Health Care Management Science (2026) 29:53
https://doi.org/10.1007/s10729-026-09789-z

Robust mask-based appointment scheduling in primary care practices

Sabrina Schmitz[1] · **Martin Comis**[1] · **Christina Büsing**[1]

Received: 27 September 2023 / Accepted: 12 August 2026
© The Author(s) 2026

Abstract
In most health care systems, a primary care physician (PCP) is both the first instance consulted by patients with medical concerns and the instance coordinating patients' continued access to medical care. Due to the PCP's pivotal role, we address challenges of a high-quality primary care service by interday appointment scheduling on a tactical decision level. In addition to patients who seek medical care by appointments, we also consider patients who complicate the PCP's schedule planning by seeking immediate care as walk-ins without prior notice or failing to show up to their appointments. We study appointment scheduling systems based on so-called masks and focus on the balanced workload of the PCP in form of the mask design problem. To account for different uncertainties in demand for treatment, we extend the mask design problem to the robust mask design and the robust multimask design problem. For all three problems, we provide a combinatorial interpretation by a network flow and design model. We develop a solution approach that combines binary search with compact formulations (of extensions) of minimum cost flow problems. Finally, we conduct an extensive case study by agent-based simulation. We evaluate the mask-based appointment scheduling systems and compare them with five appointment scheduling systems from the literature on the basis of key performance indicators such as the PCP's overtime and idle time as well as the number of unsuccessful treatment requests or rejected walk-ins.

Keywords Primary health care · Appointment scheduling · Demand uncertainty · Robust minimum cost flows

1 Introduction

1.1 Motivation

Primary care physicians (PCPs) play a crucial role in health care systems as they are often the first instance consulted by patients with medical concerns [1, 2]. PCPs' responsibilities begin with conducting an examination to determine the patients' states of health and assessing whether they themselves can provide the necessary medical care. If this is not the case, they are obliged to ensure the continuation of medical care, for example by referring the patient to a specialist [3]. As the instance facilitating and coordinating patients' access to medical care, the PCP's quality of service is not only measured on the basis of competent medical care but also by aspects such as the patient-physician continuity, timely access to medical care, and waiting times within the practice, as discussed in more detail in the following.

Patient-physician continuity is desirable as patients benefit from follow-up treatments and referrals [3, 4]. PCPs themselves benefit as they have to spend less time dealing with medical histories of new patients [1]. Clearly, patient-physician continuity may conflict with swift access to health care if patients prefer to be treated by their familiar PCP rather than by any available PCP [5]. Timely access to medical care is imperative, as it is decisive for the patients' choice of medical facility, course of illness, and medical outcome [1, 6]. For instance, long access times for PCP practices increase the likelihood of patients visiting emergency rooms [7, 8]. In the United States, patients' timely access is endangered by the clash of increasing demand for primary health care and the nationwide lack of PCPs [2, 9]. Similar conflicts arise in other countries. For instance in rural areas of Germany, mainly elderly PCPs near retirement [10] combined with fewer medical graduates who are willing to start a practice [11] threaten the prospect of timely medical care. Due to the worldwide aging population with the highest risk of chronic conditions [9, 12], PCPs' schedules are already

✉ Sabrina Schmitz
schmitz@combi.rwth-aachen.de

[1] Combinatorial Optimization, RWTH Aachen University, Im Süsterfeld 9, 52072 Aachen, Germany

Published online: 31 October 2026

overloaded [13]. Overloading in turn leads to long waiting times in the practice—one of the major reasons for patients' complaints [14].

To face the described and further challenges with providing high-quality service, an effective appointment scheduling system is required [15, 16]. However, the PCPs' schedule planning is complicated as patients do not attend their appointments or patients forgo scheduling an appointment and seek immediate care by walking into the practice without prior notice—as so-called no-shows or walk-ins. In the United States, no-show and walk-in rates range between 20% to 30% and 10% to 25% of daily primary care visits, respectively [17, 18]. German surveys reveal similar data regarding walk-ins, namely that an average of 18.8% of patients are walk-ins [19]. Compared to data for specialists with an average rate of 6.5%, this almost three times higher walk-in rate highlights the difficulty of scheduling in German primary care practices. In addition to high numbers of walk-ins and/or no-shows, the scheduling is further complicated by fluctuating demand for treatment which leads to an inefficient and imbalanced workflow of PCPs [20]. PCPs have to work overtime on one day and are underutilized on another day of the same week. In summary, an effective appointment scheduling system maintains patient-physician continuity, enables timely access, reduces waiting times, smooths the PCP workflow, and provides robustness against demand uncertainties [1].

Contributing to the efforts of determining an effective appointment scheduling system, we study the appointment scheduling problem on a tactical decision level in form of a so-called mask design problem (MDP). The MDP addresses the problem of matching the PCP's capacity and the demand for treatment of different patient types with different needs and behaviors. For this purpose, the PCP works with a so-called mask designed for their appointment schedule that specifies which time slots are intended for which patients. We consider patients who seek medial care by appointment but may not show up to their appointments as well as patients who seek medical care as walk-ins. To account for uncertain demand for treatment, we extend the MDP to the robust mask design problem (rMDP) and the robust multimask design problem (rMMDP). To that end, we model uncertain demands for treatment using potential scenarios that may occur, which we integrate into our model in a robust optimization framework. The main focus of this paper is the rMMDP for which we aim at finding multiple masks that are robust against demand uncertainties and between which PCPs can switch to flexibly react to changes in demand. For instance, if a PCP uses a mask that is tailored to low demand and notices an increase in demand, they can switch to a mask designed for that level of demand. Before discussing previous related work and go into further detail, we stress our contribution to the field.

1.2 Contribution and outline

The main contributions of this paper are summarized as follows. We define the MDP considering patients who seek medical care by appointment (but may not show up) and patients who seek medical care as walk-ins. For this problem, we present a network flow and design model using a solution approach that combines binary search and compact integer program (IP) formulations of minimum cost flow problems. Due to uncertainties in demand for treatments, we extend the MDP to the rMDP. To account for additional uncertainties and overfitted solutions, we present the rMMDP. For the rMDP and rMMDP, we extend our solution approach by robust compact minimum cost flow formulations. To analyze the resulting mask-based appointment scheduling systems, we conduct a case study using agent-based simulation. In this case study, we test on an operational decision level whether the mask-based appointment scheduling systems can withstand a complex primary health care system that includes individual patient events and characteristics such as cancellations, the willingness to wait to an appointment, and session preferences. We evaluate that the robust mask-based appointment scheduling systems lead to high-quality solutions in several key performance indicators. Compared to the appointment scheduling systems from the literature, for example, the number of patients' unsuccessful appointment requests is halved and the number of rejected walk-ins is reduced to one third. In addition, patients' access times to appointments is shortened and PCP's overtime is reduced. To the best of our knowledge, this is the first contribution that studies interday appointment scheduling by means of combinatorial and robust methods. More precisely, this is the first contribution that studies interday appointment scheduling as a robust network flow and design problem.

The paper is structured as follows. In Section 2, we outline literature of appointment scheduling in primary health care with a focus on interday appointment scheduling. In Section 3, we define a physician-patients setting, introduce mask-based interday appointment scheduling, and present the MDP. In Section 4, we model the MDP in a deterministic setting and subsequently extend it to a robust setting. In Section 5, we consider a further extension of the MDP, the rMMDP. We note that the modeling in Sections 4 and 5 is presented step-by-step in great detail. In Section 6, we present and discuss the results from our case study. In Section 7, we discuss and conclude our results.

2 Appointment scheduling in the literature

Health care services face many challenges due to the clash of increasing demand for medical care and the resources available [21, 22]. In the overview [23], Hulshof et al. identify

the general classifications of planning and control decisions of health care services. Concentrating on appointment scheduling in outpatient services, we refer to the overview of Cayirli and Veral [24] which focuses on formulating general problems and their various modelings. Factors that make appointment scheduling challenging are demonstrated by Gupta and Denton [1]. Furthermore, we refer to Ahmadi-Javid et al. [25] for a review of analytical and numerical optimization studies for outpatient appointment scheduling systems. In the following, we provide an overview about the categories into which appointment scheduling systems are classified as well as related work.

2.1 Categories of appointment scheduling systems

Appointment scheduling systems established in the literature are categorized into four different types of systems [26]: traditional, advanced access, hybrid, and carve-out appointment scheduling systems. The traditional appointment scheduling system allows all slots to be scheduled for appointments by requesting patients. Appointments may be scheduled well in advance and are often used to postpone patients requests. In contrast, the advanced access appointment scheduling system accommodates all patients on the day of their request. The core idea is to schedule same-day appointments only [16]. Robinson and Chen [27] compare these two contrasting scheduling policies regarding various performance measures. Both appointment scheduling systems show a lack of planning security with respect to the actual number of treatments. The traditional appointment scheduling system is disadvantageous for two reasons: short-term cancellations and no-shows, i.e., patients who do not show up for their appointments. Overbooking strategies have been proven to reduce the resulting negative impacts [27–33]. Further difficulties for the traditional appointment scheduling system are, for example, long access times to appointments and insufficient time for (walk-in) patients with acute or urgent concerns [34]. The advanced access appointment scheduling system becomes less efficient due to the uncertain demand for treatment. This causes waiting times for patients as well as inefficient and imbalanced working hours for PCPs [27]. The request for appointments scheduled in advance is also not considered, which particularly affects elderly and chronically ill patients [25, 26]. Furthermore, on days with high demand for treatment patients are still referred to the next session or the following day—the so-called overflow patients. Thus, same-or-next-day scheduling policies also appear in the context of advanced access systems [16] which have been proven to be successful against simple same-day scheduling policies [27]. For this reason, hybrid appointment scheduling systems are introduced that combine the traditional and the advanced access appointment scheduling systems [27, 31, 35]. Finally, the carve-out appointment scheduling system is managed as the traditional appointment scheduling system where a part of the PCP's working hours is reserved for specific procedures or urgent services [25, 26]. Both the hybrid and carve-out appointment scheduling systems become less efficient if the ratio between traditional scheduled appointments, and same-day appointments or reserved working hours for urgent services is not chosen accurately. A disadvantage and a major challenge of these systems is therefore determining the correct ratio [27]. We note that the categorization of appointment scheduling systems becomes blurred as the systems overlap. In addition, the categorization is not clearly defined due to different assumptions, settings, services, and facilities of various health care systems of different countries.

Besides the differentiation in appointment scheduling systems, studies are further categorized according to intraday and interday appointment scheduling [36]. Intraday appointment scheduling concentrates on a single day with emphasis on, for example, the reduction of patient waiting times or the patient-physician continuity. In this context, mostly same-day scheduling policies are investigated on tactical and operational decision levels. In contrast, the concept of interday appointment scheduling aims at allocating appointments for consecutive days so that the PCP's working hours are efficiently used. More precisely, on a tactical decision level, the question arises as to how many appointments are scheduled on which days in order to manage demand and take uncertainties such as walk-ins into account. Furthermore, fluctuating demand for treatment is smoothed throughout the week, and therefore, the utilization of the PCP's working hours and patient arrivals are controlled in a more efficient way. The allocation and management of the PCP's capacity is significant. This often includes the capacity ratio between same-day demand and demand that can be postponed.

In this study, we analyze an appointment scheduling system for a single PCP operating in a primary care practice that analogously operates to a carve-out system. Appointments are traditionally scheduled while ample time is reserved for walk-ins to serve their urgent needs. For designing an appointment scheduling system, we refer to studies [37, 38] that introduce a notion similar to our tactical scheduling planning. However, unlike these studies, we consider interday rather than intraday appointment scheduling and mainly focus on the balance of the PCP's utilization. As we consider interday appointment scheduling in this study, we propose further related literature in the following section that focuses on the ratio of same-day demand to demand that can be postponed.

2.2 Related work

Qu et al. [39] manage a hybrid appointment scheduling system by a two time horizon. The first time horizon is used

to schedule same-day appointments and the second to postpone demand for treatment within a maximum of one week. Regarding the expectation and variance of the number of patients seen by the PCP, the authors show that the two time horizon system is at least as good as an advanced access appointment scheduling system. Further, Qu et al. [34] investigate the ratio between same-day appointments and appointments in advance by an analytical method with the aim of maximizing the expected number of treated patients. The results show that the optimal ratio depends on the PCP's capacity, the ratio of the show-up rate of prescheduled appointments to the show-up rate of same-day appointments, and the joint distribution of demands for prescheduled and same-day appointments.

Dobson et al. [40] formulate a stochastic model to determine the optimal number of slots reserved for urgent patients. They compare the resulting carve-out appointment scheduling system with an advanced access appointment scheduling system on the basis of revenue, i.e., the average number of urgent patients who are not seen during the PCP's working hours and the average queue length for routine patients. Feldman et al. [36] develop a dynamic appointment scheduling model that considers not only appointment cancellations and no-shows but also patients' preferences. If appointments do not match their preferences, patients are not seen by the PCP. With their so-called next-n policy, Feldman et al. aim at a maximum expected revenue per day, which is determined by patients treated who are allowed to choose their appointments from n days.

In contrast to these studies, like Wiesche et al. [20], we consider patients with acute concerns who attempt scheduling an appointment within their willingness to wait. If no appointment is proposed within this time, they forgo scheduling an appointment and become walk-ins. This behavior is inspired by German practices where acute patients may walk into the practice even though there are no more appointments available. If possible, the patient is suggested a preferred time to visit the practice. However, we note that walk-ins have to expect long waiting times for their treatments as in the treatment order patients with appointments are preferred over walk-ins. In this study, we follow this strategy and postpone possible overbooking strategies to future studies. Wiesche et al. [20] determine the minimum number of appointments scheduled for a weekly profile to match capacity with demand by a mixed integer program formulation. Stochastic simulation shows that their appointment scheduling strategy balances the PCP's work hours and reduces overflow patients by offering mainly walk-in blocks and only a few appointments before and after lunch. Schacht [41] extends the contributions by seasonal and stochastic aspects. He concludes that reconfiguration of the PCP's working hours throughout the year is advantageous. Unlike studies of Wiesche et al. [20] and Schacht [41], we do not minimize the number of appointment slots to have as much capacity as possible for walk-ins. Instead, we endeavor to offer appointment slots, motivated by the German health care system where only three percent of all primary care practices do not offer appointments [19]. More precisely, we aim at determining an optimal daily ratio between appointment slots and reserved time for walk-ins to control the PCP's workload to match the PCP's capacity with demand for treatment. We also note that, unlike studies [20] and [41], we do not neglect patients whose demand for treatment is known well in advance, for instance as it is the case for chronically ill patients. Instead, we use this controllable demand to manage and balance the PCP's workload.

The key difference between previously published studies and ours is that we use combinatorial optimization and do not consider stochastic but robust optimization models. More precisely, we use a robust network flow and design model. Network flow models are well analyzed and solvable in polynomial time [42, 43]. In contrast, robust flows with consistent flow constraints under demand uncertainty are $\mathcal{NP}$-hard to compute [44]. To the best of our knowledge, we are the first who use combinatorial and robust solving methods to develop an appointment scheduling system.

3 Setting and problem definition on tactical decision level

In this section, we introduce notations and formalize the appointment scheduling problem. In Section 3.1, we introduce a deterministic physician-patients setting. In Section 3.2, we show how to control demand for treatment by masked-based interday appointment scheduling. In Section 3.3, we present the MDP.

We note that the appointment scheduling problem considered in this section is on a tactical decision level and comparable with capacity allocation. We do not schedule individual patients to appointment slots but analyze the required capacity for walk-ins, chronic, and acute patients. For this reason, we consider the model on a higher level and do not take into account all events and uncertainties, such as scheduling times, service times, and cancellations. These mostly patient individual events and uncertainties are modeled on the operational level in the case study by agent-based simulation presented in Section 6.

3.1 Deterministic physician-patients setting

Like most studies on primary health care [25], our study assumes a *single-server system*, i.e., practices with one PCP, because of the importance of patient-physician continuity. The single-server assumption holds even in joint practices, as PCPs usually provide medical care to their own panel of

patients [24, 45]. Furthermore, we assume patient loyalty: patients seek medical care by their PCP and do not abandon them, for example by switching to another PCP. We note that multiple-server systems become important if the utilization of individual PCPs is not balanced and their capacities do not meet the demand for treatment. PCPs generally operate in clinical sessions—a morning and an afternoon session. We denote the set of all sessions by $\mathcal{K}$. Concentrating on a single working week, we describe each *session* by a tuple $k = (i, j) \in \mathcal{K}$, where $i \in \{0, \dots, 4\}$ indicates the working day from Monday to Friday and a binary indicator $j \in \{0, 1\}$ represents whether the morning ($j = 0$) or afternoon ($j = 1$) session is considered. For each of these sessions, we refine the PCP's working hours by differentiating between opening hours, buffer time, and overtime. *Opening hours* determine the time span $\boldsymbol{o} \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ in which appointments are scheduled and patients may be admitted for treatments. After opening hours, we assume PCPs to use *buffer time* $\boldsymbol{b} \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ to compensate for contingencies such as delays. We refer to the time during the start of the opening hours and the end of the buffer as the PCP's *capacity*, denoted by parameter $\boldsymbol{c} := \boldsymbol{o}+\boldsymbol{b}$. PCPs who do *overtime* work even beyond the buffer time. Figure 1 illustrates an exemplary working day of a PCP who operates in two sessions.

During the PCP's opening hours, patients have two options to seek medical care in the modeled setting—either by appointment or as walk-in. Whether or not patients try to make an appointment or accept a proposed appointment depends on their health status and their *willingness to wait*. In this context, we divide patients into three categories: walk-ins, chronic patients, and acute patients.

Walk-ins have no willingness to wait, forgo scheduling an appointment, and walk into the practice during a specific session without prior notice. We specify their demand per session by parameter $\boldsymbol{d}^w \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$. *Chronic patients* request (recurring follow-up) appointments well in advance. We assume that their willingness to wait ensures flexibility in their appointment requests and we therefore specify their demand as a single weekly demand $d^c \in \mathbb{Z}_{\geq 0}$. *Acute patients* prefer an appointment and request one for a specific session, specified by demand $\boldsymbol{d}^a \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$. If they are not offered an appointment or an alternative appointment within their willingness to wait, acute patients walk into the practice during the session of their initial request and are handled like walk-ins. Motivated by the German labor law, which states that employees must provide a sick note after three days of sick leave [46], we uniformly set their willingness to wait to a maximum of three days. Thus, this is the period in which we want to guarantee an appointment for acute patients. We note that the performance of the appointment scheduling systems determines how many acute patients become walk-ins, in other words, how much controllable demand becomes uncontrollable demand. Nevertheless, there always exist acute patients who arrange appointments but do not attend them. As non-occurring demand only becomes known when the appointment would take place, we specify this missing demand per session by parameter $\boldsymbol{d}^n \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ and refer to it as no-show demand.

If patients receive treatment, the duration of the treatments is subject to variation. Nevertheless, studies have shown that the length of the treatments only marginally varies in primary care [47, 48]. Thus, we assume a uniform *anticipated service time* $t \in \mathbb{Z}_{>0}$ for all patients.

Before proposing a method to manage the patient demand, we note that our models generalize to finer demand aggregations, for example, for every 30 minutes of a session. However, as our models support on a tactical decision level and as some of the models are based on a single demand realization, finer aggregations may lead to overfitting and consequently to appointment scheduling systems that are not generally applicable.

3.2 Controlling patient demand by masked-based interday appointment scheduling

Naturally, patient demand is unevenly distributed over the sessions within a week. Addressing this issue, we manage controllable demand and consider potential uncontrollable demand by interday appointment scheduling. More precisely, we beneficially distribute chronic demand d^c over the week while we shift acute demand $\boldsymbol{d}^a$ from days with high demand to days with lower demand. Following this strategy, we account for the walk-in demand $\boldsymbol{d}^w$ which is not controllable by postponement to a later session. To address this issue, we employ a cyclic model of a single working week. We assume that the acute demand which is shifted to the following week due to scheduled appointments, e.g., from Friday to Monday, compensate for the acute demand that would, theoretically, be shifted from the previous week. To shift (part of) the acute



Fig. 1 Schematic representation of a PCP's morning and afternoon session visualizing buffer, idle- and overtime

demand $d_k^a, k \in \mathcal{K}$ to an alternative session that is within the patient's willingness to wait, we define a set of all feasible sessions as follows

$$\mathcal{K}^+_{k=(i,j)} = \left\{ k' = (i', j') \in \mathcal{K} \;\middle|\; \begin{array}{l} (i' - i) \bmod 5 \leq 2, \\ o_{k'} > 0 \end{array} \right\}.$$

For later purposes, we define a set to summarize the sessions originally have been requested by acute patients alternatively scheduled in session $k \in \mathcal{K}$ as follows

$$\mathcal{K}^-_{k=(i,j)} = \left\{ k' = (i', j') \in \mathcal{K} \;\middle|\; \begin{array}{l} -(i' - i) \bmod 5 \leq 2, \\ o_{k'} > 0 \end{array} \right\}.$$

Using these sets, we formalize the distribution of the chronic demand and the shifts of the acute demand in the following definition.

Definition 1 (Assignment) Let demands $d^c \in \mathbb{Z}_{\geq 0}$ and $\boldsymbol{d}^a \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ be given. A $(d^c, \boldsymbol{d}^a)$-*assignment* $\boldsymbol{\alpha} = (\alpha^c, \boldsymbol{\alpha}^a) = (\alpha^c, \alpha^a_{(0,0)}, \ldots, \alpha^a_{(4,1)})$ is a $(|\mathcal{K}| + 1)$-tuple of functions $\alpha^c : \mathcal{K} \to \mathbb{Z}_{\geq 0}$ and $\alpha^a_k : \mathcal{K}^+_k \to \mathbb{Z}_{\geq 0}$, $k \in \mathcal{K}$ that distribute and shift the demand across sessions, respectively, such that $\sum_{k \in \mathcal{K}} \alpha^c(k) = d^c$ and $\sum_{k' \in \mathcal{K}^+_k} \alpha^a_k(k') \leq d^a_k, k \in \mathcal{K}$ holds.

Considering the definition of an assignment, we note the following three aspects. First, an assignment has to distribute the entire chronic demand d^c and may not shift more than the acute demand d^a_k of each session $k \in \mathcal{K}$. Second, whenever an acute patient is assigned by function $\boldsymbol{\alpha}^a$ to the requested or a later occurring session, we assume that the patient will see the PCP by appointment. Conversely, this means that the number of acute patients who are not assigned, i.e., $d^a_k - \sum_{k' \in \mathcal{K}^+_k} \alpha^a_k(k')$, determines the number of acute patients who become walk-ins in session $k \in \mathcal{K}$. Third, assignments do not have to respect the PCP's capacity by definition and thus may lead to overtime. For this reason, we strive for a system by which demand is reasonably assigned—in other words, an appointment scheduling system.

In primary care practices, appointment scheduling systems are usually based on time slots. To implement time slots in our setting, we divide the time span o_k of the PCP's opening hours for each session $k \in \mathcal{K}$ into time periods of equal length $\ell \in \mathbb{Z}_{>0}$. Each *slot* is described by a tuple $s = (k, q_k)$ where $k \in \mathcal{K}$ indicates the session and $q_k \in \{1, \ldots, n_k\}$ with $n_k := \lceil \frac{o_k}{\ell} \rceil \in \mathbb{Z}_{\geq 0}$ the position. We denote the set of all slots within the weekly PCP's capacity by $\mathcal{S}$. We assume one scheduled patient per slot and, in order to reserve a reasonable amount of time, we require the slot length ℓ to be at least as long as the anticipated service time t, i.e., $\ell \geq t$. Figure 2 visualizes an exemplary schedule of a five-day working week with ten-minute slots where the PCP's practice is closed on Wednesday and Friday afternoons.

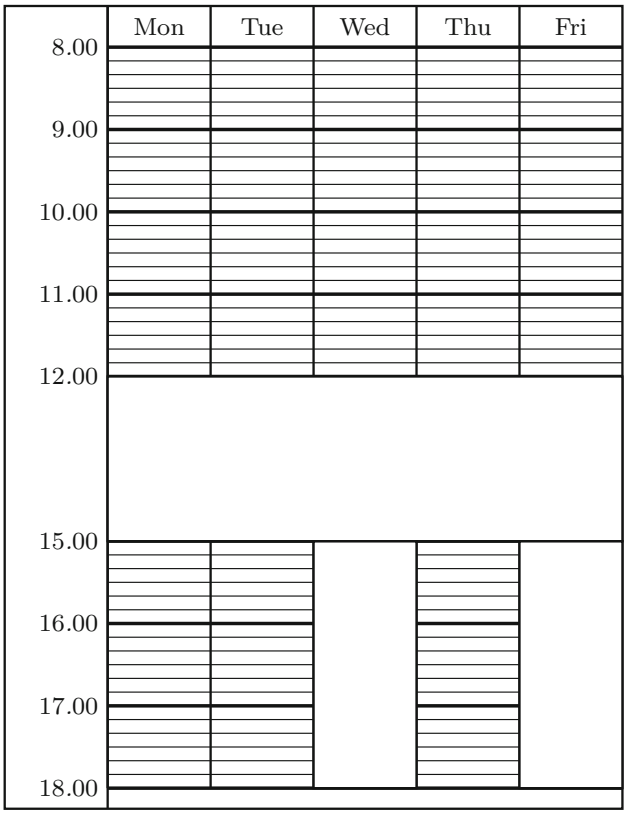

Fig. 2 Schedule defined by consecutive slots

At this point, we allow slots to have one of the following three *states*: chronic c, acute a and walk-in w. States are supposed to indicate whether a slot is reserved to schedule a chronic or acute patient or whether a slot is reserved for a potential walk-in. For simplicity's sake, we refer to slots with state 'chronic', 'acute', and 'walk-in' as chronic, acute, and walk-in slots. Using the states of slots, we can define a mask.

Definition 2 (Mask) A *mask* is a function $\mu : \mathcal{S} \to \{c, a, w\}$ that assigns each slot one state.

The aim is to assign demand, or in other words, schedule appointments on the basis of a mask. As a mask regulates how many chronic and acute patients may be scheduled and how many slots are reserved for potential walk-ins, an assignment needs to be compatible with the considered mask.

Definition 3 ($\boldsymbol{\alpha}$-compatible mask) Mask μ and $(d^c, \boldsymbol{d}^a)$-assignment $\boldsymbol{\alpha} = (\alpha^c, \boldsymbol{\alpha}^a)$ are *compatible* with each other if $\boldsymbol{\alpha}$ assigns at most as much chronic and acute demand as chronic and acute slots are available, respectively, i.e., for

$k \in \mathcal{K}$ it holds

$$\alpha^c(k) \leq n_\mu(c,k) := |\{s = (k, q_k) \in \mathcal{S} \mid \mu(s) = c\}|,$$
$$\sum_{k' \in \mathcal{K}_k^-} \alpha_{k'}^a(k) \leq n_\mu(a,k) := |\{s = (k, q_k) \in \mathcal{S} \mid \mu(s) = a\}|.$$

The set of all $(d^c, \boldsymbol{d}^a)$-assignments that are compatible with mask μ is denoted by $\mathcal{A}^\mu$. If $\boldsymbol{\alpha} \in \mathcal{A}^\mu$ holds, we say a mask μ is $\boldsymbol{\alpha}$-*compatible*.

Overall, we obtain an appointment scheduling system based on a mask and compatible assignments referred to as *mask-based appointment scheduling system*. In the next section, we design a mask that aims at a balanced utilization of the capacity of the PCP. We note that masks can be designed according to different objectives such as the distribution of certain types of patients to certain working days.

3.3 Balanced utilization by the MDP

One of the main goal PCPs strive for is an efficient workflow. This includes a balanced workload to avoid PCPs working overtime on one day while not working at full capacity on another day of the same week. To achieve this in our setting on a tactical decision level, we strive for a balanced use of the weekly capacity. We start with formal definitions for the workload per session, the weekly utilization, and the optimal workload per session in terms of a balanced utilization of the weekly capacity.

Given demands $d^c \in \mathbb{Z}_{\geq 0}$ and $\boldsymbol{d}^a, \boldsymbol{d}^w, \boldsymbol{d}^n \in \mathbb{Z}_{\geq 0}^{\mathcal{K}}$ as well as a $(d^c, \boldsymbol{d}^a)$-assignment $\boldsymbol{\alpha}$, we define the PCP's *workload* $\Phi_k^{\boldsymbol{\alpha}}$ in session $k \in \mathcal{K}$ by the anticipated working time resulting from all treatments requested in and assigned to session k minus no-shows, i.e.,

$$\Phi_k^{\boldsymbol{\alpha}} = \left(d_k^w + d_k^a - \sum_{k' \in \mathcal{K}_k^+} \alpha_k^a(k') - d_k^n + \sum_{k' \in \mathcal{K}_k^-} \alpha_{k'}^a(k) + \alpha^c(k) \right) t.$$

We note that the PCP's workload depends on demands and assignments. Recalling the cyclic model and using the weekly patient demand, we can compute the *weekly utilization* u of the PCP by

$$u = \frac{t}{C} \left(d^c + \sum_{k \in \mathcal{K}} (d_k^a + d_k^w - d_k^n) \right),$$

where $C = \sum_{k \in \mathcal{K}} c_k$ denotes the PCP's weekly capacity. We obtain the, in terms of balance, *optimal workload* per session as the product of the weekly utilization and the session's capacity, i.e., $u \cdot c_k$ for $k \in \mathcal{K}$. As the PCP's workload depends on assignments and the assignments in turn are restricted by masks, we can manage the PCP's workload by the design of a mask. Aiming at an optimal workload in every session, we evaluate masks according to the achievable workload balance.

Definition 4 Let the PCP's capacity $\boldsymbol{c} \in \mathbb{Z}_{\geq 0}^{\mathcal{K}}$ and demands $d^c \in \mathbb{Z}_{\geq 0}$, $\boldsymbol{d}^a, \boldsymbol{d}^w, \boldsymbol{d}^n \in \mathbb{Z}_{\geq 0}^{\mathcal{K}}$ be given. We define the *cost* of a mask μ as the maximum deviation of the actual from the optimal workload among all sessions, provided that the workload is subject to a $(d^c, \boldsymbol{d}^a)$-assignment $\boldsymbol{\alpha} \in \mathcal{A}^\mu$ that minimizes the cost, i.e.,

$$c(\mu) = \min_{\boldsymbol{\alpha} \in \mathcal{A}^\mu} \max_{k \in \mathcal{K}} \left| \Phi_k^{\boldsymbol{\alpha}} - u \cdot c_k \right|.$$

We conclude this section with the definition of the MDP and a summary of notations in Table 1.

Table 1 Notation table

Notation	Specification
$\mathcal{K}$	index set of sessions
$\mathcal{S}$	set of slots
$\mathcal{A}^\mu$	set of assignments suitable for mask μ
d^c	weekly chronic demand
d_k^a	acute demand per session $k \in \mathcal{K}$
d_k^w	walk-in demand per session $k \in \mathcal{K}$
d_k^n	no-show demand per session $k \in \mathcal{K}$
o_k	PCP's opening hours in session $k \in \mathcal{K}$
n_k	number of slots in session $k \in \mathcal{K}$
b_k	PCP's buffer time in session $k \in \mathcal{K}$
c_k	PCP's capacity in session $k \in \mathcal{K}$
C	PCP's weekly capacity
t	anticipated service time
ℓ	slot length
u	PCP's utilization
α^c	weekly assignment for chronic patients
α_k^a	assignment for acute pat. from session $k \in \mathcal{K}$
$\boldsymbol{\alpha}$	demand assignment
$n_\mu(c,k)$	number of chronic slots in mask μ
$n_\mu(a,k)$	number of acute slots in mask μ
$\Phi_k^{\boldsymbol{\alpha}}$	PCP's workload under assignment $\boldsymbol{\alpha}$ and demands of session $k \in \mathcal{K}$

Definition 5 (MDP) Given the PCP's capacity $\boldsymbol{c} \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ and demands $d^c \in \mathbb{Z}_{\geq 0}$, $\boldsymbol{d}^a, \boldsymbol{d}^w, \boldsymbol{d}^n \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$, the *mask design problem* aims at a mask of minimum cost.

4 Modeling the MDP

In this section, we present an extensive combinatorial interpretation of the MDP by a network flow and design model. First, we consider a model in a deterministic setting and subsequently extend this model to a robust setting.

4.1 MDP in a deterministic setting

In this section, we model the MDP by a network flow and design formulation. For this purpose, we construct a network such that a flow represents a PCP's patient flow of one week from request to treatment with or without appointment. An exemplary network of a five-day working week with one session per day is visualized in Fig. 3. The network is defined by an acyclic digraph $G = (V, A)$ with vertex set V and arc set A. Besides vertices v^1_k and v^2_k for all $k \in \mathcal{K}$, vertex set V includes multiple sources and sinks that represent the patient demand. More precisely, let $\sigma^c \in V$ be the source that models the weekly chronic demand and let $\sigma^a_k \in V$ be the source that models the acute demand per session $k \in \mathcal{K}$. Vertex $\sigma^w_k \in V$ models the demand per session $k \in \mathcal{K}$ which results from walk-ins minus no-shows. Let $\tau \in V$ be the sink which ensures that all patients are seen by the PCP. A flow originating from one of the sources only reaches sink $\tau \in V$ by using at least one of two kind of arcs. The first type of arcs $a^1_k = (v^1_k, \sigma^w_k) \in A$ represents acute and chronic patients scheduled in session $k \in \mathcal{K}$ referred to as appointment arcs. To limit the number of patients scheduled, we define upper arc capacities $\overline{\psi} : A \to \mathbb{Z}_{\geq 0}$ on the appointment arcs that are equal to the number of slots of the PCP's opening hours, i.e., $\overline{\psi}(a^1_k) = n_k$ for all $k \in \mathcal{K}$. As acute patients may be scheduled within three days after their requests, we add arcs that connect source $\sigma^a_k, k \in \mathcal{K}$ with the tail of the appointment arcs, i.e., with vertices $v^1_{k'} \in V$ for $k' \in \mathcal{K}^+_k$. Furthermore, as chronic patients may be scheduled in each session, source σ^c is also connected with vertex v^1_k for all $k \in \mathcal{K}$. The second type of arcs $a^2_k = (\sigma^w_k, v^2_k) \in A$ represents all patients seen by the PCP in session $k \in \mathcal{K}$ referred to as treatment arcs. Accordingly, the treatment arc a^2_k is connected at vertex $\sigma^w_k \in V, k \in \mathcal{K}$ in series with the appointment arc a^1_k of the same session $k \in \mathcal{K}$. In addition, source σ^a_k, $k \in \mathcal{K}$ is also connected with vertex σ^w_k, i.e., with the tail of treatment arc a^2_k, which models acute patients who become walk-ins. Finally, each treatment arc $a^2_k, k \in \mathcal{K}$ is connected with sink τ. For all vertices $v \in V$, balances

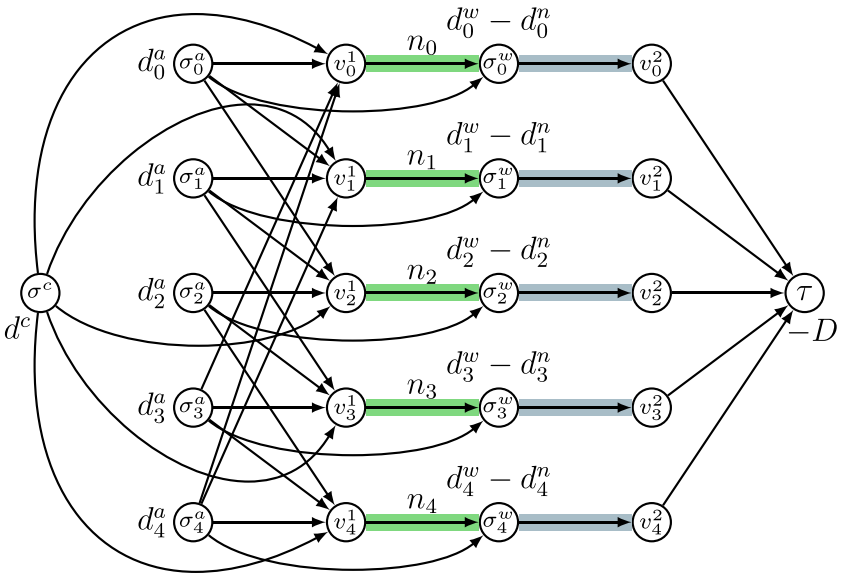

Fig. 3 Network flow model for managing patient flow

$\beta : V \to \mathbb{Z}$ with $\sum_{v \in V} \beta(v) = 0$ are defined as follows

$$\beta(v) = \begin{cases} d_k^a & \text{if } v = \sigma_k^a, k \in \mathcal{K}, \\ d_k^w - d_k^n & \text{if } v = \sigma_k^w, k \in \mathcal{K}, \\ d^c & \text{if } v = \sigma^c, \\ -\left(d^c + \sum_{k \in \mathcal{K}} \left(d_k^a + d_k^w\right)\right) & \text{if } v = \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Overall, we obtain network $N = (G, \overline{\psi}, \beta)$.

In network N, a feasible *β-flow* is defined by a function $f : A \to \mathbb{Z}_{\geq 0}$ that satisfies the flow balance constraints $\sum_{a=(v,w) \in A} f(a) - \sum_{a=(w,v) \in A} f(a) = \beta(v)$ at every vertex $v \in V$ and the capacity constraints $0 \leq f(a_k^1) \leq \overline{\psi}(a_k^1)$ on all appointment arcs $a_k^1 \in A$. We note that all flows in this study are defined to be integral. A β-flow in network N corresponds to a $(d^c, \boldsymbol{d}^a)$-assignment and a mask that are compatible with each other. We recall that the flow value on the treatment arc a_k^2 represents the number of patients seen by the PCP in session $k \in \mathcal{K}$. So, multiplying this number by the anticipated service time t results in the PCP's workload in session $k \in \mathcal{K}$. Thus, to obtain an optimal solution to the MDP, we aim at a feasible β-flow f in network N that is balanced on the treatment arcs, i.e., a flow f that minimizes the following expression

$$\max_{k \in \mathcal{K}} |t \cdot f(a_k^2) - u \cdot c_k|. \tag{1}$$

To determine such a flow, we add lower and upper arc capacities $\underline{\psi}_\omega, \overline{\psi}_\omega : A \to \mathbb{Z}_{\geq 0}$ for a parameter $\omega \in \mathbb{Z}_{\geq 0}$ on the treatment arcs $a_k^2 \in A$, $k \in \mathcal{K}$ of network N. The lower and upper arc capacities are defined such that the PCP's actual workload in session $k \in \mathcal{K}$ may deviate from the optimal workload $u \cdot c_k$ by an amount of $\omega \in \mathbb{Z}_{\geq 0}$, i.e., $\underline{\psi}_\omega(a_k^2) = \max\left\{\left\lceil \frac{1}{t}\left(u \cdot c_k - \omega\right)\right\rceil, 0\right\}$, $\overline{\psi}_\omega(a_k^2) = \left\lfloor \frac{1}{t}\left(u \cdot c_k + \omega\right)\right\rfloor$. We denote the resulting network by $N_\omega = (G, \overline{\psi}, \overline{\psi}_\omega, \underline{\psi}_\omega, \beta)$. If parameter ω is minimal, a feasible β-flow in network N_ω corresponds to a feasible flow that minimizes expression (1). Consequently, for minimum parameter ω, a feasible β-flow in network N_ω defines an optimal solution to the MDP.

Before presenting an algorithm to find the minimum parameter ω, we consider the solution values that can be realized in the MDP. The PCP's workload $\Phi_k^{\boldsymbol{\alpha}}$ is computed by the anticipated service time t times the number of treatments in session $k \in \mathcal{K}$ which can vary between zero and $D_k := d_k^w - d_k^n + \sum_{k' \in \mathcal{K}_k^-} d_{k'}^a + d^c$. Thus, the possible solution values to the MDP form the finite set $\Omega = \bigcup_{k \in \mathcal{K}} \Omega_k$ with

$$\Omega_k = \left\{|t \cdot i - u \cdot c_k| \,\middle|\, i \in \{0, 1, \ldots, D_k\}\right\}.$$

Using this, Algorithm 1 attached to Appendix A computes the optimal parameter $\omega \in \Omega$ based on the following procedure. For the sake of simplicity, binary search is used to first determine the minimum parameter i of set $\{0, 1, [3] \ldots, \max_{k \in \mathcal{K}} D_k\}$ for which a feasible β-flow exists in network $N_{\overline{\omega}}$ with $\overline{\omega} = i \cdot t$. We note that the gap between two consecutive values in set Ω_k, $k \in \mathcal{K}$ is t (if the elements of set Ω_k were ordered by size). Consequently, from the first step, we obtain an interval $(\overline{\omega} - t, \overline{\omega}]$ that contains the optimal parameter ω. In the second step, set Ω is searched for parameters that are within this interval (and smaller than $\overline{\omega}$), i.e., set $(\overline{\omega} - t, \overline{\omega}) \cap \Omega$ is searched. We note that such parameters may exist as the optimal workload $u \cdot c_k$ need not be a multiple of the anticipated service time t. If there are such parameters for which additionally feasible flows in the corresponding networks exist, the minimum parameter found so far is updated. We note that a feasible β-flow can be computed by, for example, the Minimum Mean Cycle Canceling algorithm in polynomial time [43]. Finally, Algorithm 1 returns a β-flow with minimum parameter $\omega \in \Omega$ that defines an optimal solution to the MDP as shown in the following theorem.

Theorem 1 *A mask determined by a β-flow obtained by Algorithm 1 is of minimum cost.*

The proof is moved to Appendix A. We note that the flow obtained by Algorithm 1 is not unique, as for minimum parameter $\omega \in \Omega$ an arbitrary feasible β-flow in network N_ω is determined. Using this, we integrate a second optimization stage into our network flow model that takes into account the following. Acute patients strive for an appointment. If they are offered an appointment within their willingness to wait, not only their requests are satisfied but also demand is beneficially controlled for the PCP. Considering this win-win situation, we aim at minimizing the unsuccessful appointment requests, or in other words, the number of acute patients who become walk-ins. Acute patients who become walk-ins have not been assigned by a $(d^c, \boldsymbol{d}^a)$-assignment $\boldsymbol{\alpha} = (\alpha^c, \boldsymbol{\alpha}^a)$. In total, that is $d_w^a(\boldsymbol{\alpha}^a) := \sum_{k \in \mathcal{K}} d_k^a - \sum_{k' \in \mathcal{K}_k^+} \alpha_k^a(k')$ acute patients. As assignments are restricted by masks, the number of acute patients who become walk-ins is also influenced by the design of a mask. For our network flow model, this means that we aim at finding a β-flow in network N_ω with minimum flow value on arcs $(\sigma_k^a, \sigma_k^w) \in A$, $k \in \mathcal{K}$. We can easily implement this into our model by assigning arbitrary positive cost to these arcs. Overall, we obtain a minimum cost flow model that we integrate into Algorithm 1 by determining a minimum cost

β-flow instead of an arbitrary feasible β-flow in the respective network. We refer to the model and procedure presented in this section as *Det-F*. We conclude this section with the following corollary.

Corollary 1 *Det-F is a modeling for the MDP.*

4.2 MDP in a robust setting

So far, we considered the MDP in a deterministic setting based on the assumption that the entire demand data is given. However, as acute patients contact their PCP when they suffer from acute illnesses, their demand is subject to uncertainty. Naturally, the walk-in and no-show demand is also subject to uncertainty. Only the chronic demand is regulated by recurring appointments scheduled in advance. For an effective appointment scheduling system, a mask that is robust against demand uncertainties is essential, i.e., a mask whose use achieves a balanced utilization of the PCP's capacity for all potential demands. Following the core idea of robust optimization, we aim at a version of the MDP whose solution is immune to variation in demand data. To address this issue, we model the uncertainty in acute, walk-in, and no-show demands by a set of scenarios. Our choice is a finite scenario set Λ due to the lack of historical data and the difficulty of, for example, budget uncertainty sets in terms of construction and computational complexity. In scenario $\lambda \in \Lambda$, the acute, walk-in, and no-show demands are specified by parameters $\boldsymbol{d}^a_\lambda \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$, $\boldsymbol{d}^w_\lambda \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$, and $\boldsymbol{d}^n_\lambda \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$, respectively. However, we assume that the variations in demand are limited due to reasons such as a fixed panel of patients. We require the total weekly acute, walk-in, and no-show demands to be equal for all scenarios, i.e., $\sum_{k\in\mathcal{K}} d^a_{\lambda,k} = D^a$, $\sum_{k\in\mathcal{K}} d^w_{\lambda,k} = D^w$, and $\sum_{k\in\mathcal{K}} d^n_{\lambda,k} = D^n$ for all $\lambda \in \Lambda$ for given parameters $D^a, D^w, D^n \in \mathbb{Z}_{\geq 0}$. We note that the chronic demand is still specified by parameter d^c. Given this new setting, we define a mask that is robust to uncertainties in demand.

Definition 6 (Robust Mask) A mask is called *robust* if there exists a compatible $(d^c, \boldsymbol{d}^a_\lambda)$-assignment for each scenario $\lambda \in \Lambda$. The set of all $(d^c, \boldsymbol{d}^a_\lambda)$-assignments that are compatible with mask μ is denoted by $\mathcal{A}^\mu_\lambda$ for every scenario $\lambda \in \Lambda$.

We note that the robustness of a mask is by definition independent of the walk-in and no-show demand $\boldsymbol{d}^w_\lambda$ and $\boldsymbol{d}^n_\lambda$, $\lambda \in \Lambda$. Furthermore, we note that the PCP's utilization u and the resulting optimal workload $u \cdot c_k$ in session $k \in \mathcal{K}$ is independent of the scenario considered by assumption. In contrast, the PCP's workload $\Phi^{\boldsymbol{\alpha}}_{\lambda,k}$ in session $k \in \mathcal{K}$ depends on a $(d^c, \boldsymbol{d}^a_\lambda)$-assignment $\boldsymbol{\alpha}$ and on walk-in and no-show demands $d^w_{\lambda,k}$ and $d^n_{\lambda,k}$ and thus on scenario $\lambda \in \Lambda$.

Considering the set of scenarios, we evaluate masks according to the achievable workload balance in the worst-case scenario as formalized in the following definition.

Definition 7 Let the PCP's capacity $\boldsymbol{c} \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$, demand $d^c \in \mathbb{Z}_{\geq 0}$, and demands $\boldsymbol{d}^a_\lambda, \boldsymbol{d}^w_\lambda, \boldsymbol{d}^n_\lambda \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ for all scenarios $\lambda \in \Lambda$ be given. The *cost* of a robust mask μ is defined by the worst-case maximum deviation of the actual from the optimal workload among all sessions of all scenarios, i.e.,

$$c(\mu) = \max_{\lambda\in\Lambda} \min_{\boldsymbol{\alpha}\in\mathcal{A}^\mu_\lambda} \max_{k\in\mathcal{K}} \left| \Phi^{\boldsymbol{\alpha}}_{\lambda,k} - u \cdot c_k \right|.$$

Using this, we define a robust version of the MDP as follows.

Definition 8 (rMDP) Let $\boldsymbol{c} \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ be the PCP's capacity, $d^c \in \mathbb{Z}_{\geq 0}$ be the chronic demand, and $\boldsymbol{d}^a_\lambda, \boldsymbol{d}^w_\lambda, \boldsymbol{d}^n_\lambda \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ be the acute, walk-in, and no-show demand for all scenarios $\lambda \in \Lambda$, respectively. The *robust mask design problem* aims at a robust mask of minimum cost.

In the next step, we model the rMDP by an adjusted network flow and design model in which a so-called robust flow under consistent flow constraints is sought [44]. We consider a network $N^\Lambda = (G = (V, A), \overline{\psi}, \boldsymbol{\beta})$ whose balances $\beta^\lambda : V \to \mathbb{Z}$ with $\sum_{v\in V} \beta^\lambda(v) = 0$ are given for all scenarios $\lambda \in \Lambda$, denoted by $\boldsymbol{\beta} = (\beta^1, \ldots, \beta^{|\Lambda|})$. Furthermore, we introduce two arc subsets $A^c_{\text{fix}} := \{(\sigma^c, v^1_k) \mid k \in \mathcal{K}\} \subseteq A$ and $A^a_{\text{fix}} := \{a^1_k \mid k \in \mathcal{K}\} \subseteq A$ and refer to their arcs as *fixed arcs*. In network N^Λ, a feasible *robust $\boldsymbol{\beta}$-flow* $\boldsymbol{f} = (f^1, \ldots, f^{|\Lambda|})$ is defined by feasible β^λ-flows $f^\lambda : A \to \mathbb{Z}_{\geq 0}$ for all scenarios $\lambda \in \Lambda$ that satisfy the following equal flow property

$$(F) \quad f^\lambda(a) = f^{\lambda'}(a) \text{ for all } a \in A^c_{\text{fix}} \cup A^a_{\text{fix}},\ \lambda, \lambda' \in \Lambda.$$

Analogous to Section 4.1, a robust $\boldsymbol{\beta}$-flow in network N^Λ corresponds to a solution to the rMDP. To obtain an optimal solution, we aim at a feasible robust $\boldsymbol{\beta}$-flow $\boldsymbol{f} = (f^1, \ldots, f^{|\Lambda|})$ in network N^Λ that minimizes the following expression

$$\max_{\lambda\in\Lambda} \max_{k\in\mathcal{K}} |t \cdot f^\lambda(a^2_k) - u \cdot c_k|. \tag{2}$$

By an adjusted version of Algorithm 1, we can determine the desired robust $\boldsymbol{\beta}$-flow $\boldsymbol{f}$ in network $N^\Lambda_\omega = (G, \overline{\psi}, \overline{\psi}_\omega, \underline{\psi}_\omega, \boldsymbol{\beta})$ with minimum parameter $\omega \in \Omega$, which additionally minimizes the unsuccessful treatment requests. However, we note that there generally does not exist a polynomial-time algorithm to compute a feasible robust flow on acyclic digraphs [44]. For this reason, we determine the

robust flows computed in the adjusted algorithm by means of the following compact IP formulation.

$$\min \sum_{\lambda\in\Lambda}\sum_{k\in\mathcal{K}}\sum_{(\sigma_k^a,\sigma_k^w)\in A} f_a^\lambda \tag{3}$$

$$\text{s.t.} \sum_{a=(v,w)\in A} f_a^\lambda - \sum_{a=(w,v)\in A} f_a^\lambda = \beta^\lambda(v) \quad \forall\, v\in V,\ \lambda\in\Lambda \tag{4}$$

$$0 \le f_a^\lambda \le \psi(a) \quad \forall\, a\in A, \lambda\in\Lambda \tag{5}$$

$$\underline{\psi}_\omega(a) \le f_a^\lambda \le \overline{\psi}_\omega(a) \quad \forall\, a=a_2^k\in A, \lambda\in\Lambda \tag{6}$$

$$f_a^\lambda = f_a^{\lambda'} \quad \forall\, a\in A_{\text{fix}}^a\cup A_{\text{fix}}^c,\ \lambda,\lambda'\in\Lambda \tag{7}$$

$$f_a^\lambda \in \mathbb{Z}_{\ge 0} \quad \forall\, a\in A, \lambda\in\Lambda \tag{8}$$

Let $f_a^\lambda \in \mathbb{Z}_{\ge 0}$ be the variables that model the flow on arc $a \in A$ in scenario $\lambda \in \Lambda$. The IP can be formulated as presented in Equations 3–8. Objective (3) minimizes the flow value that represents unsuccessful treatment requests. The flow balance constraints (4) and capacity constraints (5) ensure feasible flows in all scenarios. Capacity constraints (6) guarantee the desired balanced flow on the treatment arcs. The so-called consistent flow constraints (7) enforce the desired equal flow values on fixed arcs. Constraints (8) define the domain of the variables.

Overall, we obtain the following theorem.

Theorem 2 *A robust mask determined by a robust $\boldsymbol{\beta}$-flow obtained by the adjusted Algorithm 1 is of minimum cost.*

Proof Analogous to the proof of Theorem 1. □

We refer to the resulting model and procedure presented in this section as *Rob-F* and conclude with the following corollary.

Corollary 2 *Rob-F is a modeling for the rMDP.*

5 Robust multimask design problem and its modeling

In the previous section, we have considered the rMDP where the weekly distribution of the acute, walk-in, and no-show demands are uncertain, but the respective total demand remains the same in every scenario. Naturally, not only the distribution but also the total demand is uncertain and fluctuates. Under such circumstances, a robust mask may produce poor results with respect to our objective. For example, if many slots need to be reserved for appointments for acute patients in one scenario due to high demand and only a few in another scenario due to low demand, a robust mask would achieve a trade-off to minimize the worst-case scenario. This could lead to poor results for both scenarios, in contrast to a mask that is adapted to the level of demand. To address this issue, from now on we assume that the acute, walk-in, and no-show demands are specified by parameters $\boldsymbol{d}_\lambda^a \in \mathbb{Z}_{\ge 0}^{\mathcal{K}}$, $\boldsymbol{d}_\lambda^w \in \mathbb{Z}_{\ge 0}^{\mathcal{K}}$, and $\boldsymbol{d}_\lambda^n \in \mathbb{Z}_{\ge 0}^{\mathcal{K}}$ for all scenarios $\lambda \in \Lambda$, respectively, without further restrictions. In response to the new setting, we extend the rMDP and employ not only a single mask but multiple robust masks, ideally each feasible for a cluster of scenarios. To flexibly react to changes in demand, we aim at finding multiple robust masks that PCPs can switch between. For instance, if a PCP uses a mask that is tailored to low demand and notices an increase, they can switch to a mask designed for that level of demand. To facilitate the switch from a mask for lower demand to a mask for high demand, we require a nested construction as shown in the following definition. We use the notation $[n] := \{1,\dots,n\}$.

Definition 9 Masks $\mu^1,\dots,\mu^n$ are called *nested* if

(1) they have the same number of chronic slots in every session $k \in \mathcal{K}$, i.e., $n_{\mu^i}(c,k) = n_{\mu^j}(c,k)$ for all $i, j \in [n]$ and
(2) mask μ^j has at least as many acute slots as mask μ^i for $i < j$ in every session $k \in \mathcal{K}$, i.e., $n_{\mu^i}(a,k) \le n_{\mu^j}(a,k)$ for $i < j$ with $i, j \in [n]$.

With nested masks $\mu^1,\dots,\mu^n$, a PCP can always start using mask μ^1 and easily switch to another mask μ^j with $j > 1$. We note that the number of masks can be selected as required. However, if the number of masks is large, the masks will not differ significantly from each other but only by single slots. It is therefore advisable to adjust the number of masks to the fluctuation in demand. Given this new setting, we define a robust multimask.

Definition 10 (Robust Multimask) A *multimask* $\boldsymbol{\mu} = (\mu^1,\dots,\mu^n)$ is defined by a tuple of nested masks $\mu^i : \mathcal{S} \to \{c,a,w\}$, $i \in [n]$. It is called *robust* if a $(d^c, \boldsymbol{d}_\lambda^a)$-assignment exists for each scenario $\lambda \in \Lambda$ that is compatible with at least one of the masks of $\boldsymbol{\mu}$. The set of all $(d^c, \boldsymbol{d}_\lambda^a)$-assignments that are compatible with at least one mask of $\boldsymbol{\mu}$ is denoted by $\mathcal{A}_\lambda^{\boldsymbol{\mu}}$, $\lambda \in \Lambda$.

Before evaluating a robust multimask, we note that the PCP's utilization u_λ depends now on the scenario $\lambda \in \Lambda$, i.e., $u_\lambda = \frac{t}{C}\left(d^c + \sum_{k\in\mathcal{K}}(d_{\lambda,k}^a + d_{\lambda,k}^w - d_{\lambda,k}^n)\right)$.

Definition 11 Let the PCP's capacity $\boldsymbol{c} \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$, demand $d^c \in \mathbb{Z}_{\geq 0}$, and demands $\boldsymbol{d}^a_\lambda, \boldsymbol{d}^w_\lambda, \boldsymbol{d}^n_\lambda \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ for all scenarios $\lambda \in \Lambda$ be given. The *cost* of a robust multimask $\boldsymbol{\mu}$ is defined as

$$c(\boldsymbol{\mu}) = \max_{\lambda \in \Lambda} \min_{\boldsymbol{\alpha} \in \mathcal{A}^{\boldsymbol{\mu}}_\lambda} \max_{k \in \mathcal{K}} \left| \Phi^{\boldsymbol{\alpha}}_{\lambda,k} - u_\lambda \cdot c_k \right|.$$

Finally, we can define the rMMDP.

Definition 12 (rMMDP) Let $\boldsymbol{c} \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ be the PCP's capacity, $d^c \in \mathbb{Z}_{\geq 0}$ be the chronic demand, and $\boldsymbol{d}^a_\lambda, \boldsymbol{d}^w_\lambda, \boldsymbol{d}^n_\lambda \in \mathbb{Z}^{\mathcal{K}}_{\geq 0}$ be the acute, walk-in, and no-show demands for all scenarios $\lambda \in \Lambda$, respectively. The *robust multimask design problem* aims at finding a robust multimask of minimum cost.

In the following, we also model the rMMDP by a network flow and design model. Therefore, let $\pi : \Lambda \to \{\mu^1, \ldots, \mu^n\}$ be a function that allocates one mask to each scenario. In network N^Λ, a feasible *nested robust $\boldsymbol{\beta}$-flow* $\boldsymbol{f} = (f^1, \ldots, f^{|\Lambda|})$ is defined by feasible β^λ-flows $f^\lambda : A \to \mathbb{Z}_{\geq 0}$ for all scenarios $\lambda \in \Lambda$ that satisfy the flow properties presented in $(F1) - (F3)$:

$(F1)$ $f^\lambda(a) = f^{\lambda'}(a)$ for all $a \in A^c_{\text{fix}}$ and $\lambda, \lambda' \in \Lambda$,

$(F2)$ $f^\lambda(a) = f^{\lambda'}(a)$ for all $a \in A^a_{\text{fix}}$ and $\lambda, \lambda' \in \Lambda$ with $\pi(\lambda) = \pi(\lambda')$,

$(F3)$ $f^\lambda(a) \leq f^{\lambda'}(a)$ for all $a \in A^a_{\text{fix}}$ and $\lambda, \lambda' \in \Lambda$ where $\lambda \in \pi^{-1}(\mu^i)$, $\lambda' \in \pi^{-1}(\mu^j)$ with $i \leq j$.

Flow property $(F1)$ ensures that the flow value on a fixed arc contained in set A^c_{fix} is equal among all scenarios. Flow property $(F2)$ ensures that the flow value on a fixed arc contained in set A^a_{fix} is equal for all scenarios which are assigned to the same mask. Flow property $(F3)$ ensures non-decreasing flow values on fixed arcs contained in set A^a_{fix} among scenarios assigned to different masks. Analogous to the previous sections, a nested robust $\boldsymbol{\beta}$-flow in network N^Λ that minimizes the following expression

$$\max_{\lambda \in \Lambda} \max_{k \in \mathcal{K}} |t \cdot f^\lambda(a^2_k) - u_\lambda \cdot c_k|$$

corresponds to a solution to the rMMDP. In contrast to the rMDP, the solution space of the rMMDP differs among the scenarios as the PCP's utilization depends on the scenario considered. The possible solution values form the finite set $\Omega^\Lambda = \bigcup_{\lambda \in \Lambda, k \in \mathcal{K}} \Omega_{\lambda,k}$ with

$$\Omega_{\lambda,k} = \left\{ |t \cdot i - u_\lambda \cdot c_k| \,\middle|\, i \in \{0, 1, \ldots, D_{\lambda,k}\} \right\},$$

where $D_{\lambda,k} = d^w_{\lambda,k} - d^n_{\lambda,k} + \sum_{k' \in \mathcal{K}^-_k} d^a_{\lambda,k'} + d^c$. Analogous to the previous section, we use an adjusted version of Algorithm 1 to find the minimum parameter $\omega \in \Omega^\Lambda$ for which a feasible nested robust $\boldsymbol{\beta}$-flow $\boldsymbol{f}$ exists in network N^Λ_ω. We determine a feasible nested robust flow using a compact IP formulation.

$$\min \sum_{\lambda \in \Lambda} \sum_{k \in \mathcal{K}} \sum_{(\sigma^a_k, \sigma^w_k) \in A} f^\lambda_a \tag{9}$$

$$\text{s.t.} \sum_{a=(v,w) \in A} f^\lambda_a - \sum_{a=(w,v) \in A} f^\lambda_a = \beta^\lambda(v) \quad \forall\, v \in V,\ \lambda \in \Lambda \tag{10}$$

$$0 \leq f^\lambda_a \leq \psi(a) \quad \forall\, a \in A,\ \lambda \in \Lambda \tag{11}$$

$$\underline{\psi}_\omega(a) \leq f^\lambda_a \leq \overline{\psi}_\omega(a) \quad \forall\, a = a^k_2 \in A, \lambda \in \Lambda \tag{12}$$

$$f^\lambda_a = f^{\lambda'}_a \quad \forall\, a \in A^c_{\text{fix}},\ \lambda, \lambda' \in \Lambda \tag{13}$$

$$\sum_{i=1}^{n} y^i_\lambda = 1 \quad \forall\, \lambda \in \Lambda \tag{14}$$

$$f^\lambda_{a^1_k} \leq m^i_k + n_k(1 - y^i_\lambda) \quad \forall\, i \in [n],\ k \in \mathcal{K},\ \lambda \in \Lambda \tag{15}$$

$$f^\lambda_{a^1_k} \geq m^i_k - n_k(1 - y^i_\lambda) \quad \forall\, i \in [n],\ k \in \mathcal{K},\ \lambda \in \Lambda \tag{16}$$

$$m^i_k \leq m^{i+1}_k \quad \forall\, i \in [n-1],\ k \in \mathcal{K} \tag{17}$$

$$f^\lambda_a \in \mathbb{Z}_{\geq 0} \quad \forall\, a \in A,\ \lambda \in \Lambda \tag{18}$$

$$m^i_k \in \mathbb{Z}_{\geq 0} \quad \forall\, i \in [n],\ k \in \mathcal{K} \tag{19}$$

$$y^i_\lambda \in \{0, 1\} \quad \forall\, i \in [n],\ \lambda \in \Lambda \tag{20}$$

Let $f^\lambda_a \in \mathbb{Z}_{\geq 0}$ be the variables that model the flow on arc $a \in A$ in scenario $\lambda \in \Lambda$. Let $m^i_k \in \mathbb{Z}_{\geq 0}$, $i \in [n]$ be auxiliary variables that indicate the amount of flow on arcs $a^1_k \in A$. Further, let $y^i_\lambda \in \{0, 1\}$ be binary variables that determine whether a mask μ^i, $i \in [n]$ is allocated to scenario $\lambda \in \Lambda$ ($y^i_\lambda = 1$) or not ($y^i_\lambda = 0$), corresponding to function π. The IP can be formulated as presented in Equations 9–20. Objective (9) minimizes the flow value that represents unsuccessful treatment requests. The flow balance constraints (10) and capacity constraints (11) ensure feasible flows in all scenarios. Capacity constraints (12) guarantee the desired balanced flow on the treatment arcs. Constraints (13) enforce the equal flow values on arcs $a \in A^c_{\text{fix}}$ among all scenarios as required in flow property $(F1)$. Constraints (14) model that one mask is assigned to each scenario which corresponds to function π. Constraints (15) and (16) enforce the equal flow values on arcs $a \in A^a_{\text{fix}}$ (in amount of m^i_k) for two scenarios $\lambda, \lambda' \in \Lambda$ if the scenarios are assigned to the same mask, i.e., $\mu^i = \pi(\lambda) = \pi(\lambda')$, $i \in [n]$ ($y^i_\lambda = y^i_{\lambda'} = 1$), as required in flow property $(F2)$. Otherwise, they function

only as upper and lower bounds on the flows. The combination of constraints (15)–(17) enforce the non-decreasing flow values on arcs $a \in A^a_{\text{fix}}$ for two scenarios $\lambda, \lambda' \in \Lambda$ if the scenarios are assigned to different masks μ^i, μ^j with $i \leq j$, i.e., $\mu^i = \pi(\lambda)$ and $\mu^j = \pi(\lambda')$ ($y^i_\lambda = y^j_{\lambda'} = 1$), as required in flow property ($F3$). Finally, constraints (18)–(20) define the domain of the variables.

Overall, we obtain the following theorem.

Theorem 3 *A robust multimask determined by a nested robust $\boldsymbol{\beta}$-flow obtained by the adjusted Algorithm 1 is of minimum cost.*

Proof Analogous to the proof of Theorem 1. □

We refer to the extended model and procedure presented in this section as *RobMulti-F* and conclude with the following corollary.

Corollary 3 *RobMulti-F is a modeling for the rMMDP.*

6 Case study

In this section, we present a case study conducted to demonstrate the potential of the mask-based appointment scheduling systems obtained by solving MDP, rMDP, and rMMDP. In Section 6.1, we present the experimental environment for the case study. In Section 6.2, we present the setting of the implementation and computation. In Section 6.3, we discuss the results of the optimization models. In Section 6.4, we evaluate the mask-based appointment scheduling systems and compare them with some appointment scheduling systems from the literature using agent-based simulation.

6.1 Experimental environment

In this section, we present the experimental environment for the case study. This includes the physician-patients setting, the simulation model and its input data, and the appointment scheduling systems. In addition, this includes the instance generation and input parameter for the optimization models. We note that our setting is tailored to a German primary health care scenario on the basis of our knowledge and available data. We use this scenario to show the potential of the mask-based system. Naturally, the mask-based appointment scheduling systems can be tailored to different scenarios and settings as required.

6.1.1 Physician-patients setting

In primary health care, the physician-to-patient ratio is commonly used to measure medical care. Following the German Federal Joint Committee's (GFJC) guideline [49], we set the ratio of one PCP per 1607 patients as a benchmark for analyzing the appointment scheduling systems in this study. The GFJC defines an excess of 10% in the ratio as oversupply, i.e., one PCP per 1448 patients. An undersupply is defined by a 25% shortfall in the ratio, i.e., one PCP per 2011 patients. Representing these three scenarios, we analyze the performance of the appointment scheduling systems for populations of 1400, 1600, and 2000 patients. According to a study of the German Robert Koch Institute [50], 43% of female and 38% of male patients are chronically ill. As our study is gender neutral, we assume that 40% of all patients suffer from chronic illnesses. Due to the worldwide aging population with the highest risk of chronic conditions, an increase in the number of chronically ill patients is expected. Thus, we analyze a chronic proportion of 40% and 45%. Over the past ten years, surveys conducted by the National Association of Statutory Health Insurance Physicians (NASHIP) [19] have revealed that on average 18.8% of the patients holding health insurance are walk-ins that forgo scheduling an appointment with their PCPs. To cover this and slight deviations, we analyze a walk-in rate of 15%, 18%, and 20%. Furthermore, the NASHIP found that the no-show rates of 86% of German primary care practices are only up to 10% [19]. Therefore, it is not surprising that 83% of practices consider no-shows to be no problem, only a very minor problem, or a minor problem. Based on these results, we consider a uniform no-show rate of 10%. In total, we obtain 18 different configurations for the patient population considered in this case study, as summarized in Table 2.

Table 2 Patient configurations

Config. ID	# patients	chronic. ill (%)	walk-in (%)	no-show (%)
1	1400	40	15	10
2	1400	40	18	10
3	1400	40	20	10
4	1400	45	15	10
5	1400	45	18	10
6	1400	45	20	10
7	1600	40	15	10
8	1600	40	18	10
9	1600	40	20	10
10	1600	45	15	10
11	1600	45	18	10
12	1600	45	20	10
13	2000	40	15	10
14	2000	40	18	10
15	2000	40	20	10
16	2000	45	15	10
17	2000	45	18	10
18	2000	45	20	10

Based on a survey regarding the practice's opening hours conducted by The National Association of Statutory Health Insurance Funds [51], we determine the PCP's opening hours as follows. We decide the PCP to operate in a four-hour morning session from Monday to Friday and in a two-and-a-half-hour afternoon session on Mondays, Tuesdays, and Thursdays. This results in a time span $o_{(i,0)} = 240$ for $i \in \{0, \ldots, 4\}$ and $o_{(i,1)} = 150$ for $i \in \{0, 1, 3\}$. The first hour after each session is used as buffer time, i.e., $b_k = 60$ for all $k \in \mathcal{K}$. Accordingly, overtime is caused if the PCP works beyond the capacity $c_{(i,0)} = 300$ for $i \in \{0, \ldots, 4\}$ and $c_{(i,1)} = 210$ for $i \in \{0, 1, 3\}$.

6.1.2 Simulation model SiM-Care and modifications

To generate instances and to analyze the appointment scheduling systems designed in this study, we use the agent-based simulation tool 'SiM-Care' (Simulation Model for Primary Care) [52]. SiM-Care simulates the dynamics of primary health care systems and aims at supporting decision-makers in their planning, analysis, and adaption. For this purpose, it models PCPs, patients, and their interaction on an individual level. Tracking the interactions allows identifying the interdependencies between different subproblems, evaluating new planning approaches, and quantifying the effects of interventions based on multiple performance measurements. SiM-Care enables in particular the comparison and assessment of appointment scheduling systems in terms of, for example, patients' waiting times, patients' access times, or the PCPs' utilization.

SiM-Care relies on stochastic values to approximate real-world variability and control the frequency of events. This applies to aspects of illnesses as well as to patient arrivals, appointment cancellations, and service times. Table 3 lists all probabilistic model aspects of SiM-Care. We refer to the study of Comis et al. [52] for details of the parameterization of the distributions underlying the random values. For the sake of validity, SiM-Care uses real-world data to the extent available. For example, the distribution of illnesses and their characteristics are estimated on the basis of publications of health insurances and federal government agencies. All unavailable data was either empirically collected in a primary care practice, for instance the patients' service times, or where this was not possible, it was inferred as described in more detail in the next section.

Table 3 Basis for the selection of input parameters

Aspects	Distribution
frequency of acute illnesses	exponential distribution
type of acute illnesses	age-class-illness distribution
seriousness of acute illnesses	triangular distribution
duration of acute illnesses	log-normal distribution
patients' willingness to wait	Weibull distribution
patient punctuality	normal distribution
walk-in arrivals	beta distribution
service time	log-normal distribution
appointment cancellations	binomial distribution
appointment/walk-in	discrete probability distr.
show/no-show	discrete probability distr.

To simulate the needs of our primary health care setting, we make the following changes to SiM-Care. First, PCPs may reject patients due to capacity overloads in the current modeling. Rejected patients then seek care at other PCPs. As we only consider one PCP in our study, we integrate an emergency practice to SiM-Care. Accordingly, if patients are rejected by the PCP in our setting, they seek immediate care at a physician in the emergency practice. Second, patients always try to make an appointment first in the current modeling. If they succeed, they arrange appointments, which they later attend or cancel. Otherwise, they seek immediate care as walk-in within their willingness to wait. We note that the patients' willingness to wait in SiM-Care follows a Weibull distribution and depends on factors such as the illness, seriousness of the illness (depending on the patient's health condition) with the resulting expected willingness to wait, and the deviation from the expected willingness to wait due to patient's age. We integrate the feature that patients forgo making an appointment and instead seek immediate care as walk-in. Furthermore, we integrate the feature that patients do not show up to their arranged appointments. The percentages of walk-ins and no-shows can be defined by the user. Third, we assume that walk-ins are subject to the same service time log-normal distribution as patients with appointments.

Before addressing the input parameters for SiM-Care in the next section, we note that SiM-Care is a complex simulation model, much more complex than our optimization models. We have chosen SiM-Care on an operational decision level to prove that the optimization models developed on a tactical decision level can withstand complex systems.

6.1.3 Input parameters for SiM-Care

Comis et al. [52] present a case study for a primary health care system in Germany to showcase SiM-Care and its validation through expert input and empirical data. Specifically, they create a baseline scenario representing a real-world primary health care system. In this study, we conduct a case study based on this baseline scenario. In the following, we detail the most important input parameters and the parameter choices that deviate from the baseline scenario. Table 4 summarizes the selection of input parameters. For further details, we refer to the study of Comis et al. [52].

Table 4 Basis for the selection of input parameters

Attribute	Basis (Source)
PCP's	
opening hours	empirical ([51])
strategies	literature ([52–54])
Patients'	
age class	empirical ([55])
health condition	inferred ([52])
service times	empirical ([52])
Age classes	
exp. annual acute illnesses	inferred ([52])
dev. illness duration	inferred ([52])
dev. willingness to wait	inferred ([52])
availabilities	inferred ([52])
appointment cancellations	inferred ([52])
chronic patients	empirical ([50])
Families of illnesses	
characteristics	inferred ([52])
age-class illness distribution	empirical ([56])

First, we model a time period of one year (52 weeks) preceded by a warm-up period of also one year. Second, we consider randomly generated locations of the patients and physicians as performance indicators such as travel times are not important in our case study. Third, the modeled PCP and their opening hours as well as the modeled patients are as described in Section 6.1.1. Fourth, we assume the same distributions of age classes and health conditions of patients as in the baseline scenario. This includes the estimations of the data of age-class-dependent attributes for each patient as listed in Table 4. Fifth, we consider the same families of illnesses as in the baseline scenario, consisting of three chronic and four acute illnesses as listed in Tables 5 and 6, which is a subset of the 100 ICD-10 codes most frequently reported to the Association of Statutory Health Insurance Physicians Nordrhein [57].

Sixth, we apply the same estimations for the characteristics of the families of illnesses such as expected willingness to wait (cf. Tables 5 and 6), duration or the treatment frequency, as real data is protected by confidentiality and not published. The only difference is that we assume that patients

Table 5 Chronic illness families $f \in \mathcal{F}$ with respective expected willingness to wait (WTW) depending on seriousness $s \in [0, 1]$

ICD	Name	exp. WTW W_f
I10	high blood pressure	$W_f(s) = -10s + 20$
E11	diabetes	$W_f(s) = -4s + 14$
I25	ischemic heart disease	$W_f(s) = -4s + 10$

Table 6 Acute illness families $f \in \mathcal{F}$ with respective expected willingness to wait (WTW) depending on seriousness $s \in [0, 1]$

ICD	Name	exp. WTW W_f
E78	high cholesterol level	$W_f(s) = -5s + 8$
M54	back pain	$W_f(s) = -3s + 4$
Z25	vaccination	$W_f(s) = 40$
J06	cold	$W_f(s) = -2s + 2$

with acute illnesses do not require a follow-up treatment (i.e., the treatment frequency is not applicable). Seventh, we consider the same data for the age-class illness distribution estimated on the basis of the reported incidence rates of a large German health insurer [56]. Finally, we present the integrated appointment scheduling, the treatment, and the admission strategy employed by the PCP modeled in SiM-Care. The integrated appointment scheduling strategies are detailed in Section 6.1.4. The treatment strategy is defined as in the baseline scenario [52]. The strategy includes the order of treatments on a priority-first-come-first-served basis: Patients with appointment are prioritized over walk-ins and within their respective groups, patients are served in order of their arrival, i.e., first come, first served. In addition, the treatment strategy includes the PCP's consultation speed, that is, the PCP adjusts the service times whenever more than three patients await treatment. The service times are sampled from a maximum likelihood fitted log-normal distribution based on empirical service times resulting in a mean service time of 7.8 minutes. The admission strategy is defined as follows. Like in studies [53, 54], walk-ins are admitted up to a certain utilization threshold. Therefore, PCPs predict their remaining workload by multiplying the average service time of 7.8 minutes with the number of currently waiting patients and upcoming scheduled appointments. If this estimated workload is lower than the remaining duration of the current session including buffer time, walk-ins are admitted, otherwise rejected. Patients with appointments are admitted as long as they arrive before the end of the buffer time. We note that patients with appointments do not always arrive on time in SiM-Care. Patient arrivals deviate around the scheduled time of the appointment and follows a normal distribution with a mean arrival deviation of -5 minutes and standard deviation of 6 minutes such that roughly 20% of all patients are likely to arrive late for their appointments.

6.1.4 Appointment scheduling systems

In this section, we present the PCP's appointment scheduling strategies, or in other words, the appointment scheduling systems that we implement in SiM-Care for this study. Before presenting the considered appointment scheduling systems, we note that they are based on slot systems. To that end,

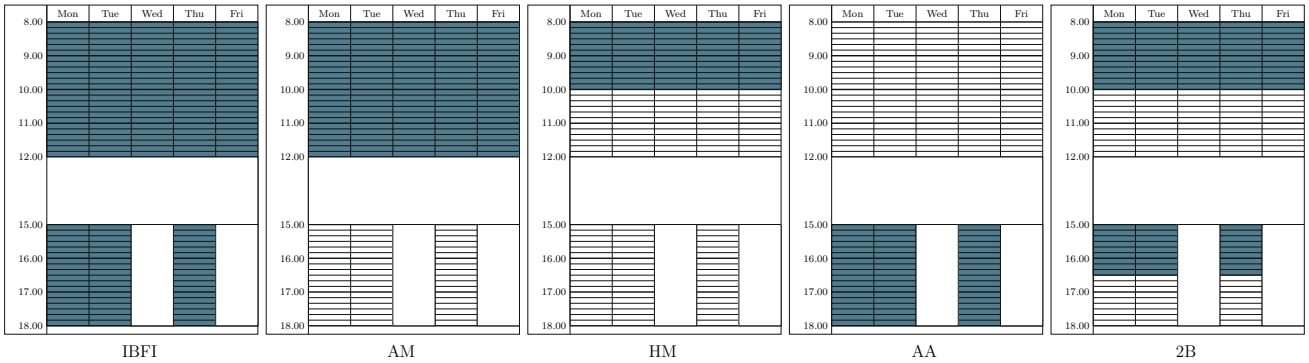

Fig. 4 Non-mask-based systems from the literature with slots ▬ reserved for appointments and slots ▭ withheld for walk-ins

the opening hours of each session are divided into slots of 10 minutes length. As the NASHIP [19] reveals that only 22% of German practices respond to no-shows with overbooking, we do not implement an overbooking strategy. This means that theoretically each slot can accommodate one appointment. However, the appointment scheduling system determines which slots are allowed to accommodate an appointment. In SiM-Care, appointments are offered to patients on a first-come-first-serve basis. Consequently, every patient is offered the earliest feasible appointment at the time of inquiry, i.e., immediately after the onset of illness. If no feasible appointment is available, the patient decides to forgo an appointment and seek immediate care as walk-in. An offered appointment is feasible for a patient if it matches with the patient's availabilities and if it is within the patient's willingness to wait. We note that acute patients may cancel their appointments if they recover before. In the following, we list the implemented appointment scheduling systems. We refer to Fig. 4 which visualizes the five appointment scheduling systems derived from the literature.

Individual-Block/Fixed-Interval (IBFI) Like Comis et al. [52], we consider the IBFI appointment scheduling system. IBFI allows all slots of each session to accommodate an appointment [24, 58]. Thus, no slots are withheld.

All Morning (AM) and Half Morning (HM) Like Balasubramanian et al. [5] and Wiesche et al. [20], we consider the AM and the HM appointment scheduling system as benchmarks. AM allows all slots of each morning session to accommodate an appointment. HM allows the first half of the slots of each morning session to accommodate an appointment. In both appointment scheduling systems, the slots of every afternoon session are withheld for treatments of walk-ins.

All Afternoon (AA) Like Balasubramanian et al. [5], we consider the AA appointment scheduling system as a benchmark. AA allows all slots of each afternoon session to accommodate an appointment. The slots of every morning session are withheld for treatments of walk-ins.

2-Block (2B) We consider the 2B appointment scheduling system introduced by Balasubramanian et al. [5]. 2B allows the slots of the first half of each session to accommodate an appointment. The slots of the second half of each session are withheld for treatments of walk-ins.

Masked-Based (M) and Robust-Mask-Based (RM) We consider the M and the RM appointment scheduling systems that are based on a mask and a robust mask obtained by solving MDP and rMDP, respectively. M and RM allow in each session the chronic and acute slots to accommodate an appointment for one chronic or acute patient, respectively. The walk-in slots of each session are withheld for treatments of walk-ins.

Multimask-Based (MM) We consider the MM appointment scheduling system based on a multimask consisting of three nested masks obtained by solving rMMDP. MM allows in each session the chronic and acute slots to accommodate an appointment for one chronic or acute patient, respectively. Every acute patient is offered the earliest feasible appointment of the first mask at the time of inquiry. If the earliest feasible appointment is not within the patient's willingness to wait, the patient is offered the earliest feasible appointment of the second and if necessary of the third mask. The walk-in slots of each session are withheld for treatments of walk-ins.

We refer to AM, HM, and AA as benchmark systems, to IBFI, AM, HM, AA, 2B as non-mask-based systems, and to M, RM, and MM as mask-based systems. Before we conclude this section, we note that the non-mask-based systems do not distinguish between slots reserved explicitly for acute or chronic patients. To make the comparison between the non-mask-based and mask-based systems fairer, we implement slots reserved for chronic patients in the non-mask-based systems in amount of the average weekly chronic demand

d^c. The chronic slots are evenly distributed among the sessions with slots that allow to accommodate an appointment. All remaining slots that allow to accommodate an appointment are reserved for acute patients. In the AM system, for example, we distribute the chronic slots across all morning sessions. The remaining slots of the morning sessions are reserved for acute patients and the slots of every afternoon session are still withheld for treatments of walk-ins. This adjustment means that all appointment scheduling systems considered in this study offer the same number of chronic slots per week.

6.1.5 Instance generation and input parameters for (robust) mask and robust multimask design problem

For the instances of MDP, rMDP, and rMMDP, demand data is crucial. In the absence of accessible empirical data on visits to primary care practices, we have to rely on simulation to obtain rough estimates. Based on a selected patient configuration presented in Section 6.1.1, we use SiM-Care [52] to simulate the number of visits to a primary care practice by acute and chronic patients as well as the numbers of walk-ins and no-shows over a one-year time horizon. We obtain the demand of 52 weeks and classify it into acute, chronic, walk-in, and no-show demand. We prepare this demand data as input to our optimization problems as explained in the following.

For the MDP, we average the demand of the 52 weeks and obtain the session-specific acute, walk-in, and no-show demands $\boldsymbol{d}^a, \boldsymbol{d}^w, \boldsymbol{d}^n$ and the weekly chronic demand d^c. For the rMDP and the rMMDP, we require different scenarios, where each scenario consists of one week's demand. For the input of both problems, we use the scenario set $\Lambda^{52} = \{\lambda_1, \ldots, \lambda_{52}\}$ that is obtained by considering the demand from each of the 52 weeks of the year as individual scenarios. As shown in the next section, the model RobMulti-F cannot optimally solve all instances to the rMMDP within the given time limit. In addition, we note that there is a risk of overfitting for the robust model Rob-F if the scenario set is too large due to the strict equal flow property (F). The same holds true for RobMulti-F due to the strict equal flow properties $(F1)$ and $(F2)$. For this reason, we use a second scenario set Λ^{13} as input for the rMDP and the rMMDP. Scenario set $\Lambda^{13} = \{\lambda_1, \ldots, \lambda_{13}\}$ is constructed by aggregating the demand of each four consecutive weeks of the 52 weeks to create 13 different scenarios. For the two scenario sets Λ^{52} and Λ^{13}, we perform the following additional processing to obtain valid inputs to the optimization problems. For the rMDP, we standardized demand so that the total acute, walk-in, and no-show demands are equal in all scenarios, i.e., $\sum_{k\in\mathcal{K}} d^a_{\lambda,k} = D^a$, $\sum_{k\in\mathcal{K}} d^w_{\lambda,k} = D^w$, and $\sum_{k\in\mathcal{K}} d^n_{\lambda,k} = D^n$ for all $\lambda \in \Lambda^{52}$, $\lambda \in \Lambda^{13}$. Furthermore, we standardize the chronic demand as it is specified by only one parameter d^c in rMDP. For the rMMDP, we only standardize the chronic demand.

Since SiM-Care relies on stochastic values, our case study includes ten independent runs. Therefore, we perform the demand generation and processing described above ten times for each patient configuration. Given the ten runs for the 18 different patient configurations, we obtain a total of 180 instances for Det-F and, due to the two different scenario sets considered, 360 instances for Rob-F and RobMulti-F. Besides the demand data, the optimization problems MDP, rMDP, and rMMDP require the following input parameters. For each of the three optimization problems, we set the uniform anticipated service time t according to the mean service time in SiM-Care, which averages 7.8 minutes. Adjusted to the anticipated service time, we determine ten-minutes slots, i.e., $\ell = 10$. The capacity and buffer time of the PCP is set as described in Section 6.1.1. For the robust model presented in Section 5, we consider multimasks that consist of three nested masks, i.e., $n = 3$, to generate a mask for weeks with low, medium, and high demand.

6.2 Implementation and computational setting

For our computational study, all computational experiments are performed on a cluster of machines running Ubuntu 20.04.6 with an Intel(R) Core(TM) $i9-9900$ CPU @ 3.10 GHz and 32 GB DDR4-Non-ECC main memory. We restrict each individual job to one physical core, 14 GB main memory, and a time of 24 hours. We implement our algorithms based on binary search and IP formulations in Java using OpenJDK 11 [59]. To solve the IP formulations, we use CPLEX 22.1 through the Java API [60]. We restrict the CPLEX optimizer to an one hour time limit and 32 GB memory limit and leave all other CPLEX parameters at their default settings.

We solve the MDP, the rMDP, and the rMMDP by using the models Det-F, Rob-F, and RobMulti-F, respectively. Each model includes two optimization stages as presented in Sections 4.1, 4.2, and 5. In theory, we can solve both stages at once by determining flows of minimum cost instead of feasible flows. In contrast, when solving the models, we consider the stages separately and sequentially to speed up the computations. In the first optimization stage, we aim at finding the minimum parameter $\omega \in \Omega$ for which a feasible flow in network N_ω exists by binary search. Once the minimum parameter $\omega \in \Omega$ is found, we proceed with the

second optimization stage. In the second optimization stage, we aim at finding a minimum cost flow, robust flow, or robust nested flow in network N_ω. We recall that there does not exist a polynomial-time algorithm to determine a robust or a robust nested flow. Therefore and for simplicity's sake, we compute all flows of all models, including the deterministic model, by IPs. While runtime is clearly not a focus of this study, we note the following regarding the solvability. The first optimization stage is performed in a few seconds for most instances and in less than three and a half hours for every instance. Consequently, for all instances an optimal solution to the first stage and thus a feasible solution to the second stage is found. The second optimization stage can be solved to optimality (CPLEX default MIP gap tolerance 10^{-4}) for 66 out of the 180 instances based on scenario set Λ^{52} within the given time limit of one hour. For the remaining 114 instances, the best solution found so far is considered. As described in Section 6.1.5, due to non-solvability within the given time limit, we additionally consider instances based on scenario set Λ^{13} as input for the optimization problems Rob-F and RobMulti-F. For scenario set Λ^{13}, all instances are solved to optimality within the given time limit.

6.3 Evaluation of the optimization models and the resulting masks

In this section, we analyze the results of the Det-F, the Rob-F, and the RobMulti-F model. Before we begin, we recall the following. The solution of each model provides us with the following three results. First, we obtain the objective value $\omega \in \mathbb{Z}_{\geq 0}$ for which the respective flow with minimum cost is found in the corresponding network. Second, we obtain the number of the unsuccessful appointment requests ($d_w^a(\cdot)$) by the cost of the flow. Third, we obtain a mask (of cost ω) and thus the number of chronic, acute, and walk-in slots ($\sum_{k\in\mathcal{K}} n_\mu(\cdot, k)$) by the flow itself. For all instances, the average results of the three models are summarized in Tables 7–8 in Appendix B.

We start by considering the objective value ω. For each model, the average objective value of the ten runs is visualized for all patient configurations and both scenario sets in Fig. 5. We recall that the results of Det-F are not computed based on the demands of the scenario sets Λ^{52} and Λ^{13}, instead the input is the average annual demand. The double visualization is only for the comparability of the models. For Rob-F and RobMulti-F, the average objective value ω is affected by the choice of the scenario set. We confirm that the results for scenario set Λ^{52} are significantly higher than those for scenario set Λ^{13}. This meets our expectations as the computation of the minimum parameter ω becomes more difficult the more scenarios are considered in the sense that the parameter must be suitable for more scenarios. Accordingly, Det-F achieves the best average objective values since the minimum parameter ω is determined for only a single demand scenario. We confirm that the patient configuration has no influence on these results. Considering the fact that the objective value ω corresponds to the cost of an optimal mask, we have to pay on average higher cost of a robust or robust nested mask compared to a non-robust mask, the so-called price of robustness [61]. RobMulti-F achieves for both scenario sets on average better objective values than Rob-F. This also meets our expectation due to the greater flexibility of a robust nested multimask compared to a robust mask. Furthermore, Rob-F and RobMulti-F have the tendency that

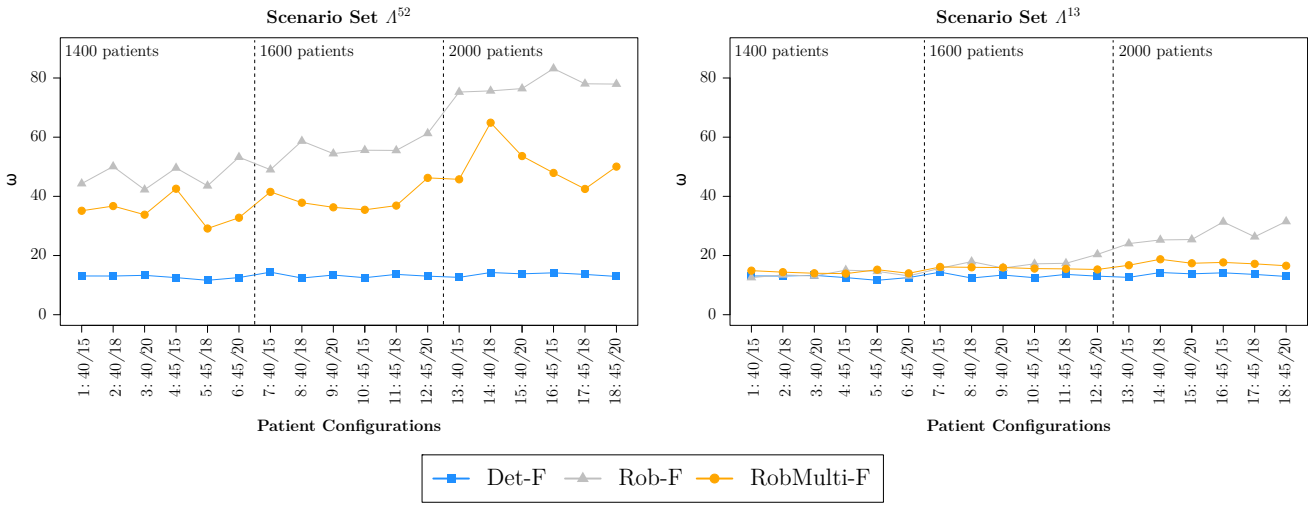

Fig. 5 Average cost of masks ω depending on the scenario set (Λ^{52}, Λ^{13}), patient configuration (ID: chronic/walk-in), and optimization model (Det-F, Rob-F, RobMulti-F)

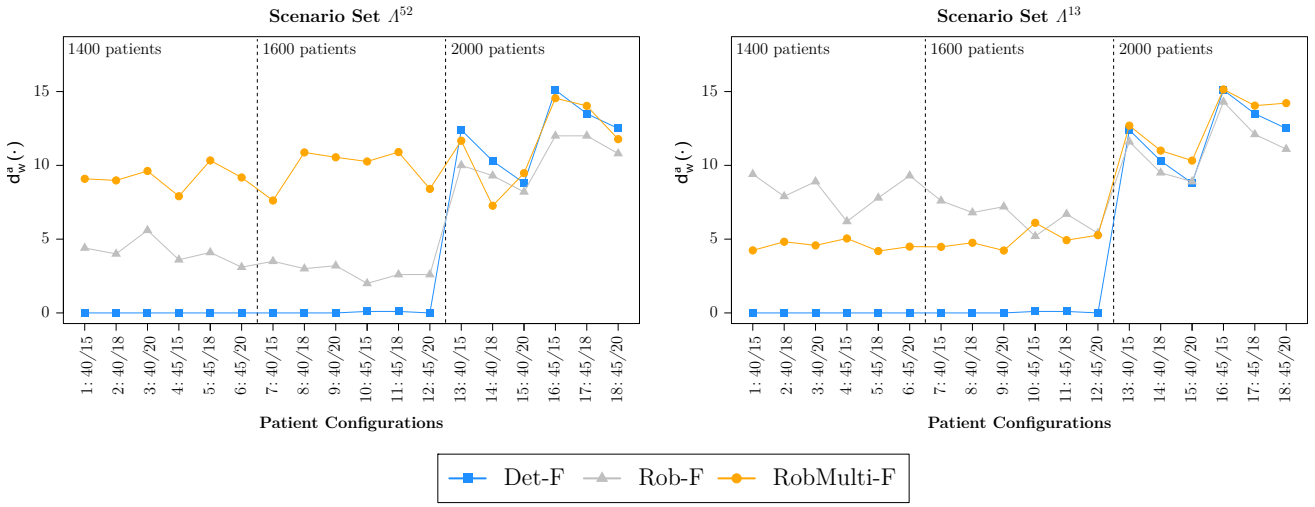

Fig. 6 Average number of unsuccessful appointment requests $d^a_\omega(\cdot)$ depending on scenario set (Λ^{52}, Λ^{13}), patient configuration (ID: chronic/walk-in), and optimization model (Det-F, Rob-F, RobMulti-F)

the average objective value ω increases (with exceptions) the more patients are considered (patient configurations $1-6$ versus $7-12$ versus $13-18$). Finally, recall that parameter ω or in other words the cost of a corresponding mask indicates the deviation of the PCP's actual workload from the optimal workload using the respective model.

In the next step, we consider the total number of unsuccessful appointment requests ($d^a_w(\cdot)$). For each model, we average the total number of unsuccessful appointment requests over the number of scenarios to obtain the average number per scenario. We visualize the average of the ten runs for all patient configurations and both scenario sets in Fig. 6. Det-F satisfies all appointment requests for patient configurations $1-6$ and misses only a maximum average of 0.1 requests for patient configurations $7-12$. In contrast, it performs significantly worse for the patient configurations $13-18$ based on the population of 2000 patients with an average of 8.8 to 15.1 unsuccessful appointment requests. For Rob-F and RobMulti-F, the average number of unsuccessful appointment requests is affected by the choice of the scenario set. For patient configurations $1-12$, the results of Rob-F are clearly lower than those of RobMulti-F for scenario set Λ^{52} while almost the opposite is the case for scenario set Λ^{13}. We explain this by the following two reasons. First, as RobMulti-F is not optimally solved for all instances based on scenario set Λ^{52} (unlike set Λ^{13}), only the best results found so far are considered for the total number of unsuccessful treatment requests. Second, the results of the objective values ω of Rob-F (cf. Fig. 5) are significantly higher for instances based on scenario set Λ^{52} than set Λ^{13}. Consequently, the solutions to Rob-F based on scenario set Λ^{52} allow a higher deviation of the actual from the optimal PCP's workload which in turn provides more flexibility for appointments so that there are fewer unsuccessful appointment requests. Like Det-F, Rob-F and RobMulti-F perform for patient configurations $13-18$ significantly worse than for all other configurations. Consistent with the results, we expect that if the total number of patients increases while the PCP's capacity remains the same, the number of unsuccessful appointment requests increases. A second result we expect is that if the walk-in rate increases, the number of patients that request an appointment decreases. Consequently, the number of unsuccessful appointment requests also decreases. However, this effect can only be observed in the results of all three models for configurations $13-18$. We conclude with the following note. The number of unsuccessful appointment requests provides us with the information that, given an allowed maximum deviation, there is theoretically no assignment of the patient demand that would satisfy every appointment requests.

Finally, we consider the structure of the masks, robust masks, and nested robust masks that result from the optimization models Det-F, Rob-F, and RobMulti-F, respectively. Before discussing the results, we note that there are in total $|\mathcal{S}| = 165$ slots within the PCP's capacity. Furthermore, we note that in all three models, d^c out of the 165 slots are assigned the state 'chronic'. Consequently, the number of available slots for state 'acute' or 'walk-in' is $165 - d^c$. In the following, we only compare the number of acute slots of the masks among all models. For each model, the total number of acute slots is averaged over the ten runs and visualized for all patient configurations and both scenario sets in Fig. 7.

First of all, we confirm that the number of slots available for state 'acute' decreases if the population of patients and thus the proportion of chronic patients increases. Moreover,

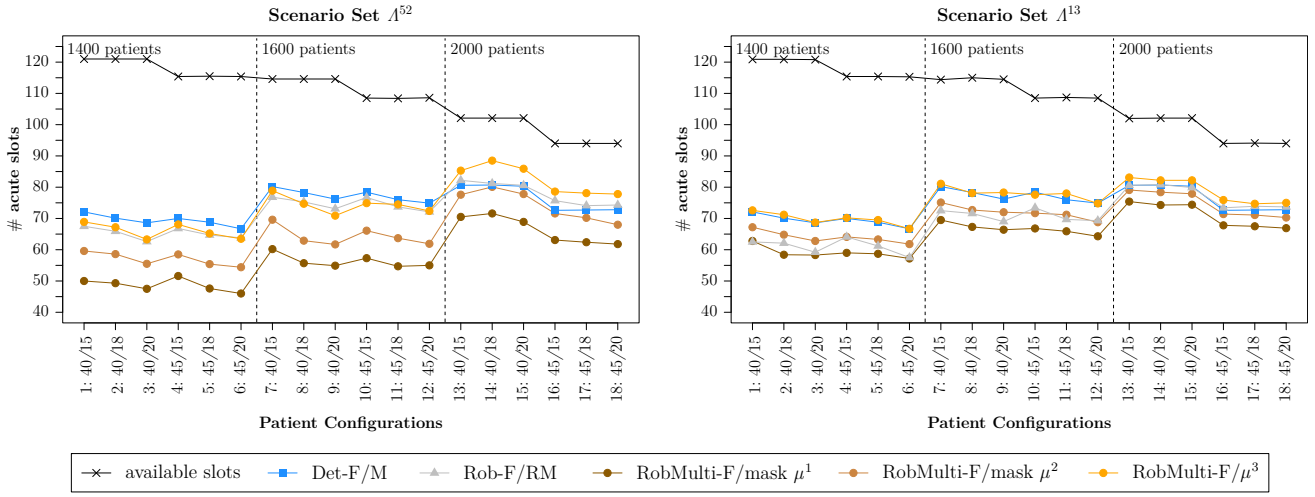

Fig. 7 Average number of acute slots of the masks depending on the scenario set (Λ^{52}, Λ^{13}), patient configuration (ID: chronic/walk-in), and optimization model (Det-F, Rob-F, RobMulti-F)

we note that none of the models assigns all available slots to the state 'acute'. Conversely, in all models there are slots assigned the state 'walk-in'. For Det-F, the average total number of acute slots is at a high level compared to the results of the other models and varies between 66.7 and 80.7, depending on the patient configuration. For Rob-F and RobMulti-F, the total numbers of acute slots are affected by the choice of the scenario set. Before discussing the differences, we confirm for RobMulti-F the resulting nested masks, denoted by μ^1, μ^2, μ^3. Considering Rob-F, we see that the total number of acute slots is higher for the robust mask based on scenario set Λ^{52} than set Λ^{13}. For RobMulti-F based on scenario set Λ^{52}, we note that the total number of acute slots of the third mask is lower than those from Det-F neglecting patient configurations $13-18$. In contrast, for RobMulti-F based on scenario set Λ^{13}, the results almost coincide with those from Det-F. Moreover, we note that the difference of the number of acute slots between the three resulting masks is smaller for RobMulti-F based on scenario set Λ^{13} than on set Λ^{52}. We explain this by the overall fewer and less diverse demand scenarios of set Λ^{13} used for the computation of the solution due to the aggregation of the scenarios. Furthermore, we see that each of the three masks based on scenario set Λ^{13} offer more acute slots than each of the corresponding masks based on set Λ^{52}. All three models have the tendency that the total number of acute slots decreases if the walk-in rate increases as seen in almost all patient configurations, namely $1-3$, $4-6$, $7-9$, $10-12$, and $16-18$.

Before concluding the section, we analyze how often each of the three resulting masks of RobMulti-F is used among all scenarios. Figure 8 visualizes the average frequency over all 10 runs of the assignment of masks μ^1, μ^2, and μ^3 to the scenarios of scenario sets Λ^{52} and Λ^{13} for all patient configurations. In general, we confirm that the average use of the masks μ^1, μ^2, and μ^3 is balanced for all patient configurations, so none of the masks is tailored exclusively for one scenario. Instead, we see that each of the three masks is assigned for a cluster of scenarios as targeted in Section 5. For scenario set Λ^{52}, we note that mask μ^3 (which includes the most acute slots) is assigned more often than mask μ^1 (which includes the fewest acute slots), compared to scenario set Λ^{13}. We explain this with the fact that mask μ^1 based on scenario set Λ^{52} includes less acute slots than mask μ^1 based on scenario set Λ^{13}. At the same time, mask μ^3 based on scenario set Λ^{52} also includes less acute slots than mask μ^3 based on scenario set Λ^{13} for patient configurations $1-12$. We conclude this section with an example of a mask, a robust mask, and a robust nested multimask resulted from the tenth run of patient configuration 8 as visualized in Fig. 9.

6.4 Evaluation by agent-based simulation

In this section, we evaluate the appointment scheduling systems presented in Section 6.1.4. We focus on the comparison of the IBFI and the 2B system with the mask-based systems M, RM, and MM and use the AM, the HM, and the AA system as a benchmark. For the RM and MM appointment scheduling system, we denote whether they are computed on scenario set Λ^{52} or Λ^{13} by RM-52 or RM-13 and MM-52 or MM-13, respectively. In the following, we address RM-52, RM-13, MM-52, and MM-13 as individual appointment scheduling systems. To evaluate the total of ten resulting appointment scheduling systems, we use some of SiM-Care's key performance indicators. For each individual key perfor-

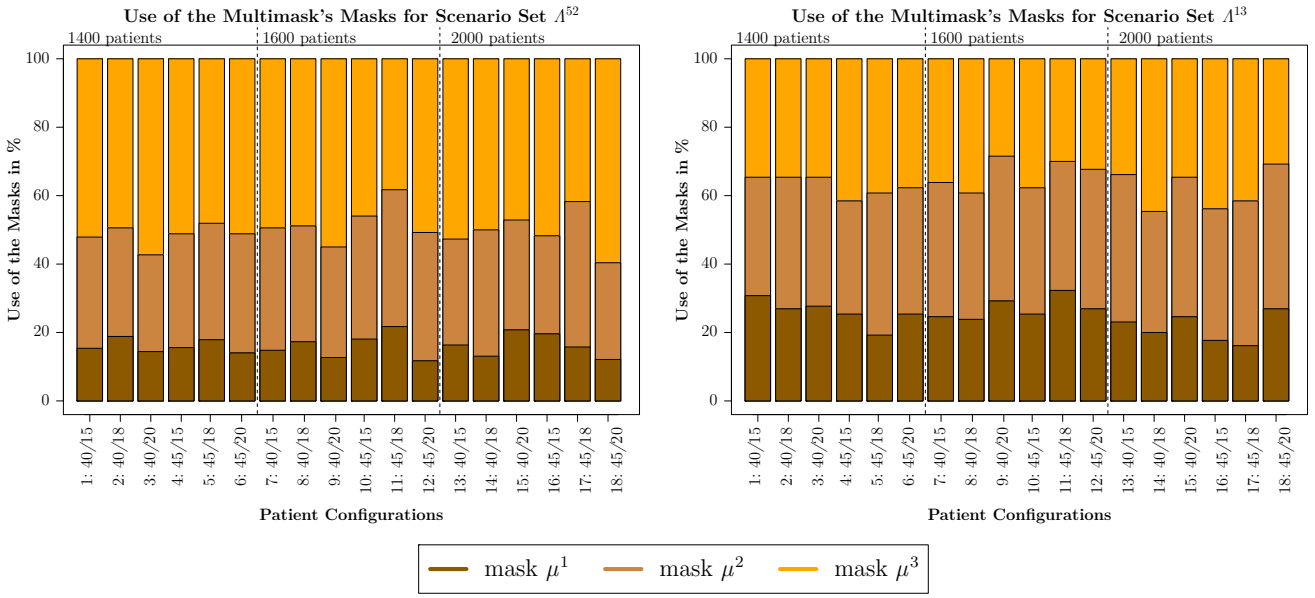

Fig. 8 Percentage use of masks μ^1, μ^2, and μ^3 depending on the scenario set (Λ^{52}, Λ^{13}) and patient configuration (ID: chronic/walk-in)

mance indicator, we summarize the results of all appointment scheduling systems in Tables 9–18 in Appendix B. In addition, for each individual key performance indicator, we visualize below the results for all appointment scheduling systems of all 180 instances by means of box plots, with the average of these results indicated by a star.

We start with evaluating three key performance indicators for the PCP: the average weekly overtime, the average weekly idle time, and the weekly maximum deviation from the actual to the optimal workload. The results of overtime and idle time are visualized in Fig. 10. We confirm that IBFI causes significantly more weekly overtime than all other appointment scheduling systems. The overtime averages 110.46 minutes. These results are consistent with our expectations as IBFI does not reserve time for potential walk-ins. The remaining appointment systems cause overtime at lower level with averages between 14.61 and 37.45 minutes. In particular, we note that the mask-based systems, which cause between 14.61 and 16.41 minutes of average overtime, perform at least 8.71 minutes better than 2B, the best non-masked-based systems with 25.12 minutes of average overtime. We confirm that MM-52 causes the least amount of average overtime at 14.61 minutes. Overall, we note that the average overtime is relatively low which can be explained by the incorporated buffers and the admission of patients up to a certain utilization threshold.

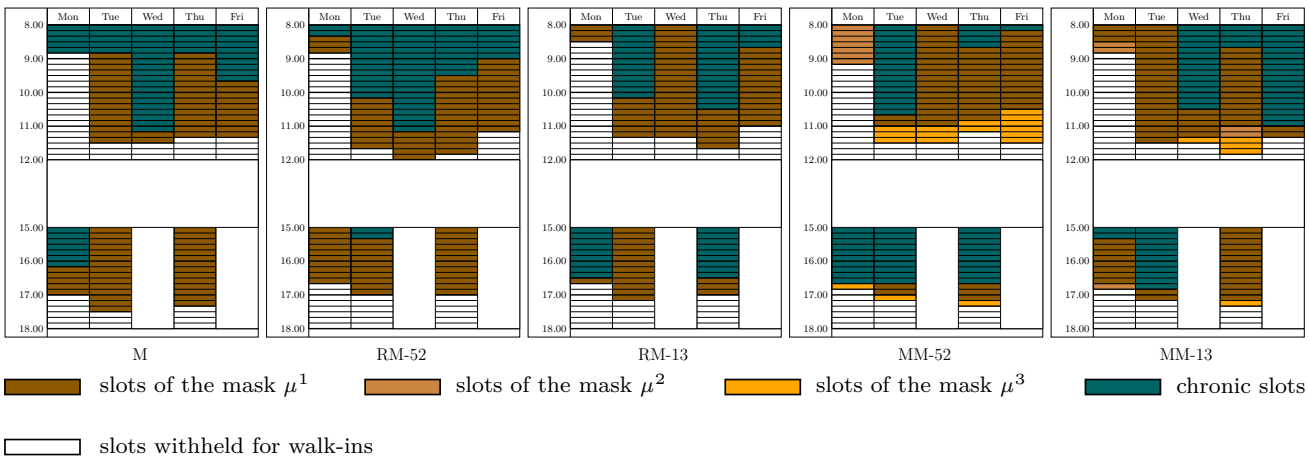

Fig. 9 The resulting mask, robust mask, and robust nested mask of the tenth run for patient configuration 8

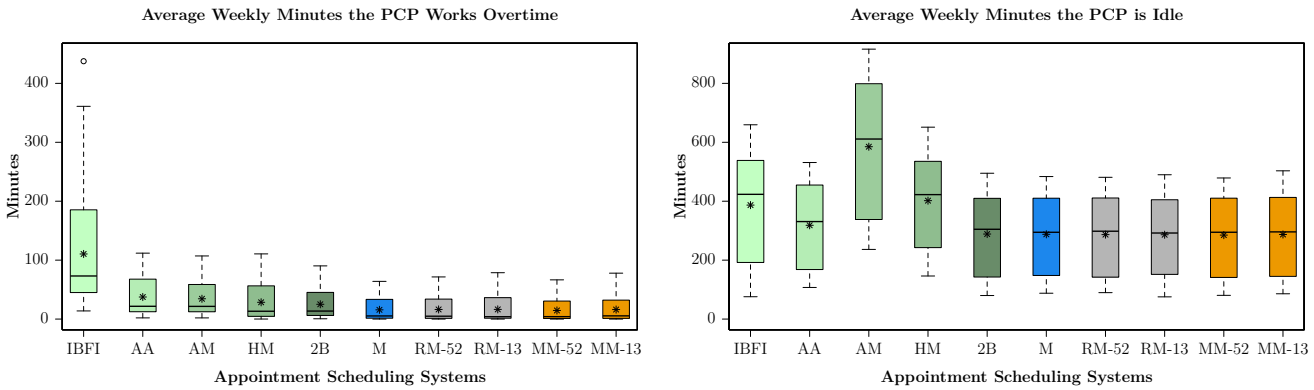

Fig. 10 PCP's average weekly overtime and idle time

Considering the idle time, we confirm that AM causes the most average weekly idle time. However, if we consider idle time in relation to overtime, we rate IBFI as the worst performing system. Overall, MM-52 achieves the lowest average weekly idle time in the amount of 285.02 minutes. In fact, all mask-based systems are at least as good as 2B, the non-mask-based system with the lowest average weekly idle time of 288.63 minutes.

Finally, across all 52 weeks, we consider the worst-case maximum deviation of the actual from the optimal PCP's workload among all sessions as visualized in Fig. 11. This corresponds to the definition of the cost of a (robust) mask for the mask-based systems. IBFI causes the largest average worst-case imbalance between the actual and optimal PCP's workload. This average imbalance is only minimally larger than that caused by AM. Based on the results of overtime and idle time, these results are consistent with our expectations. We further note that HM causes with one instance the maximum worst-case deviation of 0.8072, i.e., the PCP works 0.8072 times the optimal workload more or less in a session.

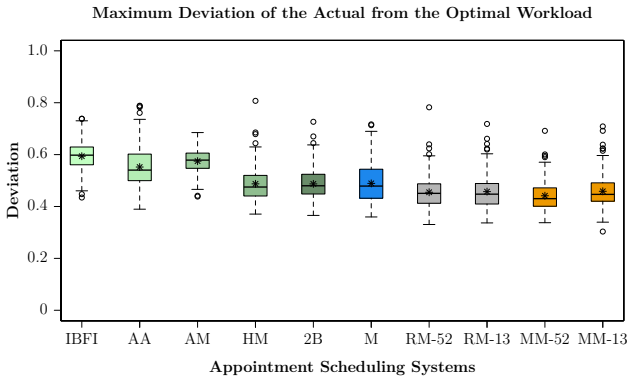

Fig. 11 Worst-case deviations of the 52 weeks of all instances, i.e., in each case the worst-case deviation of the actual from the optimal PCP's workload among all sessions

If we consider the non-mask-based systems, we confirm that 2B performs best on average. However, 2B performs on average worse than all mask-based systems except M. Among all appointment scheduling systems, MM-52 causes the smallest average worst-case deviation of 0.4413. In conclusion, compared to the appointment scheduling systems from the literature, the mask-based systems achieve to balance the utilization of the PCP 's capacity as targeted in Section 3.3.

In the next step, we consider seven key performance indicators for patients. First, we discuss the average access time to an appointment of acute and chronic patients as visualized in Fig. 12. As expected, IBFI ensures the shortest average access time to an appointment for acute and chronic patients. For instance, for acute patients the maximum and average access times to an appointment are 1.49 and 1.22 days, respectively. Clearly, the benchmark systems AA, AM, HM perform the worst as the slots that allow to accommodate appointments are limited to specific sessions. We confirm that the mask-based systems ensure a faster average access time to an appointment, ranging from 2.30 to 2.57 days, than the 2B system with 2.83 days. However, the average access time to an appointment increases for chronic patients if a mask-based system is used compared to 2B. We explain this by the fact that the non-mask-based systems equally distribute appointment slots for chronic patients among all sessions. If the appointment slots for chronic patients are spread only over a few sessions, as it is the case for the mask-based systems, this may not meet patient preferences and might lead to longer access times. In conclusion, the mask-based systems ensure for all instances a maximum access time to an appointment for acute and chronic patients of three and four days, respectively. This meets the objectives formulated in Section 3.1. At this point, we emphasize that these short access times are guaranteed, although the patients' willingness to wait in SiM-Care is not limited to three days as assumed in the optimization models. In this study, the average willingness to

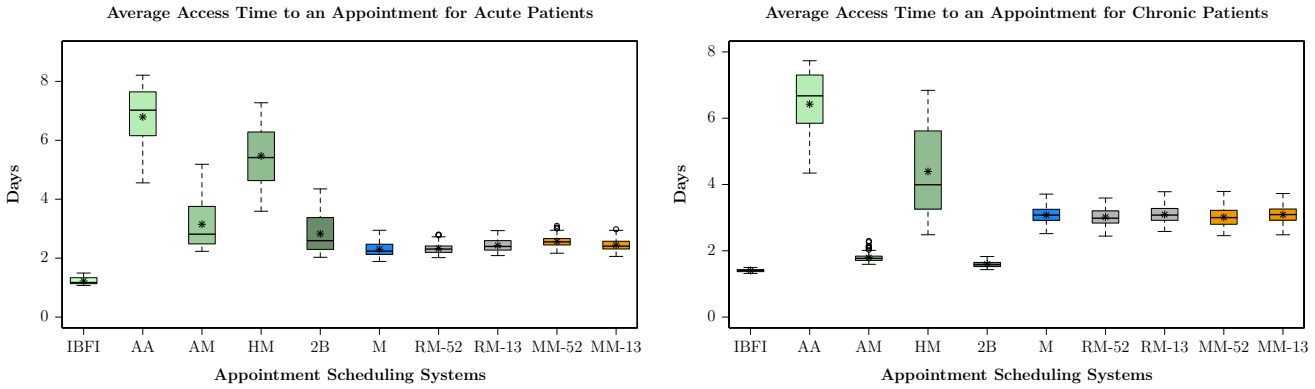

Fig. 12 Average Access time to an appointment for acute and chronic patients

wait for appointments of acute and chronic patients has been 9.39 and 12.43 days in SiM-Care, respectively. We note that the average willingness to wait of acute patients has been only 6.49 days if the illness 'vaccination' with the expected willingness to wait of 40 days (cf. Table 6) is omitted.

Second, we consider the total number of unsuccessful appointment requests within a year as visualized in Fig. 13. As IBFI allows all slots to accommodate an appointment, it clearly ensures the smallest number of unsuccessful appointment requests. The results of the benchmark systems AA, AM, and HM confirm that the number of unsuccessful appointment requests increases if appointment slots are restricted to specific sessions or possibly if too few slots exist that allow to accommodate an appointment. We underline that the mask-based systems perform significantly better than 2B. M causes a maximum of 3631 unsuccessful treatment requests which is the maximum number among all mask-based systems and still lower than 4064.61 which is the average number of 2B. Furthermore, MM-52 causes with 2076.05 the smallest average number of unsuccessful treatment requests among all mask-based systems and the second smallest average number among all appointment scheduling systems. Overall, we assess that the number of unsuccessful appointment requests is minimized as targeted in the second optimization stage of the optimization models.

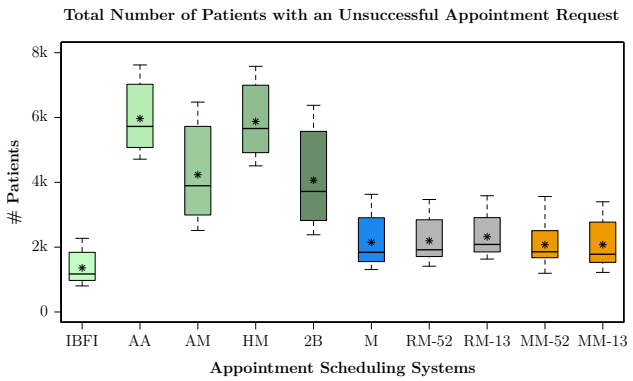

Fig. 13 Total number of patients with an unsuccessful appointment request

Third, we consider the total number of rejected walk-ins within a year, i.e., patients that are not seen by the PCP, and how many of them previously had an unsuccessful appointment request. We note that patients with an appointment are always admitted as long as they arrive before the end of the buffer. Furthermore, we note that the admission strategy in SiM-Care does not favor walk-ins who previously had an unsuccessful appointment request over walk-ins who had not requested an appointment. Accordingly, it is random how many patients are rejected twice, once for an appointment request and once for the treatment itself. Nevertheless, it is interesting to evaluate how many patients are rejected twice by the (appointment scheduling) system. The results are visualized in Fig. 14. As IBFI does not reserve slots for potential walk-ins, clearly, it is the system with the highest average total number of rejected walk-ins, namely 203 walk-ins. At the same time, IBFI is the system with the highest average number of rejected walk-ins who previously requested an appointment, even though it is the systems with the fewest average unsuccessful treatment requests. It is also noticeable that the benchmark systems AA, AM, HM and even 2B reject almost three times as many patients on average than the mask-based systems. More precisely, 2B rejects 83 walk-ins on average while the mask-based systems M, RM-52, RM-13, MM-52, and MM-13 reject 28, 27, 25, 22, 24 walk-ins on average, respectively. Furthermore, the maximum and average numbers for the rejected walk-ins who previously had an unsuccessful appointment request are significantly lower for the mask-based systems compared to the non-mask-based systems. Overall, the mask-based systems cause fewer

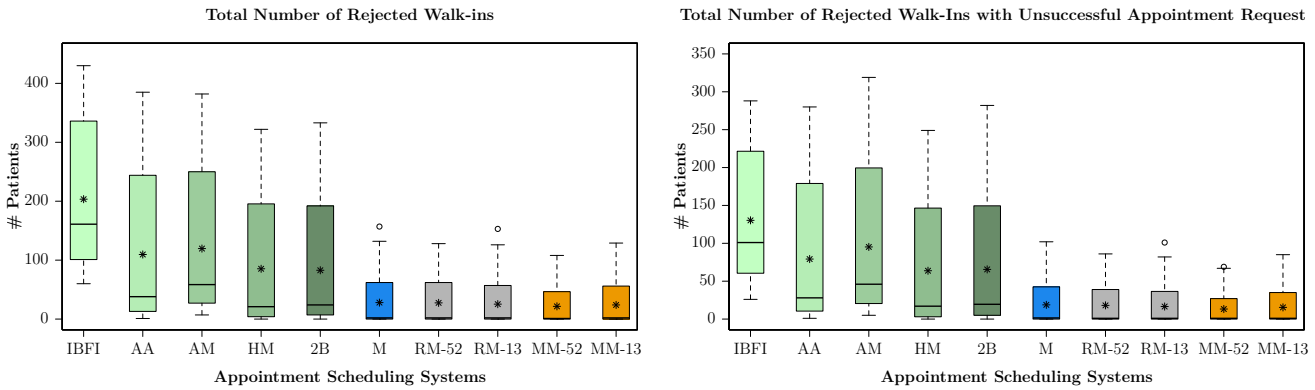

Fig. 14 Total number of rejected walk-ins and how many of them previously had an unsuccessful appointment request

unsuccessful appointment requests, fewer rejected walk-ins, and thus fewer walk-ins who have previously requested an appointment, compared to all other appointment scheduling systems.

Finally, we consider the average waiting times for treatments in the practice for patients with appointments and walk-ins. The results are visualized in Fig. 15. As the treatment strategy is defined so that patients with appointments are always prioritized over walk-ins in the treatment order, waiting times for walk-ins are significantly higher than those from patients with appointments. Considering the waiting times for patients with appointments first, we recall that roughly 20% of patients arrive late for their appointments in SiM-Care. We see that the benchmark systems AA, AM, HM perform the best with an average waiting time ranging between 3.35 and 3.57 minutes. The average waiting times caused by the mask-based systems are comparable to the average waiting time of 5.64 minutes caused by IBFI, they are only a maximum of 0.29 minutes longer. Neglecting the benchmark systems, we confirm that 2B causes with 5.44 minutes the lowest average waiting times for patients with appointments. Considering the waiting times of walk-ins next, we see that the benchmark systems perform again the best with an average waiting time ranging between 39.64 and 52.88 minutes. The mask-based systems cause the next lowest average waiting times for walk-ins ranging between 67.72 and 72.06 minutes. IBFI and 2B provide with 101.76 and 72.26 minutes the worst average waiting time for walk-ins, respectively. We note that MM-52, which causes the maximum waiting time among all mask-based systems, causes still 8.38 minutes less waiting time than the average waiting time caused by IBFI. Furthermore, we note that the maximum waiting time which a walk-in has to wait among all appointment scheduling strategies is caused by IBFI with 123.81 minutes. We conclude this section and refer to the next section where we summarize the results of the mask-based systems and assess them among each other.

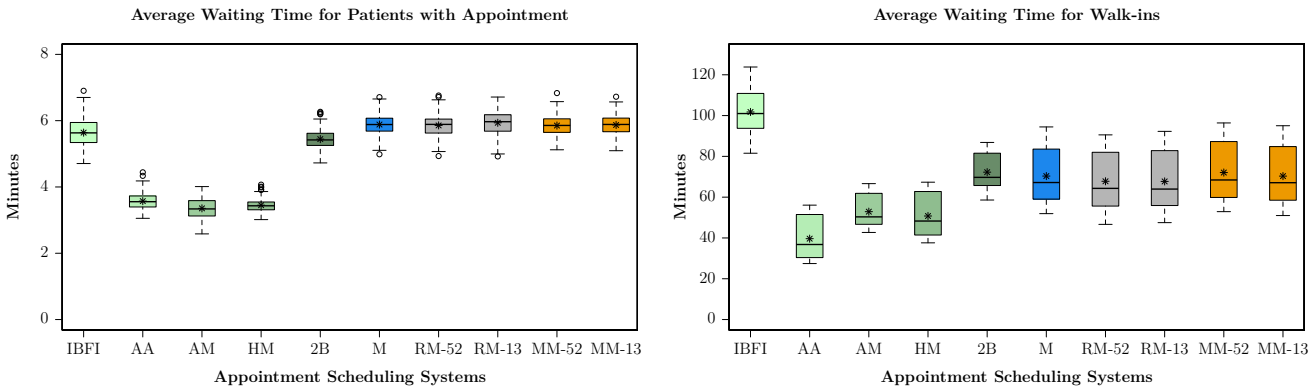

Fig. 15 Average waiting time for treatment within the practice of patients with appointment and walk-ins

7 Discussion and conclusion

In this paper, we managed a PCP's demand for treatment considering three different types of patients by interday appointment scheduling on a tactical decision level. We introduced the MDP and provided a combinatorial interpretation by a network flow and design model. Furthermore, we extended the MDP to the rMDP to obtain a robust setting in which we could account for uncertainties in the weekly distributions of demand. To avoid overfitted solutions and to prevent uncertainties not only in the weekly distribution but also in the total demand, we introduced the rMMDP. For all three problems, we developed a solution approach that combined binary search with compact IP formulations of extension of minimum cost flow problems. We analyzed the optimization models and discussed their results. Finally, we conducted an extensive case study by agent-based simulation in which we evaluated the resulting mask-based appointment scheduling systems and compared them with five adjusted appointment scheduling systems from the literature.

Our study proves on the one hand the potential of the mask-based appointment scheduling systems and on the other hand that they can more than compete with the presented appointment scheduling systems from the literature. The mask-based appointment scheduling systems improve the balanced utilization of the PCP, identifiable by the minimized overtime, idle time, and worst-case deviation of the actual from the optimal workload. Furthermore, they ensure the (timely) access to medical care which includes the following three aspects. First, more patients may receive an appointment so that there are fewer unsuccessful appointment requests. Second, the timely access to an appointment is ensured for acute patients within less than three days and for chronic patients within less than four days. Third, fewer patients are rejected and consequently more patients are admitted for treatment. We point out that the timely access comes at the expenses of waiting times within the practice. To overcome this deficit, future work needs to investigate the order of slots with specific states, especially the positioning of the walk-in slots. Since the walk-in slots are always positioned at the end of a session in the presented appointment scheduling systems, it is not surprising that walk-ins have long waiting times to their treatments. In addition to the positioning of the walk-in slots, walk-ins' waiting times can be beneficially controlled by an adjusted treatment strategy that does not always prioritize patients with appointments over walk-ins.

Like studies from the literature, our study shows that finding an optimal ratio of scheduled appointments and time for walk-ins in each session is preferable to session that are exclusively for walk-ins. We show that appointments can be used beneficially to distribute the demand for treatment. Compared to studies from the literature, our study achieves remarkable numbers in terms of rejected patients or, in other words, the number of treatments without explicitly targeting this in the objective function. Furthermore, our study provides an appointment scheduling system that is robust against small fluctuations in demand for treatment and additionally adapts to bigger change in demand for treatments.

While this study represents a valuable contribution for interday appointment scheduling in primary care practices, we must not overlook the major limitations of our case study that stem from assumptions and open up directions for further research. The lack of empirical demand data forced us to use a simulation model which can provide rough estimates at best. Furthermore, we could not sufficiently clarify on which scenario set the mask-based appointment scheduling systems work better. The question could be answered depending on available demand data and available time for computation. If the available data allows, future research may also include the analysis of different uncertainty sets. For instance, the rMDP and rMMDP could be adjusted such that a mask is robust within a certain uncertainty budget. A further interesting aspect may be to analyze the influence on the results of variation in the input parameters of the optimization models such as patients' service times and willingness to wait. We note that the weekly cyclic model must be adjusted to represent a wider range of patients' willingness to wait. In a cyclic model based on five working days, a willingness to wait of six days would correspond to a willingness to wait of only one day.

Overall, we obtain the following four key findings of this study. First, the distinction in slots reserved for specific patient types is beneficially for the demand management. Second, taking into account information like high walk-in rates in specific sessions has a major impact on the resulting appointment scheduling system and should thus be considered in interday appointment scheduling. Third, a robust nested multimask may be advantageous over a robust mask as observable in key performance indicators such as the number of patients with an unsuccessful appointment request, the number of rejected walk-ins, and the number of rejected patients who previously had an unsuccessful appointment request. We also confirm better results regarding the maximum deviation of the actual from the optimal workload of the PCP. As a result, multiple robust masks that PCPs can switch between allow to respond to changes in demand. On the one hand, PCPs are able to better meet treatment requests of patients, and on the other hand, they achieve a more balanced utilization of their weekly capacity. In future research, however, we need to investigate the circumstances under which the masks of a multimask are activated, i.e., when patients are offered appointments of the second and

third mask of the multimask, to fully realize the potential of a multimask. More precisely, we need to analyze and determine at which increase in demand for treatment the next higher mask is activated. Naturally, a multimask can only facilitate the scheduling if the PCP's capacity matches the demand for treatment. Fourth, our robust and combinatorial network flow and design models are applicable and effective methods to manage demand in primary care practices. More precisely, they are valid modelings for interday appointment scheduling on a tactical decision level. Furthermore, they can be easily adjusted to different settings and withstand a much more complex simulation model on an operational level. We conclude this paper with the confidence that our mask-based appointment scheduling systems meet the effectiveness described in the introduction at the beginning of the paper: The mask-based appointment scheduling systems maintain patient-physician continuity, enable timely access, reduce waiting times, smooth the PCP workflow, and provide robustness against demand uncertainties.

Appendix A: Algorithm 1

In this section, we present the omitted Algorithm 1 and the proof of Theorem 1.

Algorithm 1

Input: Network $N = (G, \overline{\psi}, \beta)$
Output: Feasible β-flow in network N_ω with minimum $\omega \in \Omega$
Method:

Set $\omega_{\text{LB}} = 0$ and $\omega_{\text{UB}} = \max_{k \in \mathcal{K}} D_k$
while $\omega_{\text{LB}} \neq \omega_{\text{UB}}$ **do**
  Compute $\widetilde{\omega} = \lfloor \frac{\omega_{\text{LB}} + \omega_{\text{UB}}}{2} \rfloor$
  Set $\overline{\omega} = \widetilde{\omega} \cdot t$
  if there exists a feasible β-flow $f_{\overline{\omega}}$ in network $N_{\overline{\omega}}$ **then**
    Update $\omega_{\text{UB}} = \widetilde{\omega}$ and set $f = f_{\overline{\omega}}$
  else
    Update $\omega_{\text{UB}} = \widetilde{\omega} + 1$
  end if
end while
Set $\overline{\omega} = \omega_{\text{UB}} \cdot t$
for $k \in \mathcal{K}$ **do**
  Compute $i_1 = \left\lfloor \frac{\overline{\omega} + u \cdot c_k}{t} \right\rfloor$ and $\omega_{i_1} = |t \cdot i_1 - u \cdot c_k|$
  Compute $i_2 = \left\lceil \frac{u \cdot c_k - \overline{\omega}}{t} \right\rceil$ and $\omega_{i_2} = |t \cdot i_2 - u \cdot c_k|$
  for $\omega \in \{\omega_{i_1}, \omega_{i_2}\}$ **do**
    if $\omega < \overline{\omega}$ **then**
      if there exists a feasible β-flow f_ω in network N_ω **then**
        Update $\overline{\omega} = \omega$ and set $f = f_\omega$
      end if
    end if
  end for
end for
return $\overline{\omega}$ and β-flow f

Proof (of Theorem 1) Let $\omega \in \Omega$ be the parameter and f be the β-flow in network N_ω computed by Algorithm 1. Let $\boldsymbol{\alpha}$ and μ be the corresponding $(d^c, \boldsymbol{d}^a)$-assignment and mask, respectively. Assume there exists a mask μ^* with less cost, i.e., $c(\mu^*) < c(\mu) = \omega$. There exists a $(d^c, \boldsymbol{d}^a)$-assignment $\boldsymbol{\alpha}^* \in \mathcal{A}^{\mu^*}$ that minimizes the cost of mask μ^*. Based on assignment $\boldsymbol{\alpha}^*$, we can construct a feasible β-flow f^* in network N. We note that each assignment determines a unique flow in network N. There exists a parameter $\omega^* < \omega$ such that f^* is a feasible β-flow in network N_{ω^*}. This contradicts to the correctness of Algorithm 1.

Appendix B: Computational results

In this section, we present the computational results of the optimization problems as well as the average results of the evaluation by simulation.

Table 7 Average computational results of the mask models based on scenario set Λ^{52} for all patient configurations

C-ID	Patient Config.	objective value/ deviation ω			# unsuc. appointment requests $d_w^a(\cdot)$			# acute slots $\sum_{k\in\mathcal{K}} n_\mu(r,k)$			# chronic slots $\sum_{k\in\mathcal{K}} n_\mu(c,k)$	
		Det-F	Rob-F	RobM-F	Det-F	Rob-F	RobM-F	Det-F	Rob-F	RobM-F	Det-F	Rob(M)-F
1	1400/40/15	13.09	44.3	35.13	0	4.4	9.09	72.1	67.5	68.9	43.8	44
2	1400/40/18	13.08	50.09	36.72	0	4	8.98	70.1	65.9	67.2	43.8	44
3	1400/40/20	13.3	42.25	33.8	0	5.6	9.62	68.6	62.6	63.3	44.1	44
4	1400/45/15	12.54	49.59	42.57	0	3.6	7.91	70	66.8	68.1	49.9	49.6
5	1400/45/18	11.63	43.57	29.13	0	4.1	10.33	68.8	64.7	65.2	49.7	49.5
6	1400/45/20	12.55	53.24	32.77	0	3.1	9.18	66.7	63.8	63.5	49.7	49.6
7	1600/40/15	14.38	49.01	41.51	0	3.5	7.62	80.2	76.8	78.9	50.6	50.4
8	1600/40/18	12.39	58.68	37.85	0	3	10.87	78.3	75.3	74.7	50.9	50.4
9	1600/40/20	13.4	54.44	36.29	0	3.2	10.55	76.2	73.1	70.9	50.6	50.4
10	1600/45/15	12.51	55.59	35.45	0.1	2	10.26	78.4	76.7	74.9	56.2	56.5
11	1600/45/18	13.63	55.52	36.87	0.1	2.6	10.9	76	73.7	74.5	56.2	56.6
12	1600/45/20	13.03	61.28	46.25	0	2.6	8.4	74.9	72.1	72.4	56.3	56.4
13	2000/40/15	12.62	75.25	45.75	12.4	10	11.67	80.6	82.2	85.3	63.5	62.9
14	2000/40/18	14.25	75.65	64.9	10.3	9.3	7.26	80.7	81.2	88.5	63.2	62.9
15	2000/40/20	13.81	76.43	53.62	8.8	8.2	9.47	80.3	80.6	85.9	63.3	62.9
16	2000/45/15	14.15	83.19	47.92	15.1	12	14.55	72.6	75.7	78.6	71.3	71
17	2000/45/18	13.61	78.04	42.5	13.5	12	14.03	72.7	74.1	78.1	71.3	71
18	2000/45/20	12.93	77.94	50.05	12.5	10.8	11.78	72.8	74.3	77.8	71	71

Table 8 Average computational results of the mask models based on scenario set Λ^{13} for all patient configurations

C-ID	Patient Config.	objective value/ deviation ω			# unsuc. appointment requests $d_w^a(\cdot)$			# acute slots $\sum_{k\in\mathcal{K}} n_\mu(r,k)$			# chronic slots $\sum_{k\in\mathcal{K}} n_\mu(c,k)$	
		Det-F	Rob-F	RobM-F	Det-F	Rob-F	RobM-F	Det-F	Rob-F	RobM-F	Det-F	Rob(M)-F
1	1400/40/15	13.09	12.55	14.89	0	9.4	4.24	72.1	62.5	72.6	43.8	44.1
2	1400/40/18	13.08	13.44	14.38	0	7.9	4.82	70.1	62.1	71.2	43.8	44.1
3	1400/40/20	13.3	13.08	14.01	0	8.9	4.58	68.6	59.2	68.7	44.1	44.2
4	1400/45/15	12.54	15.06	13.88	0	6.2	5.05	70	64.1	70.2	49.9	49.6
5	1400/45/18	11.63	14.66	15.22	0	7.8	4.19	68.8	61.2	69.5	49.7	49.6
6	1400/45/20	12.55	13.1	13.98	0	9.3	4.49	66.7	57.5	66.7	49.7	49.7
7	1600/40/15	14.38	15.7	16.14	0	7.6	4.48	80.2	72.5	81.1	50.6	50.6
8	1600/40/18	12.39	17.94	16.01	0	6.8	4.75	78.3	71.6	78.1	50.9	50
9	1600/40/20	13.4	15.77	15.94	0	7.2	4.23	76.2	69	78.3	50.6	50.5
10	1600/45/15	12.51	17.19	15.59	0.1	5.20	6.11	78.4	73.4	77.6	56.2	56.5
11	1600/45/18	13.63	17.37	15.49	0.1	6.7	4.93	76	69.7	78	56.2	56.3
12	1600/45/20	13.03	20.38	15.29	0	5.4	5.27	74.9	69.3	74.8	56.3	56.5
13	2000/40/15	12.62	24.04	16.72	12.4	11.6	12.69	80.6	80.6	83.1	63.5	63
14	2000/40/18	14.25	25.27	18.73	10.3	9.5	11	80.7	80.8	82.2	63.2	62.9
15	2000/40/20	13.81	25.4	17.39	8.8	8.9	10.32	80.3	79.9	82.2	63.3	62.9
16	2000/45/15	14.15	31.36	17.67	15.1	14.3	15.15	72.6	73.4	75.9	71.3	71
17	2000/45/18	13.61	26.34	17.19	13.5	12.1	14.05	72.7	73.9	74.7	71.3	70.9
18	2000/45/20	12.93	31.51	16.55	12.5	11.1	14.22	72.8	73.7	75	71	71

Table 9 Average overtime of the PCP among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	44.5	8.94	10.8	3.58	3.97	1.03	1.48	1.08	1.79	1.1
2	1400/18/40	33.11	11.78	9.8	4.56	3.83	0.67	1.36	1.03	0.87	1.13
3	1400/20/40	38.11	10.35	10.82	2.97	4.49	0.89	0.53	1.75	0.8	0.94
4	1400/15/45	51.51	9.62	12.07	5.09	5.23	0.51	0.85	1.98	1.3	0.83
5	1400/18/45	35.51	9.91	11.74	4.03	7.91	1.37	1.32	1.29	0.73	1.28
6	1400/20/45	43.23	11.92	11.15	3.87	6.01	1.89	2.04	0.77	1.59	1.88
7	1600/15/40	71.59	19.77	21.26	12.49	8.8	4.51	3.8	4.27	3.07	3.52
8	1600/18/40	76.6	24.1	21.03	11.48	13.33	4.87	6.45	4.51	4.06	4.14
9	1600/20/40	65.85	20.76	23.99	11.1	14.79	4.53	3.99	3.15	4.43	4.64
10	1600/15/45	79.89	21.23	20.01	15.63	14.01	7.82	5.68	4.75	5.57	7.41
11	1600/18/45	74.77	23.16	20.11	14.66	17.48	7.06	4.52	5.21	4.49	7.23
12	1600/20/45	69.33	23.56	23.69	14.63	12.57	7.46	7.1	5.55	5.39	6.75
13	2000/15/40	230.58	73.68	58.51	58.11	57.28	37.02	35.85	40.69	31.4	36.02
14	2000/18/40	223.14	72.76	70.6	59.29	53.71	38.16	36.15	37.91	40.92	36.1
15	2000/20/40	191.12	67.5	60.99	58.68	47.1	33.38	33.88	45.02	34.83	32.62
16	2000/15/45	228.53	87.09	75.92	73.05	58.47	43.75	53.3	48.08	44.84	51.16
17	2000/18/45	196.25	85.66	80.3	81.83	66.37	42	44.74	44.63	37.58	49.23
18	2000/20/45	234.6	92.3	77.01	77.62	56.76	45.41	50.74	43.66	39.32	44.76

Table 10 Average idle time of the PCP among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	598.57	472.03	855.76	600.24	463.63	457.56	436.44	457.76	447.39	443.26
2	1400/18/40	596.71	481.35	842.46	603.98	453.59	457.49	440.61	431.17	443.24	447.23
3	1400/20/40	578.41	489.51	853.53	580.28	447.63	443.3	455.68	444.62	441.34	447.98
4	1400/15/45	544.92	455.52	788.36	536.56	409.73	423.84	412.41	409.2	415.66	416.64
5	1400/18/45	548.12	454.73	792.67	542.12	417.72	398.01	403.42	414.27	425.1	416.22
6	1400/20/45	543.05	462.35	821.55	538.34	413.63	423.99	423.03	403.15	405.22	417.95
7	1600/15/40	442.85	355.63	651.36	429.31	327.4	324.12	306.28	300.98	311.82	318.24
8	1600/18/40	443.69	361.72	627.65	425.87	306.66	299.78	312.71	308.63	312.56	314.14
9	1600/20/40	437.44	341.04	654.25	443.58	316.36	317.06	312.27	314.02	300.18	313.83
10	1600/15/45	408.82	317.17	580.49	402.03	290.73	276.68	293.71	282.1	288.53	278.61
11	1600/18/45	411.21	306.34	592.34	402.21	297.95	281.6	276.15	276.19	281.17	281.98
12	1600/20/45	396.44	317.76	589.04	406.7	291.87	285.89	291.49	284.78	280.72	278.43
13	2000/15/40	190.5	160.29	332.88	230.67	138	136.68	148.37	141.52	136.06	156.43
14	2000/18/40	171.77	167.46	336.8	224.06	138.11	138.89	141.02	166.8	144.65	135.72
15	2000/20/40	186.04	160.94	337.35	233.07	131.11	137.36	141.73	142.19	141.24	149.86
16	2000/15/45	156.15	141.93	301.67	208.09	127.94	127.24	120.71	127.13	120.09	116.31
17	2000/18/45	154.77	149.6	297.07	215.34	108.49	125.64	125.18	126.63	118.42	116.62
18	2000/20/45	153.49	134.54	274.28	211.39	114.83	123.57	118.23	120.57	117.05	118.22

Table 11 Average maximum deviation of the actual from the optimal PCP's workload among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	0.64	0.56	0.6	0.56	0.5	0.5	0.46	0.44	0.46	0.46
2	1400/18/40	0.62	0.58	0.6	0.5	0.48	0.47	0.46	0.42	0.45	0.45
3	1400/20/40	0.64	0.6	0.61	0.49	0.51	0.48	0.42	0.47	0.43	0.48
4	1400/15/45	0.63	0.58	0.6	0.5	0.49	0.42	0.4	0.43	0.46	0.44
5	1400/18/45	0.61	0.56	0.6	0.48	0.52	0.44	0.46	0.45	0.42	0.44
6	1400/20/45	0.61	0.57	0.59	0.47	0.52	0.51	0.46	0.43	0.44	0.45
7	1600/15/40	0.61	0.56	0.58	0.51	0.47	0.51	0.46	0.43	0.43	0.45
8	1600/18/40	0.62	0.63	0.58	0.5	0.51	0.44	0.51	0.48	0.43	0.48
9	1600/20/40	0.62	0.56	0.6	0.48	0.49	0.49	0.42	0.44	0.42	0.44
10	1600/15/45	0.59	0.57	0.6	0.52	0.49	0.49	0.48	0.46	0.45	0.45
11	1600/18/45	0.59	0.55	0.57	0.48	0.54	0.48	0.43	0.48	0.43	0.49
12	1600/20/45	0.59	0.58	0.57	0.49	0.48	0.5	0.5	0.46	0.46	0.46
13	2000/15/40	0.55	0.51	0.56	0.45	0.47	0.57	0.47	0.49	0.46	0.45
14	2000/18/40	0.57	0.48	0.53	0.47	0.49	0.52	0.45	0.48	0.45	0.46
15	2000/20/40	0.56	0.52	0.55	0.46	0.45	0.54	0.44	0.49	0.47	0.46
16	2000/15/45	0.56	0.54	0.54	0.46	0.46	0.49	0.47	0.47	0.43	0.46
17	2000/18/45	0.52	0.48	0.54	0.47	0.46	0.46	0.47	0.45	0.44	0.45
18	2000/20/45	0.56	0.53	0.53	0.5	0.43	0.46	0.45	0.46	0.42	0.47

Table 12 Average access time to an appointment for acute patients among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	1.11	5.17	2.4	6.1	2.19	2.12	2.25	2.36	2.51	2.28
2	1400/18/40	1.11	5.08	2.34	6.13	2.13	2.06	2.23	2.33	2.45	2.32
3	1400/20/40	1.11	5.02	2.32	6.22	2.1	2.06	2.32	2.36	2.53	2.32
4	1400/15/45	1.15	6.33	2.54	4.56	2.34	2.17	2.21	2.3	2.48	2.34
5	1400/18/45	1.13	6.26	2.48	4.54	2.28	2.17	2.22	2.34	2.5	2.4
6	1400/20/45	1.14	6.18	2.43	4.43	2.22	2.19	2.23	2.35	2.52	2.33
7	1600/15/40	1.16	6.45	2.81	4.17	2.58	2.2	2.24	2.31	2.51	2.32
8	1600/18/40	1.16	6.45	2.75	4.3	2.51	2.11	2.22	2.28	2.56	2.33
9	1600/20/40	1.14	6.44	2.7	4.36	2.49	2.08	2.19	2.32	2.54	2.31
10	1600/15/45	1.2	7.34	2.98	4.85	2.74	2.23	2.27	2.36	2.52	2.41
11	1600/18/45	1.19	7.43	2.87	4.92	2.67	2.27	2.28	2.4	2.54	2.4
12	1600/20/45	1.19	7.42	2.86	4.93	2.65	2.27	2.28	2.43	2.57	2.42
13	2000/15/40	1.36	7.68	3.91	6.13	3.45	2.57	2.46	2.57	2.66	2.65
14	2000/18/40	1.33	7.64	3.75	6.04	3.38	2.47	2.42	2.5	2.55	2.56
15	2000/20/40	1.31	7.73	3.65	6.05	3.36	2.47	2.35	2.48	2.63	2.58
16	2000/15/45	1.46	7.9	4.77	6.91	4.04	2.72	2.6	2.79	2.78	2.7
17	2000/18/45	1.41	7.84	4.55	6.89	3.96	2.61	2.55	2.67	2.71	2.71
18	2000/20/45	1.4	7.91	4.58	6.94	3.79	2.63	2.54	2.65	2.67	2.65

Table 13 Average access time to an appointment for chronic patients among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	1.41	4.72	1.7	2.8	1.52	3.17	3.17	3.12	3.05	3.15
2	1400/18/40	1.41	4.66	1.71	2.82	1.52	3.07	3.17	3.15	2.95	3.15
3	1400/20/40	1.41	4.67	1.71	2.83	1.52	3.01	3.21	3.08	3.05	3.16
4	1400/15/45	1.4	5.93	1.75	3.33	1.55	3.11	2.97	2.98	3.03	3.01
5	1400/18/45	1.4	5.93	1.75	3.33	1.55	3.1	2.97	2.99	2.97	3.23
6	1400/20/45	1.4	5.94	1.75	3.32	1.55	3.11	3.03	2.91	2.91	3.02
7	1600/15/40	1.39	6.05	1.71	3.44	1.54	3.16	3.09	3.04	3.19	3.08
8	1600/18/40	1.4	6.04	1.71	3.44	1.54	3	3.11	3.01	3.1	3.17
9	1600/20/40	1.4	6.03	1.71	3.45	1.54	2.93	2.97	3.06	3.01	3.13
10	1600/15/45	1.4	7.04	1.79	4.54	1.6	3.09	3.05	3.14	2.93	3.11
11	1600/18/45	1.4	7.05	1.79	4.6	1.6	3.18	3.01	3.11	3.01	3.19
12	1600/20/45	1.4	7.05	1.79	4.58	1.61	3.15	3.05	3.21	3.12	3.12
13	2000/15/40	1.39	7.32	1.82	5.66	1.61	3.18	2.99	3.15	3.07	3.15
14	2000/18/40	1.39	7.34	1.82	5.62	1.62	3.07	2.95	3.1	3.09	3.05
15	2000/20/40	1.39	7.34	1.81	5.59	1.63	3.11	2.87	3.11	3.15	3.18
16	2000/15/45	1.42	7.49	1.99	6.55	1.76	3.02	2.92	3.21	2.97	2.87
17	2000/18/45	1.42	7.51	1.98	6.58	1.77	2.93	2.87	3.15	2.85	2.93
18	2000/20/45	1.42	7.5	2.02	6.56	1.76	3.01	2.92	3.17	2.8	2.91

Table 14 Average total number of patient with an unsuccessful appointment request among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	973.1	5140.8	2909.7	4841.5	2704.9	1557.7	1717.5	1903.6	1651.7	1544.1
2	1400/18/40	903.3	4981.8	2719.2	4699.7	2546.2	1463.1	1622.6	1789.6	1547	1455.8
3	1400/20/40	878.2	4818.5	2615	4607.4	2468.5	1455.8	1642.3	1807.5	1604.7	1429.2
4	1400/15/45	1017.1	5119.2	3182.7	5028.4	3022.1	1576.1	1745.6	1874.1	1638.1	1595.7
5	1400/18/45	942.6	4980.2	2958.4	4876.9	2817.2	1491.5	1667	1825.5	1565.5	1481.3
6	1400/20/45	934.6	4838.3	2840.8	4790.3	2715.7	1480.3	1592.1	1847	1580.5	1474.7
7	1600/15/40	1207.5	5900.3	3956.4	5878	3712	1940.8	1971.8	2141.8	1887	1795.7
8	1600/18/40	1135.3	5696.8	3737.7	5619.6	3556.7	1767.9	1858.5	2036.9	1839.3	1753.9
9	1600/20/40	1094.9	5566.4	3614.3	5455.1	3450.4	1726.3	1827.8	1993.9	1901.2	1699
10	1600/15/45	1275	5902.9	4217.5	5840.7	4050.4	1949.9	1981.8	2104.4	1985.5	1955.2
11	1600/18/45	1191.1	5713.6	3968.4	5672.4	3830.1	1851.8	1950.8	2108.3	1820.8	1730.1
12	1600/20/45	1176.6	5560.7	3882.2	5507.9	3714.8	1802.1	1876.9	1963.1	1814.9	1774.7
13	2000/15/40	1988.1	7448.7	5981.3	7420.2	5814.5	3086.7	3044.7	3113.5	2823.9	3032
14	2000/18/40	1845.2	7169.2	5709.2	7130.3	5567.6	2903.9	2853.9	2856.7	2482.6	2764.7
15	2000/20/40	1744.6	7001	5557.6	6930.8	5403.2	2827.8	2681.2	2791.9	2463.4	2643.4
16	2000/15/45	2222.6	7408	6342.7	7380.2	6149.8	3490.9	3337.1	3449.2	3121.5	3224.3
17	2000/18/45	2001.2	7155.4	6116.7	7107	5936.5	3193.1	3165.4	3155.7	2889.5	3090.3
18	2000/20/45	1956.6	6992.6	5904.3	7000.4	5702.4	3129.1	3006.9	3011	2751.8	2934.7

Table 15 Average total number of rejected walk-ins among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	83.9	8.2	20.8	2.5	2.2	0.1	0	0.2	0.5	0
2	1400/18/40	82.8	11	19.9	3.5	3.5	0.2	0.2	0	0.3	0.9
3	1400/20/40	86.1	9	20.3	3.9	8	0.1	0.2	0.1	0.1	0.3
4	1400/15/45	87.9	10.1	24.1	2.4	6.3	0.1	0.5	1.1	0.1	0
5	1400/18/45	97.1	9.9	18.6	5.5	5.8	0.3	0.4	0.4	0.3	0.2
6	1400/20/45	99	10.8	20.6	1.6	5	0.1	0.2	0.3	0.2	0.3
7	1600/15/40	154.4	28.5	59.5	18.4	22.4	2.1	2	1.1	0.8	1.3
8	1600/18/40	156.6	35.4	54	23.5	31.5	1.9	1.5	2.2	0.8	1.1
9	1600/20/40	156.7	34.3	60.6	16	20.9	3.6	1.7	1	2.2	1.8
10	1600/15/45	162.6	39.1	56.9	29.3	25.4	3	2.4	2.7	2.8	2
11	1600/18/45	171.2	47.2	56.4	22.7	28.8	2.8	2.7	4.2	2.2	2.2
12	1600/20/45	172.2	39.2	65.4	27.1	23.2	3.1	3	2.1	2.6	3.5
13	2000/15/40	335.3	246.9	258.8	193	200.8	65.2	60.4	55.8	47.2	61.5
14	2000/18/40	345.8	248.5	253.3	215.7	183.5	68.7	64.8	52.9	55.4	54.5
15	2000/20/40	345.3	239.1	266.7	192.1	200	66.7	59.8	68.1	47.3	54.2
16	2000/15/45	367.7	323.6	308.6	251.7	236	95.5	98.6	89.5	84	82.6
17	2000/18/45	371.4	319.1	302.7	264.3	250.6	88.3	98.2	83.5	69.6	82.5
18	2000/20/45	386.7	312.8	286.1	263	236.9	101.1	97.8	90.8	72.3	82.4

Table 16 Average total number of rejected walk-ins who previously had an unsuccessful appointment request among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	51.7	7.1	16.8	2	1.7	0.1	0	0.2	0.3	0
2	1400/18/40	47.2	8.9	15.9	2.7	2.7	0	0.2	0	0.2	0.5
3	1400/20/40	44.6	6.3	14.8	3	5.6	0.1	0	0	0	0.3
4	1400/15/45	57.7	7	19.5	2.2	4.8	0.1	0.3	0.8	0	0
5	1400/18/45	57.1	7.8	12.7	4.3	4.8	0.2	0.3	0.2	0.1	0.1
6	1400/20/45	54.2	8.7	14.1	1.1	3.6	0.1	0.1	0.3	0.2	0.2
7	1600/15/40	103.7	22.4	47.7	15.6	18.2	1.4	1.3	0.7	0.5	1
8	1600/18/40	91.2	27.9	42.7	17.7	25.2	1	0.9	1.2	0.7	0.6
9	1600/20/40	86.5	26.7	44.3	12.5	14.9	2.2	1.1	0.5	1.4	1
10	1600/15/45	113.6	30.5	45.3	23.5	20	2.3	2.1	2	1.7	1.4
11	1600/18/45	102.6	34.5	42.8	18.2	22.1	1.6	1.7	2.1	1.1	1.3
12	1600/20/45	100.7	29.3	52.1	20.6	16.4	1.8	1.3	0.7	1.6	2.3
13	2000/15/40	230.5	185.2	214	149.5	163.1	45.8	43	39.6	30.3	41.4
14	2000/18/40	227.2	183.7	199.4	163	145.1	46.7	43.4	37.1	31.4	35.3
15	2000/20/40	214.8	169	207.9	141.9	152.9	41.8	35.6	42	27.3	32.6
16	2000/15/45	264.3	233.7	256.1	192.5	197.7	69.5	69.6	63.3	57.7	56.6
17	2000/18/45	251.9	226	245	191.3	199	59.5	63.4	51.2	42.5	53.2
18	2000/20/45	246.1	211	222.6	186.5	182.2	63.3	61.1	55.8	41.6	49.9

Table 17 Average waiting time for treatment within the practice for patients with appointments among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	5.21	3.58	3.07	3.32	5.31	5.6	5.54	5.75	5.63	5.63
2	1400/18/40	5.11	3.56	3.07	3.33	5.36	5.58	5.71	5.61	5.57	5.6
3	1400/20/40	5.3	3.6	3.01	3.35	5.4	5.54	5.67	5.59	5.6	5.55
4	1400/15/45	5.31	3.65	3.12	3.41	5.25	5.82	5.54	5.77	5.62	5.72
5	1400/18/45	5.3	3.38	3.14	3.43	5.55	5.7	5.78	5.79	5.65	5.8
6	1400/20/45	5.38	3.65	3.02	3.35	5.65	5.84	5.66	5.76	5.66	5.63
7	1600/15/40	5.55	3.49	3.36	3.47	5.48	5.8	5.99	5.97	5.85	6
8	1600/18/40	5.69	3.65	3.37	3.48	5.53	6.06	5.69	5.99	5.68	5.99
9	1600/20/40	5.54	3.55	3.22	3.41	5.46	5.83	5.99	5.94	6	5.81
10	1600/15/45	5.73	3.63	3.32	3.56	5.47	6.11	5.74	5.93	5.85	5.84
11	1600/18/45	5.7	3.61	3.39	3.56	5.39	6.07	5.99	5.95	5.86	5.93
12	1600/20/45	5.46	3.73	3.37	3.59	5.44	6.03	5.9	6.12	5.98	6.05
13	2000/15/40	6.03	3.55	3.62	3.47	5.45	5.99	6.09	6.12	5.92	5.83
14	2000/18/40	6	3.52	3.65	3.44	5.37	6.02	5.94	6.06	5.89	6.07
15	2000/20/40	5.97	3.46	3.67	3.5	5.58	5.9	6.05	6.35	6.25	6.05
16	2000/15/45	5.92	3.54	3.74	3.41	5.54	6.02	6.09	6.02	6.17	5.96
17	2000/18/45	6.13	3.59	3.65	3.51	5.44	5.94	6.02	6.01	5.99	6.2
18	2000/20/45	6.16	3.58	3.6	3.5	5.34	6.06	6.02	6.11	6.17	5.99

Table 18 Average waiting time for treatment within the practice for walk-ins among the ten runs

C-ID	Config.	IBFI	AA	AM	HM	2B	M	RM-52	RM-13	MM-52	MM-13
1	1400/15/40	96.68	28.95	46.27	39.52	62.31	57.13	54.34	54.29	57.76	57.17
2	1400/18/40	89.59	29.32	45.5	39.63	63.72	56.58	53.67	53.17	58.85	57.17
3	1400/20/40	87.19	29.49	45.39	40.37	64.99	56.31	51.52	51.35	55.8	55.14
4	1400/15/45	97.41	29.89	45.61	41.25	64.39	59.46	56.01	56.33	59.06	59
5	1400/18/45	91.31	30.32	45.6	40.38	65.16	59	55.47	55.5	59.64	58.48
6	1400/20/45	91.17	30.14	45.11	41.28	65.73	58.62	53.85	55.33	59.44	57.38
7	1600/15/40	102.93	35.72	50.29	47.51	68.09	68.02	64.31	64.07	66.83	65.48
8	1600/18/40	98.63	35.85	49.69	47.85	70.21	67.14	63.25	61.48	66.26	65.09
9	1600/20/40	95.64	35.91	50.18	47.04	69.15	65.23	62.2	61.87	64.6	64.73
10	1600/15/45	103.08	37.53	50.08	48.8	69.74	69.38	65.92	65.71	70.96	68.45
11	1600/18/45	101.16	37.6	49.89	48.92	69.8	68.91	67.13	66.01	70.75	68.39
12	1600/20/45	97.45	37.61	50.76	49.13	69.51	66.02	66.31	65.24	69.23	68.17
13	2000/15/40	116.22	51.66	61.52	62.22	82.43	84.36	83.82	83.46	87.87	85.68
14	2000/18/40	113.39	51.83	62.08	62.93	81.92	86.15	81.92	81.8	89.16	83.93
15	2000/20/40	111.06	51.4	62.41	62.46	82.08	84.74	80.83	82.89	88.34	85.51
16	2000/15/45	113.05	53.46	64.03	64.4	83.83	86.03	87.22	85.83	91.54	89.22
17	2000/18/45	111.85	53.44	64.12	65.12	84.2	86.14	86.25	87.53	91.86	88.44
18	2000/20/45	113.92	53.43	63.28	64.87	83.44	86.96	86.23	87.03	89.07	88.43

Acknowledgements This study was supported by Freigeist-Fellowship of Volkswagen Stiftung, the German research council (DFG) Research Training Group 2236 UnRAVeL, and the German research council (DFG) Collaborative Research Centres SFB 1481 (Sparsity and Singular Structures).

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. IIE Trans 40(9):800–819
2. Balasubramanian H, Muriel A, Wang L (2012) The impact of provider flexibility and capacity allocation on the performance of primary care practices. Flex Serv Manuf J 24(4):422–447
3. O'Malley AS, Cunningham PJ (2009) Patient experiences with coordination of care: the benefit of continuity and primary care physician as referral source. J Gen Intern Med 24:170–177
4. Anderson R, Barbara A, Feldman S et al (2007) What patients want: a content analysis of key qualities that influence patient satisfaction. J Med Pract Manage 22(5):255–261
5. Balasubramanian H, Biehl S, Dai L, Muriel A (2014) Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments. Health Care Manag Sci 17(1):31–48
6. Wolfe A (2001) Institute of medicine report: Crossing the quality chasm: A new health care system for the 21st century. Policy Politics & Nursing Practice 2(3):233–235
7. Rust GS, Ye J, Baltrus PA, Daniels E, Adesunloye BA, Fryer GE (2008) Practical barriers to timely primary care access: impact on adult use of emergency department services. Arch Intern Med 168(15):1705–10
8. Peng Y, Xiuli Q, Shi J (2014) A hybrid simulation and genetic algorithm approach to determine the optimal scheduling templates for open access clinics admitting walk-in patients. Comput Ind Eng 72:282–296
9. Mann E, Schuetz B, Rubin-Johnston E (2010) Remaking primary care: A framework for the future (new england healthcare institute report)
10. Statistische Informationen aus dem Bundesarztregister Kassenärztliche Bundesvereinigung (2017)
11. Rüdiger Jacob J, Kopp, and S Schultz. Berufsmonitoring medizinstudenten, (2014) Ergebnisse einer bundesweiten Befragung. KBV, Kassenärztliche Bundesvereinigung, Berlin, p 2015
12. Alemayehu B, Warner KE (2004) The lifetime distribution of health care costs. Health Serv Res 39(3):627–642
13. Bundesvereinigung K
14. Huang X-M (1994) Patient attitude towards waiting in an outpatient clinic and its applications. Health Serv Manage Res 7(1):2–8
15. Harper PR, Gamlin HM (2003) Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. OR Spectrum 25(2):207–222
16. Murray M, Berwick DM (2003) Advanced access: reducing waiting and delays in primary care. JAMA 289(8):1035–1040
17. Wang S, Liu N, Wan G (2020) Managing appointment-based services in the presence of walk-in customers. Manage Sci 66(2):667–686
18. Cayirli T, Veral E, Rosen H (2008) Assessment of patient classification in appointment system design. Prod Oper Manag 17(3):338–353
19. The Associations of Statutory Health Insurance Physicians
20. Wiesche L, Schacht M, Werners B (2016) Strategies for interday appointment scheduling in primary care. Health care management science, pages 1–16
21. Bodenheimer T (2006) Primary care – will it survive? N Engl J Med 355(9):861–864 (**PMID: 16943396**)
22. Colwill JM, Cultice JM, Kruse RL (2008) Will generalist physician supply meet demands of an increasing and aging population? Health Aff 27(3):w232–w241 (**PMID: 18445642**)
23. Hulshof PJH, Kortbeek N, Boucherie RJ, Hans EW, Bakker PJM (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
24. Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. Prod Oper Manag 12(4):519–549
25. Ahmadi-Javid A, Jalali Z, Klassen KJ (2017) Outpatient appointment systems in healthcare: A review of optimization studies. Eur J Oper Res 258(1):3–34
26. Wang W-Y, Gupta D (2011) Adaptive appointment systems with patient preferences. Manuf Service Oper Manag 13(3):373–389
27. Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. Manuf Service Oper Manag 12(2):330–346
28. Kim S, Giachetti RE (2006) A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. IEEE Transactions on systems man and cybernetics-Part A Systems and humans 36(6):1211–1219
29. LaGanga LR, Lawrence SR (2007) Clinic overbooking to improve patient access and increase provider productivity. Decis Sci 38(2):251–276
30. LaGanga LR, Lawrence SR (2012) Appointment overbooking in health care clinics to improve patient service and clinic performance. Prod Oper Manag 21(5):874–888
31. Cayirli T, Gunes ED (2014) Outpatient appointment scheduling in presence of seasonal walk-ins. J Oper Res Soc 65(4):512–531
32. Zheng B, Yoon SW, Khasawneh MT (2015) An overbooking scheduling model for outpatient appointments in a multi-provider clinic. Oper Res Health Care 6:1–10
33. Huang Y, Zuniga P (2012) Dynamic overbooking scheduling system to improve patient access. J Oper Res Soc 63(6):810–820
34. Qu X, Rardin RL, Williams JAS, Willis DR (2007) Matching daily healthcare provider capacity to demand in advanced access scheduling systems. Eur J Oper Res 183(2):812–826
35. Qu X, Rardin RL, Williams JAS (2012) A mean-variance model to optimize the fixed versus open appointment percentages in open access scheduling systems. Decis Support Syst 53(3):554–564
36. Feldman J, Liu N, Topaloglu H, Ziya S (2014) Appointment scheduling under patient preference and no-show behavior. Oper Res 62(4):794–811
37. Gullhav AN, Christiansen M, Nygreen B, Aarlott MM, Medhus JE, Skomsvoll JF, Østbyhaug PO (2018) Block scheduling at magnetic resonance imaging labs. Oper Res Health Care 18:52–64
38. Hoseini B, Cai W, Abdel-Malek L (2018) A carve-out model for primary care appointment scheduling with same-day requests and no-shows. Oper Res Health Care 16:41–58

39. Qu X, Rardin RL, Williams JAS (2011) Single versus hybrid time horizons for open access scheduling. Comput Ind Eng 60(1):56–65
40. Dobson G, Hasija S, Pinker EJ (2011) Reserving capacity for urgent patients in primary care. Prod Oper Manag 20(3):456–473
41. Schacht M (2018) Improving same-day access in primary care: Optimal reconfiguration of appointment system setups. Oper Res Health Care 18:119–134
42. Ahuja RK, Magnanti TL, Orlin JB (1988) Network flows
43. Korte B, Vygen J (2012) Combinatorial optimization, vol 2. Springer
44. Christina Büsing AMCA, Koster, and Sabrina Schmitz, (2022) Robust minimum cost flow problem under consistent flow constraints. Ann Oper Res 312(2):691–722
45. American Academy of Family Physicians (2019) Primary care
46. Bundesministerium der Justiz und für Verbraucherschutz,
47. Mechanic D, McAlpine DD, Rosenthal M (2001) Are patients' office visits with physicians getting shorter? N Engl J Med 344(3):198–204
48. Tai-Seale M, McGuire TG, Zhang W (2007) Time allocation in primary care office visits. Health Serv Res 42(5):1871–1894
49. Federal Joint Commitee
50. Gesundheit in Deutschland aktuell 2010
51. GKV-Spitzenverband
52. 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, pages 1–28
53. Kim S-H, Chan CW, Olivares M, Escobar G (2015) Icu admission control: An empirical study of capacity allocation and its implication for patient outcomes. Manage Sci 61(1):19–38
54. Qu X, Peng Y, Shi J, LaGanga L (2015) An mdp model for walk-in patient admission management in primary care clinics. Int J Prod Econ 168:303–320
55. Information und Technik Nordrhein-Westfalen (2016) Zensus 2011: Vielfältiges Deutschland Statistische Ämter des Bundes und der Länder
56. Grobe T, Dörning H, Schwartz F (2011) Barmer gek arztreport 2011
57. Kassenärtzliche Vereinigung Nordrhein (2018) Die 100 häufigsten ICD-10-Schlüssel und Kurztexte (nach Fachgruppen) - 4. Quartal 2018:1–125
58. Klassen KJ, Rohleder TR (1996) Scheduling outpatient appointments in a dynamic environment. J Oper Manag 14(2):83–101
59. Oracle (2018) Open Java Development Kit
60. IBM (2018) IBM CPLEX Optimization Studio 22.1.0
61. Bertsimas D, Sim M (2004) The price of robustness. Oper Res 52(1):35–53

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.