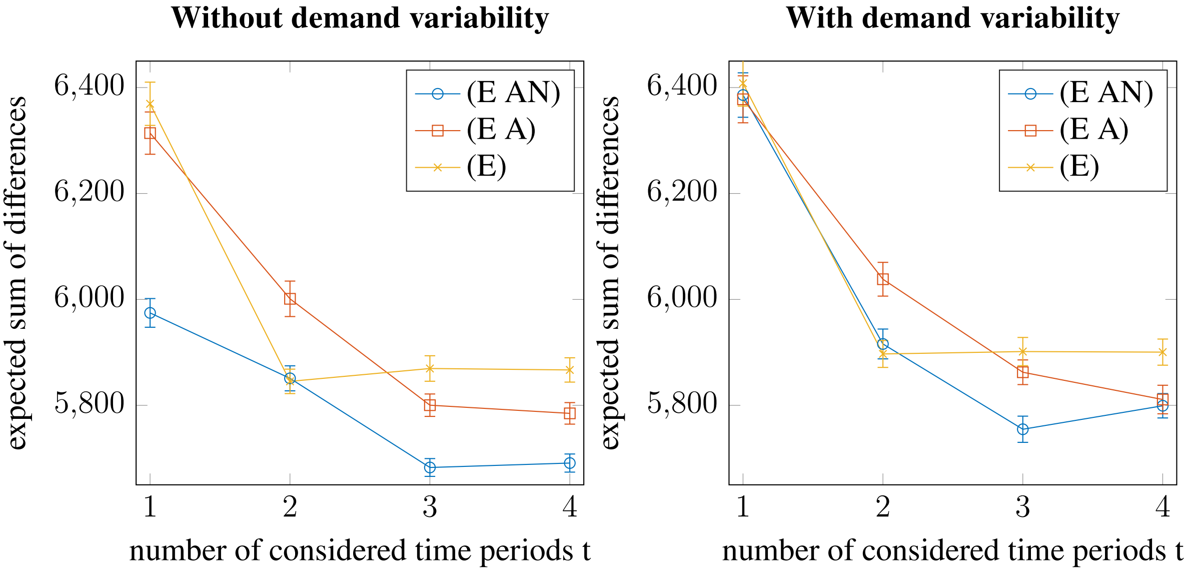

**Figure 5.6:** Simulated expected sum of differences for the three ILPs and different values of $t$ in Scenario 1

In Figure 5.6, the expected sum of differences (with and without considering exterior demand variability) dependent on the ILP and the number of considered periods is shown for Scenario $1$. The error bars represent the $95$ percent confidence interval after $2,000$ simulation replications.

Let us first compare the simulation results without exterior demand variability with the deterministic results. We see a similar behavior of the curves with a somewhat less steep decrease in the expected sum of differences in the non-deterministic case. We further observe that the expected sum of differences $5,691$ for ILP (E AN) and $t = 4$ is $1,101$ visits higher than the sum of differences $4,590$ for ILP (E AN) and $t = 4$ in the deterministic case. The question is if this difference is mainly due to the non-avoidable variance of the workload or if it is due to the modeling of the ILPs using expected panels instead of building a stochastic ILP.

To answer this question, we simulate the panel evolution starting with the panels of the deterministic model for the periods $0$ to $9$ to determine the mean absolute deviation, i.e., the expected absolute difference between the workload and the expected workload after one period. Due to the triangle inequality, the absolute difference between workload and capacity is smaller than the absolute difference between capacity and expected workload and the absolute difference between workload and expected workload. The first term is minimized in the deterministic model, and the second one is the mean absolute deviation. Specifically, we computed $980$ as the sum of the mean absolute deviations over the periods $0$ to $9$. Hence, an approximate lower bound for the expected sum of differences in the non-deterministic case is given by $4,590+980 = 5,570$. This estimated value is only off by $121$ visits or $2$ percent. This indicates that the difference between the deterministic and the non-deterministic results is mainly due to the inherent variability. The efficacy gap that could be reduced by a stochastic model formulation is very small in this case.

Comparing the graphs with and without considering exterior demand variability, we see slightly increased values for the expected sum of differences when we consider demand variability. For example, we observe a difference of $108$ visits or $2$ percent for the expected sum of differences for ILP (E AN) and $t = 4$. On the one hand, this shows a generally small influence of the exterior demand variability. It also indicates that our method to determine the upper bounds for the number of patients to be added in a period is effective. However, we can not say if this already small difference can be reduced further using a different method to handle demand variability.

The average number of models solved each period when considering $t = 4$ periods is $8$ for ILP (E AN) and $4$ for ILP (E A). Hence, we achieve good results with a small effort.

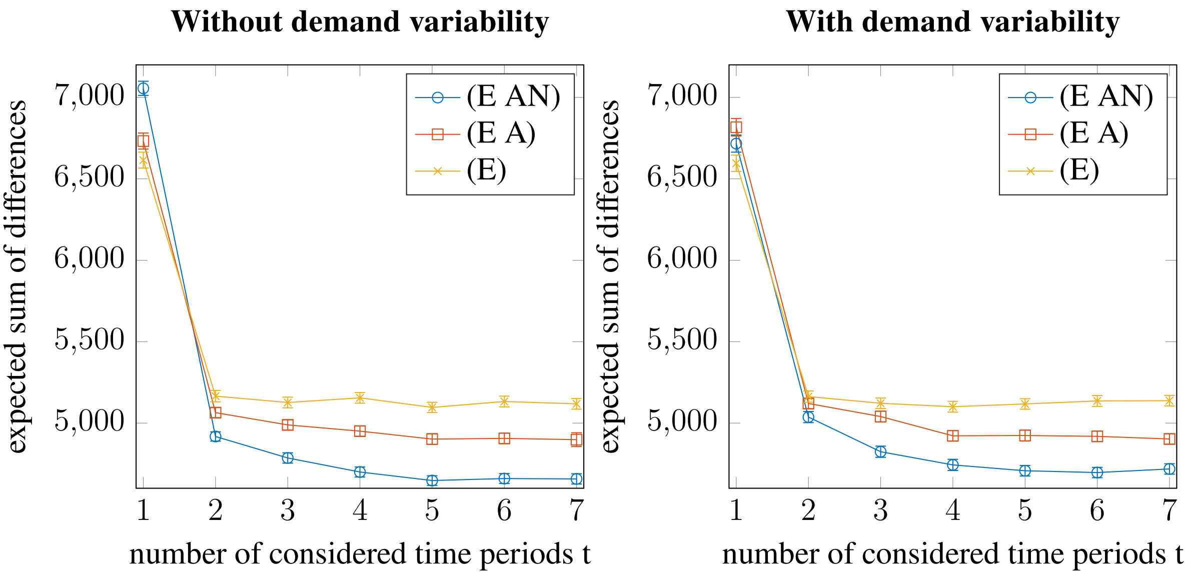

**Figure 5.7:** Simulated expected sum of differences for the three ILPs and different values of $t$ in Scenario 2

In Figure 5.7, the expected sum of differences (with and without considering demand variability) dependent on the ILP and the number of considered periods is shown for Scenario 2. The error bars represent the 95 percent confidence interval after $1,000$ simulation replications.

We again see a very similar behavior between the curves for the expected sum of difference and the curves from Figure 5.3 for the sum of differences. Again, for big values of $t$, due to the considered uncertainty, the values for the expected sum of difference increase by approximately $1,400$ visits compared to not considering any uncertainty. Moreover, comparing the graphs with and without considering exterior demand variability, we see slightly increased values by maximal 60 visits for the expected sum of differences when considering demand variability for big values of $t$. The average number of models solved each period when considering $t = 7$ periods is 8.8 for ILP (E AN) and 4.5 for ILP (E A).

In Figure 5.8, the expected sum of differences (with and without considering exterior demand variability) dependent on the ILP and the number of considered periods is shown for Scenario 3. Here, we omit the first two summands of the sum of differences due to their high variability. In the first two periods, no new patients are accepted in any setting. The error bars represent the 95 percent confidence interval after $6,000$ simulation replications.

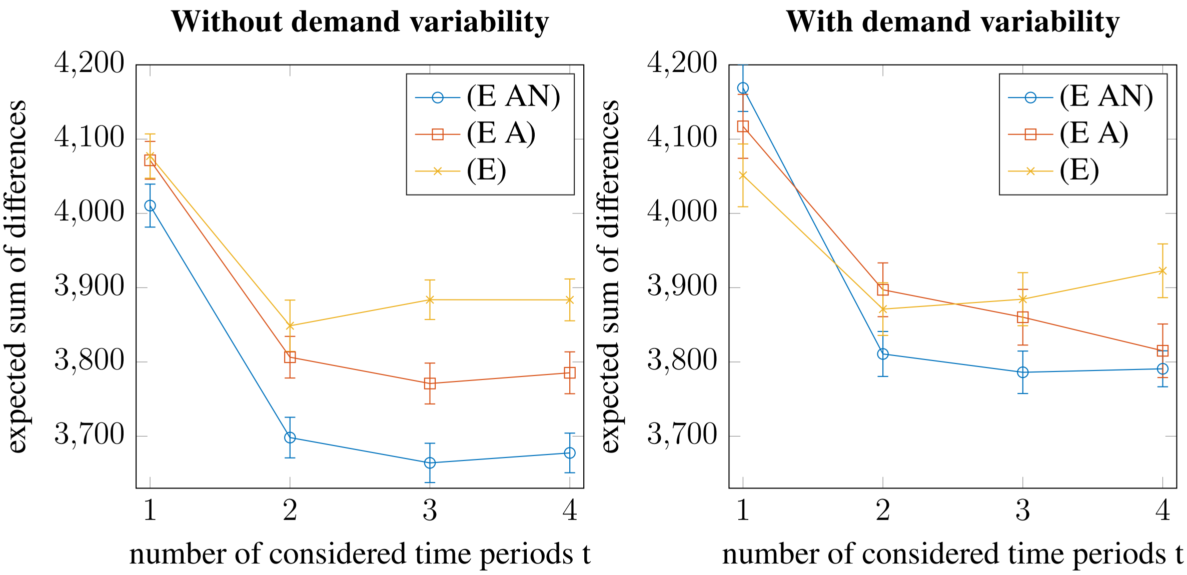

**Figure 5.8:** Simulated expected sum of differences for the three ILPs and different values of $t$ in Scenario 3

In Figure 5.8, the expected sum of differences (with and without considering exterior demand variability) dependent on the ILP and the number of considered periods is shown for Scenario 3. Here, we omit the first two summands of the sum of differences due to their high variability. In the first two periods, no new patients are accepted in any setting. The error bars represent the 95 percent confidence interval after $6,000$ simulation replications.

Again, the curves in the deterministic and non-deterministic cases show similar behavior. However, due to the missing first two periods, it is difficult to compare the deterministic case and the non-deterministic case without exterior demand variability. The big difference of $3,128$ visits between the expected sum of differences and the sum of differences for ILP (E AN) and $t = 4$ steams from the variance of the workload as before, but also from the differences between the panels in period $2$ since we omit periods $0$ and $1$. Comparing the graphs with and without considering demand variability, we experience slightly increased values by maximal $113$ visits for the expected sum of differences when we consider demand variability for big values of $t$. The average number of models solved each period when considering $t = 4$ periods is $10.4$ for ILP (E AN) and $5.2$ for ILP (E A).

We see that the three considered ILPs are beneficial in a stochastic environment for all three presented scenarios. Considering more than one future period still has the most significant effect on decreasing the expected sum of differences. However, the expected sum of differences is bigger than in the deterministic case. Based on one specific example, we show that this discrepancy is probably mainly due to the non-avoidable variance. Therefore, the effect of using potential stochastic ILPs or different methods to handle the exterior demand variability is likely to be small. We further see that the impact of considering new patient demand variability is relatively small in general. No optimization can help if the total demand is too low to reach the target capacity. If we experience high new patient demand, we are more flexible in deciding whom to accept. However, the benefit of a detailed classification of new patients is small, even smaller than in the deterministic case.

These findings together indicate that a perfect demand forecast is not necessary for our programs' usefulness. We will benefit even if we take a few future periods into account with a non-perfect demand forecast and no new patient classification.

## 5.6 Conclusion and outlook

In this paper, we present deterministic integer linear programs (ILPs) that decide on the intake of new patients into physician panels while taking into account the future panel development. The primary objective is to minimize the deviation between the expected panel workload and the physician's capacity for the current and future periods. To the best of our knowledge, this article is the first work that classifies patients by the number of visits to the physician, takes the temporal panel evolution into consideration, and that decides on the intake of new patients into existing family practice panels.

Our numerical experiments show that we can significantly lower the deviation between workload and capacity when we consider several future periods instead of one in the optimization. The deviation can be decreased a little bit further by using a detailed classification of new patients. The benefits of the developed ILPs persist even in an uncertain environment, taking as few as two to three future periods into account and without a detailed patient classification. Further, the numerical experiments show that the demand variability of new patients has a small impact on the results. Therefore, we believe that the presented models can help physicians manage their patient panels to balance supply and demand in practice.

We are aware that classifying new patients may be an ethical problem. We show that the classification has a significant benefit additional to the consideration of several future periods, but we do not imply that this benefit has to be exploited, especially in health care. However, in other application areas, the classification of new customers may be justifiable. In fact, the decision on the detail of new patient classification can be understood as a trade-off between fairness, i.e., little to no classification, and effectiveness, i.e., a significant further decrease of the sum of differences.

We find that the improvement when using potential stochastic ILPs is likely to be small which confirms the validity of using deterministic ILPs. For the case of stochastic demand of new patients with known expected values, we further propose an algorithm to define upper bounds on the number of patients in a patient class to be accepted in a period through solving the ILP several times with different demand inputs. Instead of excessive re-optimization, this approach only needs a low number of ILP runs per period. The results are close to the results in the case of deterministic demand of new patients, which suggests a small influence of the exterior demand uncertainty and shows the validity of our algorithm.

Another result of our work is the finding that the current panel size does not adequately describe workload. We further need to know the time frame in which we count the number of patients that have been seen by the physician. Analyzing our data, we saw that it is reasonable to use more extended time frames than the often-used two years. Further, the average workload of a single panel patient is needed to determine the panel's total workload.

We presented our work to the problem owners. They were very interested in the descriptive analysis of their data specifically the panel composition. They were impressed with the results and think that the model produces valuable results that are relevant in real-life. They believe that the model as it is now would probably be most useful for group practices. They further suggest to apply the model in case of practice transfers. For future research, they propose to explore the macro perspective, i.e., the development of patient visit behavior and the patient practice assignment in a whole region. Concerning implementation, the problem owners remark that the definition of a visit probably has to be refined and that the integration of the model into the practice administrative software is an issue that has to be resolved.

Further, we are aware that data collection is an issue. For example, Scenario 1 from Section 5.4, where a physician starts a new practice is a difficult use case in practice because the data needed to make the decisions is not yet available at the time of the decision making. Therefore, it would be interesting to analyze data from several practices to see if the patient visit behavior is similar, e.g., if the workload of new patients tends to increase over the first periods.

In the future, we plan to collect more data to test the ILPs not considered in the numerical experiments here. In particular, we want to investigate further static patient attributes that influence the number of visits in a period and the ILPs for the (re-)design of several physician panels at once, including the variance objective. Working with the actual time requirement of patients measured in minutes instead of counting visits would be another interesting path of future research. We also plan on exploring further application areas for our model, besides health care.

# 6 Extensions to the panel management models

In the following, we illustrate further reflections and extensions to Chapter 5. We start by presenting an ILP that considers additional resources besides the physician. Another ILP considers several physicians with possibly overlapping panels. We further introduce a mixed integer linear program (MILP) to define capacity values. Then, we shortly explain how the capacities in the panel management ILPs can be defined using the indirect queuing model. Finally, we describe another real-world appointment data set from a two-physician urology practice which includes information on the patients' gender and present selected numerical experiments.

## 6.1 Further panel management integer linear programs

In this first extension to the panel management programs, we consider different resources simultaneously, e.g., the physician and the laboratory. Here, a visit category defines an expected number of visits to the physician and an expected number of visits to the laboratory in a period. In the objective function, we minimize the sum of the sum of deviations between workload and capacity for the first resource and the sum of deviations between the second resource's workload and capacity. We name the ILP (E SR AN) following the taxonomy used in Chapter 5 adding $SR$ for several resources. Further, we use the superscripts $p$ and $l$ to specify the parameters belonging to our exemplary resources physician and laboratory. We especially define an expected number of visits to the physician $f_j^p$ and an expected number of visits to the laboratory $f_j^l$ for each visit category $j \in M$. This, in turn, leads to two different expected workloads $o_{kij}^p$ and $o_{kij}^l$ of a patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods.

$$(E\ SR\ AN) \qquad \text{min} \sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( o^{p}_{kij} v_{0ij} + \sum_{h=0}^{k-1} o^{p}_{(k-h-1)ij} w_{hij} \right) - c^{p}_{k} \right| + \sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( o^{l}_{kij} v_{0ij} + \sum_{h=0}^{k-1} o^{l}_{(k-h-1)ij} w_{hij} \right) - c^{l}_{k} \right| \tag{6.1}$$

s.t.

$$w_{kij} \leq d_{kij} \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M \tag{6.2}$$

$$w_{kij} \in \mathbb{Z}^{+}_{0} \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M \tag{6.3}$$

The ILP (E SR AN) allows, for example, to differentiate between visits with and without physician contact and can therefore level the physician and the non-physician staff capacities. However, note that we probably need to aggregate more to have enough data to compute the transition probabilities in the data preparation phase, even more so when considering more than two resources.

Considering several physicians, often there is no clear distinction between their panels. Patients do not only visit their primary care physician (PCP) but also other physicians belonging to the same practice, e.g., if their PCP is on holiday. Some patients switch several times between physicians. To represent these cases in our programs, we extend the original ILP (E SP AN). A new patient can add to several physicians' workloads based on a known distribution from historical data. To this end, we add the patient to one designated physician and one visit category. However, we assume that the patient will also consume other physicians' capacities using the parameters $\alpha_{ab}$ for $a, b \in A$ where we define $\alpha_{ab}$ as the proportion of time that a patient added to the panel of physician $b$ is, in fact, treated by physician $a$. Of course, we need to have $\sum_a \alpha_{ab} = 1$ for all $b \in A$. Again, we name the ILP (E SPD AN) following the taxonomy used in Chapter 5 adding D for distributed. Note that we minimize the sum of the sum of differences over all physicians instead of using a min-max approach as in ILP (E SP AN).

$$(E\ SPD\ AN) \qquad \text{min} \sum_{a \in A} \sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( \sum_{b \in A} o^{b}_{kij} \alpha_{ab} v^{b}_{0ij} + \sum_{b \in A} \sum_{h=0}^{k-1} o^{b}_{(k-h-1)ij} \alpha_{ab} w^{b}_{hij} \right) - c^{a}_{k} \right| \tag{6.4}$$

s.t.

$$\sum_{a \in A} w^{a}_{kij} \leq d_{kij} \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M \tag{6.5}$$

$$w^{a}_{kij} \in \mathbb{Z}^{+}_{0} \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M, a \in A \tag{6.6}$$

The ILP (E SPD AN) allows for more flexibility than the ILP (E SP AN) and probably captures the reality in group practices better. However, if no patient-physician assignment is given in a real setting, we need to define assignment rules based on data. For example, we could decide to assign the patient to the physician whom the patient visited the most in a defined time frame. We could break ties through using a random assignment.

What exactly is the difference between the two ILPs (E SR AN) and (E SPD AN)? In ILP (E SR AN) there are no separate panels. Hence the transition probabilities do not depend on the resources, and there is no patient- resource assignment upon admission. The expected workload $o^b_{kij}$ of a patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods for resource $b$ differs between resources only due to the difference in the expected number of visits $f^b_j$ to resource $b$ per period for a patient belonging to visit category $j \in M$. Whereas in ILP (E SPD AN) the different $o^b_{kij}$ values already result from the different transition probabilities that are physician-dependent.

In the two ILPs (E SR AN) and (E SPD AN), we decide to minimize the sum of the sum of differences for the different resources or physicians instead of using a min-max approach. When using a min-max approach that results in an objective value significantly bigger than zero, the potential for minimizing the other sums of differences is not exploited. If one still wanted to apply a min-max approach, one should optimize a second time using the sum approach and constraining each sum of differences by the beforehand determined min-max objective plus a potential deviation.

## 6.2 A mixed integer linear program to define the physicians' capacities

The ILPs presented so far aim to minimize the deviation between workload and capacity. However, we know that, in reality, the physician will have to handle any mismatch by adjusting the planned capacity. Therefore, it is interesting to determine capacity definitions that result in only minor deviations between workload and capacity. We explain how this can be done via solving an MILP in an exemplary scenario where we use the ILP (E AN) as a basis. Imagine a practice that wants to lower its capacity over several periods, e.g., due to a sabbatical of one physician, and then go back to its original capacity. To match the workload, the practice needs to adjust its capacities before and after the capacity change

We define that from the $t$ periods considered in the optimization, the capacity should be equal to a lower capacity value $L$ for $r$ consecutive periods and after that equal to an upper capacity value $U$ for a maximum number of consecutive periods. We further allow for a maximum deviation of the sum of differences $g$. Hence, we define the capacities $c_k$ as decision variables with $c_k \in [L, U]$ for $k \in \{1, \dots, t\}$. We further introduce binary decision variables $b_k$ and $e_k$ for $k \in \{1, \dots, t\}$ that should be equal to one if the capacity $c_k$ in period $k \in \{1, \dots, t\}$ equals the lower capacity value or the upper capacity value, respectively. We name the resulting MILP (C AN) with C for capacity.

$$(CAN) \qquad \max \sum_{k=1}^{t} e_k \tag{6.7}$$

s.t.

$$\sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left( o_{kij} v_{0ij} + \sum_{h=0}^{k-1} o_{(k-h-1)ij} w_{hij} \right) - c_k \right| \leq g \tag{6.8}$$

$$w_{kij} \leq d_{kij} \qquad \forall k \in \{0, \dots, t-1\}, i \in N, j \in M \tag{6.9}$$

$$\sum_{k=0}^{t} b_k = r \tag{6.10}$$

$$b_k \geq b_{k-1} - \frac{1}{r} \sum_{l=0}^{k-1} b_l \qquad \forall k \in \{2, \dots, t\} \tag{6.11}$$

$$e_k \geq e_{k-1} \qquad \forall k \in \{2, \dots, t\} \tag{6.12}$$

$$e_k \leq \frac{1}{r} \sum_{l=0}^{k-1} b_l \qquad \forall k \in \{1, \dots, t\} \tag{6.13}$$

$$c_k \leq b_k L + (1 - b_k) U \qquad \forall k \in \{1, \dots, t\} \tag{6.14}$$

$$c_k \geq e_k U \qquad \forall k \in \{1, \dots, t\} \tag{6.15}$$

$$c_k \in [L, U] \qquad \forall k \in \{1, \dots, t\} \tag{6.16}$$

$$w_{kij} \in \mathbb{Z}_0^+ \qquad \forall k \in \{0, \dots, t-1\}, i \in N, j \in M \tag{6.17}$$

The objective (6.7) maximizes the number of periods the capacity equals the upper capacity value. This objective minimizes the number of periods it takes to lower the capacity and then increase it again. The first constraint (6.8) bounds the sum of absolute differences between workload and capacity by $g$. The second set of constraints (6.9) ensures that we can not add more patients to the panel than there are available. The third constraint (6.10) fixes the number of periods the capacity equals the lower capacity value to $r$. The set of constraints (6.11) makes sure that the capacity equals the lower capacity value if the capacity equaled the lower capacity value in the preceding period and if there have not yet been $r$ consecutive periods with a capacity equal to the lower capacity value. The set of constraints (6.12) ensures that the capacity equals the upper capacity value if the capacity equaled the upper capacity value in the preceding period. The set of constraints (6.13) ensures that the capacity can not reach the upper capacity value in a period if there have not yet been $r$ periods with capacities equal to the lower capacity value before. The set of constraints (6.14) links the binary variables $b_k$ to the capacity variables $c_k$, ensuring that the capacity equals the lower capacity value for all periods $k \in \{1, \ldots, t\}$ with $b_k = 1$. The set of constraints (6.15) links the binary variables $e_k$ to the capacity variables $c_k$, ensuring that the capacity equals the upper capacity value for all periods $k \in \{1, \ldots, t\}$ with $e_k = 1$. The last two sets of constraints are the domain constraints.

Here again, we can, of course, alter the considered patient attributes, especially for new patients. Using more new patient attributes may result in a smaller number of periods it takes to reach the upper capacity value after the capacity changes.

## 6.3 Combining the panel management models with the indirect queueing models

In the following, we explain how the panel management ILPs and the indirect queueing model can be combined theoretically. In the integer linear programs to manage the intake of new patients into a physician panel over time presented in Chapter 5, we aim to match the physician's capacity with the demand from panel patients per period. In the following, we specifically focus on appointment demand coming from one physician panel. Then, due to the variability of the problem as described in 5.2, the physician's capacity in the panel management model should not be set to the time that the physician is available. Instead, it should be defined as the maximal workload, e.g., the number of appointments the physician can manage in a period given her time availability such that a predefined service level is met. In the following, when we talk about the physician's capacity, we mean this maximal workload. This section goes into greater detail and explains how this capacity can be determined. We focus on the case of a service level based on the indirect waiting time distribution. Then, a variant of the indirect queueing model presented in Chapter 3 can be applied.

So far, in our integer linear programs for panel management, we assume a demand from panel patients independent of the physician's capacity. This assumption might be justified in cases where the workload is small compared to the physician's capacity or where the physician will adapt her capacity and work overtime to serve all demand from panel patients regardless of the capacity defined beforehand. Otherwise, the physician's capacity will have an impact on the demand from panel patients because the forming of long backlogs might influence the demand arrival pattern and lead to rejections of appointment requests. Hence, we should include these effects in our models, especially when defining parameters from data. Again, the idea here is to combine a variant of the indirect queueing model presented in Chapter 3 with the integer linear programs for panel management.

In the following, we assume that we can observe all appointment requests from the panel. An appointment request can be rejected or accepted. An accepted appointment request can lead to a patient canceling the appointment, showing up to the appointment, or not showing up to the appointment. Concerning early cancelations, we handle them as if the patient did not request an appointment. We consider late cancelations that do not allow the released slot to be booked again to be no-shows. Last but not least, we assume that no-shows leave the physician idle in the booked appointment slot. These assumptions are in line with the requirements of the indirect queueing model.

In the following, we apply variants of the indirect queueing model with increasing complexity. Contrary to Chapter 5, we count appointment requests of patients instead of realized appointments in the panel management programs. This is because the appointment request rate $\lambda$ is a needed input for the indirect queueing model. So fixing a capacity with a known relationship to $\lambda$, we can directly apply the indirect queueing model. Using a capacity with respect to the realized appointments as in Chapter 5 has the disadvantage that there is no simple connection between the defined capacity and $\lambda$, which makes it challenging to apply the indirect queueing model. Nevertheless, we have to be aware that counting appointment requests instead of realized appointments leads to a time shift of demand because the request happens before the appointment. However, this influence is not too significant, assuming, for example, a period of a year and not too long backlogs.

### 6.3.1 Constant appointment request rate

In the simplest case, we assume an appointment request rate $\lambda$ (excluding rescheduling requests) independent of the indirect queue length.

Suppose the probability of rescheduling $\rho$ is also independent of the queue length. In that case, we can approximate and decide to integrate the rescheduling requests into the appointment requests using, for example, $\lambda + \frac{1}{T}\rho$ as the appointment request rate and setting the rescheduling probability to zero. The error in the indirect waiting time distribution is likely to be small, as we saw in Section 4.4.7. In this case, we count the number of appointment requests - including rescheduling requests - per patient in a period to define the visit categories in the panel management programs.

The physician then decides on the mean daily availability for every considered period $k$, i.e., on a mean number of offered daily slots $\mu_k$ and the queue capacity $K$. We then compute the results of our indirect queueing model using increasing values of $\lambda_k$ to determine the maximal value of $\lambda_k$ such that a service level fixed beforehand is achieved. The capacity $c_k$ of the physician in the panel management programs in a period $k$ corresponds to the total number of appointment requests per year, which we deduce as $\lambda_k$ multiplied by the number of days the physician works in period $k$.

In the next step, we consider a queue length-dependent rescheduling probability. In this case, to define the model parameters from data, we count the number of appointment requests, excluding rescheduling requests, per patient in a period to define the visit categories in the panel management programs. Further, we define the function $\rho(\cdot)$, where $\rho(i)$ is the probability of rescheduling of a patient departing from the indirect queue leaving behind $i$ patients in the system. Here, we assume that the function $\rho$ will not change over the upcoming periods.

Then, given the mean number of offered daily slots $\mu_k$, the queue capacity $K$, and the rescheduling probability function $\rho$, we again approximately compute the maximal appointment request rate $\lambda_k$ in period $k$ such that a service level fixed beforehand is achieved. We then again compute the capacity $c_k$ of the physician in the panel management programs in a period $k$ as $\lambda_k$ multiplied by the number of days the physician works in period $k$.

### 6.3.2 Non-constant appointment request rate

The problem becomes more complex if we assume an appointment request rate function $\lambda(\cdot)$ dependent on the queue length. The issue that arises is the definition of the visit categories. Assuming that the number of appointment requests of a panel patient depends on the capacity offered, we need to find a way to normalize the appointment request demand to build meaningful visit categories and then compute transition probabilities.

We can again decide between integrating the rescheduling requests into the appointment request rate or not. Either way, we need to make an assumption on the appointment request function $\lambda(\cdot)$. Let us, for example, assume a linear function $\lambda(i) = mi + n$ with $i$ being the queue length. For example, for $m > 0$, we will see more appointment requests as the indirect queue becomes longer. We know the total number of appointment requests (in- or excluding rescheduling requests) in the documented periods from data. The number of requests there would have been in a period in case of no backlogs, i.e., for $i = 0$, is given as $n$ times the number of days the physician worked. We then define the normalizing factor $\tau$ as the number of appointment requests in the case of $i = 0$ divided by the number of actual appointment requests. Next, we transfer the idea to a single panel patient and define the visit categories of a panel patient based on the appointment requests the patient would have made in a period of $i = 0$ by multiplying the number of actual appointment requests made in the period with $\tau$.

Here, another problem arises. The number of actual appointment requests is an integer. Therefore, the visit categories also represent integer numbers of visits or a range of integer numbers of visits together with an expected number of visits. However, in general, multiplying the number of actual visits by $\tau$ does not yield an integer value. To solve this issue, we partition a patient with a resulting non-integer visit number $x_1$ with weight $\lceil x_1 \rceil - x_1$ into $\lfloor x_1 \rfloor$ number of visits and with weight $x_1 - \lfloor x_1 \rfloor$ into $\lceil x_1 \rceil$ number of visits. The resulting weights are then divided between the possibly aggregated visit categories if necessary. To then define the transition probabilities, we need to know the number of visits of the considered patient in the next period. There again, we normalize the number of actual visits to $x_2$ and partition the patient between two potentially different visits categories. To define the empirical transition distributions, we then assume a transition from $\lfloor x_1 \rfloor$ to $\lfloor x_2 \rfloor$ with weight $(\lceil x_1 \rceil - x_1)(\lceil x_2 \rceil - x_2)$, a transition from $\lfloor x_1 \rfloor$ to $\lceil x_2 \rceil$ with weight $(\lceil x_1 \rceil - x_1)(x_2 - \lfloor x_2 \rfloor)$, a transition from $\lceil x_1 \rceil$ to $\lceil x_2 \rceil$ with weight $(x_1 - \lfloor x_1 \rfloor)(x_2 - \lfloor x_2 \rfloor)$, and a transition from $\lceil x_1 \rceil$ to $\lfloor x_2 \rfloor$ with weight $(x_1 - \lfloor x_1 \rfloor)(\lceil x_2 \rceil - x_2)$.

Further, we need to define the parameters $m$ and $n$ of the function $\lambda$. The only information available is the capacity $c_k$ of the physician in the period $k$. Hence, we need to define functions $m(\cdot)$ and $n(\cdot)$ dependent on the capacity. As before, for the parameter $n$, we assume that $n(c_k)$ is equal to $c_k$ divided by the number of days the physician plans to work in period $k$. For the parameter $m$, the relationship with the capacity is unclear and needs to be determined from data. It is also possible that $m$ does not change significantly over the periods and can be considered constant.

Lastly, given the mean number of offered daily slots $\mu_k$, the queue capacity $K$ and possibly the rescheduling probability function $\rho$, we again approximately determine the maximal capacity $c_k$ resulting in an appointment request rate function $\lambda_k$ with $\lambda_k(i) = m(c_k)i + n(c_k)$ in period $k$ such that a service level fixed beforehand is achieved.

## 6.4 Appointment data from an urology group practice

The real-world appointment data used in Chapter 5 unfortunately did not contain all necessary information to test every ILP presented in that chapter. Therefore, we use another data set here to test more ILPs from Chapter 5 as well as the ILPs (E SR AN), (E SPD AN) and (C AN). We have real-world appointment data from a urology practice run by two physicians. In contrast to the data set presented in Chapter 5 we have information on the gender of patients. We further observe that most panel patients can be assigned to a principal physician.

We have complete appointment data for the years 2014 – 2017. Every row in the data set corresponds to a visit to the practice and contains the following details: date, time of day, planned duration time, booking date, category, patient number, gender, age, visit type, resources, and waiting status. The attribute category indicates the physician, another staff member of the practice, or diagnostics (laboratory). However, for some visits, the category is empty. The patient number is the practice-intern assigned patient-ID. The visit type shortly refers to the kind of visit, e.g., consultation with a specific physician, injection, surgery, etc. The resources entry lists the resources (including physicians) used during the visit, the first on the list being the primary resource. The waiting time status is updated from "planned" to "in treatment" to "finished" for an attended appointment. A no-show is indicated by a "no status" waiting status.

We exclude all visit entries with either missing patient number, gender, or age for our subsequent analysis. We further exclude visits with a visit date corresponding to a Saturday or Sunday and all visits with a waiting time status "no status." We note that the physician indicated in the "category"-column is not always mentioned in the corresponding "resources"-column and also does not always correspond with the value in the "visit type"-column and vice versa. The "resources"-column was determined to be the relevant one since it was assumed to represent the actual resources employed for the particular patient visit, while the assigned values in the "category"- and the "visit type"-columns are assumed to represent the original assignment which was done in advance. As explained above, from the "resources"-column, the first resource listed is assumed to be the primary resource. If it is one of the physicians, the corresponding visit was classified as a "physician visit" assigned to that particular physician. All the other visits (with no physician as the primary resource) were classified as "other visits."

Since most appointments had a planned length of $10$ minutes, we decide to count the number of visits and not go into detail on the actual appointment lengths. We count $8,379$ patients who visited the practice $43,992$ times in those $4$ years after data selection. As in Chapter 5, we decide on using a period length of one year. Now, having a data set of $4$ years, we define that a patient who does not show up for $3$ years in a row left the panel and that a patient who did not visit the practice during the last $3$ years is new. The urology practice also assigns ascending patient numbers to new patients. Hence, we can determine how often a new patient, in our definition, is, in fact, a panel patient that did not show for $3$ years. Indeed, in $2017$, $599$ of the $1310$ new patients, according to our definition, have visited the practice before. At the end of $2017$, we find $7030$ panel patients ($5337$ males and $1698$ females) with an average number of visits of $1.53$ ($1.69$ for males and $1.04$ for females).

We include three zero visits categories in our models. Visit category $0$ contains all patients that did not visit this year and not in the last three years. By definition, patients in this visit category will stay in this visit category for all future periods and are not considered part of the panel anymore. Visit category $1$ contains all patients that did not visit this year and the last two years but visited the practice three years ago. Visit category $2$ is defined similarly. We group several numbers of visits such that a visit category consists of at least $5\%$ of all made observations in the $4$ considered years. This grouping results in a total of $9$ visit categories. For every visit category, we then compute the average number of visits to define $f_j$, i.e., the expected number of visits of a patient in visit category $j$. For a patient of visit category $j$, we further set $f_j^p$ and

$f_j^l$ to the average number of visits to the physicians per period and the average number of other visits (e.g., laboratory) per period, respectively.

Because the length of a period is a year, the age categories reflect the age in years. We work with $102$ age categories reflecting the ages $0$ to $101$. A patient of age $101$ automatically leaves the panel in the following period. Figure 6.1 shows the panel composition by age for the $2017$ panel.

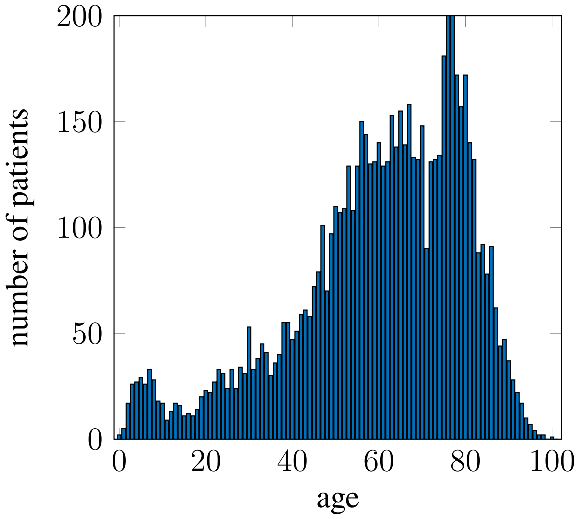

**Figure 6.1:** Number of panel patients per age category in 2017

For the numerical experiments we assume a constant exterior demand $d_{kij}$ in period $k \in \{0, \ldots, t-1\}$ according to age and visit category $i \in N$ and $j \in M$ equivalent to the composition of accepted new patients in $2017$. Note that the real exterior demand in $2017$ may have been higher. Unfortunately, we do not have records of rejected exterior demand. Figure 6.2 shows the number of accepted new patients by age category in $2017$.

Most patients mainly visit one of the two physicians. However, many patients still visit both physicians. Therefore, we refrain from testing ILP (E SP AN) and instead test (E SPD AN). Nonetheless, we need to define a principal physician for every patient based on the data. We either assign the physician that the patient has explicitly visited more or, if no unambiguous assignment is possible that way, the physician that the patient has visited first in the considered period of 2014 – 2017. If the patient has not explicitly visited any of the two physicians in the period (so the patient only had "other visits," e.g., only diagnostics), the primary physician is assigned randomly (with a 50:50-probability). The total number of panel patients in $2017$, which is $7030$, can be divided into $3872$ patients assigned to physician one and $3158$ to physician two with an average number of visits in $2017$ of $1.18$ for physician one and $1.96$ for physician two. For ILP (E SPD AN) we further need to define the parameters $\alpha$ indicating the share of visits of a patient assigned to a physician made to the other physician. We see that patients assigned to physician one visit the second physician in $8.2$ percent of cases and patients assigned to physician two visit physician one in $7.6$ percent of cases.

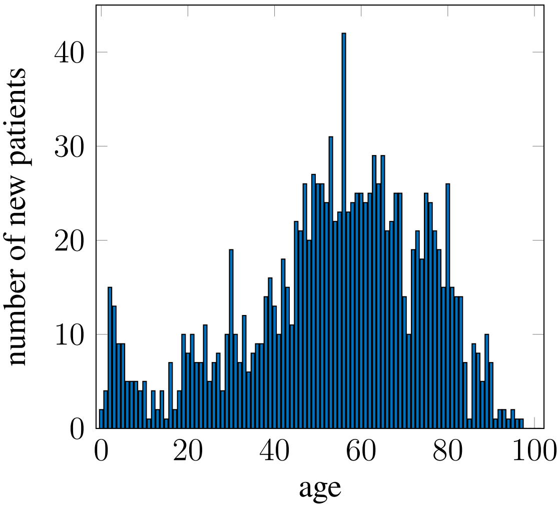

**Figure 6.2:** Number of new patient per age category admitted in 2017

To compute the transition probabilities $q_{ijl}$ that a patient who belonged to age category $i \in N$ and visit category $j \in M$ last period belongs to the visit category $l \in M$ this period, we consider the data from the current age category as well as from the surrounding age categories together. We use at least $5$ age categories and even more age categories with a maximal number of $11$ such that for every age category, there is at least one visit category with at least $30$ observations, i.e., transitions from one year to the next. This procedure was done for the whole panel and for subgroups of the panel corresponding to males, females, patients assigned to the first, and patients assigned to the second physician. However, we did not combine gender and physician assignment because this shrinks the data basis too much. Note that if there are few observations to define a transition distribution, the variance will be artificially low, not representing reality.

For this data set, for some combinations of age category and visit category $i, j$ (and possibly gender or physician category $b$), the number of total observations is zero. Therefore, we need to exclude those combinations in the exterior demand and the starting panel. We also observe transitions into a category combination for which no further transitions can be defined. Here, we manually adjust such observations by setting the number of observations for this transition to zero and add those observations to the next smaller visit category. Figure 6.3 shows the average number of visits per age category.

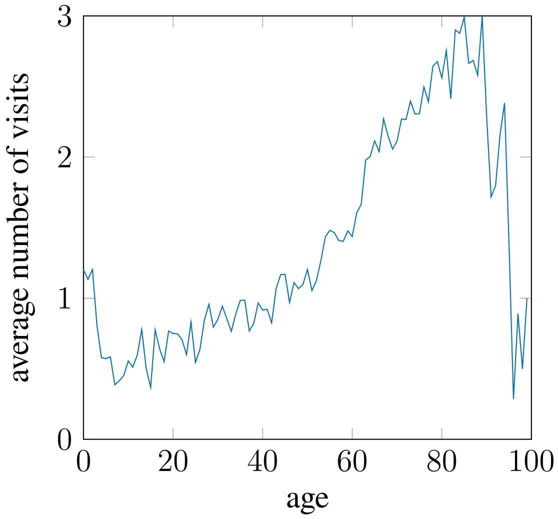

**Figure 6.3:** Average number of visits per age category

Based on the transition probabilities and the average numbers of visits per period, we can determine $o_{kij}$, the expected workload of a single patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods as well as $u_{kij}$, the variance of the workload of a single patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods.

Using the panel of 2017 as the starting panel, we find that the variance of the workload after one period is $29,133.66$, which translates into a standard deviation of $170.69$. The standard deviation corresponds to approximately $1.5$ percent of the starting panel's expected workload after one period.

Compared to the other data set from Chapter 5, we observe significantly fewer visits per panel patient. We further see that a random new patient's expected workload is not increasing in the first periods after admission to the panel, as it was the case in the other data set.

## 6.5 Selected numerical experiments

Next, we present some selected numerical experiments. To this end, we come up with a 10-period scenario based on the real-world data set from Section 6.4 where the benefit of using the ILP models becomes apparent. In this scenario, we first lower the capacity over some periods before elevating it again.

As before, we opt to use a constant exterior demand equal to the accepted new patients in 2017. We observe that this demand is not enough to maintain the workload of the 2017 panel. Hence, we use a starting panel with a smaller workload, i.e., $8,500$, corresponding to $0.8$ times the workload of the 2017 panel. We set the capacities for the periods $k = 0, 1, 2, \ldots, 9$ as follows:

$6,397$; $5,500$; $5;500$; $8,500$; $8,500$; $\ldots$; $8,500$. The first capacity value allows reaching the lower capacity value $5,500$ without a deviation between expected workload and capacity in the first periods by not accepting any demand in the first period. For the ILPs (E SR AN) and (E SPD AN), we need to define individual capacities for the two physicians or the resources, respectively. Here, we split the total capacity such that the proportions remain equal to the proportions of the workloads in $2017$.

As before, we present numerical experiments in a deterministic setting, i.e., with deterministic exterior demand and expected future panels. Again, we stop calculations whenever we reach an absolute MIP gap equal to the number of considered periods in the optimization. We solve an ILP once per period and use the resulting panel from applying the decision of the first period as a starting panel for the next period. We consider different values of $t$ for the number of considered periods in the optimization and different levels of detail concerning the classification of (new) patients. We compare the sum of differences, i.e., the sum of the absolute values of differences between workload and capacity over the $10$ periods of the scenario.

We start with experiments where we include the patient attribute gender. Here, we also present simulation experiments to include uncertainty. Next, we analyze the ILPs with several resources and with several physicians. Then, we investigate the objective of minimizing variance. Last but not least, we show results using the ILP to define capacities. Unfortunately, we cannot present numerical experiments for the combination of the panel management ILPs with the indirect queueing models since the data set does not allow to define the relevant parameters with enough precision.

### 6.5.1 Including the patient attribute gender

Here, we compare the ILPs (E), (E A), (E AN) with the corresponding ILPs including the attribute gender, i.e., (E G), (E G G), (E G A), (E G AG) and (E G ANG). Figure 6.4 shows the sum of differences for the $8$ ILPs with a varying number of considered periods $t$. We see that the sum of differences decreases with the number of considered periods $t$ for all ILPs until it stabilized between $t = 3$ to $t = 5$ periods. Considering $t = 3$ periods covers most of the decrease in the sum of difference for all ILPs except ILP (E G), which needs one more period. Comparing the sums of differences for $t = 5$, we observe that the more detailed the classification of (new) patients, the lower the sum of differences. Here, it is especially interesting to compare the ILPs that only differ in the inclusion of the patient attribute gender for the (new) patient classification. For example, consider the ILPs (E), (E G) and (E G G). Using the ILP (E G) where panel patients are classified by gender (but not new patients) in contrast to ILP (E) results in a drop of the sum of difference by $3.4$ percent. If we also classify new patients by gender using ILP (E G G), we gain another $1.3$ percent. Comparing the foremost best ILP (E AN) with the ILP (E G ANG), which classifies new and panel patients by gender in addition to age and number of visits, we observe an $8.5$ percent decrease in the sum of differences. The results show that including gender as a patient attribute has a significant impact for the considered data set and scenario.

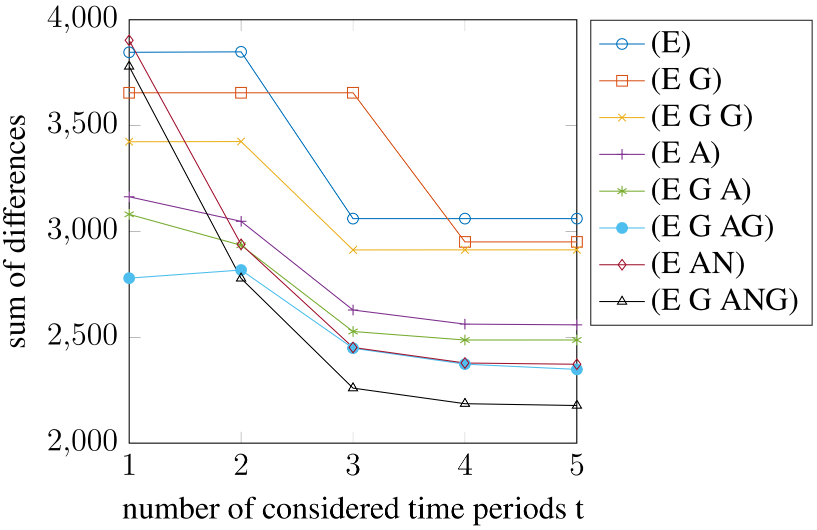

**Figure 6.4:** Sum of differences for the $8$ ILPs and different values of $t$

In Figure 6.5, we depict the simulation results with demand variability, i.e.,the expected sum of differences for different values of $t$ computed as explained in Section 5.5.2.

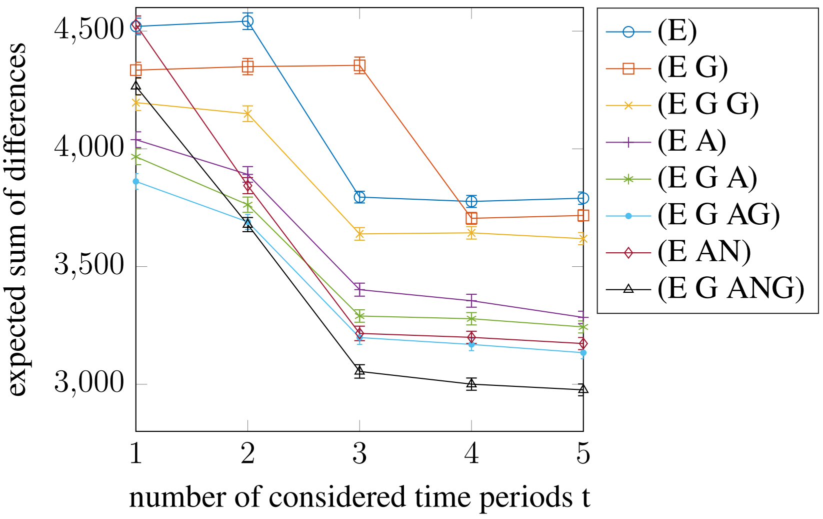

**Figure 6.5:** Expected sum of differences for the $8$ ILPs and different values of $t$

Comparing the graphs to the deterministic results shown in Figure 6.4, we see a very similar behavior with the simulated values being circa $800$ bigger than the deterministic values for $t = 5$. Again, this shows that the benefit of applying the models, including the hierarchy between the model variants, remains in an uncertain environment. In summary, we see that for this scenario, in general, $t = 3$ periods are enough (excluding (E G)) to capture the majority of the decreasing potential. However, we see that the clear hierarchy of model variants is not yet given for $t = 1$ or $t = 2$ periods. Therefore, if one were to apply the model with few periods, it is advisable to apply several model variants to find the best solution. Also, a non-expected relation between two or more model variants may indicate that the chosen number of considered periods $t$ in the optimization is not enough.

Besides the expected sum of differences, we also investigate the sum of variances of the workloads over all $10$ periods. The sum of variances also decreases for increasing values of $t$ for all model variants, as shown in Figure 6.6. With higher values of $t$, we do not only improve the match between workload and capacity but also decrease the uncertainty in the workload itself. However, there seems to be no apparent relationship between the model variants and the sum of variance values.

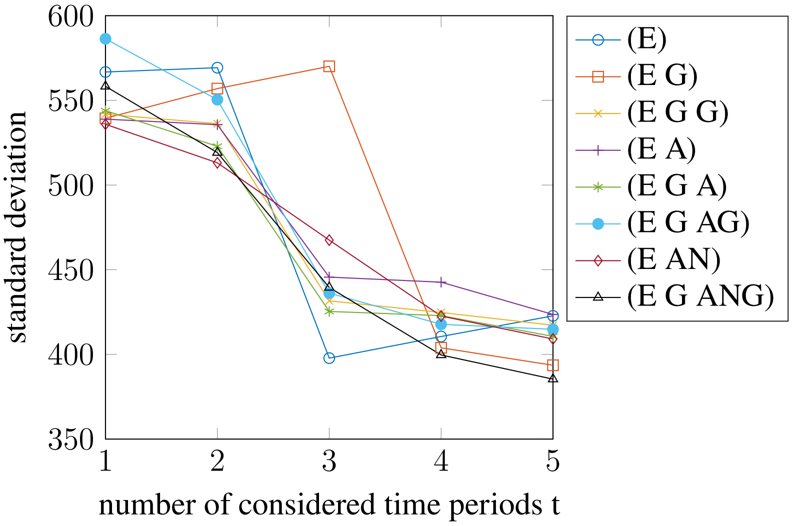

**Figure 6.6:** Standard deviation of the expected sum of workloads for the $8$ ILPs and different values of $t$

As explained in Section 5.5.2, in the simulation, we solve an ILP several times to define upper bounds for the number of patients to be admitted for each patient class in a period. The average number of models that have to be solved per period in the simulation for each model variant is shown in Figure 6.7. There is no clear relationship between the average number of models solved and the number of considered periods in the optimization. However, as expected, the more new patient attributes are considered, the more models need to be solved. Fortunately, the average number of models solved seems not proportional to the number of new patient classes. For example, the ILP (E G G) considers two patient classes (females and males), the ILPs (E G AG) and (E G ANG) consider $2 \cdot 102 = 204$ and $2 \cdot 9 \cdot 102 = 1836$ new patient classes, respectively. Hence, the number of patient classes is approximately $100$ and $1000$ times bigger.

The average number of models solved only increases from $1.8$ to $5.5$ to $9$. This indicates that our Algorithm 1 as defined in Section 5.5.2 is also useful for model variants that consider many patient classes.

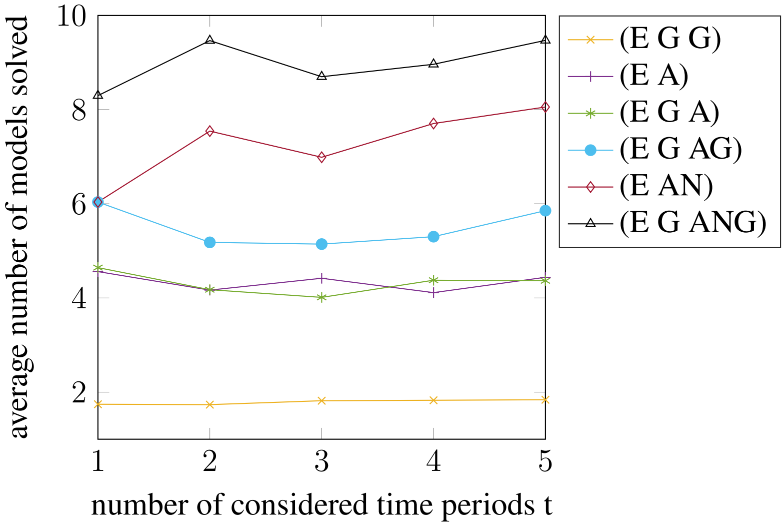

**Figure 6.7:** Average number of models solved for the 6 ILPs and different values of $t$

### 6.5.2 Considering several resources

Here, we consider two resources: the physicians and the laboratory (i.e., other visits). In Figure 6.8, we see the expected number of visits $f_j^b$ per period to resource $b$ of a patient of visit category $j$. Interestingly, for smaller visit categories, most visits are visits to the physicians, whereas, for bigger visit categories, patients visit the laboratory more often than the physicians.

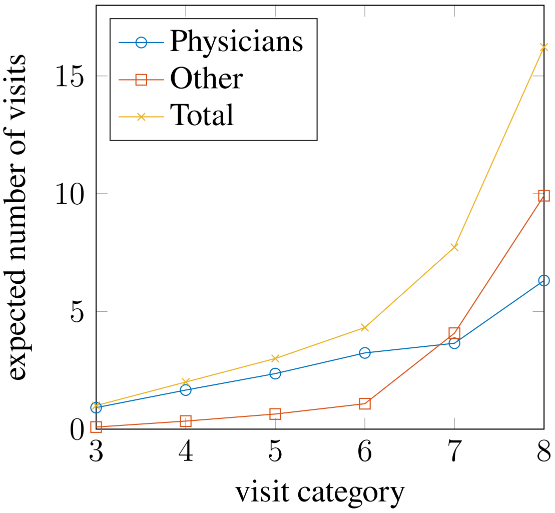

**Figure 6.8:** Expected number of visits per visit category and resource

We compare results for the ILPs (E SR), (E SR A) and (E SR AN). In Figure 6.9 we depict the sum of differences for the three ILPS with a varying number of considered periods $t$. We see that it decreases with $t$ for all ILPs until it stabilized at $t = 3$ for ILP (E SR) and $t = 4$ for ILPs (E SR A) and (E SR AN). Again the lowest sum of differences is reached using ILP (E SR AN) with $t = 5$.

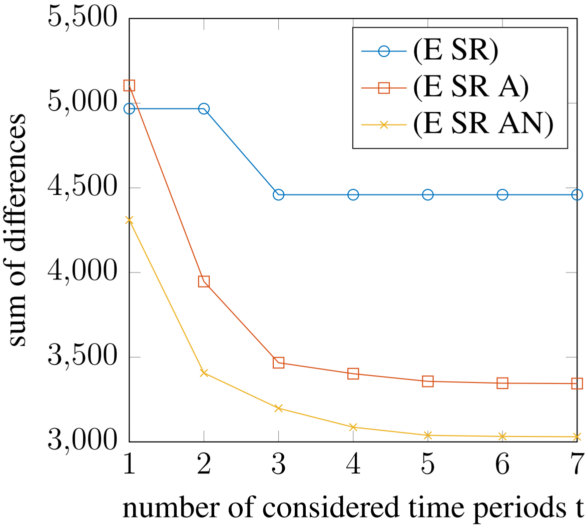

**Figure 6.9:** Sum of differences for the 3 ILPs and different values of $t$

In the considered ILPs, we sum up the deviation between workload and capacity of the physicians and the laboratory (here, i.e., other visits), assuming equal importance of both resources. If the physician is to be assumed the central resource, one could also choose to set the laboratory capacities as decision variables.

### 6.5.3 Including patient physician assignment

Here, we compare the ILPs (E SPD), (E SPD A), and (E SPD AN), which besides the admission decision, also decide on the patient-physician assignment considering the three levels of new patient classification. The results can be seen in Figure 6.10. Again $t = 3$ periods in the optimization are enough to capture the majority of the possible decrease in the sum of differences. Including age as a new patient attribute results in a decrease of 25 percent compared to using no new patient classification. Another 11.2 percent can be gained by including the number of visits as a new patient attribute. Comparing the sum of differences with those computed in Section 6.5.1, we see a significate decrease of 7.8 percent for ILP (E G ANG) compared to ILP (E SPD AN) for $t = 5$. Note that the objective function for the ILPs considering the physician assignment measures the deviations between workload for both physicians separately and then adds them up.

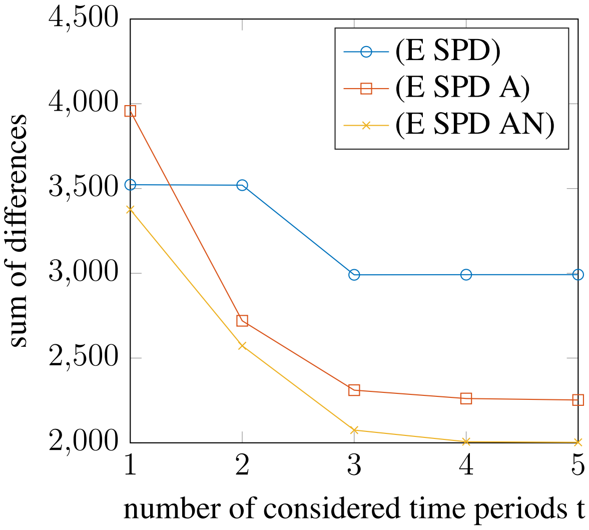

**Figure 6.10:** Sum of differences for the three ILPs and different values of $t$

Even if not realistically achievable it is still interesting to consider ILPs with patient-physician assignment that use the objective function of the ILPs without patient-physician assignment, i.e., that measure the deviation of the workload from the capacity for both physicians together.

Hence, instead of minimizing

$$\sum_{a\in A}\sum_{k=1}^{t}\left|\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(\sum_{b\in A}o_{kij}^{b}\alpha_{ab}v_{0ij}^{b}+\sum_{b\in A}\sum_{h=0}^{k-1}o_{(k-h-1)ij}^{b}\alpha_{ab}w_{hij}^{b}\right)-c_k^a\right|$$

we minimize

$$\sum_{k=1}^{t}\left|\sum_{a\in A}\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(\sum_{b\in A}o_{kij}^{b}\alpha_{ab}v_{0ij}^{b}+\sum_{b\in A}\sum_{h=0}^{k-1}o_{(k-h-1)ij}^{b}\alpha_{ab}w_{hij}^{b}\right)-\sum_{a\in A}c_k^a\right|.$$

We add a T in the ILP denotation to indicate that we minimize the deviation between the total workload from the total capacity. The results can be seen in Figure 6.11. The sum of differences for ILP (ET SPD AN) for $t = 5$ is $39.6$ percent lower than the sum of differences of ILP (E G ANG) and $34.5$ percent lower than the sum of difference of ILP (E SPD AN).

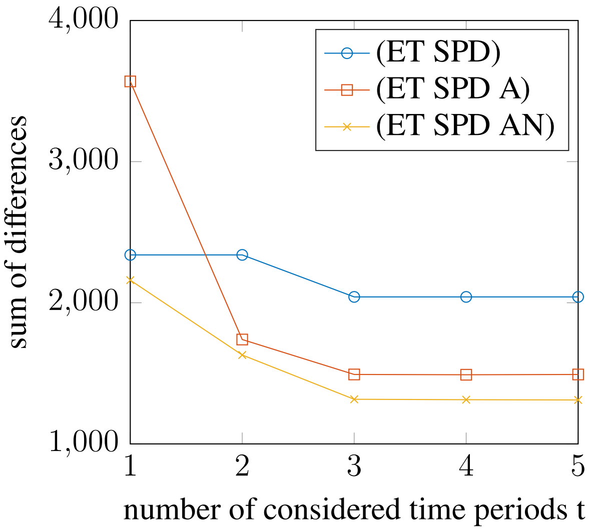

**Figure 6.11:** Sum of differences for the three ILPs and different values of $t$

We conclude that the additional decision of the patient-physician assignment has a high potential in decreasing the sum of differences. If enough data is available, we believe that considering gender and patient-physician assignment together will probably result in an additional decrease.

### 6.5.4 Minimizing variance

In this section, we focus on the objective to minimize the workload variance. We explained before that, in general, using a min-max approach does not exploit every decreasing potential. Therefore, we do not work with ILP (SD SP AN) as defined in Chapter 5 but instead, we use the model formulation of ILP (E SPD AN) as a basis and work with:

$$(SD\ SPD\ AN) \qquad \min \sum_{a\in A}\sum_{k=1}^{t}\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(u^a_{kij}v^a_{0ij}+\sum_{h=0}^{k-1}u^a_{(k-h-1)ij}w^a_{hij}\right) \tag{6.18}$$

s.t.

$$\sum_{a\in A}\sum_{k=1}^{t}\left|\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(o^a_{kij}v^a_{0ij}+\sum_{h=0}^{k-1}o^a_{(k-h-1)ij}w^a_{hij}\right)-c^a_k\right|\leq s \tag{6.19}$$

$$\sum_{a\in A}w^a_{ijk}=w_{kij} \tag{6.20}$$

$$w^a_{kij}\in\mathbb{Z}^+_0 \qquad \forall k\in\{0,\dots,t-1\}, i\in N, j\in M, a\in A \tag{6.21}$$

In the following, we will again present results for the ILPs (SD SPD), (SD SPD A), and (SD SPD AN), which differ in the classification of new patients. To achieve meaningful comparisons, it is essential to define the bound $s$ in constraint (6.19). Therefore, we first solve ILP (E SPD AN) and set $s$ to the obtained objective plus a small buffer ($10$ in this case). Subsequently, we solve ILP (SP SPD AN), where we transfer the ILP (E SPD AN) solution as a MIP start and allow for a MIP gap of $5$ percent. We apply the decisions of the first period to the panels and again solve ILP (E SPD AN) to define the bound $s$ and so forth. As an output, we compute the sum variances over all $10$ considered periods and both physicians. In Figure 6.12, the resulting sums of variances are shown for the ILPs that minimize the expected deviation between workload and capacity and those that minimize variance. We see that there is no potential for variance optimization in case of no new patient classification since solving both ILPs (E SPD) and (SD SPD) results in the same sum of variances values. However, classifying new patients by age and number of visits or by age only allows minimizing variance for a considered number of optimization periods $t$ of at least $3$. Considering age as a new patient attribute and using $t = 7$ periods, we lower the sum of variances from $166,245$ to $159,809$ by $3.8$ percent. This decrease corresponds to a $2$ percent decrease in the standard deviation. Considering the age and the number of visits for $t = 7$, the sum of difference decreases from $169,540$ to $158,110$ by $6.7$ percent. The corresponding decrease of the standard deviation is $3,4$ percent.

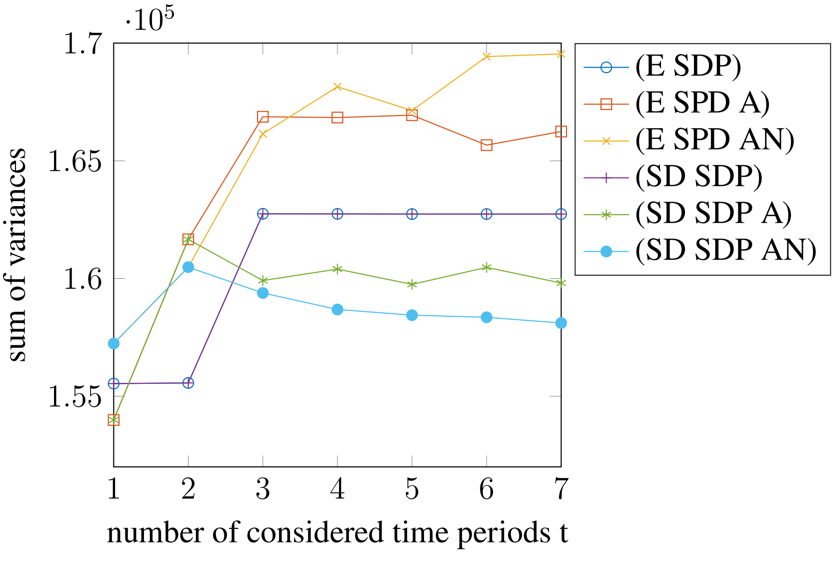

**Figure 6.12:** Sum of variances for the 6 ILPs and different values of $t$

The improvements in the variance are relatively small in this numerical experiment. However, note that in the results, the sum of differences for both ILP types, e.g., for ILP (E SPD AN) and ILP (SD SPD AN), differ by at most $30$. The values for the ILPs that minimize the expected deviation between workload and capacities can be seen in Section 6.5.3 in Figure 6.10. Bigger improvements can probably be achieved when we allow for bigger bounds $s$. In summary, we see that it is reasonable to minimize variance if at least one new patient attribute is considered.

### 6.5.5 Defining capacities

Let us consider a modification of the scenario defined in Section 6.5. Suppose we start with a panel having $0.8$ times the workload of the $2017$ panel and aim to lower the capacity to $5,500$ for two consecutive periods and increase the capacity in the subsequent periods to $8,500$ and keep it there. Again we use a constant exterior demand equal to the accepted new patients in $2017$. We consider a total of $t = 10$ periods. We use the ILPs (C), (C A), and (C AN) to define the capacities. Here, we set $g$, i.e., the allowed deviation between workload and capacity, to $300$. Solving the three ILPs leads to the objective values $4$, $4$, and $5$, indicating that using a detailed new patient classification allows conducting the capacity changes in fewer periods given the bound on the workload capacity deviation. The resulting capacity definitions for the $10$ periods are depicted in Figure 6.13.

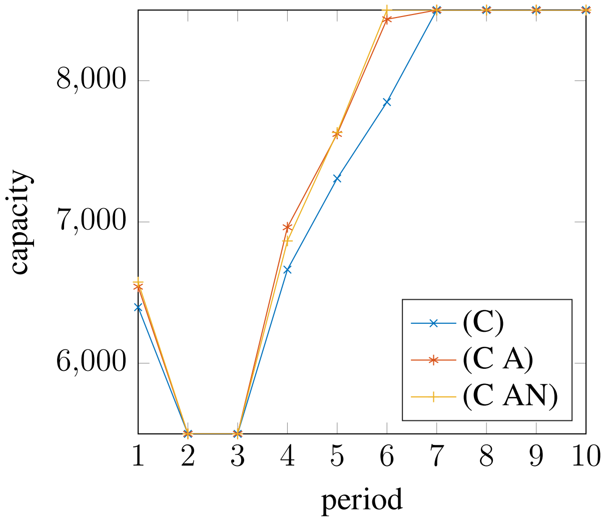

**Figure 6.13:** Capacities for the $10$ periods as defined by the three ILPs

## 6.6 Conclusion and outlook

In this chapter, we investigated theoretical extensions to the panel management ILPs from Chapter 5, presented a second real-world appointment data set, and conducted further numerical experiments.

We learned that the panel management ILPs can be combined with an indirect queueing model to help define the capacities in the panel management ILPs based on the actual physician availability and capture the possible influence of the capacity decision on the appointment demand from panel patients.

We saw that the basic modeling idea, i.e., defining the expected deviation between workload and capacity and its variance via linear functions using age and number of visits as core panel patient attributes, can easily be applied and extended to many situations. Here, we showcased ILPs that consider different resources with their own capacity restrictions simultaneously as well as separated but potentially overlapping panels for different physicians. We further presented an ILP that decides on the capacity values.

The real-world appointment data set from a urology group practice includes information on gender and allows for a patient-physician assignment. The data set is not ideal for showcasing the usefulness of the models because we observe few visits per panel patient per period, and we see that a random new patient's expected workload is not increasing in the first periods after admission to the panel. Hence, if the capacities do not change significantly, there is no need to look several periods ahead when deciding on accepting new patient demand.

However, in a scenario with significant changes in the capacity, the ILPs help lower the deviation between workload and capacity. The more periods considered in the optimization and the more detailed the (new) patient classification, the better. Here, especially, we saw that including the patient attribute gender for panel patients and new patients yields better results. We could further show that also the ILPs considering several resources and physicians, show the expected improvement, where we especially see that the additional decision on the patient-physician assignment provides additional improvement potential.

For the models with one panel and one resource, we also conducted a simulation and showed once again that the benefit of applying the ILPs in an uncertain environment remains. We further saw that the workload's variance decreases alongside the deviation between workload and capacity for an increasing number of considered periods in the optimization. Also, the number of models that need to be solved does not increase too much for model variances considering many new patient classes. Hence, the simulation can still be conducted in a reasonable amount of time.

Our numerical experiments further indicate that there is potential to minimize variance after minimizing the expected deviation if considering at least one new patient attributed and several periods in the optimization. However, here we have to be aware that some age and visit category combinations show a low variance due to few observations, which probably influences the results. Finally, we saw that a detailed patient classification helps to undertake a faster capacity change.

For future research, one could develop even more ILPs and investigate in more detail how to deal with multiple objectives, e.g., through using a weighted sum approach or lexicographic optimization first using a min-max approach and then a min-sum approach.

Further, we wish to gather even more appointment data sets. Here, we would like to test the combination of panel management ILPs with indirect queueing models. Further, this would allow us to develop theoretical transition distributions instead of relying on the empirical distributions and, therefore, help investigate the variance optimization further. Having a data set over a more extended period, e.g., $10$ years, would also make it possible to validate the assumption that we can predict the visit behavior of new patients based on the historical visit behavior of panel patients.

# 7 Conclusion and outlook

In this thesis focusing on tactical demand and capacity management for medical practices, we first investigated the whole spectrum of decisions that need to be taken by a practice manager when opening and then running a medical practice. Here, we put particular emphasis on the need for relevant data to define model parameters to be able to apply Operations Research models. We realized that there is still a considerable gap between the academic, theoretical, and the practical world. One, necessary data to define model parameters is often not available or only in an aggregated format containing errors and uncertainties. Two, modelers often produce complex models without considering their practical application and the provider's actual goals or data availability. Further, often a blueprint of how to transform raw data into model parameters is missing and information on when and how to update the parameters. Therefore, to strive for more application of models in practice, we need to educate healthcare managers about the potential of models and their need for data. Further, we as modelers need to focus on the providers' actual needs, data availability, and parameters definition based on data.

Next, an analytical queueing model was built with the motivation to investigate relationships between the appointment request rate, the physician's capacity, and the resulting distribution of indirect waiting times. The main mathematical contribution is the flexibility of the queueing model because it can represent many different settings by integrating queue length-dependent parameters such as the appointment request rate, the no-show probability, and the rescheduling probability. We further extended the model by considering queue length-dependent and random service times. The queueing model delivers exact queue length distributions for a linear queue length-dependent total arrival rate together with a queue length-dependent rescheduling probability and random service times. Moreover, we showed how the appointment request rate of a physician with panel patients could be modeled even including different individual arrival rates or a mix of panel and non-panel patients.

We showed that it is enough to study different total arrival rate functions to cover the main modeling effects for practical purposes. However, due to the assumption that patients always book the next available appointment, it is difficult to define model parameters from real-world data where this assumption often does not hold. Still, using synthetic but realistic parameters, we were able to conduct a vast number of numerical experiments to study the general behavior of the queueing system.

Then, we presented a framework for building integer linear programs (ILPs) to decide on the intake of new patients into a physician panel over time. The main objective was to minimize the sum of deviations between the expected panel workload and the physician's capacity over all considered periods. We also explained how those ILPs could be applied to decide on accepting or rejecting patient requests to enter the panel online. Our main contributions are considering the future panel development while making these decisions and an algorithm to define upper bounds on the number of patients in a patient class to be accepted in a period by solving the ILP several times with different demand inputs in the first period. Our numerical experiments show that we can significantly lower the deviation between workload and capacity when considering several future periods instead of one in the optimization. The deviation can be decreased further by using a detailed classification of new patients. In contrast to the queueing model, we applied the most straightforward approach using parameters that can easily be defined on real-world data, even adjustable to the level of data detail. We found that the improvement when using a more sophisticated model that uses the same data input is likely to be small. Hence, we believe that our framework can be applied and deliver benefits in practice.

Finally, we commented on how both modeling approaches can be combined. In fact, the queueing model can be used to define the capacities in the panel management ILPs based on the average daily appointment capacity, the booking horizon, the types of functions for the appointment requests, and the rescheduling probability. Here, we can even integrate the possible influence of the capacity decision on the appointment demand from panel patients.

For future research, we would like to investigate further how to support planning for integrated care. For a medical practice setting with physicians with different specialties, we could extend our ideas on panel management to include patient classes for co-morbidities. Then we could define the corresponding needed capacities of the different specialties involved per patient class. In the ILP, we could once again manage the intake of new patients and necessary capacities per specialty.

Another idea to extend the panel management programs is to split a period into several parts and match capacities for those. In this case, we still consider age and visit transitions from one period to another but include a distribution of visits over the parts of a period for every patient class. This way, we can, for example, account for seasonal effects.

Concerning the queueing model, we would like to find real-world data suitable for defining the model parameters. We would also like to test the combined approach of panel management with the queueing model. In any case, we want to investigate the applicability of the queueing model in more realistic settings, such as for seasonal demand oscillations, through producing appointment data via a simulation that can then be used to define model parameters.

Our research so far has focused on single medical practices. For future research, we would like to investigate how the models could support the planning of health care provision in a whole region, e.g., supporting the decision on how many practices of a given type are needed to serve a population and where they should be located.

Moreover, we should continue to include the influence of provider decisions on the demand composition and volume into models, similar to our ideas of combining the two presented model approaches.

Further, we want to explore other application areas for our models besides health care. Here, service providers where people tend to go to more than once come to mind, for example, cosmetics, hairdresser, or citizens' service. Another application could be in workforce planning, e.g., when we decide on hiring new personal, the future development of the workforce should be taken into account. We could investigate actual waiting lines for the queueing model, where the total arrival rate depends on the queue length because customers are discouraged from joining an already long queue.

We also would like to research other approaches for tactical planning in combination with operational planning for medical practices. Here, we could investigate the tradeoff between good template schedules and access to care. Assuming a physician with panel patients, we can collect data on an individual patient level and use a detailed classification of patients to predict their relevant parameters for scheduling. The adverse effects of using many patient classes in template scheduling on the actual assignment of patients to time slots could be attenuated through using several overlapping template schedules for one session. The decision on which template schedule to use should then be taken during the session and depended on the actual patient demand.

# A Appendix

## A.1 Further simulation results for the queueing model

### A.1.1 Realistic simulation for a constant individual arrival rate

Using $15$ replications, a total number of $100,000$ simulated patients with an initial transient phase of $20,000$ patients for $N = 2300$, we obtain $\widehat{W} = 0.32$ and a $95\%$ confidence interval of: $[0.31, 0.33]$. Again, the results are not far off from the analytically determined expected indirect waiting time of $0.35$ days. The same holds for the indirect waiting time distributions.

For $N = 2540$, we use $15$ replications, an initial lequeue length of $200$, a total number of $200,000$ simulated patients with an initial transient phase of $20,000$ patients and obtain $\widehat{W} = 10.17$ and a $95\%$ confidence interval of: $[10.01, 10.34]$. The estimated expected indirect waiting time and the analytically determined expected indirect waiting time are close to each other. However, the simulated queue length distribution, which can be seen in Figure A.1 shows a smaller variance compared to the analytically determined one.

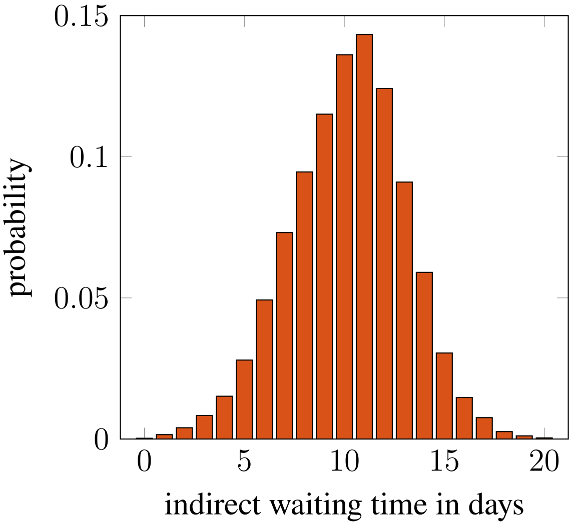

**Figure A.1:** Simulated indirect waiting time distribution for $N = 2540$ for a constant individual arrival rate

For $N = 2800$, we use $15$ replications, an initial queue length of $400$, a total number of $100,000$ simulated patients with an initial transient phase of $20,000$ patients and obtain $\widehat{W} = 19.63$ and a $95\%$ confidence interval of: $[19.62, 19.63]$. The estimated expected indirect waiting time and the estimated indirect waiting time distribution are very similar to the analytical values.

### A.1.2 Realistic simulation for a non-constant individual arrival rate

We limit ourselves to investigate the panel sizes $N = 2300$ and $2360$. Additionally, we implement another difference to the queueing model. Patients change their individual appointment request rate based on the queue length. In the queueing model, this happens immediately. Hence we assume that all patients know about the state of the queueing system at all times. In the simulation, we assume - more realistically - that patients only change their individual appointment request rate when they leave the queue after service, i.e., at points in time when they can observe the state of the indirect queue.

For $N = 2300$, using $15$ replications, a total number of $100,000$ simulated patients with an initial transient phase of $20,000$ patients we obtain $\widehat{W} = 0.34$ and a $95\%$ confidence interval of: $[0.33, 0.35]$.

For $N = 2360$, using $15$ replications, an initial queue length of $400$, a total number of $100,000$ simulated patients with an initial transient phase of $30,000$ patients we obtain $\widehat{W} = 19.58$ and a $95\%$ confidence interval of: $[19.577, 19.583]$.

### A.1.3 Results for rescheduling and rejected patients for a non-constant individual arrival rate

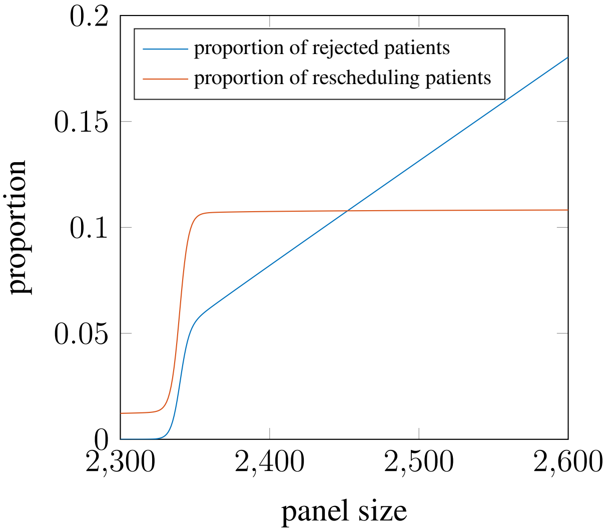

**Figure A.2:** Proportion of rescheduling and rejected patients for a non-constant arrival rate

In Figure A.2, we show the proportion of rescheduling and rejected patients for a non-constant individual arrival rate for different panel sizes. We see a similar behavior of the curves compared to the case of a constant total arrival rate as depicted in Figure 4.16.

### A.1.4 Results for the same-day appointment probability and the physician utilization for a constant individual arrival rate

For a constant individual arrival $\eta$ rate with a total arrival rate $\lambda(k) = \eta(N - k)$, we show the results for the same-day appointment probability and the physician utilization in Figures A.3 and A.4. Here, when we vary the capacity, we keep a fixed panel size of $N = 2540$.

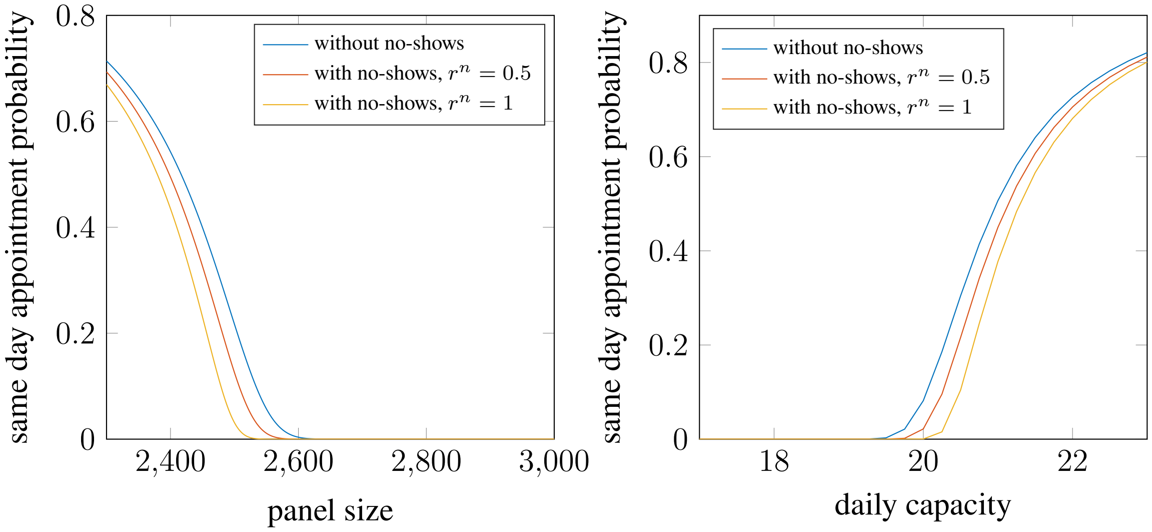

**Figure A.3:** Same day appointment probability dependent on the panel size and on the daily capacity (with $N = 2540$) for a constant individual arrival rate

In comparison to the results for a constant total arrival rate in Figure 4.18, we notice that the same-day appointment probability decreases less steeply for larger panel sizes. For the physician utilization, we notice a similar behavior of the curves to Figure 4.19. However, the curves lie closer together, indicating a lesser influence of no-shows on the physician utilization value. For $r^n = 0$, the maximal utilization value of $0.9729$ is reached for a panel size of $2529$ and an indirect waiting time of $2.09$ days. For $r^n = 0.5$, the maximal utilization of $0.9714$ is reached for a panel size of $2500$ and an indirect waiting of $2.17$ days. Moreover, for $r^n = 1$, the maximal utilization of $0.9693$ is reached for a panel size of $2468$ and an expected indirect waiting time of $2.25$ days.

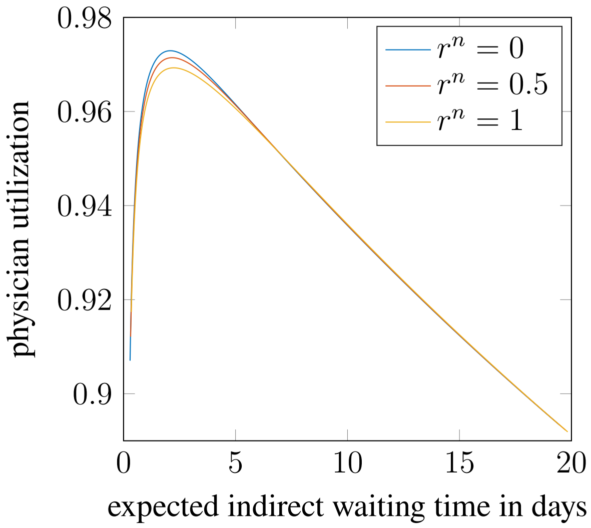

**Figure A.4:** Physician utilization dependent on the indirect waiting time in days for a constant individual arrival rate

### A.1.5 Results for the same-day appointment probability and physician utilization for a non-constant individual arrival rate

We plot the same-day appointment probability as well as the physician utilization for the case of a non-constant individual arrival rate with a total arrival rate $\lambda(k) = \eta(k)(N - k)$. Our results can be seen in Figures A.5 and A.6. When we vary the capacity for the same-day appointment probability, we keep a fixed panel size of $N = 2340$. The same-day appointment probability drops even faster when compared to Figure 4.18 with increasing panel sizes.

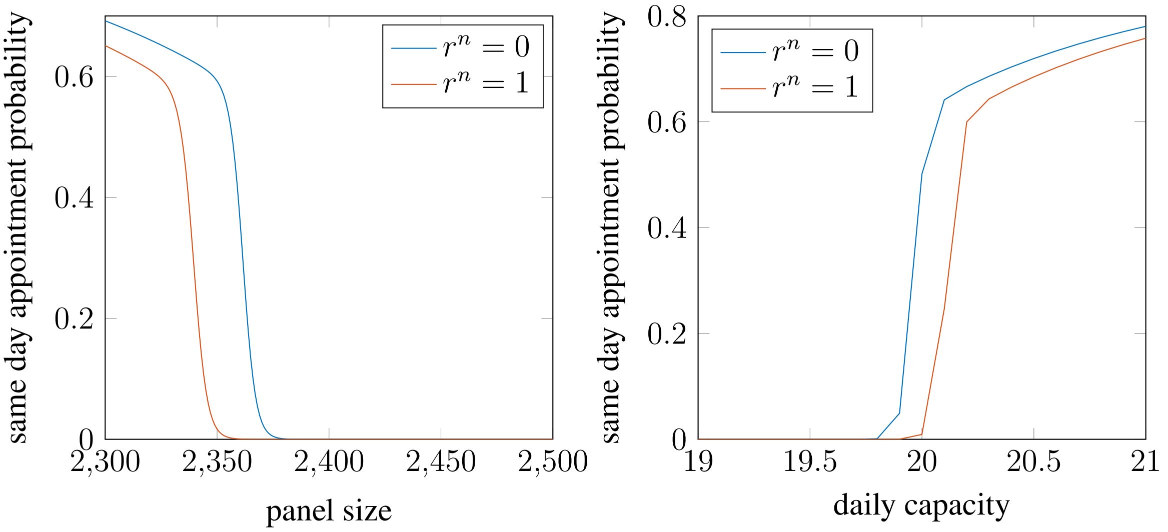

**Figure A.5:** Same day appointment probability dependent on the panel size and on the daily capacity (with $N = 2340$) for a non-constant individual arrival rate

The physician utilization reaches only a value of around $0.93$ in the maximum. Further, there is only a slight difference between the curves for $r^n = 0$ and $r^n = 1$. For $r^n = 0$, the maximal utilization value of $0.93$ is reached for a panel size of $2352$ and an indirect waiting time of $1.09$ days, and for $r^n = 1$ the maximal utilization of $0.9305$ is reached for a panel size of $2330$ and an expected indirect waiting of $1.21$ days.

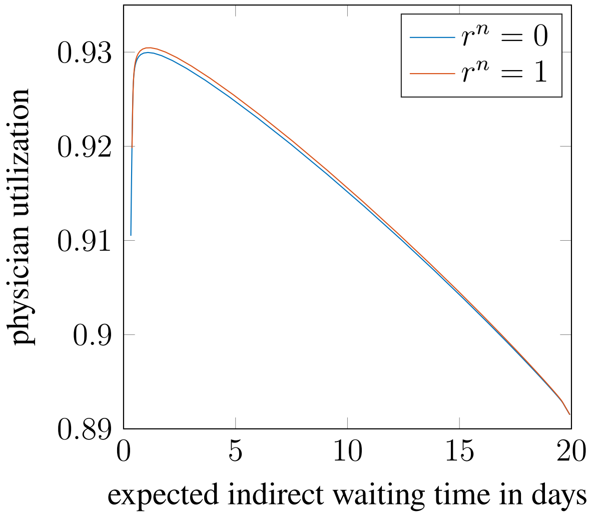

**Figure A.6:** Physician utilization dependent on the indirect waiting time in days for a non-constant individual arrival rate

### A.1.6 Sensitivity analysis for a non-constant individual arrival rate

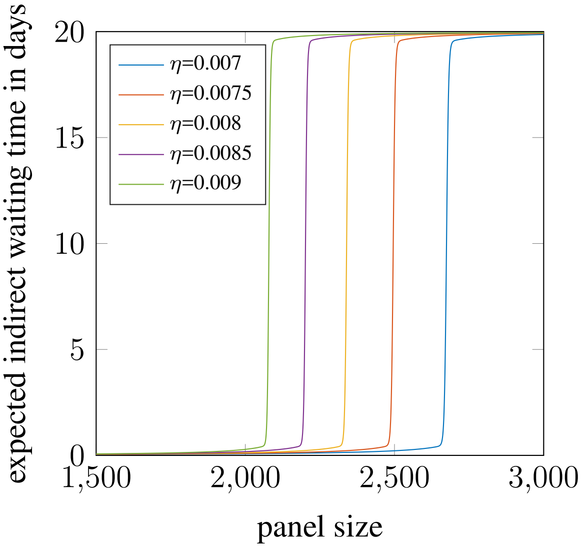

**Figure A.7:** Expected indirect waiting time in days dependent on the panel size for different values of the individual arrival rate for a non-constant individual arrival rate

We plot the expected indirect waiting time in days dependent on the panel size for different values of $\eta = \delta$ for a non-constant individual arrival rate in Figure A.7. We see very similar results as for a constant total arrival rate in Figure 4.22.

### A.1.7 Results for using variable appointment offerings for a non-constant individual arrival time

For a constant individual arrival rate, we investigate if we can alter the indirect waiting time distribution for $N = 2340$ using a variable number of offered appointment slots per day. If we assume a constant number of appointment slots offered per day, namely $20$, we find an expected indirect waiting of $10.7$ days. Let us assume that the physician offers $19$ time slots per day as long as the queue length is less than $50$. If the queue length lies between $50$ and $150$, the physician offers $20$ slots, and for a queue length of $150$ or more, the physician offers $21$ slots. Putting this into the model, we find an expected indirect waiting time of $1.9$ days. The expected number of appointment slots offered per day is $19.27$, which means that the physician works less on average. At the same time, patients experience less indirect waiting time, similar to the case of a constant total arrival rate. Figure A.8 shows the indirect waiting time distribution. Again, in comparison to Figure 4.13, we obtain a queue length distribution with a much smaller variance.

We validate our results via simulation. Using $15$ replications, an initial queue length of $40$, a total number of $200,000$ simulated patients with an initial transient phase of $30,000$ patients we obtain $\widehat{W} = 1.74$ and a $95\%$ confidence interval of: $[1.71, 1.77]$.

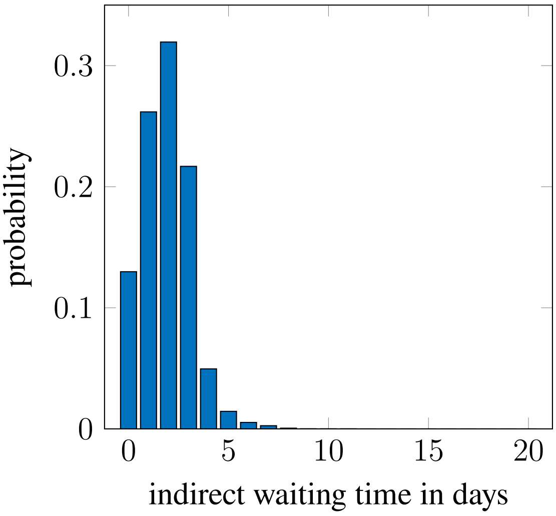

**Figure A.8:** Indirect waiting time distribution for $N = 2340$ with a variable number of offered appointment slots per day for a non-constant individual arrival rate

# List of figures

1.1 Patient request modes . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Schematic interaction between panel/demand management, capacity management and the appointment backlog queue . . . . . . . . . . . . . . . . . . . 6

2.1 Illustration of planning problems for ambulatory care practices . . . . . . . . . . . 14

2.2 Possible patient demand behavior . . . . . . . . . . . . . . . . . . . . . . . . . . 24

4.1 Average number of patients seen per day dependent on the number of appointments booked in the physician schedule . . . . . . . . . . . . . . . . . . . 91

4.2 Number of appointment bookings per day including and excluding rescheduling patients dependent on the number of appointments booked in the physician schedule and the distribution of indirect waiting time in days . . . . . . . . . . . . 92

4.3 Probability of being a no-show or a rescheduling patient given the indirect waiting time in days . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92

4.4 No-show probability dependent on the queue length . . . . . . . . . . . . . . . . 100

4.5 Runtime of the Java implementation dependent on the queue capacity for N=2344 with a constant total arrival rate . . . . . . . . . . . . . . . . . . . . . . 102

4.6 Queue length distributions at departure times calculated using the Java and the Matlab implementation for a constant total arrival rate . . . . . . . . . . . . . . . 103

4.7 Moving average of $\overline{L}_{\bullet m}$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104

4.8 Calculated and simulated queue length distributions for $N = 2300$, $N = 2344$ and $N = 2400$ for a constant total arrival rate . . . . . . . . . . . . . . . . . . 105

4.9 Calculated and simulated indirect waiting time distributions for panel sizes $N = 2300$ and $N = 2400$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107

4.10 Section of the queue length distributions for panel size $N = 2550$ using the binomial and the Poisson distribution for arrivals during a time interval for a constant individual arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . . 108

4.11 Queue length and indirect waiting time distributions for $N = 2300$, $N = 2344$ and $N = 2400$ for a constant total arrival rate . . . . . . . . . . . . . . . . . . 109

4.12 Indirect waiting time distributions for $N = 2300$, $N = 2540$ and $N = 2800$ for a constant individual arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . 109

4.13 Indirect waiting time distributions for $N = 2300$, $N = 2340$ and $N = 2360$ for a non-constant individual arrival rate . . . . . . . . . . . . . . . . . . . . . . . . 110

4.14 Comparison of the different queue length distributions for panel size $N = 2345$ for a constant total arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . . 111

4.15 Comparison of the different queue length distributions for panel size $N = 2550$ for a constant individual arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . 112
4.16 Proportion of rescheduling and rejected patients for a constant total and individual arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
4.17 Expected indirect waiting time in days dependent on the panel size and the daily capacity (with $N = 2344$) for a constant total arrival rate . . . . . . . . . . . 114
4.18 Same-day appointment probability dependent on the panel size and on the daily capacity (with $N = 2344$) for a constant total arrival rate . . . . . . . . . . . . . 115
4.19 Physician utilization dependent on the indirect waiting time in days for a constant total arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116
4.20 Expected indirect waiting time in days dependent on the panel size and on the daily capacity (with $N = 2540$) for a constant individual arrival rate . . . . . . . . 116
4.21 Expected queue length dependent on the panel size and on the daily capacity (with $N = 2340$) for a non-constant individual arrival rate . . . . . . . . . . . . 117
4.22 Expected indirect waiting time in days dependent on the panel size for different values of the individual arrival rate for a constant total and individual arrival rate . 118
4.23 Panel size and rate of rejected patients dependent on the queue capacity such that the expected indirect waiting time remains approximately constant for a constant total arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118
4.24 Panel size and rate of rejected patients dependent on the queue capacity such that the expected indirect waiting time remains approximately constant for a constant individual arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
4.25 Comparison of the effective arrival rate for $N = 2344$ with the function $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ and the resulting queue lengths distributions using the original model, the effective arrival rate as the total arrival rate, and $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ as the total arrival rate for a constant total arrival rate . . . . . . . . . . . . . . . . 120
4.26 Comparison of the effective arrival rate for $N = 2540$ with the function $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ and the resulting queue lengths distributions using the original model, the effective arrival rate as the total arrival rate, a cut version of the effective arrival rate as the total arrival rate and $\lambda(\cdot) + \frac{1}{T}\rho(\cdot - 1)$ as the total arrival rate for a constant individual arrival rate . . . . . . . . . . . . . . . . . . . 121
4.27 Comparison of the queue length distributions for different linear total arrival rates such that $200m + n = 20$ . . . . . . . . . . . . . . . . . . . . . . . . . . 122
4.28 Comparison of the queue length distributions for different linear total arrival rates such that $m = -0.008$ . . . . . . . . . . . . . . . . . . . . . . . . . . 122
4.29 Comparison of the expected indirect waiting time dependent on the panel size for different arrival distributions for a constant total arrival rate . . . . . . . . . . . 123
4.30 Analytical and simulated indirect waiting time distribution for N=2344 with a variable number of offered slots per day for a constant total arrival rate . . . . . . . 124
4.31 Analytical and simulated indirect waiting time distribution for $N = 2540$ and a variable number of appointment slots offered per day for a constant individual arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125

4.32 Analytical and simulated indirect waiting time distribution for $N = 2540$ and a random number of appointment slots offered per day for a constant individual arrival rate . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126
4.33 Comparison of the queue length distributions for two subgroups $N_1 = N_2 = 1270$ with individual arrival rates $\eta_1 + \eta_2 = 0.008$ . . . . . . . . . . . 127
4.34 Analytical and simulated indirect waiting time distribution for two subgroups $N_1 = N_2 = 1270$ with individual arrival rates $\eta_1 = 0.006$ and $\eta_2 = 0.01$ . . . . . . 128
4.35 Comparison of the queue length distributions for different panel sizes $N$ and arrival rates of non-panel patients $r^{np}$ for a constant individual arrival rate for panel patients . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 129
4.36 Analytical and simulated indirect waiting time distribution for $N = 2300$ and $r^{np} = 1.6$ for a constant individual arrival rate . . . . . . . . . . . . . . . . . . 129
4.37 Panel size $N_1$ dependent on the daily capacity $\mu_1$ and proportion of rejected patients dependent on the panel size $N_1$ for a constant individual arrival rate . . . . 130

5.1 Number of panel patients by age and by number of visits per year in 2014 . . . . . 152
5.2 Sum of differences for the three ILPs and different values of $t$ in Scenario 1 . . . . 156
5.3 Sum of differences for the three ILPs and different values of $t$ in Scenario 2 . . . . 157
5.4 Sum of differences for the three ILPs and different values of $t$ in Scenario 3 . . . . 157
5.5 Simulation flowchart . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 159
5.6 Simulated expected sum of differences for the three ILPs and different values of $t$ in Scenario 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 161
5.7 Simulated expected sum of differences for the three ILPs and different values of $t$ in Scenario 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 163
5.8 Simulated expected sum of differences for the three ILPs and different values of $t$ in Scenario 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 164

6.1 Number of panel patients per age category in 2017 . . . . . . . . . . . . . . . . . 176
6.2 Number of new patient per age category admitted in 2017 . . . . . . . . . . . . . 177
6.3 Average number of visits per age category . . . . . . . . . . . . . . . . . . . . . 178
6.4 Sum of differences for the 8 ILPs and different values of $t$ . . . . . . . . . . . . . 180
6.5 Expected sum of differences for the 8 ILPs and different values of $t$ . . . . . . . . 180
6.6 Standard deviation of the expected sum of workloads for the 8 ILPs and different values of $t$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 181
6.7 Average number of models solved for the 6 ILPs and different values of $t$ . . . . . . 182
6.8 Expected number of visits per visit category and resource . . . . . . . . . . . . . 182
6.9 Sum of differences for the 3 ILPs and different values of $t$ . . . . . . . . . . . . . 183
6.10 Sum of differences for the three ILPs and different values of $t$ . . . . . . . . . . . 184
6.11 Sum of differences for the three ILPs and different values of $t$ . . . . . . . . . . . 185
6.12 Sum of variances for the 6 ILPs and different values of $t$ . . . . . . . . . . . . . 186
6.13 Capacities for the 10 periods as defined by the three ILPs . . . . . . . . . . . . . 187

A.1 Simulated indirect waiting time distribution for $N = 2540$ for a constant individual arrival rate . . . 195
A.2 Proportion of rescheduling and rejected patients for a non-constant arrival rate . . . 196
A.3 Same day appointment probability dependent on the panel size and on the daily capacity (with $N = 2540$) for a constant individual arrival rate . . . 197
A.4 Physician utilization dependent on the indirect waiting time in days for a constant individual arrival rate . . . 198
A.5 Same day appointment probability dependent on the panel size and on the daily capacity (with $N = 2340$) for a non-constant individual arrival rate . . . 198
A.6 Physician utilization dependent on the indirect waiting time in days for a non-constant individual arrival rate . . . 199
A.7 Expected indirect waiting time in days dependent on the panel size for different values of the individual arrival rate for a non-constant individual arrival rate . . . 199
A.8 Indirect waiting time distribution for $N = 2340$ with a variable number of offered appointment slots per day for a non-constant individual arrival rate . . . 200

# List of tables

2.1 Available information for booked appointments in the four data sets . . . . . . . . 32

3.1 Comparison of the queueing models from literature to model indirect waiting time . 42
3.2 Basic model notation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45
3.3 Model features and extensions . . . . . . . . . . . . . . . . . . . . . . . . . 46

4.1 Model features and extensions . . . . . . . . . . . . . . . . . . . . . . . . . 99
4.2 Basic parameter settings from Green and Savin (2008) . . . . . . . . . . . . . . . 101

5.1 Overview of the different models . . . . . . . . . . . . . . . . . . . . . . . . 145
5.2 Basic model notation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 147
5.3 Additional model notation . . . . . . . . . . . . . . . . . . . . . . . . . . . 148

# Bibliography

Abernathy, W. J. & Hershey, J. C. (1972). A spatial-allocation model for regional health-services planning. *Operations Research*, *20*(3), 629–642, https://doi.org/10.1287/opre.20.3.629.

Abouee-Mehrizi, H. & Baron, O. (2016). State-dependent M/G/1 queueing systems. *Queueing Systems*, *82*, 121–148, https://doi.org/10.1007/s11134-015-9461-y.

Ahmadi-Javid, A., Jalali, Z., & Klassen, K. J. (2017a). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, *258*(1), 3–34, https://doi.org/10.1016/j.ejor.2016.06.064.

Ahmadi-Javid, A. & Ramshe, N. (2020). A stochastic location model for designing primary healthcare networks integrated with workforce cross-training. *Operations Research for Health Care*, *24*, 100226, https://doi.org/10.1016/j.orhc.2019.100226.

Ahmadi-Javid, A., Seyedi, P., & Syam, S. S. (2017b). A survey of healthcare facility location. *Computers and Operations Research*, *79*, 223–263, https://doi.org/10.1016/j.cor.2016.05.018.

Arnolds, I. & Nickel, S. (2013). Multi-period layout planning for hospital wards. *Socio-Economic Planning Sciences*, *47*(3), 220–237, https://doi.org/10.1016/j.seps.2013.02.001.

Arnolds, I. & Nickel, S. (2015). Layout planning problems in health care. In M. V. Eiselt H. (Ed.), *Applications of location analysis* (pp. 109–152). Springer, https://doi.org/10.1007/978-3-319-20282-2_5.

Association of Statutory Health Insurance Physicians of Rhineland-Palatinate (2016). *Versorgungsatlas Rheinland-Pfalz 2016 (Atlas of care 2016)*. Technical report, Association of Statutory Health Insurance Physicians of Rhineland-Palatinate, Mainz, https://kv-rlp.de/fileadmin/user_upload/Downloads/Institution/Engagement/Versorgungsforschung/KVRLP_Versorgungsatlas_2016.pdf.

Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., & Stahl, J. (2010). Improving clinical access and continuity through physician panel redesign. *Journal of general internal medicine*, *25*(10), 1109–15, https://doi.org/10.1007/s11606-010-1417-7.

Benitez, G. B., Da Silveira, G. J. C., & Fogliatto, F. S. (2019). Layout Planning in Healthcare Facilities: A Systematic Review. *Health Environments Research and Design Journal*, *12*(3), 31–44, https://doi.org/10.1177/1937586719855336.

Burke, E. K., De Causmaecker, P., Berghe, G. V., & Van Landeghem, H. (2004). The State of the Art of Nurse Rostering. *Journal of Scheduling*, *7*(6), 441–499, https://doi.org/10.1023/B:JOSH.0000046076.75950.0b.

Büsing, C., Comis, M., Schmidt, E., & Streicher, M. (2021). Robust strategic planning for mobile medical units with steerable and unsteerable demands. *European Journal of Operational Research*, *295*(1), 34–50, https://doi.org/10.1016/j.ejor.2021.02.037.

Cayirli, T. & Veral, E. (2009). Outpatient Scheduling in Health Care: A Review of Literature. *Production and Operations Management*, *12*(4), 519–549, https://doi.org/10.1111/j.1937-5956.2003.tb00218.x.

Cheang, B., Li, H., Lim, A., & Rodrigues, B. (2003). Nurse rostering problems—a bibliographic survey. *European Journal of Operational Research*, *151*(3), 447–460, https://doi.org/10.1016/S0377-2217(03)00021-3.

Cho, D. & Cattani, K. (2018). The Patient Patient: The Performance of Traditional versus Open-Access Scheduling Policies. *Decision Sciences*, *50*(4), 756–785, https://doi.org/10.1111/deci.12351.

Comis, M., Cleophas, C., & Büsing, C. (2021). Patients, primary care, and policy: Agent-based simulation modeling for health care decision support. *Health Care Management Science*, https://doi.org/10.1007/s10729-021-09556-2.

Dantas, L. F., Fleck, J. L., Cyrino Oliveira, F. L., & Hamacher, S. (2018). No-shows in appointment scheduling – a systematic literature review. *Health Policy*, *122*(4), 412–421, https://doi.org/10.1016/j.healthpol.2018.02.002.

Daskin, M. S. & Dean, L. K. (2005). Location of Health Care Facilities. In *Operations Research and Health Care* (pp. 43–76). Boston: Kluwer Academic Publishers, https://doi.org/10.1007/1-4020-8066-2_3.

De Bruecker, P., Van den Bergh, J., Beliën, J., & Demeulemeester, E. (2015). Workforce planning incorporating skills: State of the art. *European Journal of Operational Research*, *243*(1), 1–16, https://doi.org/10.1016/j.ejor.2014.10.038.

Dreiher, J., Comaneshter, D. S., Rosenbluth, Y., Battat, E., Bitterman, H., & Cohen, A. D. (2012). The association between continuity of care in the community and health outcomes:

A population-based study. *Israel Journal of Health Policy Research*, *1*(1), 21, https://doi.org/10.1186/2045-4015-1-21.

Drezner, Z. & Hamacher, H. W., Eds. (2002). *Facility Location Application and Theory*. Berlin, Heidelberg: Springer.

Erhard, M., Schoenfelder, J., Fügener, A., & Brunner, J. O. (2018). State of the art in physician scheduling. *European Journal of Operational Research*, *265*(1), 1–18, https://doi.org/10.1016/j.ejor.2017.06.037.

Ernst, A., Jiang, H., Krishnamoorthy, M., & Sier, D. (2004). Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, *153*(1), 3–27, https://doi.org/10.1016/S0377-2217(03)00095-X.

Farahani, R. Z., SteadieSeifi, M., & Asgari, N. (2010). Multiple criteria facility location problems: A survey. *Applied Mathematical Modelling*, *34*(7), 1689 – 1709, https://doi.org/10.1016/j.apm.2009.10.005.

Feldman, J., Liu, N., Topaloglu, H., & Ziya, S. (2014). Appointment Scheduling Under Patient Preference and No-Show Behavior. *Operations Research*, *62*(4), 794 – 811, https://doi.org/10.1287/opre.2014.1286.

Gallucci, G., Swartz, W., & Hackerman, F. (2005). Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric services (Washington, D.C.)*, *56*(3), 344–6, https://doi.org/10.1176/appi.ps.56.3.344.

Garcia, J. M., Brun, O., & Gauchard, D. (2002). Transient analytical solution of M/D/1/N queues. *Journal of Applied Probability*, *39*(4), 853–864, https://doi.org/10.1239/jap/1037816024.

German Federal Ministry of Health (2021a). Hausarztsystem (family doctor system). Online: Accessed 2021-07-05, https://www.bundesgesundheitsministerium.de/hausarztsystem.html.

German Federal Ministry of Health (2021b). Medizinische Versorungszentren (medical care centers). Online: Accessed 2021-07-05, https://www.bundesgesundheitsministerium.de/themen/krankenversicherung/ambulante-versorgung/medizinische-versorgungszentren.html.

Graber-Naidich, A., Carter, M. W., & Verter, V. (2015). Primary care network development: the regulator's perspective. *Journal of the Operational Research Society*, *66*(9), 1519–1532, https://doi.org/10.1057/jors.2014.119.

Green, L. V. & Savin, S. (2008). Reducing Delays for Medical Appointments: A Queueing Approach. *Operations Research*, *56*(6), 1526–1538, https://doi.org/10.1287/opre.1080.0575.

Green, L. V., Savin, S., & Murray, M. (2007). Providing timely access to care: What is the right patient panel size? *Joint Commission Journal on Quality and Patient Safety*, *33*(4), 211–218, https://doi.org/10.1016/S1553-7250(07)33025-0.

Gröne, O., Garcia-Barbero, M., & WHO European Office for Integrated Health Care Services (2001). Integrated care: a position paper of the WHO European Office for Integrated Health Care Services. *International journal of integrated care*, *1*(e21) https://pubmed.ncbi.nlm.nih.gov/16896400/.

Gu, W., Wang, X., & McGregor, S. E. (2010). Optimization of preventive health care facility locations. *International Journal of Health Geographics*, *9*(1), 17, https://doi.org/10.1186/1476-072X-9-17.

Guagliardo, M. F. (2004). Spatial accessibility of primary care: concepts, methods and challenges. *Int J Health Geogr*, *3*(1), 3, https://doi.org/10.1186/1476-072x-3-3.

Güneş, E. D. & Nickel, S. (2015). Location Problems in Healthcare. In G. Laporte, S. Nickel, & F. da Gama (Eds.), *Location Science* (pp. 555–579). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-319-13111-5_21.

Güneş, E. D., Yaman, H., Çekyay, B., & Verter, V. (2014). Matching patient and physician preferences in designing a primary care facility network. *Journal of the Operational Research Society*, *65*(4), 483–496, https://doi.org/10.1057/jors.2012.71.

Gupta, D. & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, *40*(9), 800–819, https://doi.org/10.1080/07408170802165880.

Helber, S., Böhme, D., Oucherif, F., Lagershausen, S., & Kasper, S. (2016). A hierarchical facility layout planning approach for large and complex hospitals. *Flexible Services and Manufacturing Journal*, *28*, 5–29, https://doi.org/10.1007/s10696-015-9214-6.

Hillsman, E. L. (1980). *Multiobjective location planning for primary medical services in rural Iowa*. Technical report, Oak Ridge National Laboratory, USA.

Hodgson, M. J., Laporte, G., & Semet, F. (1998). A Covering Tour Model for Planning Mobile Health Care Facilities in SuhumDistrict, Ghama. *Journal of Regional Science*, *38*(4), 621–638, https://doi.org/10.1111/0022-4146.00113.

Hulshof, P. J. H., Kortbeek, N., Boucherie, R. J., Hans, E. W., & Bakker, P. J. M. (2012). Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems*, *1*(2), 129–175, `https://doi.org/10.1057/hs.2012.18`.

Izady, N. (2015). Appointment Capacity Planning in Specialty Clinics: A Queueing Approach. *Operations Research*, *63*(4), 916–930, `https://doi.org/10.1287/opre.2015.1391`.

Jack, E. P. & Powers, T. L. (2009). A review and synthesis of demand management, capacity management and performance in health-care services. *International Journal of Management Reviews*, *11*(2), 149–174, `https://doi.org/10.1111/j.1468-2370.2008.00235.x`.

Jamali, N., Leung, R. K., & Verderber, S. (2020). A review of computerized hospital layout modelling techniques and their ethical implications. *Frontiers of Architectural Research*, *9*(3), 498–513, `https://doi.org/10.1016/j.foar.2020.01.003`.

Klute, B., Homb, A., Chen, W., & Stelpflug, A. (2019). Predicting Outpatient Appointment Demand Using Machine Learning and Traditional Methods. *Journal of Medical Systems*, *43*(9), `https://doi.org/10.1007/s10916-019-1418-y`.

Kuiper, A., de Mast, J., & Mandjes, M. (2021a). The problem of appointment scheduling in outpatient clinics: A multiple case study of clinical practice. *Omega*, *98*, 102–122, `https://doi.org/10.1016/j.omega.2019.102122`.

Kuiper, A., Mandjes, M., de Mast, J., & Brokkelkamp, R. (2021b). A flexible and optimal approach for appointment scheduling in healthcare. *Decision Sciences*, `https://doi.org/10.1111/deci.12517`.

Kuo, Y.-H., Balasubramanian, H., & Chen, Y. (2020). Medical appointment overbooking and optimal scheduling: tradeoffs between schedule efficiency and accessibility to service. *Flexible Services and Manufacturing Journal*, *32*(1), 72–101, `https://doi.org/10.1007/s10696-019-09340-z`.

Laan, C., van de Vrugt, M., Olsman, J., & Boucherie, R. J. (2018). Static and dynamic appointment scheduling to improve patient access time. *Health Systems*, *7*(2), 148–159, `https://doi.org/10.1080/20476965.2017.1403675`.

Law, A. M. (2007). *Simulation Modeling and Analysis*. New York: McGraw-Hill, 4 edition.

Leeftink, G., Martinez, G., Hans, E. W., Sir, M. Y., & Pasupathy, K. S. (2021). Optimising the booking horizon in healthcare clinics considering no-shows and cancellations. *International Journal of Production Research*, `https://doi.org/10.1080/00207543.2021.1913292`.

Leichsenring, K. (2004). Developing integrated health and social care services for older persons in Europe. *International Journal of Integrated Care*, *4*(e10), https://doi.org/10.5334/ijic.107.

Liu, N., van de Ven, P. M., & Zhang, B. (2019). Managing Appointment Booking Under Customer Choices. *Management Science*, *65*(9), 4280–4298, https://doi.org/10.1287/mnsc.2018.3150.

Liu, N. & Ziya, S. (2014). Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management*, *23*(12), 2209–2223, https://doi.org/10.1111/poms.12200.

Maarsingh, O. R., Henry, Y., Van De Ven, P. M., & Deeg, D. J. (2016). Continuity of care in primary care and association with survival in older people: A 17-year prospective cohort study. *British Journal of General Practice*, *66*(649), 531–539, https://doi.org/10.3399/bjgp16X686101.

Margolius, D., Gunzler, D., Hopkins, M., & Teng, K. (2018). Panel size, clinician time in clinic, and access to appointments. *Annals of Family Medicine*, *16*(6), 546–548, https://doi.org/10.1370/afm.2313.

Marshall, E. G., Clarke, B., Burge, F., Varatharasan, N., Archibald, G., & Andrew, M. K. (2016). Improving continuity of care reduces emergency department visits by long-term care residents. *Journal of the American Board of Family Medicine*, *29*(2), 201–208, https://doi.org/10.3122/jabfm.2016.12.150309.

Marx, R., Drennan, M. J., Johnson, E. C., Hirozawa, A. M., Tse, W. M., & Katz, M. H. (2011). Assessing and increasing patient panel size in the public sector. *Journal of Public Health Management and Practice*, *17*(6), 506–512, https://doi.org/10.1097/PHH.0b013e318211393c.

Murray, M. & Berwick, D. M. (2003). Advanced Access. *Journal of the American Medical Association*, *289*(8), 1035–1040, https://doi.org/10.1001/jama.289.8.1035.

Murray, M., Davies, M., & Boushon, B. (2007). Panel size: How many patients can one doctor manage? *Family Practice Management*, *14*(4), 44–51, https://aafp.org/fpm/2007/0400/p44.html.

National Association of Statutory Health Insurance Physicians (2018). *Berufsmonitoring Medizinstudierende 2018 (Job monitoring of medicine students 2018)*. Technical report, National Association of Statutory Health Insurance Physicians, Berlin, https://kbv.de/media/sp/Berufsmonitoring_Medizinstudierende_2018.pdf.

National Association of Statutory Health Insurance Physicians (2021). Disease Management Programme (disease management programs). Online: Accessed 2021-07-05, `https://kbv.de/html/dmp.php`.

Nguyen, T. B. T., Sivakumar, A. I., & Graves, S. C. (2018). Capacity planning with demand uncertainty for outpatient clinics. *European Journal of Operational Research*, *267*(1), 338–348, `https://doi.org/10.1016/j.ejor.2017.11.038`.

Nova Scotia Health Authority (2020). Collaborative family practice teams. Online: Accessed 2020-11-25, `http://cfpt.nshealth.ca/`.

Ozen, A. & Balasubramanian, H. (2013). The impact of case mix on timely access to appointments in a primary care group practice. *IIE Transactions*, *16*(2), 101–18, `https://doi.org/10.1007/s10729-012-9214-y`.

Panagiotis Mitropoulosa, Ioannis Mitropoulos, I. G. (2013). Combining dea with location analysis for the effective consolidation of services in the health sector. *Computers and Operations Research*, *40*(9), 2241 – 2250, `https://doi.org/10.1016/j.cor.2012.01.008`.

Parker, B. R. & Srinivasan, V. (1976). A consumer preference approach to the planning of rural primary health-care facilities. *Operations Research*, *24*(5), 991–1025.

Pérez-Gosende, P., Mula, J., & Díaz-Madroñero, M. (2021). Facility layout planning. An extended literature review. *International Journal of Production Research*, *59*(12), 3777–3816, `https://doi.org/10.1080/00207543.2021.1897176`.

Pong, R. & Pitblado, J. (2005). *Geographic distribution of Physicians in Canada: beyond how many and where*. Technical report, Canadian Institute for Health Information, Ottawa, `https://secure.cihi.ca/free_products/Geographic_Distribution_of_Physicians_FINAL_e.pdf`.

Raffoul, M., Moore, M., Kamerow, D., & Bazemore, A. (2016). A primary care panel size of 2500 is neither accurate nor reasonable. *Journal of the American Board of Family Medicine*, *29*(4), 496–499, `https://doi.org/10.3122/jabfm.2016.04.150317`.

Rais, A. & Viana, A. (2011). Operations Research in Healthcare: a survey. *International Transactions in Operational Research*, *18*(1), 1–31, `https://doi.org/10.1111/j.1475-3995.2010.00767.x`.

Reuter-Oppermann, M., Nickel, S., & Steinhäuser, J. (2019). Operations research meets need related planning: Approaches for locating general practitioners' practices. *PLOS ONE*, *14*(1), e0208003, `https://doi.org/10.1371/journal.pone.0208003`.

Reuter-Oppermann, M., Rockemann, D., & Steinhäuser, J. (2017a). A GIS-based decision support system for locating primary care facilities. In S. Za, M. Drăgoicea, & M. Cavallari (Eds.), *Exploring Services Science* (pp. 210–222). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-319-56925-3_17.

Reuter-Oppermann, M., van den Berg, P. L., & Vile, J. L. (2017b). Logistics for Emergency Medical Service systems. *Health Systems*, *6*(3), 187–208, https://doi.org/10.1057/s41306-017-0023-x.

Riens, B., Erhart, M., & Mangiapane, S. (2012). *Arztkontakte im Jahr 2007 - Hintergründe und Analysen (Patient physician contacts in 2007 - settings and analyses)*. Technical report, Zentralinstitut für die kassenärztliche Versorgung in der Bundesrepublik Deutschland, https://www.versorgungsatlas.de/fileadmin/ziva_docs/14/Arztkontakte_Bericht_2012-02-15.pdf.

Rismanchian, F. & Lee, Y. H. (2017). Process Mining–Based Method of Designing and Optimizing the Layouts of Emergency Departments in Hospitals. *HERD: Health Environments Research and Design Journal*, *10*(4), 105–120, https://doi.org/10.1177/1937586716674471.

Samorani, M. & Ganguly, S. (2016). Optimal Sequencing of Unpunctual Patients in High-Service-Level Clinics. *Production and Operations Management*, *25*(2), 330–346, https://doi.org/10.1111/poms.12426.

Schacht, M. (2018). Improving same-day access in primary care. *Operations Research for Health Care*, *18*, 119–134, https://doi.org/10.1016/j.orhc.2017.09.003.

Schulz, M., Czihal, T., Bätzing-Feigenbaum, J., & Von Stillfried, D. (2016). Zukünftige relative Beanspruchung von Vertragsärzten - Ein Projektion nach Fachgruppen für den Zeitraum 2020 bis 2035 (Future relative demands for statutory health insurance physicians). *Versorgungsatlas-Bericht*, *16*(2), https://doi.org/10.20364/VA-16.02.

Schuurman, N., Berube, M., & Crooks, V. A. (2010). Measuring potential spatial access to primary health care physicians using a modified gravity model. *The Canadian Geographer / Le Géographe canadien*, *54*(1), 29–45, https://doi.org/10.1111/j.1541-0064.2009.00301.x.

Shin, Y. W. (2000). Transient distributions of level dependent quasi-birth-death processes with linear transition rates. *Korean Journal of Computational and Applied Mathematics*, *7*(1), 83–100, https://doi.org/10.1007/BF03009929.

Srinivas, S. & Ravindran, A. R. (2018). Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework. *Expert Systems with Applications*, *102*, 245–261, `https://doi.org/10.1016/j.eswa.2018.02.022`.

Stiglic, G. & Kokol, P. (2005). Intelligent patient and nurse scheduling in ambulatory health care centers. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference* (pp. 5475–5478). `https://doi.org/10.1109/IEMBS.2005.1615722`.

Takács, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press.

Takagi, H. (1993). *Queueing Analysis: A Foundation of Performance Evaluation, Vol. 2: Finite systems*. Queueing Analysis: A Foundation of Performance Evaluation. Elsevier Science Publisheres B.V.

Tien, J. M. & El-Tell, K. (1984). A quasihierarchical location-allocation model for primary health care planning. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-14*(3), 373–380, `https://doi.org/10.1109/TSMC.1984.6313229`.

Urban, R. M. J. (2019). Redesigning Panel Configurations: A Case Study in Primary Care. Dalhousie University, `https://dalspace.library.dal.ca/handle/10222/76836`.

Vahdatzad, V. & Griffin, J. (2016). Outpatient clinic layout design accounting for flexible policies. In *2016 Winter Simulation Conference (WSC)* (pp. 3668–3669). `https://doi.org/10.1109/WSC.2016.7822391`.

Van Servellen, G., Fongwa, M., & Mockus D'Errico, E. (2006). Continuity of care and quality care outcomes for people experiencing chronic conditions: A literature review. *Nursing and Health Sciences*, *8*(3), 185–195, `https://doi.org/10.1111/j.1442-2018.2006.00278.x`.

Vanberkel, P. T., Litvak, N., Puterman, M. L., & Tyldesley, S. (2018). Queuing network models for panel sizing in oncology. *Queueing Systems*, *90*, 291–306, `https://doi.org/10.1007/s11134-018-9571-4`.

Vos, L., Groothuis, S., & van Merode, G. G. (2007). Evaluating hospital design from an operations management perspective. *Health Care Management Science*, *10*(4), 357–364, `https://doi.org/10.1007/s10729-007-9034-7`.

Welch, P. D. (1981). On the problem of the initial transient in steady-state simulation. *IBM Watson Research Center*.

Welch, P. D. (1983). The statistical analysis of simulation results. *The computer performance modeling handbook*, *22*, 268–328.

Wolinsky, F. D., et al. (2010). Continuity of care with a primary care physician and mortality in older adults. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *65 A*(4), 421–428, https://doi.org/10.1093/gerona/glp188.

Yanık, S. & Bozkaya, B. (2020). A review of districting problems in health care. In R. Z. Ríos-Mercado (Ed.), *Optimal Districting and Territory Design* (pp. 31–55). Cham: Springer International Publishing, https://doi.org/10.1007/978-3-030-34312-5_3.

Zacharias, C. & Armony, M. (2017). Joint Panel Sizing and Appointment Scheduling in Outpatient Care. *Management Science*, *63*(11), 3978–3997, https://doi.org/10.1287/mnsc.2016.2532.

Zacharias, C. & Yunes, T. (2020). Multimodularity in the Stochastic Appointment Scheduling Problem with Discrete Arrival Epochs. *Management Science*, *66*(2), 744–763, https://doi.org/10.1287/mnsc.2018.3242.

Zander, A. (2017). Modeling Indirect Waiting Times with an M/D/1/K/N Queue. In *Proceedings of the Second KSS Research Workshop : Karlsruhe, Germany, February 2016. Ed.: P. Hottum*, volume 69 of *KIT Scientific Working Papers* (pp. 110–119).: Karlsruher Institut für Technologie (KIT), https://doi.org/10.5445/IR/1000076542.

Zander, A. & Mohring, U. (2016). Dynamic Appointment Scheduling with Patient Time Preferences and Different Service Time Lengths. In *Applied Operational Research: 8th International Conference on Applied Operational Research, Proceedings, Rotterdam, The Netherlands, 28th - 30th June 2016. Ed.: K. Sheibani*, volume 8 of *Lecture Notes in Management Science* (pp. 72–77).: ORLAB Analytics, https://doi.org/10.5445/IR/1000077793.

Zander, A., Nickel, S., & Vanberkel, P. (2021). Managing the intake of new patients into a physician panel over time. *European Journal of Operational Research*, *294*(1), 391–403, https://doi.org/10.1016/j.ejor.2021.01.035.

Zonderland, M. E. (2021). Theoretical and Practical Aspects of Outpatient Clinic Optimization. In M. E. Zonderland, R. J. Boucherie, E. W. Hans, & N. Kortbeek (Eds.), *Handbook of Healthcare Logistics* (pp. 25–36). Springer International Publishing, https://doi.org/10.1007/978-3-030-60212-3_3.

Zonderland, M. E. & Boucherie, R. J. (2021). A Survey of Literature Reviews on Patient Planning and Scheduling in Healthcare. In M. E. Zonderland, R. J. Boucherie, E. W. Hans, & N. Kortbeek (Eds.), *Handbook of Healthcare Logistics* (pp. 17–23). Springer International Publishing, https://doi.org/10.1007/978-3-030-60212-3_2.