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

Anne Zander[a,*], Stefan Nickel[a], Peter Vanberkel[b]

[a] *Karlsruhe Institute of Technology, Department of Economics and Management,*
Kaiserstraße 89, 76133 Karlsruhe, Germany

[b] *Dalhousie University, Department of Industrial Engineering,*
PO BOX 15000 Halifax, NS B3H 4R2, Canada

Abstract

This article focuses on balancing supply and demand for physicians and panel patients on a tactical level. Here, patients are part of the physician's panel if they visit the physician somewhat regularly. We propose a taxonomy of deterministic integer linear programs that decide on the intake of new patients into panels over time, taking into account the future panel development. Objectives are to minimize the deviation between the expected panel workload and the physician's capacity and minimize the panel workload variance over time. We classify panel patients with respect to age and the number of visits in a period and assume a transition probability from one visit category to another from one period to the next. We can include stationary patient attributes to classify patients and consider several physicians together. The programs work with different aggregation levels for the demand of new patients concerning the patient attributes. We conduct numerical experiments with parameter definitions based on real-world data. In a simulation, we consider the transition between visit categories and the new patients' demand to be stochastic. We define upper bounds on the number of patients in a patient class to be accepted in a period through solving the programs several times with different demand inputs. Even in this uncertain environment, we can significantly reduce the expected differences between workload and capacity over time, taking into account several future periods instead of just the following period. Using a detailed classification of new patients decreases the expected differences further.

Keywords: OR in health services, Integer programming, Multi-period, Physician panel management, Access to care

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[*] Corresponding author

Email addresses: anne.zander@kit.edu (Anne Zander), stefan.nickel@kit.edu (Stefan Nickel), peter.vanberkel@dal.ca (Peter Vanberkel)

1. Introduction

Rural regions in Germany face a shortage of medical care provided by office-based physicians. One reason for this is demographic change. Even though the population is decreasing in total, the increasing number of older patients who need more medical attention implies an increasing demand for medical care for most medical specialties of office-based physicians (Schulz et al., 2016). This situation is further aggravated due to the physicians themselves getting older and stopping work. For example, in the federal state Rhineland-Palatinate, the percentage of general practitioners over 60 years old increased from 14 percent in 2005 to 38 percent in 2015 (Kassenärztliche Vereinigung Rheinland-Pfalz).

Every year the German National Association of Statutory Health Insurance Physicians surveys medical students on a number of work-related topics (Kassenärztliche Bundesvereinigung). The results of the survey show that there are not enough medical students interested in specializing in the field of general medicine. Those interested can only replace 53 percent of the health care provision in the future in reference to the year 2009. Further, a third of the medical students do not want to work in locations with less than $10,000$ inhabitants. In addition, the preference for working as an employee in a group practice increases. The fraction of medical students who prefer to work in group practices rather than in a practice managed by them alone has risen from 39.9 percent in 2010 to 50.6 percent in 2018. For more than 90 percent of medical students, work-life balance is of high importance. More than 80 percent of the medical students value flexible working hours as very important.

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

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

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

In this work, we focus on matching supply and demand for physicians and patients on a tactical level. We assume that the most significant share of demand comes from patients that have visited before and from whom we know the visiting history. We consider equal-sized periods and use age and the number of visits in the last period to classify patients and build a distribution for the number of visits in the following period. In the basic integer linear program, we decide whether or not to accept external demand per period to balance the physician's working hours with the expected workload produced by panel patients for all considered periods. Further programs allow us to consider group practices with several physicians and the integration of constraints including the standard deviation of the workload produced by the panel.

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

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

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. In section 3, we present our integer linear programs (ILPs). In section 4, we describe a real-world data set from a general practitioner group practice and define model parameters. We report on our numerical experiments in Section 5. Here, we present deterministic results as well as results from a simulation where the ILPs are solved to guide decisions in a stochastic environment. Section 6 provides concluding remarks and defines lines of future research.

2. Literature Review

We investigate the following literature streams relevant to our research on matching supply and demand for practices and physicians on a tactical level: panel size, case mix and workload distribution, forecasting of appointment demand and continuity of care.

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

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

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

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

In practice, determining a manageable panel size is not enough. The physician also has to define which patients belong to her panel to determine the current panel size. That is easier said than done. On the one hand, patients leaving the panel do not always say that they will not visit again. So, the physician has to wait some time before she can remove those patients from the panel. On the other hand, some patients remain with the same physician but visit rarely. Therefore, the physician might wrongfully be inclined to remove those, too. There is no standard definition of how to determine the panel size. Some use the number of patients seen by the physician in the last two years (Margolius et al., 2018; Marx et al., 2011) others the number of patients in the last 18 months (Raffoul et al., 2016; Murray & Berwick, 2003; Murray et al., 2007). Of course, the panel sizes change dependent on the considered time frame and are not comparable with each other. Therefore, it is not surprising that only one-third of family physicians in the U.S. can estimate their patient panel size (Peterson et al., 2015). We will see in the data section that the considered time frame indeed has a significant impact on the panel size when experiencing a high proportion of patients that visit the physician rarely. We, therefore, argue that panel size as the only measure of workload is not enough. At least the corresponding average appointment request rate should be indicated. Therefore, to estimate the workload produced by a panel, we count patients and categorize them with respect to the number of visits per period.

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

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

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

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

3. Integer Linear Programs for Panel Management

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

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

We can also calculate the variance of the workload that a patient contributes to the next period. Both the expected workload and the variance of the workload in future periods of a panel patient can be represented by linear functions, which can be used to build linear programs.

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

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

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

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

$$p_{kijl} = \sum_{h=0}^{m-1} q_{(i+k-1)hl} p_{(k-1)ijh} \quad \forall i \in N \backslash \{n-1\}, j \in M, k \in \{1, \ldots, n-i-1\}, l \in M. \tag{2}$$

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

$$o_{kij} = \begin{cases} \sum_{l=0}^{m-1} f_l p_{kijl} & \forall i \in N, j \in M, k \in \{0, \ldots, n-i-1\}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

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

$$u_{0ij} = 0 \quad \forall i \in N, j \in M, \tag{4}$$

$$u_{1ij} = \sigma_{ij}^2 \quad \forall i \in N \backslash \{n-1\}, j \in M, \tag{5}$$

$$u_{kij} = \begin{cases} \sum_{l=0}^{m-1} p_{(k-1)ijl} \sigma_{(i+k-1)l}^2 & \forall i \in N \backslash \{n-1\}, j \in M, k \in \{1, \ldots, n-i-1\}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Now, let us consider a single panel. We define the number of patients in the starting panel (period $k = 0$) as v_{0ij} for age category $i \in N$ and visit category $j \in M$. At the end of a period $k \in T \backslash \{n\}$ a decision is taken to impanel a number of patients w_{kij} belonging to age category $i \in N$ and visit category $j \in M$.

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

$$\sum_{j=0}^{m-1} \sum_{i=0}^{n-1} f_j v_{kij} = \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left(o_{kij} v_{0ij} + \sum_{h=0}^{k-1} o_{(k-h-1)ij} w_{hij} \right). \tag{7}$$

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

$$\sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left(u_{kij} v_{0ij} + \sum_{h=0}^{k-1} u_{(k-h-1)ij} w_{hij} \right). \tag{8}$$

In the following, we will present some examples of integer linear programs (ILPs) using the linear terms (7) and (8). We differentiate our examples with respect to the main component in the objective, the patient attributes used to categorize panel patients besides age and number of visits per period, the number of physicians considered, and the patient attributes used to categorize new patients. In the objective function, we minimize the deviation of the expected workload from the capacity, or we minimize the standard deviation of the workload using a summation approach or a min-max approach over the considered periods and physicians. The expected workload of a panel might match the capacity in all the considered periods. Still, when the actual workload is realized, the difference between workload and capacity might be substantial due to a high variance in the workload. Hence, it is reasonable to consider the standard deviation of the workload. As an example, we consider gender as an additional patient attribute besides age and number of visits. New patients can be categorized with respect to a subset of the attributes of panel patients (age, number of visits per period, and gender) or even to aggregated attributes. For example, it might not be possible or desirable to classify new patients according to age and number of visits. Maybe, new patients should only be categorized by age but measured in larger time units than defined by the length of a period. Hence, we present an example were we categorize patients with respect to age groups, where one age group contains several age categories. Table 1 gives an overview of the different models we are going to present. Based on this classification of models other models can easily be generated.

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

Table 1: Overview of the different models

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

$$(E\ AN) \qquad \min \sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left(o_{kij} v_{0ij} + \sum_{h=0}^{k-1} o_{(k-h-1)ij} w_{hij} \right) - c_k \right| \tag{9}$$

s.t.

$$w_{kij} \leq d_{kij} \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M \tag{10}$$

$$w_{kij} \in \mathbb{Z}_0^+ \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M \tag{11}$$

The objective function (9) minimizes the sum over all periods of the absolutes values of the differences between the workload of the panel and the capacity of the physician. The first set of constraints (10) assures that we can not add more patients of a certain age and a certain visit category in a period than there is demand of such patients in this period. The third set of constraints (11) forces the decision variables to be non-negative integers. We do not explicitly linearize the absolute values in the objective function (9) here but note that this can be done using standard methods. Table 2 summarizes the parameters and variables used so far.

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

Table 2: Basic model notation

For our next ILP we group several age categories together resulting in a set $E = \{0, \ldots, e-1\}$ of age groups where each age group consists of several age categories. We define the decision variables x_{kg} as the number of added patients of age group $g \in E$ in period $k \in \{0, \ldots, t-1\}$. We assume that we know the probability that a patient of a certain age group is a patient of a certain age category. Hence, we further define d_{kg} as the exterior demand in period $k \in \{0, \ldots, t-1\}$ according to age group $g \in E$ and b_{kig} as the probability that a random patient of age group $g \in E$ belongs to age category $i \in N$ in period $k \in \{0, \ldots, t-1\}$. For example, we can assume for the demand $d_{kg} = \sum_{i \in g} \sum_{j \in M} d_{kij}$ using the demand parameters from the basic ILP (E AN). Then, b_{kig} can be defined as $b_{kig} = \frac{\sum_{j \in M} d_{kij}}{d_{kg}}$. We further assume that we know the probability r_{kij} that a patient of a age category $i \in N$ belongs to the visit category $j \in M$ in period $k \in \{0, \ldots, t-1\}$. Then, r_{kij} can be defined as $r_{kij} = \frac{d_{kij}}{\sum_{j \in M} d_{kij}}$. To obtain our new model formulation, we substitute the first set of constraints (10) with $x_{kg} \leq d_{kg}, \forall k \in \{0, \ldots, t-1\}, g \in E$ and eliminate w_{hij} in ILP (E AN) via setting $w_{hij} = r_{hij} b_{hig} x_{hg}$ where g is the age group such that $i \in g$. We obtain the following ILP:

$$(E\ AA) \qquad \min \sum_{k=1}^{t} \left| \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left(o_{kij} v_{0ij} + \sum_{h=0}^{k-1} o_{(k-h-1)ij} r_{hij} b_{hig} x_{hg} \right) - c_k \right| \tag{12}$$

s.t.

$$x_{kg} \leq d_{kg} \qquad \forall k \in \{0, \ldots, t-1\}, g \in E \tag{13}$$

$$x_{kg} \in \mathbb{Z}_0^+ \qquad \forall k \in \{0, \ldots, t-1\}, g \in E \tag{14}$$

We summarize the model notation in addition to the basic notation in Table 3.

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

Table 3: Additional model notation

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

$$(E\ G\ ANG) \qquad \min \sum_{k=1}^{t} \left| \sum_{b \in \{f,m\}} \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} \left(o^b_{kij} v^b_{0ij} + \sum_{h=0}^{k-1} o^b_{(k-h-1)ij} w^b_{hij} \right) - c_k \right| \tag{15}$$

s.t.

$$w^b_{kij} \leq d^b_{kij} \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M, b \in \{f, m\} \tag{16}$$

$$w^b_{kij} \in \mathbb{Z}_0^+ \qquad \forall k \in \{0, \ldots, t-1\}, i \in N, j \in M, b \in \{f, m\} \tag{17}$$

Now, imagine a group practice with several physicians who all have their own patient panel. In this case, when accepting a new patient we also have to decide to which physician the patient should be assigned. Hence, we use physician dependent parameters and variables. In order to balance the workload between the physicians we can use a min-max optimization approach. Let A be the set of physicians. Then, we obtain yet another ILP:

$$(E\ SP\ AN) \qquad \min z \tag{18}$$

$$\text{s.t.}$$

$$\sum_{k=1}^{t}\left|\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(o^a_{kij}v^a_{0ij}+\sum_{h=0}^{k-1}o^a_{(k-h-1)ij}w^a_{hij}\right)-c^a_k\right| \leq z \qquad \forall a \in A \tag{19}$$

$$\sum_{a\in A} w^a_{kij} \leq d_{kij} \qquad \forall k \in \{0,\ldots,t-1\}, i \in N, j \in M \tag{20}$$

$$w^a_{kij} \in \mathbb{Z}^+_0 \qquad \forall k \in \{0,\ldots,t-1\}, i \in N, j \in M, a \in A \tag{21}$$

So far, all presented ILPs minimize the difference between workload and capacity. Now, we give an example where we minimize variance. As before, imagine a group practice with several physicians who all have their own patient panel. As a first step, using an ILP for the whole practice without differentiation based on the physicians, we decide on patients to be accepted on the practice level, the result being w_{kij} for $i \in N$, $j \in M$ and $k \in \{0,\ldots,t-1\}$. In a second step, we minimize the maximal variance of the physicians' expected workloads while constraining for a small difference between workload and capacity for every physician. This approach yields a fair allocation of patients to physicians in the sense that the variability of the panel demand is similar for each physician. For example, the physicians then experience similar distributions of overtime and idle time. Let s^a be a small constant dependent on $a \in A$. We obtain:

$$(SD\ SP\ AN) \qquad \min z \tag{22}$$

$$\text{s.t.}$$

$$\sum_{k=1}^{t}\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(u^a_{kij}v^a_{0ij}+\sum_{h=0}^{k-1}u^a_{(k-h-1)ij}w^a_{hij}\right) \leq z \qquad \forall a \in A \tag{23}$$

$$\sum_{k=1}^{t}\left|\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}\left(o^a_{kij}v^a_{0ij}+\sum_{h=0}^{k-1}o^a_{(k-h-1)ij}w^a_{hij}\right)-c^a_k\right| \leq s^a \qquad \forall a \in A \tag{24}$$

$$\sum_{a\in A} w^a_{ijk} = w_{kij} \tag{25}$$

$$w^a_{kij} \in \mathbb{Z}^+_0 \qquad \forall k \in \{0,\ldots,t-1\}, i \in N, j \in M, a \in A \tag{26}$$

Of course, many more versions and extensions are possible. For example, our model can also be used to redesign panels in a group practice. To this end, we model patients eligible for reassignment as new patients in an ILP such as ILP (E SP AN) or ILP (SD SP AN). Instead of minimizing variance, we could also opt to minimize the maximal overflow frequency (on the period level) as in Ozen & Balasubramanian (2013). But, note that this modeling leads to a non-linear objective function. Further, we could easily include weights for the considered periods in the optimization to discount periods that lie further into the future.

In our ILPs, we decide on the number of new patients differentiated by categories and groups to impanel for all considered t periods. However, the ILPs presented should be solved at least once every period. Therefore, we will only act on decisions for the first period. The decisions for future periods are merely taken to account for the future panel development. Further, every time a program is solved, the starting panel can be adjusted. In this way, we can, for example, account for patients that have left the panel.

Until now, we have assumed that we know the exact demand for this period and the upcoming periods. In reality, we might have an estimation for the expected demand during a period based on historical data or based on demographic data on the region. To better adapt to the uncertainty in the current period, we can quickly adjust the programs such that we can solve them several times during a period or even every time a request to enter the panel occurs. To this end, we constrain that the already accepted demand during the current period is added to the panel. Moreover, we adjust the remaining demand until the end of the current period. However, the optimization still relies on the demand forecast for the remaining part of the current period and future periods. We will show in Section 5 that using our ILPs to guide the acceptance decisions of new patients is still meaningful even if only the expected future demand with our without patient classification is known.

Note that we could also use the actual time a patient spends with his physician, e.g., measured in minutes, to define visit categories. The issue here is that such data is often not available or reliable. Hence, we work with the number of visits.

4. Defining Model Parameters Using Real-World Data

To test our models, we use a real-world data set from a group practice for general medicine that is run by 3 physicians. One of them is working full time and the other two practice approximately 50 percent part-time. They care for their patients together meaning that they have one panel for the whole practice. We have data from the years 2010 to 2014. Every row in the data represents a chargeable service for the patient (this does not necessarily mean a patient-physician contact) on a specific day. Examples for such service charges are general treatment charges, laboratory charges, charges for chronically ill patients, emergency charges, and blood pressure measurement charges. We assume that every service produces workload for the practice (for the physicians or the other staff of the practice). We cannot determine the time, personal, and equipment requirements for the different services from the data. Therefore, we simplify by assuming the same time required for every service partitioned proportionally between the physicians.

Following the terminology of our model, we refer to every service as a visit in the following. The critical columns of the data for us are the practice-specific patient-ID, the birthdate of the patient, and the date of the service. Unfortunately, there is no column indicating gender or any other stationary patient attribute. We count $7,472$ patients who visited the practice $220,710$ times in those 5 years.

We decide on using a period length of one year. We believe this is a good choice since a year is long enough to level out seasonal effects between periods. Hence, it is reasonable to look at the number of visits last period to predict the number of visits next period instead of taking into account several preceding periods. As we are interested in the time evolution of the panel over the years, we need to define who belongs to the panel. Having a data set of 5 years, we define that a patient who does not show up for 4 years in a row left the panel and that a patient who did not visit the practice during the last 4 years is new. In reality, we may have panel patients that do not show up for 4 years in a row without leaving the panel. In our model, those patients are then counted as new patients when they show up again. The practice assigns ascending patient-IDs to new patients. Therefore, we know that patients with patient-IDs smaller than the current highest patient-ID must have visited the practice before. Hence, we can determine how often a new patient, in our definition, is, in fact, a panel patient that did not show for 4 years. Indeed, in 2014, 179 of the 756 new patients, according to our definition, have visited the practice before.

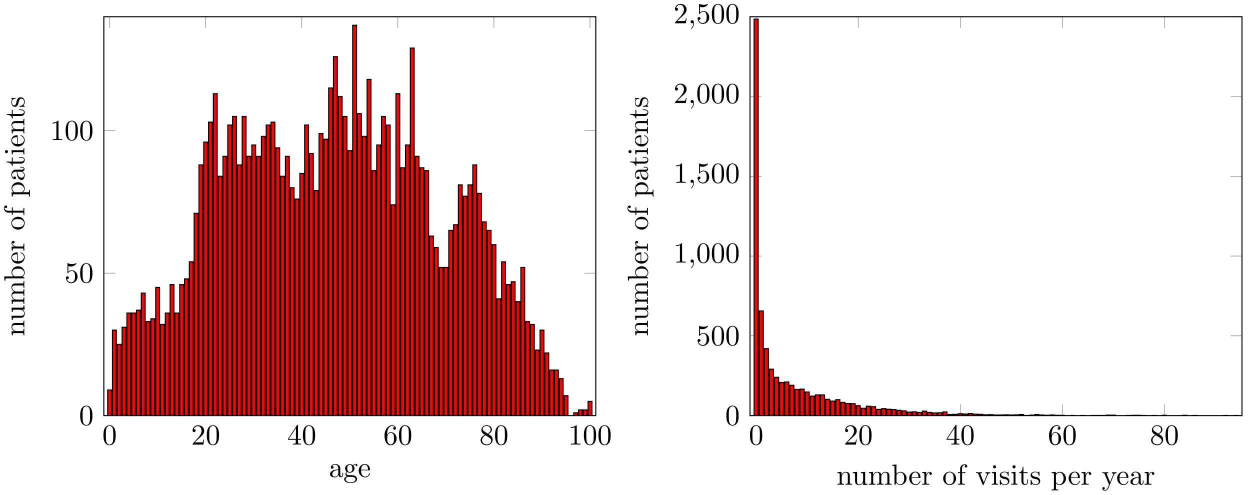

Figure 1: Number of panel patients by age and by number of visits per year in 2014

In 2014, we find 6864 patients in the panel with an average number of visits of 7.11. Figure 1 shows the number of panel patients by age and by the number of visits for 2014. Note that many panel patients, i.e., 2485 out of 6864, do not show in 2014. To come back to our argument that panel size as an only measure of workload is not enough, we investigate the resulting panel sizes if we define a different time frame for belonging to the panel. Considering only patients that visited in 2014 yields 4379 patients. Looking back 2 or 3 years yields 5423 or 6273, respectively. Note that the average number of visits for panel patients in 2014 differs for the different time frames. It is 7.11 for our definition of looking back 4 years, and it is 11.15 when we only consider patients that visited this year.

To reflect the definition of new patients and leaving patients in our model, we include 4 zero visits categories. Visit category 0 contains all patients that did not visit this year and not in the last three years. By definition, patients in this visit category will stay in this visit category for all future periods and are not considered part of the panel anymore. Visit category 1 contains all patients that did not visit this year and the last two years, but visited the practice three years ago. Visit categories 2 and 3 are defined similarly. There are little patients with a very high number of visits per year. Hence, we group several numbers of visits such that a visit category consists of at least 5% of all made visits in the 5 considered years. This grouping results in a total of 17 visit categories.

Because the length of a period is a year, the age categories reflect the age in years. We decide to work with 100 age categories reflecting the ages 0 to 99. A patient of age 99 automatically leaves the panel in the following period.

For the numerical experiments we assume a constant exterior demand d_{kij} in period $k \in \{0, \dots, t-1\}$ according to age and visit category $i \in N$ and $j \in M$ equivalent to the composition of accepted new patients in 2014. Note that the real exterior demand in 2014 may have been higher. Unfortunately, we do not have records of rejected exterior demand.

To show that age and number of visits last period are predictors for the number of visits this period, we perform regression analyses. Using $17,648$ observations and including ages between 20 and 80, the coefficient of determination for an age-dependent average number of visits per year is 0.937 with an additional number of 0.21 visits per year. However, looking at individual patients, using 21811 observations of all age categories, a regression model including age and number of visits last year to predict the number of visits next year yields an adjusted coefficient of determination value of 0.631. A regression model considering the number of visits last period as an only explanatory variable yields an adjusted coefficient of determination value of 0.627. Hence, even if there is a strong correlation between age and number of visits on average the influence of age at an individual patient level is small compared to the influence of the number of visits last period.

To compute the transition probabilities q_{ijl} that a patient who belonged to age category $i \in N$ and visit category $j \in M$ last period belongs to the visit category $l \in M$ this period, we group age categories together such that there are at least 3000 patients in every considered age group. Even with the data aggregation used, for some age groups and visit categories, there are few patients (< 100) to define the transition probabilities. This problem appears mainly for the bigger visit categories and very young or old patients. Therefore, and also to smooth the tails of the transition distributions, we perform small data changes to keep the data plausible. For example, if there is a positive probability to transition from a visit category j to l and $l+2$, we add a virtual patient such that there is also a positive probability to transition from j to $l+1$. We add (or remove) 286 virtual patients based on $21,918$ real patients.

Based on the transition probabilities and the average numbers of visits per period, we can determine o_{kij}, the expected workload of a single patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods as well as u_{kij}, the variance of the workload of a single patient of age category $i \in N$ and visit category $j \in M$ in $k \in T$ periods. Using the panel of 2014 as the starting panel, we find that the variance of the workload after one period is $247,710$, which translates into a standard deviation of 497.7. Hence, the standard deviation corresponds to approximately 1 percent of the starting panel's expected workload after one period.

We do not fix specific capacities c_k, $k \in \{1, \dots, t\}$ here. We will still orientate ourselves on the number of visits in the year 2014 for the numerical experiments assuming a good match between demand and supply for the year 2014.

5. Numerical Experiments

In this section, we present numerical experiments in a deterministic setting, i.e., with deterministic exterior demand and expected future panels and later on experiments using a simulation to include uncertainty. In the following, when solving an ILP, we stop calculations whenever we reach an absolute MIP gap of t, i.e., the number of considered periods in the optimization, to speed up computation time. The MIP gap corresponds to a maximal average deviation of 1 visit from the optimal solution for every considered period. To solve our models, we use CPLEX 12.8 on an Intel 1.9 GHz PC with 16 GB. For $t < 10$, the computation time is maximal 0.2 seconds setting all parameters to those determined in Section 4. Particularly, the starting panel is set to the panel in 2014. The capacity, as well as the demand, are assumed to be the same over all considered periods. Higher values of t are neither practical (in terms of demand forecasting) nor necessary. In fact, as we will see later in the numerical experiments, considering two to four future periods already yields a significant improvement compared to considering only the next period. Solving the model without accepting MIP gaps can take a very long time, even for small values of t. However, due to the inherent variance of the workload, it is unnecessary to solve the models to optimality.

In our analyses, we consider three ILPs for a single panel, the first one being ILP (E AN), where we consider the exterior demand on the level of age and number of visits. In ILP (E A), we categorize new patients according to age only (similar to the presented ILP (E AA) but without age groups). In ILP (E), we do not categorize new patients. We just count them. We further consider three scenarios with a length of 10 periods. In a scenario, we solve an ILP once per period. The panel that results from applying the decisions of period 0 of the ILP solution is used as the starting panel for the next period. Note that those resulting panels are expected panels and therefore contain non-integer values. For every scenario, we then compare the three ILPs with a varying number of considered periods t in the optimization. In the following, we consider using $t = 1$ periods in the optimization as the base case. For example, using

ILP (E) with $t = 1$ reduces to determine the expected workload of the current panel in the next period and to divide the difference between capacity and this expected workload by the expected workload of a new patient to determine the number of new patients that should be taken in.

The first scenario considers a physician who starts a new practice working part-time with 25 percent. In the second scenario, we consider a physician working part-time with 25 percent, who increases her working hours to 50 percent. In the third scenario, we consider a group practice of three full time working physicians that reduces its capacity to two full time working physicians. Note that we choose those three scenarios in order to illustrate the benefits of considering more than one period at a time.

We define the parameters for all scenarios as determined in Section 4 apart from the starting panel and the capacity. In particular, we assume the same demand for all considered periods. Depending on the ILP, we aggregate the demand data accordingly, as described in Section 3. The panel in 2014 from Section 4 corresponds to two full time working physicians. For Scenario 1, we start with an empty panel and use a capacity of $6,074$ visits per period, which corresponds to 1/8 of the total visits in 2014. Scenario 2 uses a capacity of $6,074$ for the first 4 periods and from there on a capacity of $12,148$. Scenario 3 uses a capacity of $48,593$. To simulate the three scenarios later, we need integer-valued starting panels for Scenarios 2 and 3. Therefore, we randomly remove patients from or add patients to the 2014 panel. We continue until the panel exhibits the necessary workload of 1/8 or 3/2 of the workload of the 2014 panel. For the third scenario, we further let the panel with 3 physicians age for 5 periods with no intake of new patients. The resulting panel is the starting panel for Scenario 3. This approach shortens an otherwise long time where no patient is added.

The objectives of our ILPs minimize the sum over all considered periods t of the absolutes values of the differences between the workload of the panel and the capacity of the physician, in short, the sum of differences. To compare the ILPs and to analyze the influence of t, we use the sum over all 10 considered periods of the differences as an output.

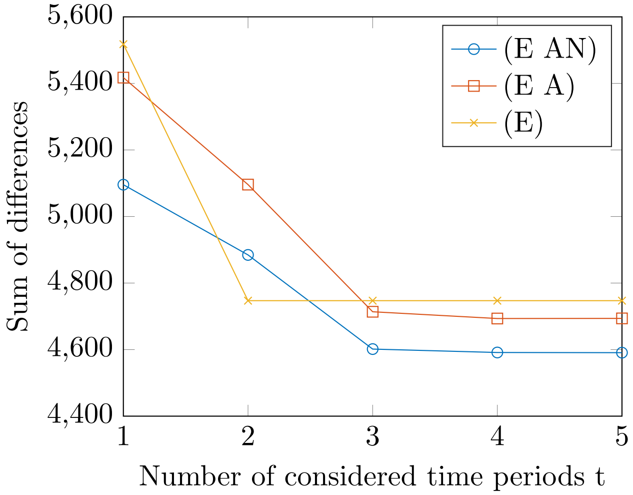

Figure 2: Sum of differences for the three ILPs and different values of t in Scenario 1

In Figure 2, we compare the sum of differences for the three ILPs with a varying number of considered periods t for Scenario 1. We see that the sum of differences decreases with the number of considered periods t for all three ILPs. The sum of differences decreases substantially from $t = 1$ to $t = 2$ and from $t = 2$ to $t = 3$ for ILP (E AN) and ILP (E A). The sum of differences of ILP (E) decreases from $t = 1$ to $t = 2$ and then remains on that level. We see no consistency in the relationship between the sum of differences of the 3 ILPs for $t = 1$ and $t = 2$. However, for $t \geq 3$, we observe that the sum of differences of ILP (E AN) is lower than the sum of differences of ILP (E A), which in turn is lower than the sum of differences of ILP (E). Considering values of $t > 5$ does not lead to significant improvements in the sum of differences. In total, we see that the lowest sum of differences $4,591$ is reached for ILP (E AN) using $t = 5$ periods. This value is 10 percent lower than the sum of differences of ILP (E AN) using $t = 1$ periods and 17 percent lower compared to the sum of differences of ILP (E) using $t = 1$ periods. Further, comparing the sum of differences for $t = 5$, we see a decrease of 1 percent between ILP (E) and ILP (E A) and a decrease of 2 percent between ILP (E A) and ILP (E AN).

How can we interpret those findings? Analyzing o_{kij}, i.e., the expected workload of a single patient of age category $i \in N$ and visit category $j \in M$ in k periods, defined by the data presented in Section 4, we realize that the expected number of visits of a random newly impaneled patient increases over the first periods. Therefore, adding a lot of new patients to the panel simultaneously without considering future effects leads to an overload in future periods. In our scenario, looking $t = 3$ periods into the future is already enough to foresee those effects and to lower the sum of differences significantly. We further see that being able to classify new patients by age or even by age and number of visits is beneficial to decrease the sum of differences further than without the classification of new patients. However, of course, there is a lower bound for the sum of differences. The demand per period, which corresponds to $2,515$ visits per period, is not enough to reach the capacity of $6,075$ visits in the first two periods even though the patients added in the first period require more visits in the second period.

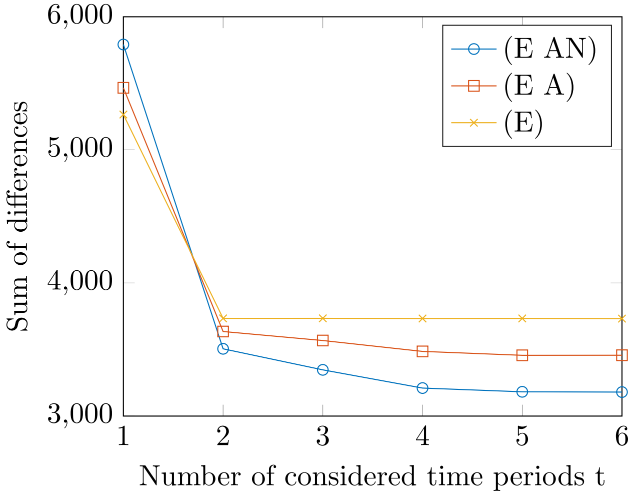

Figure 3: Sum of differences for the three ILPs and different values of t in Scenario 2

In Figure 3 for Scenario 2 we see a similar behavior to Scenario 1. Again, the sum of differences decreases with the number of considered periods t. The slop of the decrease is flattening faster for ILP (E) than for ILP (E AN) and ILP (E A). The lowest sum of differences 3180 is reached by ILP (E AN), considering $t = 6$ periods. This value is 45 percent lower compared to the sum of differences of ILP (E AN) using $t = 1$ periods. Further, comparing the sum of differences for $t = 6$, we see a decrease of 7 percent between ILP (E) and ILP (E A) and a decrease of 8 percent between ILP (E A) and ILP (E AN).

The behavior of the curves in this scenario can again be explained by a high intake of new patients that produce overload in future periods for small values of t. Again, classifying new patients helps to lower the sum of differences further. Depending on the classification it is enough to consider two to four future periods to lower the sum of differences significantly. The sum of differences can not decrease more because the change from the first capacity to the next one takes two to three periods due to the limited demand.

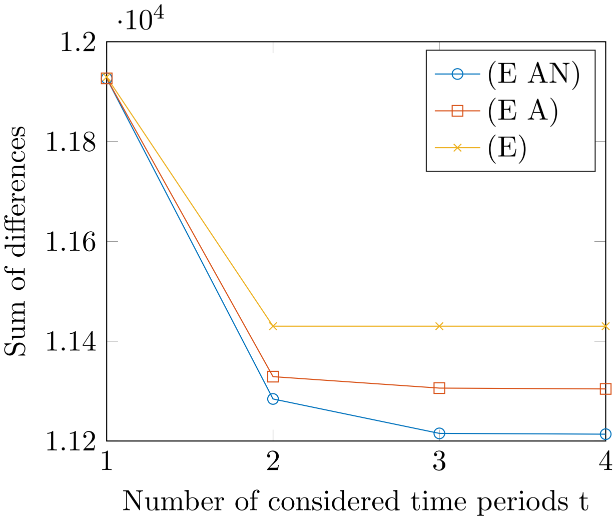

Figure 4: Sum of differences for the three ILPs and different values of t in Scenario 3

In Figure 4 we again see a similar curve behavior for Scenario 3 when compared to Scenario 1 and 2. Nevertheless, the interpretation is different. To lower the workload of the panel, new patients are not accepted in the first periods. However, once the workload reaches the new capacity, the demand is not enough to compensate for the decrease in workload over the following periods if we use $t = 1$. Looking $t = 2$ periods into the future is enough to foresee that effect for ILP (E A) and ILP (E). For ILP (E AN), looking 3 periods into the future decreases the sum of difference again slightly compared to $t = 2$. Again, classifying new patients helps to lower the sum of differences further.

Because it takes the first two periods to decrease the panel's workload close to the desired capacity, we experience a high sum of difference. However, to compare the results with our simulation results later on, we omit the first two periods in the resulting sum of differences. This reduction is possible because the differences for period 1 and 2 are the same for all considered values of t. For $t = 4$, this leads to sums of differences of 549, 640, and 765 for the ILPs (E AN), (E A), and (E), respectively. For $t = 1$, this leads to sums of differences of $1,262$, $1,262$, and $1,264$ for the ILPs (E AN), (E A), and (E), respectively. Here, comparing those results for ILP (E AN) between $t = 1$ and $t = 4$, we experience a reduction of 44 percent. Further, comparing the sum of differences for $t = 4$, we see a decrease of 16 percent between ILP (E) and ILP (E A) and a decrease of 14 percent between ILP (E A) and ILP (E AN).

For all three scenarios, in a deterministic setting, we see that considering the future panel evolution decreases the deviation between workload and capacity over time. Here, we find that it not necessary to consider numerous future periods, but instead, two to four periods suffice. Classifying new patients helps decrease the deviation further but does not have as much impact as considering future periods. Here, the decision on the detail of new patient classification can be understood as a trade-off between fairness, i.e., little to no classification, and effectiveness, i.e., a significant further decrease of the sum of differences.

Unfortunately, we are not able to test the other ILPs presented in Section 3 due to data unavailability.

5.1. Simulation

To show that our deterministic ILPs are still useful in a stochastic setting, we built a discrete event simulation in AnyLogic 8. We simulate the requests to enter the panel, the acceptance decisions concerning those requests, and the resulting panel evolution from period to period.

In a first step, we consider the stochastic transition of patients from one visit category to the next; the demand remains deterministic, i.e., the exterior demand to enter the panel is known for all considered periods. In a second step, we include stochastic demand. Hence, the rates of Poisson distributions for the patient demand of every patient class is known for all considered periods.

We simulate the three scenarios combined with the three ILPs and a varying number of considered periods t in the optimization, as presented before in Section 5. At the beginning of a period, we decide on upper bounds for the number of patients to be added this period for each patient class. When a patient arrives requesting to join the panel, she is accepted if the current number of new patients that have been accepted in her patient class is below or equal to the defined upper bound; otherwise, the patient is rejected.

Again, the main output is the sum of the absolute differences between the workload of the panel and the capacity for the 10 considered periods, which is now a random variable. To obtain statistically relevant results, we run numerous replications for each considered setting, i.e., a combination of scenario and ILP.

Only considering the panel evolution, i.e., the stochastic transition of patients from one visit category to another, we solve the ILP once at the beginning of every period. We use the solution of our ILP for the first period to define the number of patients to be added per patient class in the considered period, just as we did in our first numerical experiments.

Including demand variability in the simulation, we could solve the ILP using the expected number of patients per class as the demand input. Then, we could use the solution of the first period to define upper bounds on the number of patients to be added for each patient class for the current period. The problem with this approach is that if more patients arrive than expected, they will never be accepted even if it would be beneficial. One possibility to overcome this problem would be to resolve the ILP whenever a patient arrives who would be rejected. This approach leads to a high number of ILPs that we need to solve during a period. In practice, this is feasible because one ILP can be solved in less than a second. However, in the simulation, the high number of ILPs that need to be solved leads to very long, not feasible run times. We decided to take a different approach and solve the ILP several times only at the beginning of a period with different demand inputs to obtain upper bounds on the number of patients to be accepted for each patient class for the considered period.

For the ILP version with only one patient class, i.e., where we only count patients but do not classify them, we run the ILP with no exterior demand intake restriction in the first period. However, we constrain the intake of new patients for the following periods by the expected demand. This way, the solution of the ILP for the first period yields an upper bound on the number of patients to be added this period, taking the expected demand in future periods into account. For the two ILPs with patient classes, the approach is similar. We always constraint the intake of new patients by the expected demand for all periods except the first one. For the first period, we determine the 99 percent quantile of every Poisson distribution for every patient class. We use this quantile as the exterior demand input in the first period. The solution for the first period then yields the number of patients to be added if the demand is high.

In general, we will see some patient classes where the solution value is bigger than the expected demand. In this case, we will fix the upper bounds for those patient classes to the solution value. Of course, in general, the upper bound will not be reached when the demand is realized, which means that the ILP solution contains more workload from those patient classes than what will be available. In turn, the other patient classes are under-represented. Therefore, we solve the model again, constraining the exterior demand intake for those patient classes where we already set an upper bound by the expected demand. We set the demand input for the other patient classes to the 99 percent quantile. Again, we will probably have some patient classes where the solution value exceeds the expected demand. In those cases, we fix the upper bound to the solution value, and solve the model again, constraining the exterior demand intake by the expected demand for those patient classes. We continue until the solution value is below or equal to the expected demand for all patient classes. For those classes where we have not yet set the upper bound, we set the upper bound to the solution value of this last ILP.

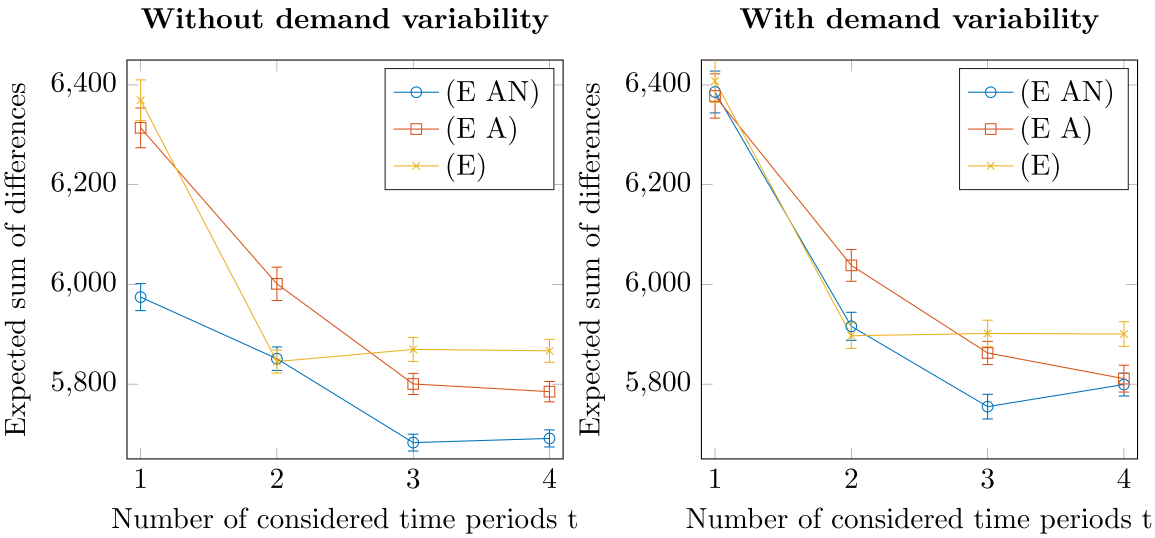

Figure 5: Simulated expected sum of differences for the three ILPs and different values of t in Scenario 1

In Figure 5, the expected sum of differences (with and without considering exterior demand variability) dependent on the ILP and the number of considered periods is shown for Scenario 1. The error bars represent the 95 percent confidence interval after $2,000$ simulation replications.

Let us first compare the simulation results without exterior demand variability with the deterministic results. We see a similar behavior of the curves with a somewhat less steep decrease in the expected sum of differences in the non-deterministic case. We further observe that the expected sum of differences $5,691$ for ILP (E AN) and $t = 4$ is $1,101$ visits higher than the sum of differences $4,590$ for ILP (E AN) and $t = 4$ in the deterministic case. The question is if this difference is mainly due to the non-avoidable variance of the workload or if it is due to the modeling of the ILPs using expected panels instead of building a stochastic ILP.

To answer this question, we simulate the panel evolution starting with the panels of the deterministic model for the periods 0 to 9 to determine the mean absolute deviation, i.e., the expected absolute difference between the workload and the expected workload after one period. Due to the triangle inequality, the absolute difference between workload and capacity is smaller than the absolute difference between capacity and expected workload and the absolute difference between workload and expected workload. The first term is minimized in the deterministic model, and the second one is the mean absolute deviation. Specifically, we computed 980 as the sum of the mean absolute deviations over the periods 0 to 9. Hence, an approximate lower bound for the expected sum of differences in the non-deterministic case is given by $4,590 + 980 = 5,570$. This estimated value is only off by 121 visits or 2 percent. This indicates that the difference between the deterministic and the non-deterministic results is mainly due to the inherent variability. The efficacy gap that could be reduced by a stochastic model formulation is very small in this case.

Comparing the graphs with and without considering exterior demand variability, we see slightly increased values for the expected sum of differences when we consider demand variability. For example, we observe a difference of 108 visits or 2 percent for the expected sum of differences for ILP (E AN) and $t = 4$. On the one hand, this shows a generally small influence of the exterior demand variability. It also indicates that our method to determine the upper bounds for the number of patients to be added in a period is effective. However, we can not say if this already small difference can be reduced further using a different method to handle demand variability.

The average number of models solved each period when considering $t = 4$ periods is 8 for ILP (E AN) and 4 for ILP (E A). Hence, we achieve good results with a small effort.

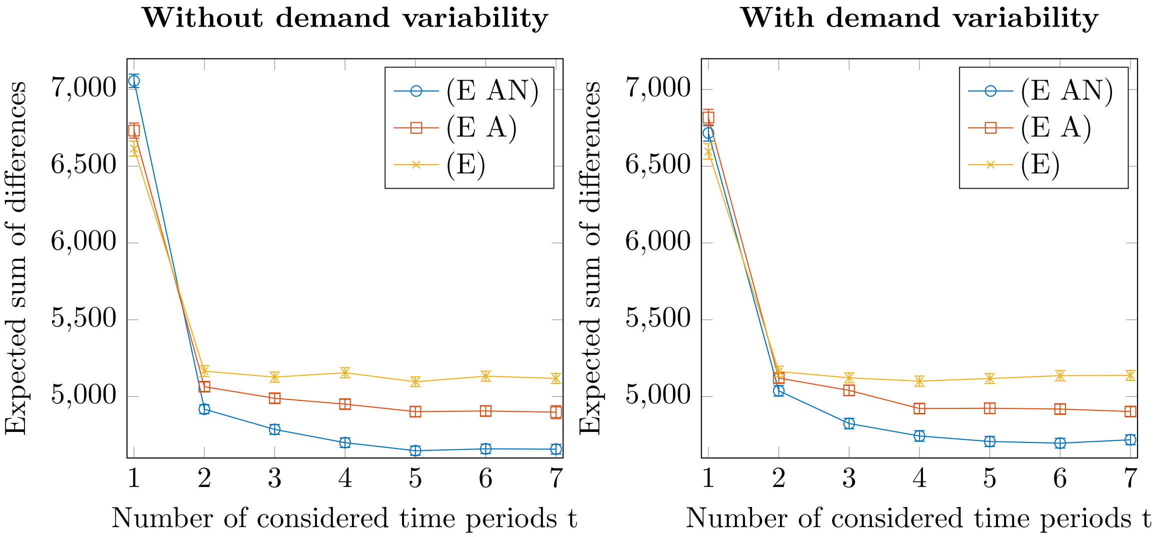

Figure 6: Simulated expected sum of differences for the three ILPs and different values of t in Scenario 2

In Figure 6, the expected sum of differences (with and without considering demand variability) dependent on the ILP and the number of considered periods is shown for Scenario 2. The error bars represent the 95 percent confidence interval after $1,000$ simulation replications.

We again see a very similar behavior between the curves for the expected sum of difference and the curves from Figure 3 for the sum of differences. Again, for big values of t, due to the considered uncertainty, the values for the expected sum of difference increase by approximately 1400 visits compared to not considering any uncertainty. Moreover, comparing the graphs with and without considering exterior demand variability, we see slightly increased values by maximal 60 visits for the expected sum of differences when considering demand variability for big values of t. The average number of models solved each period when considering $t = 7$ periods is 8.8 for ILP (E AN) and 4.5 for ILP (E A).

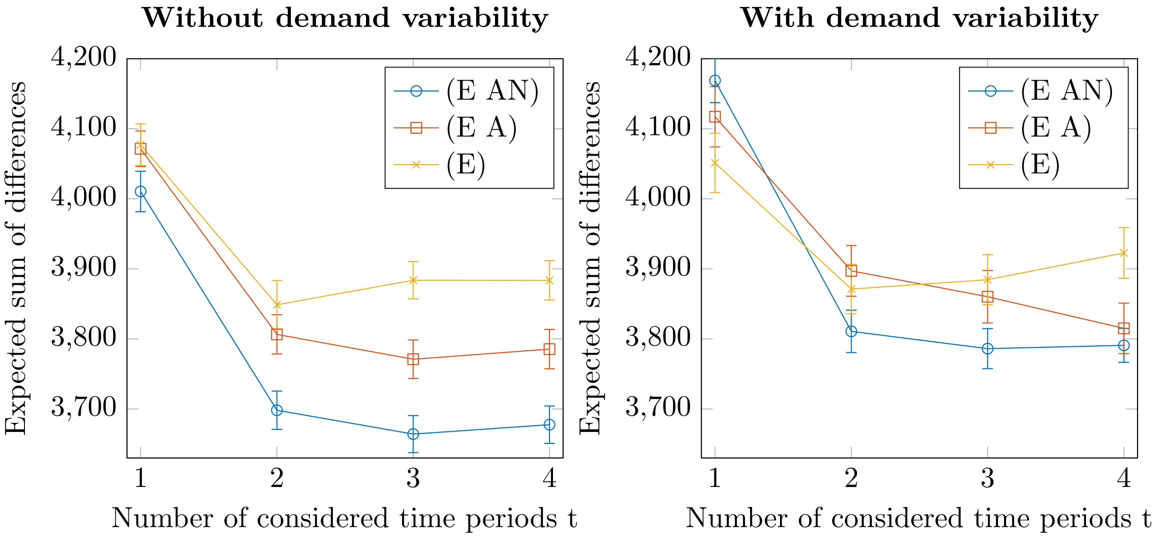

Figure 7: Simulated expected sum of differences for the three ILPs and different values of t in Scenario 3

In Figure 7, the expected sum of differences (with and without considering exterior demand variability) dependent on the ILP and the number of considered periods is shown for Scenario 3. Here, we omit the first two summands of the sum of differences due to their high variability. In the first two periods, no new patients are accepted in any setting. The error bars represent the 95 percent confidence interval after $6,000$ simulation replications.

Again, the curves in the deterministic and non-deterministic cases show similar behavior. However, due to the missing first two periods, it is difficult to compare the deterministic case and the non-deterministic case without exterior demand variability. The big difference of $3,128$ visits between the expected sum of differences and the sum of differences for ILP (E AN) and $t = 4$ steams from the variance of the workload as before, but also from the differences between the panels in period 2 since we omit periods 0 and 1. Comparing the graphs with and without considering demand variability, we experience slightly increased values by maximal 113 visits for the expected sum of differences when we consider demand variability for big values of t. The average number of models solved each period when considering $t = 4$ periods is 10.4 for ILP (E AN) and 5.2 for ILP (E A).

We see that the three considered ILPs are beneficial in a stochastic environment for all three presented scenarios. Considering more than one future period still has the most significant effect on decreasing the expected sum of differences. However, the expected sum of differences is bigger than in the deterministic case. Based on one specific example, we show that this discrepancy is probably mainly due to the non-avoidable variance. Therefore, the effect of using potential stochastic ILPs or different methods to handle the exterior demand variability is likely to be small. We further see that the impact of considering new patient demand variability is relatively small in general. No optimization can help if the total demand is to low to reach the target capacity. If we experience high new patient demand, we are more flexible in deciding whom to accept. However, the benefit of a detailed classification of new patients is small, even smaller than in the deterministic case.

These findings together indicate that a perfect demand forecast is not necessary for our programs' usefulness. We will benefit even if we take a few future periods into account with a non-perfect demand forecast and no new patient classification.

6. Conclusion and Outlook

In this paper, we present deterministic integer linear programs (ILPs) that decide on the intake of new patients into physician panels while taking into account the future panel development. The primary objective is to minimize the deviation between the expected panel workload and the physician's capacity for the current and future periods. To the best of our knowledge, this article is the first work that classifies patients by the number of visits to the physician, takes the temporal panel evolution into consideration, and that decides on the intake of new patients into existing family practice panels.

Our numerical experiments show that we can significantly lower the deviation between workload and capacity when we consider several future periods instead of one in the optimization. The deviation can be decreased a little bit further by using a detailed classification of new patients. The benefits of the developed (ILPs) remain even in an uncertain environment, taking into account as few as two to three future periods into account and without a detailed patient classification. Further, the numerical experiments show that the demand variability of new patients has a small impact on the results. Therefore, we believe that the presented models can help physicians manage their patient panels to balance supply and demand in practice.

We are aware that classifying new patients may be an ethical problem. We show that the classification has a significant benefit additional to the consideration of several future periods, but we do not imply that this benefit has to be exploited, especially in health care. However, in other application areas, the classification of new customers may be justifiable. In fact, the decision on the detail of new patient classification can be understood as a trade-off between fairness, i.e., little to no classification, and effectiveness, i.e., a significant further decrease of the sum of differences.

We find that the improvement when using potential stochastic ILPs is likely to be small which confirms the validity of using deterministic ILPs. For the case of stochastic demand of new patients with known expected values, we further propose a method to define upper bounds on the number of patients in a patient class to be accepted in a period through solving the ILP several times with different demand inputs. Instead of excessive re-optimization, this approach only needs a low number of ILP runs per period. The results are close to the results in the case of deterministic demand of new patients, which suggests a small influence of the exterior demand uncertainty and shows the validity of our method.

Another result of our work is the finding that the current panel size does not adequately describe workload. We further need to know the time frame in which we count the number of patients that have been seen by the physician. Analyzing our data, we saw that it is reasonable to use more extended time frames than the often-used two years. Further, the average workload of a single panel patient is needed to determine the panel's total workload.

We are aware the data collection is an issue. For example, Scenario 1 from Section 4, where a physician starts a new practice is a difficult use case in practice because the data needed to take the decisions is not yet available at the time of the decision taking. Therefore, it would be interesting to analyze data from several practices to see if the patient visit behavior is similar, e.g., if the workload of new patients tends to increase over the first periods.

In the future, we plan to collect more data to test the ILPs not considered in the numerical experiments here. In particular, we want to investigate further static patient attributes that influence the number of visits in a period and the ILPs for the (re-)design of several physician panels at once, including the variance objective. Working with the actual time requirement of patients measured in minutes instead of counting visits would be another interesting path of future research. We also plan on exploring further application areas for our model, besides health care.

References

Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., & Stahl, J. (2010). Improving clinical access and continuity through physician panel redesign. *Journal of general internal medicine*, *25*, 1109–15. https://doi.org/10.1007/s11606-010-1417-7.

Dreiher, J., Comaneshter, D. S., Rosenbluth, Y., Battat, E., Bitterman, H., & Cohen, A. D. (2012). The association between continuity of care in the community and health outcomes: A population-based study. *Israel Journal of Health Policy Research*, *1*, 21. https://doi.org/10.1186/2045-4015-1-21.

Green, L. V., & Savin, S. (2008). Reducing Delays for Medical Appointments: A Queueing Approach. *Operations Research*, *56*, 1526–1538. https://doi.org/10.1287/opre.1080.0575.

Green, L. V., Savin, S., & Murray, M. (2007). Providing timely access to care: What is the right patient panel size? *Joint Commission Journal on Quality and Patient Safety*, *33*, 211–218. https://doi.org/10.1016/S1553-7250(07)33025-0.

Izady, N. (2015). Appointment Capacity Planning in Specialty Clinics: A Queueing Approach. *Operations Research*, *63*, 916–930. https://doi.org/10.1287/opre.2015.1391.

Kassenärztliche Bundesvereinigung (2018). *Berufsmonitoring Medizinstudierende 2018*. Technical Report Kassenärztliche Bundesvereinigung Berlin.

Kassenärztliche Vereinigung Rheinland-Pfalz (2016). *Versorgungsatlas Rheinland-Pfalz 2016*. Technical Report Kassenärztliche Vereinigung Rheinland-Pfalz Mainz.

Liu, N., & Ziya, S. (2014). Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management*, *23*, 2209–2223. https://doi.org/10.1111/poms.12200.

Maarsingh, O. R., Henry, Y., Van De Ven, P. M., & Deeg, D. J. (2016). Continuity of care in primary care and association with survival in older people: A 17-year prospective cohort study. *British Journal of General Practice*, *66*, 531–539. https://doi.org/10.3399/bjgp16X686101.

Margolius, D., Gunzler, D., Hopkins, M., & Teng, K. (2018). Panel size, clinician time in clinic, and access to appointments. *Annals of Family Medicine*, *16*, 546–548. https://doi.org/10.1370/afm.2313.

Marshall, E. G., Clarke, B., Burge, F., Varatharasan, N., Archibald, G., & Andrew, M. K. (2016). Improving continuity of care reduces emergency department visits by long-term care residents. *Journal of the American Board of Family Medicine*, *29*, 201–208. https://doi.org/10.3122/jabfm.2016.12.150309.

Marx, R., Drennan, M. J., Johnson, E. C., Hirozawa, A. M., Tse, W. M., & Katz, M. H. (2011). Assessing and increasing patient panel size in the public sector. *Journal of Public Health Management and Practice*, *17*, 506–512. https://doi.org/10.1097/PHH.0b013e318211393c.

Murray, M., & Berwick, D. M. (2003). Advanced Access. *JAMA*, *289*, 1035–1040. https://doi.org/10.1001/jama.289.8.1035.

Murray, M., Davies, M., & Boushon, B. (2007). Panel size: How many patients can one doctor manage? *Family Practice Management*, *14*, 44–51.

Ozen, A., & Balasubramanian, H. (2013). The impact of case mix on timely access to appointments in a primary care group practice. *IIE Transactions*, *16*, 101–18. https://doi.org/10.1007/s10729-012-9214-y.

Peterson, L. E., Cochrane, A., Bazemore, A., Baxley, E., & Phillips, R. L. (2015). Only one third of family physicians can estimate their patient panel size. *Journal of the American Board of Family Medicine*, *28*, 173–174. https://doi.org/10.3122/jabfm.2015.02.140276.

Raffoul, M., Moore, M., Kamerow, D., & Bazemore, A. (2016). A primary care panel size of 2500 is neither accurate nor reasonable. *Journal of the American Board of Family Medicine*, *29*, 496–499. https://doi.org/10.3122/jabfm.2016.04.150317.

Riens, B., Erhart, M., & Mangiapane, S. (2012). *Arztkontakte im Jahr 2007 - Hintergründe und Analysen*. Technical Report Zentralinstitut für die kassenärztliche Versorgung in der Bundesrepublik Deutschland.

Schulz, M., Czihal, T., Bätzing-Feigenbaum, J., & Von Stillfried, D. (2016). Zukünftige relative Beanspruchung von Vertragsärzten - Ein Projektion nach Fachgruppen für den Zeitraum 2020 bis 2035. *Versorgungsatlas-Bericht*, *16*. https://doi.org/10.20364/VA-16.02.

Van Servellen, G., Fongwa, M., & Mockus D'Errico, E. (2006). Continuity of care and quality care outcomes for people experiencing chronic conditions: A literature review. *Nursing and Health Sciences*, *8*, 185–195. https://doi.org/10.1111/j.1442-2018.2006.00278.x.

Vanberkel, P. T., Litvak, N., Puterman, M. L., & Tyldesley, S. (2018). Queuing network models for panel sizing in oncology. *Queueing Systems*, *90*, 291–306. https://doi.org/10.1007/s11134-018-9571-4.

Wolinsky, F. D., Bentler, S. E., Liu, L., Geweke, J. F., Cook, E. A., Obrizan, M., Chrischilles, E. A., Wright, K. B., Jones, M. P., Rosenthal, G. E., Ohsfeldt, R. L., & Wallace, R. B. (2010). Continuity of care with a primary care physician and mortality in older adults. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *65 A*, 421–428. https://doi.org/10.1093/gerona/glp188.

Zacharias, C., & Armony, M. (2017). Joint Panel Sizing and Appointment Scheduling in Outpatient Care. *Management Science*, *63*, 3978–3997. https://doi.org/10.1287/mnsc.2016.2532.