Capacity Planning for Heterogeneous Patient Populations in Primary Care and Specialty Networks

CAPACITY PLANNING FOR HETEROGENEOUS PATIENT POPULATIONS IN PRIMARY CARE AND SPECIALTY NETWORKS

A Dissertation Presented

by

PRASHANT MECKONI

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2023

Mechanical & Industrial Engineering

CAPACITY PLANNING FOR HETEROGENEOUS PATIENT POPULATIONS IN PRIMARY CARE AND SPECIALTY NETWORKS

A Dissertation Presented

by

PRASHANT MECKONI

Approved as to style and content by:

Hari Balasubramanian, Chair

Ana Muriel, Member

Senay Solak, Member

Stephen S. Nonnenmann, Graduate Program Director
Mechanical & Industrial Engineering

ACKNOWLEDGMENTS

I am extremely grateful to my advisor, Prof Hari Balasubramanian, who unquestionably supported me throughout my PhD journey. He has been extremely generous with his time, guidance and motivation. He would start every meeting appreciating me for the work done and the efforts put in. I humbly thank my committee members, Prof Ana Muriel and Prof Senay Solak, for giving numerous insights and perspectives to my work, making my dissertation richer.

I am thankful to Prof Chaitra Gopalappa for the early guidance on how to conduct research and for getting me excited to pursue my PhD. I am grateful for her support and guidance over all these years. I am grateful to Prof James Smith for his teachings and his book on Queuing Networks. I appreciate his patience and faith in me when I was still struggling in my first semester of graduate school. I would also like to thank Prof Bernd F. Schliemann for being the best TA supervisor, and John Caranci from the Science & Engineering Library for his daily motivations and wisdom on organizing knowledge.

Some teachers leave a big impact on your and make you fall in love with a topic. For me, Prof L Ganapathy from NITIE, was that teacher. I thank him for introducing me to Operations Research.

I am grateful to the Debian GNU/Linux project for making available all the software I needed for my research work. A big thanks to Richard Stallman and all the GNU Emacs developers for the beautiful text editor that I used every day for my research.

ABSTRACT

CAPACITY PLANNING FOR HETEROGENEOUS PATIENT POPULATIONS IN PRIMARY CARE AND SPECIALTY NETWORKS

MAY 2023

PRASHANT MECKONI

B.E., K.J. SOMAIYA COLLEGE OF ENGINEERING, UNIVERSITY OF MUMBAI

P.G.D.I.E., NATIONAL INSTITUTE OF INDUSTRIAL ENGINEERING, MUMBAI

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Hari Balasubramanian

Access to primary care has a direct impact on morbidity and mortality, and is strongly influenced by indirect waiting time: the delay between the requested and allotted appointment day. We present a new modeling framework to describe the heterogeneous appointment seeking patterns of a primary care patient panel. Specifically, we define a patient-level stochastic process parameterized to reflect the diversity of primary care visit rates in the United States which includes realistic features such as recurring appointments (with detailed time-stamps related to each appointment), cancellations and no-shows. We then model the superposition of the stochastic processes of the panel of patients, using a simulation framework over a long time horizon to quantify: (a) the distribution of daily appointments from a capacity planning point of view, and (b) the distribution of delays for different patient classes in a closed loop queueing system.

From the capacity planning viewpoint, we estimate the distribution of daily appointments, and show that the variability of the distribution can be significantly reduced by heuristics that intelligently use patient flexibility regarding the day of the appointment. From the

viewpoint of delays, we demonstrate that in a first-come, first-served system, patients who need the most frequent appointments suffer the greatest delays, motivating the need to reserve slots for high-visit patient classes. Our simulation model of recurring visits for the fixed panel also shows that primary care practices can operate at lower capacity than the mean demand with relatively low delays, since the fixed patient panel works in a closed loop. These insights are not possible using analytical queueing networks and aggregate level single-period models that have been used so far in the panel size literature.

To further understand the inequity in delay, we model the primary care appointment system as a Discrete Time Markov Chain (DTMC). While our primary care access models discussed above focus on day-level delay (more meaningful in practice), we show the equivalence between slot-level delay and day-level delay in the DTMC. We derive an analytical expression for delay in terms of the patient's probability of daily visit. We show that conditions for monotone mapping of the probability of visit to delay are intractable and give numerical results that support monotonicity.

In our last chapter, we expand our scope beyond primary care to include specialty care networks. Using patient-level longitudinal data from the Medical Expenditure Panel Survey (MEPS), we model the sequence of appointments with multiple specialty types and the time intervals between such appointments as a Markov Renewal Process (MRP). We use comorbidity count to model patient heterogeneity class and extract the MRP parameters for each class. Next, we adapt the steady state results for a MRP to provide an analytical expression of the expected fill-rate of the appointment requests by specialty and patient class. Our analytical results demonstrate that patients with higher comorbidity count typically have a lower fill-rate—because of shorter lead time between appointments—thereby necessitating either overtime or reserved slots to ensure timely access. We further simulate appointment seeking patterns of a nationally representative panel of patients in the specialty network and estimate the distribution of daily appointment requests for each specialty. Similar to the primary care case, we show that heuristics that leverage patient flexibility regarding the day of the appointment can reduce variability in appointment requests for each specialty.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

INTRODUCTION

One of the four overarching goals of the *Healthy People 2030* (Office of Disease Prevention and Health Promotion, n.d.-b) is to "eliminate health disparities, achieve health equity, and attain health literacy to improve the health and well-being of all". It follows from a similar goal of *Healthy People 2020*—"Achieve health equity, eliminate disparities, and improve health for all groups" (Office of Disease Prevention and Health Promotion, n.d.-a). Objective AHS-04 of Healthy People 2030 — 'Reduce the proportion of persons who are unable to obtain or delayed in obtaining necessary medical care' is specifically intended to eliminate or at least reduce barriers to healthcare. It has a target to reduce this proportion from 4.1% in 2017 to 3.3% in 2030 (Office of Disease Prevention and Health Promotion, n.d.-c).

While these delays appear to be small in proportion, they have a compounding impact in terms of both health outcomes and costs, generally due to disease progression. Effect of such delays in healthcare are immediately visible from long waiting queues in emergency rooms, and it is essential to know their reasons in order to eliminate delays, or at least reduce them. Patients suffering from multiple chronic conditions benefit from continuity of the healthcare provider, since only 44% patients report that different doctors they see are "coordinated at all times", while 30% patients report "Not coordinated at all or only some of the time" (O'Malley & Cunningham, 2009). Cook et al., 2020 have quantified the benefit of continuity and access in primary care. Primary care providers who had improved access to primary care provider had improved the continuity, reduced discontinuity, and decreased emergency room events, whereas those providers with worsening access had decreased provider continuity and increased emergency room events. Many practitioners of Operations Research believe that despite challenges in stakeholders' beliefs, policy and legislation, and operational economic framework accessibility can be addressed using Operations Research methods (Linda V. Green, 2008).

In this dissertation, we:

1. analyze capacity planning for multi-class primary care appointment scheduling by appointment scheduling heuristics,
2. demonstrate interventions for reducing inequity in delay in getting appointments,
3. fundamentally analyze primary care appointment delays for patients with most healthcare needs,
4. provide understanding on specialty network outpatient referral patterns for a heterogeneous patient population,
5. estimate distribution of the appointment demand distribution and the best capacity using patient flexibility.

Throughout the dissertation we use data from the Medical Expenditure Panel Survey conducted by the Agency for Healthcare Research and Quality. This survey follows households for two years in order to understand medical care usage and expenses for families and individuals. This dissertation uses nationally representative population for analysis using the data from the survey.

The dissertation outline consists of three chapters.

In chapter 1, we use heterogeneous patient panel for primary care to analyze appointment scheduling heuristics, and to provide understanding and interventions in inequity in access to healthcare as measured by delay in appointment availability. We model the heterogeneous panel behavior. We first compare performance of simple online appointment scheduling heuristics with offline know-it-all optimization models, to reduce daily appointment variance. We then determine the patterns in delay in appointments based on patient health and show its sensitivity with daily appointment capacity. We provide simple intervention to reduce inequity in delay.

In chapter 2, we model our simulation as a Markov chain to determine analytical reasons behind the differential delay seen in chapter three. The appointment calendar, when used as a random variable can represent the state of the system in a Markov chain. This modeling allows us to use properties of the Markov chain to understand the system behavior better. We are particularly interested in understanding the expected delay in appointments for the different classes of patients. Computational complexities for analyzing Markov chains restrict our model size and structure, yet it allows us to generalize the results for larger

model sizes. The intractability of the analytical model justifies the use of simulation models from chapter 1.

In chapter 3, we model appointments to a outpatient multi-specialty network for a nationally representative heterogeneous population based on their comorbidity count. We provide a data-driven approach to parameterize the outpatient visits as a Markov renewal process (MRP). We analytically derive expected fill-rate analysis based on patient class and referral specialty from the stochastic process. This fill-rate analysis can give specialty providers and estimate on the last-minute appointment requests from different patient-classes and referral specialties. We use the MRP to simulate a regional population's appointment demand and the corresponding aggregate capacity needed when a scheduling heuristic is used.

CHAPTER 1

MODELING RECURRING APPOINTMENTS FOR HETEROGENEOUS PATIENT PANELS IN PRIMARY CARE

1.1 Introduction

Patients consider their Primary Care Provider (PCP) as a source of first-contact care, as a coordinator of referrals to specialists, and as a healthcare provider who knows about all their medical problems (Grumbach et al., 1999; O'Malley & Cunningham, 2009). The availability of primary care providers is directly linked to reduced mortality rate and improved health outcomes (Macinko, Starfield, & Shi, 2007; Starfield, Shi, & Macinko, 2005). Yet timely access to primary care still remains a concern. Rust et al., 2008 analyzed the 2005 National Health Interview Survey and found that 33% of the patients visiting the emergency room "couldn't get an appointment soon enough". Cheung, Wiler, Lowe, and Ginde, 2012 expanded the same survey data from 1999 to 2009 by including ten times as many individuals, to show that among patients who needed the emergency room, 60.9% patients with Medicaid and 26.6% patients with private insurance could not get an appointment with their medical provider soon enough.

The relationship between a PCP and her patients can last for years, sometimes even decades. This relationship is formalized by the idea of a *panel* which refers to the patients to whom the PCP provides holistic care on an ongoing basis. The primary care panel sizing problem has been an active area of research in the operations research literature. Specifically, research related to optimal primary care panel size has studied the balance between timely access for patients and practice sustenance. Larger panels will improve the providers' utilization, reduce staff idle time thus improving revenue and profitability but come at the cost of increased delays, staff burnout and reduced patient satisfaction. Smaller panels will result in the converse.

In this paper, we provide a new discrete-time stochastic modeling framework to study the panel size problem in primary care. The framework considers several realistic features of primary care delivery, including: (1) widely differing/heterogeneous visit patterns among patients, (2) recurring primary care appointments, (3) granular details of the appointment scheduling process. We elaborate of these features next.

Visit Heterogeneity: Our model of visit heterogeneity follows a common pattern observed in the United States and in other countries. Specifically, while the majority of the population is relatively healthy and needs little or no primary care visits, a small fraction requires frequent visits in short intervals. This pattern mirrors the distribution of medical expenditures observed in the United States (Emily M. Mitchell, 2019, 2020, 2021). As an example, in 2018 the top 1% spenders accounted for 21% of overall healthcare expenditure, while the bottom 50% accounted for only 3% of healthcare expenditure (Emily M. Mitchell, 2021). The most frequent users of primary care are typically also patients with the highest costs. They generally have two or more chronic conditions, which may also be called *multiple chronic conditions* (MCC) or *multimorbodity*. Ozen and Balasubramanian, 2013 showed, using data from the Mayo Clinic, that the mean number of appointments increases with the increase in the chronic condition count. The CDC estimates for 2018 show that only 48.2% of US adults have no chronic conditions, while 27.2% have MCC (Boersma, Black, & Ward, 2020). One study based in Scotland's primary care shows the strong mortality link for patients with multiple chronic conditions that have missed appointments, especially when mental health conditions are also considered (McQueenie, Ellis, McConnachie, Wilson, & Williamson, 2019). The authors conclude "existing primary healthcare appointment systems are ineffective" for such patients.

Recurring Visits: A key feature of our framework is that we model *recurring PCP appointments for each panel patient*. Our focus on recurring appointments is motivated by patient level longitudinal data in the Medical Expenditure Panel Survey (MEPS). For example, fig. 1.1 shows actual recurring PCP appointment dates for three sample patients who had eight PCP visits in a two-year period (2010-2011). Recurring PCP appointments are very common in primary care panel patients, yet they have not played a role in existing

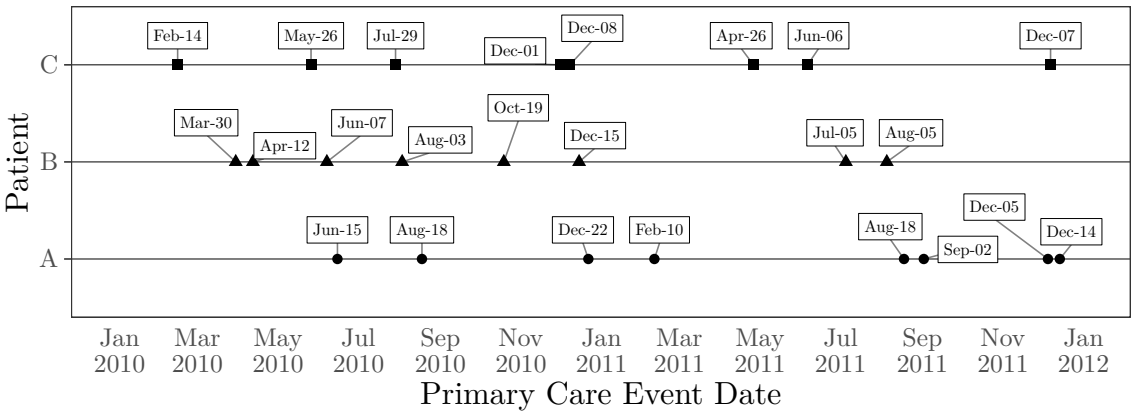


Figure 1.1: Primary care visits for three patients, each of who had recurring visits over a 2-year period, 2010-2011. Data from the Medical Expenditure Panel Survey.

panel size models. In our modeling framework, each panel patient follows a stochastic process that includes repeated visits at a rate specific to that patient.

Details of the Appointment Scheduling Process: Models used in the panel size literature have thus far assumed simplified appointment behaviors for analytical tractability. In contrast, out patient-level stochastic process for recurring appointments includes the following realistic features:

- Time points such as the day on which appointment request originated, the precise day for which it was requested, and the day it was actually scheduled.
- Follow-up requests for the next appointment that originate when the current appointment concludes as well as appointments that originate independently of prior appointments.
- Flexibility in the days that a patient desires an appointment. We model this feature based on the fact that patients who make a request well in advance, i.e., longer lead times, typically have greater flexibility regarding the days on which they can schedule an appointment.
- Reserved slots for specific patient classes, intended to mitigate the higher than average delays experienced by them.

- Cancellations that occur as a function of the appointment lead times. Specifically, appointments have a higher probability of being cancelled if they had longer lead times. We also include cancellations that occur on the day prior to the appointment day, and these serve as no-shows in our model.

In modeling how recurring appointments arising from a diverse patient panel are booked in the physician's calendar, we infer broader patterns from two different perspectives: (1) capacity planning, from the perspective of primary care providers, and (2) differences in the distribution of delays among panel patients.

In the capacity planning perspective, the practice controls day to day variability of appointments booked using heuristics that take advantage of patient flexibility regarding the day the appointment is scheduled. We benchmark the performance of the heuristics to a globally optimal integer program. We identify a simple allocation heuristic in the patient's flexibility window that significantly reduces the day to day variability in booked slots, thereby reducing the probability of both idle time and overtime.

In the second perspective, which focuses on quantifying delays, our modeling framework is a discrete time closed-loop queuing system where the calling population are the patients in the panel. In this model, we demonstrate the impact of *appointment lead time*—the difference between the requested day and the day the request originated—on patient delays. Specifically, we show the negative impact of a first-come, first-serve advance booking system, commonly used in practice, on the patients who have the greatest need for PCP appointments. We also demonstrate the impact of reserving appointments for high-need patients.

To the best of our knowledge, such realistic and practical details of the appointment scheduling process have not been studied in the panel size literature before. Panel size models have largely focused on aggregate models such as single period newsvendor like frameworks or $M/D/1/K/K$ and $M/M/1/K/K$ analytical queueing frameworks. In these models analytical insights are possible only because appointment behaviors are significantly simplified. While our detailed modeling framework is not analytically tractable and requires simulation instead, we nevertheless generate new insights not possible in prior single period and Markovian queueing models. Since we model the longitudinal appointment behavior of

each panel patient to infer broader system level patterns, our model has the flavor of an agent based simulation that includes heuristics, optimization and the principles of queueing. Furthermore, our test cases are based on nationally representative patterns observed in the MEPS survey that are representative of the US demographic.

The rest of the paper is organized as follows. In section 1.2 we provide a literature review on similar work related to primary care panel size modeling. In section 1.3, we describe panel heterogeneity and the patient-level stochastic process for recurring appointments. In section 1.4, we present a method for estimating the distribution of daily appointments and describe easy heuristics for appointment allocation that minimize the variability of this distribution. In section 1.5, we describe for appointment systems simulated with a strict limit on provider capacity and expected delays. In section 1.6, we expand on the modeling of no-shows and cancellations. In section 1.7, we describe the experimental setup and in section 1.8 we report results of our computational experiments. In section 1.9 we briefly discuss our conclusions.

1.2 Literature Review

Literature most relevant to our study lies at the intersection of panel size models and appointment scheduling in outpatient care. Appointment scheduling is a vast and still-growing area of research, which has necessitated comprehensive reviews such as Cayirli and Veral, 2003, Gupta and Denton, 2008, and Ahmadi-Javid, Jalali, and Klassen, 2017. Of relevance to our study is the difference between the direct waiting time, which is more of an inconvenience from spending time in the waiting room on the day (maximum direct waits in outpatient care rarely exceed an hour) of the appointment, and the indirect waiting time, which is the delay between the requested appointment and the actual appointment. Gupta and Denton, 2008 emphasize that indirect waiting time—measured typically in days or weeks or months—is arguably more critical in primary care since it will have a significant outcome on patient health and safety. Examples include delayed detection of conditions such as diabetes, high blood pressure, and certain cancers which lead to increases in disease severity and complications in treatment. In our study indirect wait time and its variation among panel patients is an important outcome measure.

Within the appointment scheduling literature, primary care panel size models, represent a smaller and more focused subset of this literature, and we restrict our focus on these studies in our review. The earliest studies on panel size use a deterministic framework. Some studies like those by Murray, Davies, and Boushon, 2007 provide deterministic approach to an optimal panel size that is simple and yet powerful for quick estimates. The authors take the number of visits that can be provided per year and divide it by expected annual visits per patient to obtain the panel size. The procedure is simple enough for primary care providers to do it on their own, without requiring any help from experts or consultants. However, this method overlooks the stochastic nature of visits that often results in demand-supply mismatch due to the impact of variability.

Linda V Green, Savin, and Murray, 2007 use a stochastic demand model to study the link between panel sizes and timely access to the primary care provider. They use a proxy measure called *overflow frequency* for timely access. Overflow frequency is the probability that the physician's daily demand exceeds available daily capacity. High values of overflow frequencies are likely to result in longer delays for patients. The authors assume that each patient in the panel of N patients has a probability p of requesting an appointment on any given day. If patient appointment requests are independent of one another, panel demand follows a binomial distribution with parameters N and p. If the daily capacity of the physician is known, then the probability of overflow can be easily calculated using the complement of the CDF of the binomial distribution. By varying the panel size, the probability parameter (estimated from historical data) and capacity, the feasibility of panel sizes can be tested. This is an example of a single period model of panel size.

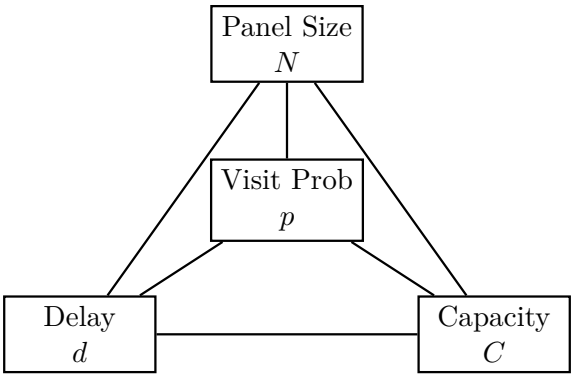
Queuing models are able to extend this model of arrivals (patient demand per unit time for appointments) and services (number of patients seen per unit time) to model backlogs (queue lengths) or waiting times in infinitely many time periods. Linda V Green and Savin, 2008 use M/D/1/K and as M/M/1/K queuing models with no-shows dependent on the patient's backlog at the time of appointment booking, to show the expected backlog as a function of the patient panel size. Zander, 2017 extends the queue to the M/D/1/K/N model by limiting the panel size and restricting new appointments to patients not already in the appointment queue to compare the backlog for various panel sizes. Liu and Ziya,

2014 show a way to determine the optimal panel size by examining appointment no-shows with a Poisson arrival process. They formulate reward maximization problems to come up with best panel size for different no-show probabilities under various appointment capacities. Overbooking is used as a policy to improve utilization in the presence of no-shows.

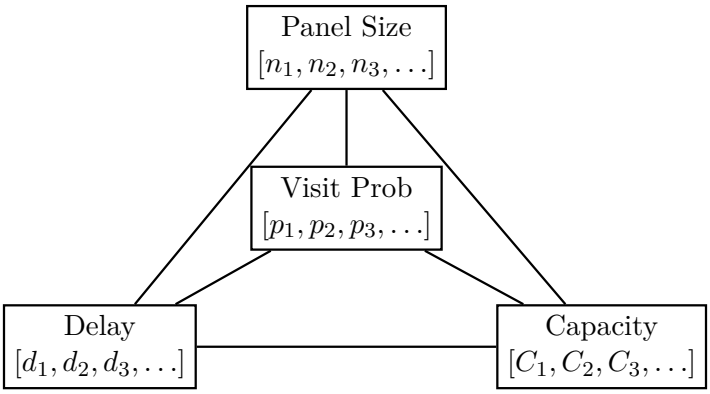
Zacharias and Armony, 2016 propose a queueing based analytical framework to study the interplay between panel size and appointment capacity. They study the problem at two different time-scales: they characterize the patient appointment arrival with binomial distribution as a GI/D/1 queue with balking and the in-clinic queue as GI/GI/1 with no-shows. The combination of these two queues is then used to maximize a net-profit based reward in order using the decision variables of appointment capacity and arrival rate dependent on panel size. The authors use many settings including advanced access to determine the optimal panel size.

We now turn to studies that explicitly model heterogeneous panels, which are sometimes referred to as *case mix* models. A central argument in these papers is that the homogeneous patient panels (each patient visiting at the same rate) are not realistic. Ozen and Balasubramanian, 2013 and Balasubramanian, Banerjee, Denton, Naessens, and Stahl, 2010 are two early examples heterogeneous panel size models. Both investigate heterogeneity in the context of at multi-physician practices to balance workloads of physicians. In both these studies, age, gender and the number of simultaneous chronic conditions (comorbidities) were used as predictors of appointment request rates. Additionally, since the models considered multi-physician practices, the question of how best to redesign panels optimally—through stochastic optimization and heuristic methods—was central to both papers. By redesigning panels, that is by changing patient-physician assignments in a group practice, each physician's panel demand can be brought in balance with the physician's available capacity and this in turn can minimize imbalances in timely access across physicians.

More recently, Harrington, Rubin, and Bai, 2021 balance workload from patient panels on existing providers in group practices and uses new hires to take up extra patient workload such that the cost of reallocating patients to different physicians in reduced. They provide evidence of linear relationship between the overflow frequency as used in Ozen and Balasubramanian, 2013 with physician utilization, which allows modeling the problem as a mixed integer



(a) Current Models



(b) Proposed Models

Figure 1.2: Differences in existing models in the panel size literature. Section 1.2 considers the average or homogeneous patient parameters that represent the entire panel, which are typical in queueing models. Section 1.2 groups similar patients together, uses different parameters for different patient groups and sometimes considers multiple providers with different capacities.

programming problem. Patient panels are partitioned in clusters and represent age, gender and annual visits to the PCP. They also use the number of chronic conditions to represent patient clusters.

Zander, Nickel, and Vanberkel, 2021 describes a model to decide if new patients can to be admitted to the patient panel handled by single or multi-physician practice by minimizing the absolute difference between the physician workload and the demand for care in number of visits. The authors use combinations of gender, age, and number of annual visits to represent heterogeneity in the panel and allow decision at that level. A novelty of the model is that patients progress to different classes with time. Deterministic scenarios use the

expected visits or its standard deviation and are modeled using mixed integer programming. Stochastic scenarios generate the demand for visits and progression to different classes are modeled using simulations. This demand is used in the mixed integer programming model for decision to admit new patients. The authors use patient visit data from a group practice in Germany to provide numerical analysis.

We also note that while heterogeneity has primarily been used to model differences in visit rates Gupta and Wang, 2008; J. Wang and Fung, 2015; W.-Y. Wang and Gupta, 2011 use patient preferences to represent heterogeneity in the patient panel.

Finally, we discuss two papers that consider recurring visits. Vanberkel, Litvak, Puterman, and Tyldesley, 2018 makes the case for recurring visits to Oncology practices to determine the patient panel size using analytical network queuing models for both stationary and non-stationary cases. Patient arrivals are segregated by their introduction to the practice as new patients, recurring visit by health status, and inactive patients who may discontinue visits. The authors provide analysis for new practices trying to start a new practice by allowing new patients, and for existing practices that try to balance new patients with existing patients. Bavafa, Savin, and Terwiesch, 2019 provide analyses on optimal revisit interval, patient preference and flexibility towards using e-visits towards need based partial replacement of office based visits, panel sizing, physician capacity and their revenue model, and the related health outcome. In their model, the authors build expressions based on revisit interval that reflect the average cost of a patient's visits to their physician, the expected revenue of a physician under fee-for-service and capitation schemes, and the impact of e-visits on the costs and revenues. They analyze conditions including panel size and visit intervals under which the physician and the patients can aim for their own optimal rewards. They use patient health status to represent heterogeneity as "healthy" and "sick" in patient panels.

In summary, we observe (broadly speaking) two classes of models of panel size in the literature: queueing models and single period models that balance each physician's demand with capacity. The models can be both homogeneous, presenting only an analysis of average behavior, as well as heterogeneous, allowing for differences in patient visit rates and involving multiple providers (see Figure 1.2 for a visual summary). While both classes of models have

provided high-level insights on the link between panel size and capacity planning, neither approach considers the precise dynamics of how patients seek primary care appointments. In queuing models, appointment requests are aggregated as arrival rates and patients join a queue until their service begins. However, in practice appointments are often booked in advance, and the lead time (how far ahead the appointment was requested), which queueing models and other aggregate models do not consider, plays an important role in both capacity planning as well as delay distributions. In contrast to prior papers, we simulate the stochastic progression of recurring appointments (including time points such as appointment origin, day of request, day scheduled) for every panel patient in time. Thus, our model has a flavor of an agent-based simulation coupled with optimization and queuing dynamics, which allows us to quantify aggregate patterns related to capacity planning and delay distributions. In our results we demonstrate insights that would not have been possible using queueing and single period models.

1.3 Heterogeneous Panel Model

We consider a panel of N patients associated with a primary care physician. This panel is represented as a set H. We partition the panel into disjoint classes indicated by subscript j. Each partition H_j represents the set of patients with similar healthcare needs. We use the number of annual visits to the healthcare provider as a surrogate measure of these healthcare needs. A patient of class j needs N_j primary care appointments each year. Each patient $k \in H_j$ has the same probability p_j of requesting an appointment on any given day, where $p_j = N_j/D$, assuming D workdays in a year. Thus, each patient class is homogeneous, while the panel is heterogeneous. The number of patients in class H_j is $n_j = |H_j|$, $\sum_j n_j = |H| = N$ and $H_j \cap H_{j'} = \emptyset$ for all $j \neq j'$.

For all evaluations in this paper, we use the panel composition as shown in table 1.1. Our panels consist of 20 classes of patients. They are based on total annual visits observed per individual in the 2011 Medical Expenditure Panel Survey. Since MEPS is a nationally representative survey, the primary care visit patterns observed can be reliably assumed to follow the US demographic. The p_j values in the table are calculated based on annual visits and assuming 250 workdays in a year. This distribution of annual visits for individuals

in a nationally representative sample of 2000 individuals was first discussed in Rossi and Balasubramanian, 2018. Analysis of the nationally representative samples in other years of the survey also reveals identical patterns. While we test various panel sizes in this paper, for each panel size value, we retain the proportions of individuals in each class. The panel composition in table 1.1 shows a great deal of variation in visits: 56.5% of patients need one or less visits per year, while less than 3.2% of patients need 12 visits or more per year. There are 15 individuals who need 20 or more visits in a year; for the purposes of our study, we combine them into a single group.

For each patient in the heterogeneous panel of size N, our model explicitly considers the longitudinal pattern of seeking PCP appointments. To describe this behavior for each patient, we establish the following notation. Suppose the $(i-1)^{\text{th}}$ PCP appointment for patient k was on day $a_{(k,i-1)}$. Let $r_{(k,i)}$ be the requested day for appointment number i for patient $k \in H_j$, and suppose that the request for this appointment originated on $o_{(k,i)}$, where $a_{(k,i-1)} \leq o_{(k,i)} < r_{(k,i)}$.

Further, we assume that the probability for the request for next appointment is made on some day later than the previous appointment is p_b. So, with probability $(1-p_b)$, the request for the next appointment is made immediately after the prior appointment is completed, i.e. $o_{(k,i)} = a_{(k,i-1)}$. This occurs in many situations where a follow-up appointment is scheduled in the PCP office after the consultation is complete. With probability p_b, the request the request arises on any day leading up to $r_{(k,i)}$, i.e. $a_{(k,i-1)} < o_{(k,i)} < r_{(k,i)}$. The precise day on which the request originates is assumed to follow a discrete uniform distribution with each day from $a_{(k,i-1)} + 1$ to $r_{(k,i)} - 1$ having an equal probability. This case reflects the situation where the patient does not anticipate a follow-up after seeing the PCP on $a_{(k,i-1)}$ but later on day $o_{(k,i)}$ experiences symptoms that lead to the request on $r_{(k,i)}$. Due to the lack of data, we have assumed the value of p_b as 0.5. Note that in our model, $o_{(k,i)} < r_{(k,i)}$: the origin of the request is always less than the day of the request. Same day appointments are not explicitly considered since the granularity of our model is one day. However, our model does allow patients to request an appointment the very next day; these requests are reasonable proxy for same-day requests.

In our model, we first generate the day of the next request using eq. (1.1).

Table 1.1: Distribution of patients by frequency of visits in a year

MEPS Data		Input for Simulation			
Annual visits $\widehat{j}$	Number of patients $\widehat{n_j}$	Probability of visit on a day $p(j) = \frac{\widehat{j}}{250}$	Number of patients n_j	Proportion of panel $\frac{n_j}{\sum_j n_J}$	Class j
0	687	—	—	—	—
1	442	0.002[a]	1129[b]	0.565	1
2	271	0.008	271	0.136	2
3	173	0.012	173	0.087	3
4	111	0.016	111	0.056	4
5	68	0.02	68	0.034	5
6	51	0.024	51	0.026	6
7	39	0.028	39	0.02	7
8	36	0.032	36	0.018	8
9	22	0.036	22	0.011	9
10	22	0.04	22	0.011	10
11	15	0.044	15	0.008	11
12	15	0.048	15	0.008	12
13	10	0.052	10	0.005	13
14	4	0.056	4	0.002	14
15	7	0.06	7	0.004	15
16	3	0.064	3	0.002	16
17	5	0.068	5	0.003	17
18	3	0.072	3	0.002	18
19	1	0.076	1	0.001	19
20+	15	0.08	15	0.008	20
TOTAL	2000		2000	1	

[a] Daily visit probability $p_1 = 0.002 = \frac{0.5}{250}$ to accommodate merging of patients having 0 and 1 annual visits.
[b] Class 1 merges patients having 0 and 1 annual visits.

$$r_{(k,i)} := a_{(k,i-1)} + X_j, \quad k \in H_j. \tag{1.1}$$

where X_j follows a Geometric distribution with parameter p_j and $\mathbb{E}[X_j] = 1/p_j$. While any distribution could be used for X_j, we chose the Geometric based on the histograms (see Appendix section 1.10) for intervals between successive PCP appointments for individuals surveyed in the Medical Expenditure Panel Survey (MEPS). Next, the origin day for i^{th} request of patient k is determined using eq. (1.2).

$$\begin{aligned} o_{(k,i)} := a_{(k,i-1)} + X_b X_{u_{(k,i)}}, \\ \text{where } X_b \sim \text{Bern}(p_b), \\ \text{and } X_{u_{(k,i)}} \sim \text{Unif}\left\{0, r_{(k,i)} - a_{(k,i-1)} - 1\right\}. \end{aligned} \tag{1.2}$$

There is an important nuance related to $o_{(k,i)}$. When $o_{(k,i)} > a_{(k,i-1)}$, the origin and request days are future scheduled events in the simulation event calendar; the practice remains unaware of when the next appointment request originates until the simulation moves to $o_{(k,i)}$.

Finally, the i^{th} appointment which originated on $o_{(k,i)}$ is scheduled on $a_{(k,i)}$ which may or not be the same as the $r_{(k,i)}$. The exact day that the appointment is scheduled depends on whether whether the model is (a) uncapacitated with patients having some flexibility around the requested date (discussed in section 1.4) or (b) has capacity constraints with patients having to experience delays (discussed in section 1.5). Some of the relevant timepoints are indicated in fig. 1.3.

While we have described the patient-level stochastic process for recurring appointments, this process needs to be reconciled to a time horizon. Suppose there are T days in the time horizon, where T is very large, in the scale of years, since the panel is expected to stay with the PCP for many years. Let $t = 1, 2, \ldots, T$ denote a day in the horizon. The simulation is initialized by randomly generating the first appointment request $r_{(k,1)}$ from eq. (1.1) by assuming $a_{(k,0)} = 0$ (i.e. a dummy appointment day 0) for all patients. We count the number of appointments on day t as A_t as shown in eq. (1.3)

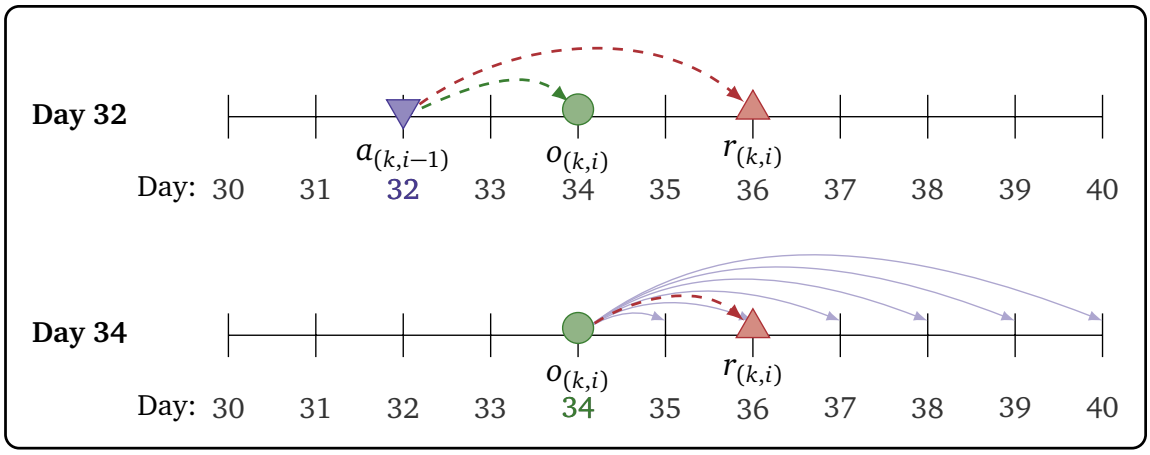


Figure 1.3: Relevant time-points in patient-level stochastic process related to the scheduling of the next appointment. The first panel shows that at $a_{(k,i-1)} = 32$, the request and origin dates, $r_{(k,i-1)} = 36$ and $o_{(k,i-1)} = 34$, are generated in the simulation event calendar. The practice is not aware at this time that a request will arise from the patient on day 34. On Day 34, the practice receives the patient's request, and determines where the next appointment should be scheduled. The blue arrows indicate that days on which the next appointment can be scheduled.

$$A_t := \sum_{k \in H} \mathbb{1}_{a_{(k,i)}=t} \tag{1.3}$$

1.4 Capacity planning

Given the patient-level, heterogeneous appointment request process described in the previous section, we now develop a methodology to estimate *the distribution of daily appointment slots* such that all appointment requests arising from the panel during the time horizon are satisfied without delay. Since we are interested in quantifying the daily demand distribution, we assume in this section that there is no constraint on the PCP capacity. The purpose of estimating the demand distribution is to allow the PCP to adequately plan their daily capacity using newsvendor-like models that balance the probability of the PCP going idle (under-utilization) with the probability of overtime (i.e. the PCP working beyond their designated capacity level). Included in our modeling framework are heuristics and an optimization model that allow a PCP to reduce their day to day variation in appointment slots by optimally using patient flexibility related to the day the appointment is scheduled.

One simple technique to estimate the distribution of daily appointment slots is to assume that each patient appointment request to be satisfied on the day it was requested—that is,

$a_{(k,i)} = r_{(k,i)}$. With $A_t, t = 1, 2, \ldots, T$ calculated assuming $a_{(k,i)} = r_{(k,i)}$, we can obtain the distribution of daily appointment slots used in each day of the time horizon.

The downside of such an approach is that it exposes the primary care physician to significant day to day variability, similar to the kind of variability seen in urgent care centers and emergency rooms. A more practical and realistic scenario, given that patients requests are typically non-urgent, is that a patient has some *flexibility* with regard to the requested appointment date. Specifically, the patient is satisfied so long as the appointment is scheduled in a time window centered around the request date. Let $\delta_{(k,i)}$ be the flexibility (in days) associated with request $r_{(k,i)}$. The appointment can then be scheduled on a day given by eq. (1.4).

$$\begin{aligned} r_{(k,i)} - \delta_{(k,i)} \leq a_{(k,i)} \leq r_{(k,i)} + \delta_{(k,i)} \\ l_{(k,i)} \leq a_{(k,i)} \leq u_{(k,i)} \end{aligned} \tag{1.4}$$

where we use $l_{(k,i)} \coloneqq r_{(k,i)} - \delta_{(k,i)}$ and $u_{(k,i)} \coloneqq r_{(k,i)} + \delta_{(k,i)}$ for the lower and upper limits for $a_{(k,i)}$.

The introduction of flexibility gives the practice a method reduce the day to day variability. While there is no limit on how many daily PCP slots are available, patient flexibility allows the practice to *spread the appointment slots more evenly in the time-horizon*, thereby reducing variability. The flexibility $\delta_{(k,i)}$ can take many forms, but the the most realistic scenario would be to base the flexibility on the urgency of the appointment requested. This urgency is reflected by the number of days in advance that the appointment is requested – i.e. the appointment lead time. Thus, $\delta_{(k,i)} \propto r_{(k,i)} - o_{(k,i)}$: the shorter the lead time, the higher the urgency and the smaller the flexibility $\delta_{(k,i)}$.

While the above relationship certainly holds in practice, precise data on how patient flexibility changes in relation to lead time is typically not collected by practices. Flexibility is often informally expressed in a patient's conversations with the scheduler at the time the appointment is booked, and are difficult to capture quantitatively. In the absence of such data, in our model we assume $\delta_{(k,i)}$ follows eq. (1.5).

$$\delta_{(k,i)} := \min\left(\left\lfloor \frac{r_{(k,i)} - o_{(k,i)} - 1}{5} \right\rfloor, 7\right) \tag{1.5}$$

Equation (1.5) implies that flexibility increases by a day for each week (5 working days per week) of lead time, and the maximum flexibility is capped at 7 working days. For example, if the difference between the requested and origin dates is 12 (i.e. if the request was made two weeks and two days in advance), the patient will have a flexibility of 2 days, while those appointments requested within 5 days are urgent requests with no flexibility. We note that other (less or more restrictive) flexibility-lead time relationships can be incorporated into the model.

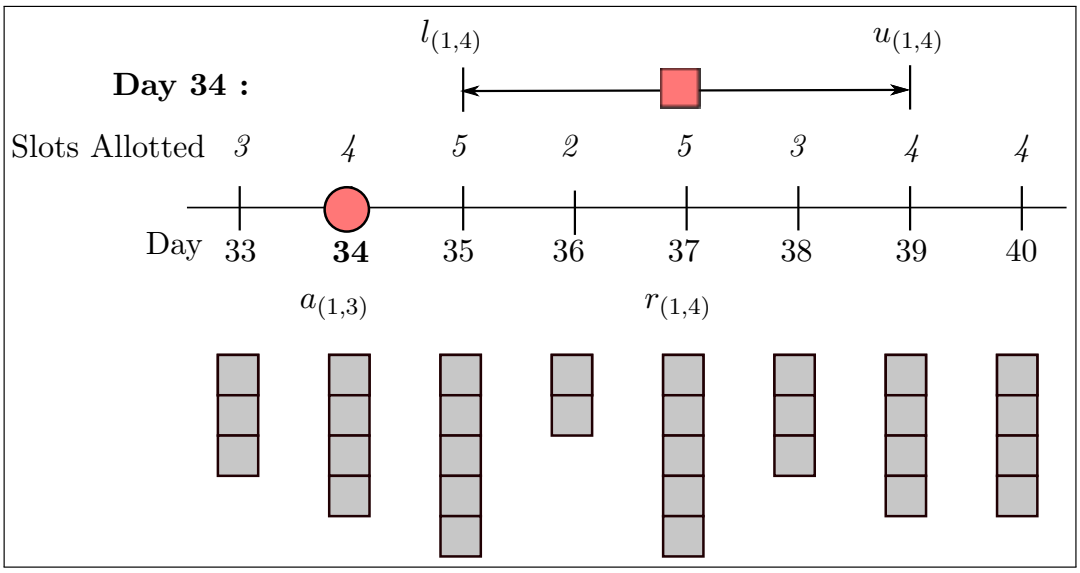
1.4.1 Heuristics for Assigning Appointment Requests

The introduction of flexibility in appointment dates implies that the scheduler has to decide at $o_{(k,i)}$, i.e. the day the request originated, where in the window $\left[l_{(k,i)}, u_{(k,i)}\right]$ the request must be scheduled. Since the PCP is interested in reducing the day-to-day variability in daily appointment slots, the best strategy is to schedule the requested appointment on that day which has the lowest total booked slots. Since multiple days in the window can have the same number of booked slots, the scheduled can choose the *earliest such day* in $\left[l_{(k,i)}, u_{(k,i)}\right]$ with the lowest slots booked. We call this the *First Minimum* heuristic. We consider this heuristic assuming we want to allocate most of the appointments as early as possible (to reduce potential under-utilization), while balancing the number of appointments on all days.

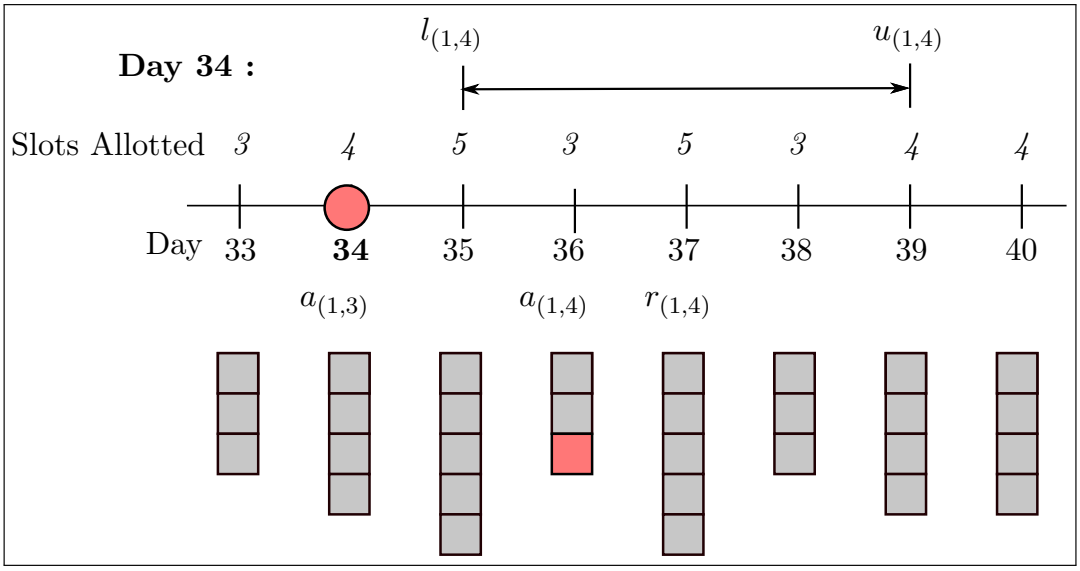
The first-minimum heuristic is illustrated in fig. 1.4 where a patient makes an appointment request (fig. 1.4a), the scheduler allocates a slot to that request (fig. 1.4b). After that, another patient makes an appointment request (fig. 1.4c) which is allotted a slot by the scheduler (fig. 1.4d).

In addition, to First Minimum, two other simple heuristics that can be used by schedulers are described below.

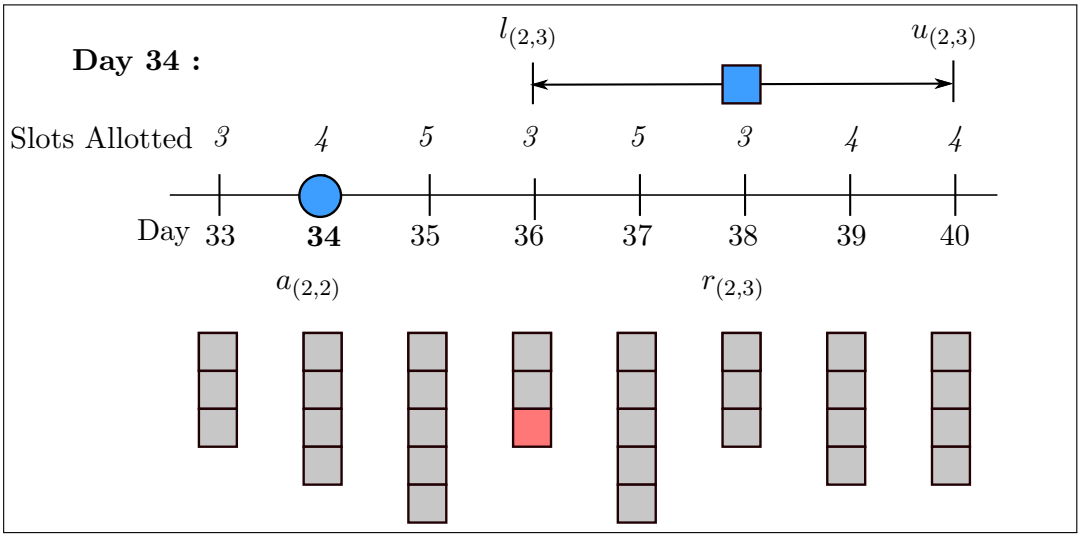
Last Minimum The day with the minimum number of appointments in the interval $\left[l_{(k,i)}, u_{(k,i)}\right]$ is allotted. Ties are broken by selecting the latest of such days. We consider this assuming we want to keep open earlier slots for frequently visiting



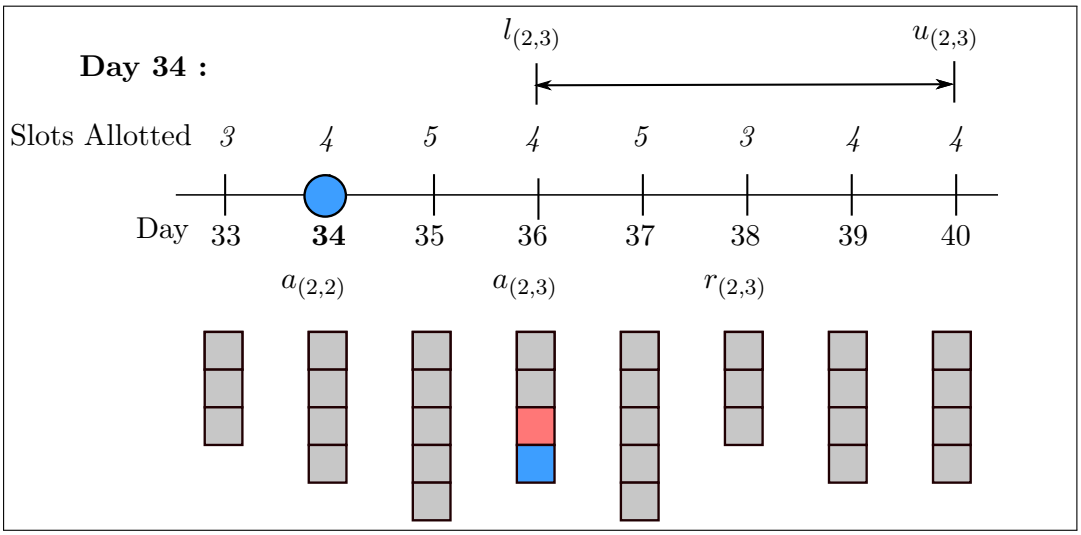
(a) At the end of the third appointment of Patient-1 on day 34 shown by $a_{(1,3)}$, she requests for a next (fourth) appointment on day 37 denoted by $r_{(1,4)}$. She has a flexibility δ of 2 days, so her lower limit for the appointment is on day 35 denoted by $l_{(1,4)}$ and her upper limit is on day 39 denoted by $u_{(1,4)}$.



(b) The scheduler looks up for the minimum slots allotted for other appointments on each of the days from day 35 to day 39. She finds that day 36 has the minimum number of slots allotted so far (two slots). This is the only day which has two slots. She allocates the request $r_{(1,4)} = 37$ on day $a_{(1,4)} = 36$. There are now three slots allotted on day 36.



(c) Patient-2 completes her second appointment $a_{(2,2)}$ on day 34 after Patient-1. She makes a request for her third appointment $r_{(2,3)}$ for day 38. Patient-2 also has flexibility δ of 2 days. Her lower limit $l_{(2,3)}$ for the appointment is on day 36 and the upper limit $u_{(2,3)}$ is on day 40.



(d) The scheduler looks up for minimum slots allotted for appointments on each of the days from 36 to day 40. She finds that day 36 and day 38 have the minimum number of slots allotted so far (three slots). She selects the earliest day of the candidate days for breaking the tie. She allocates the request $r_{(2,3)} = 38$ on day $a_{(2,3)} = 36$.

Figure 1.4: Simulation illustrated using the first-minimum heuristic.

patients that have less flexibility, while balancing the number of appointments on all days.

Uniform Random Any day in the interval $\left[l_{(k,i)}, u_{(k,i)}\right]$ is chosen with a same (uniform) probability. We consider this heuristic to see if random allocation can work better.

For each of the heuristics, we can obtain $A_t, t = 1, 2, \ldots, T$, which gives the distribution of daily appointments, as well as $\mu_{\text{sim}} = \max_t A_t$, the minimum capacity needed to satisfy every request without delay.

1.4.2 Offline Optimization: A Mixed Integer Programming Approach for Assigning Appointment Requests

We also introduce an *offline optimization* approach that assumes perfect information about each patient's appointment request dates $r_{(k,i)}$ in the time horizon (generated using a particular heuristic) and the flexibility $\delta_{(k,i)}$ associated with each request date. Instead of minimizing the maximum assigned slots in a window corresponding a particular patient request (which First Minimum and Last Minimum achieve), the integer program determines the *globally optimal* assignment of requests in the entire time horizon, $t = 1, 2, \ldots, T$ with the objective of minimizing the maximum slots assigned to each day in the horizon. Such a minimax approach helps reduce day to day variability of scheduled slots in the entire horizon. While the integer programs perfect information assumption is unrealistic, its principal role in this study is to provide a benchmark on the performance of each heuristic. The optimization formulation is provided below.

$$\text{minimize} \quad \mu_{\text{opt}} \tag{1.6a}$$

$$\text{subject to} \quad \sum_{(k,i)} a_{(k,i),t} \leq \mu_{\text{opt}} \qquad \forall t, \tag{1.6b}$$

$$\sum_{t=l_{(k,i)}}^{u_{(k,i)}} a_{(k,i),t} = 1 \qquad \forall (k,i), \tag{1.6c}$$

$$\mu_{\text{opt}} \geq 0, \tag{1.6d}$$

$$a_{(k,i)} \in \{0,1\} \quad \forall (k,i) \tag{1.6e}$$

The binary decision variable $a_{(k,i),t}$ determines the day t within the interval $\left[l_{(k,i)}, u_{(k,i)}\right]$ on which the appointment related to request $r_{(k,i)}$ is allotted, as shown in eq. (1.4). Each appointment request is allotted exactly once as shown in eq. (1.6c). The sum of all appointments for each day t is restricted by the capacity μ_{opt}, as shown in eq. (1.6b). The integer program seeks to minimize μ_{opt}.

We can determine the daily sum of allotted appointments using eq. (1.7) from the appointments allotted in the optimal solution.

$$a_t^* = \sum_k \sum_i a_{(k,i),t} \quad \forall t \tag{1.7}$$

Recall that the integer program uses the $r_{(k,i)}$ and $\delta_{(k,i)}$ values specific to each heuristic ($r_{(k,i)}$ values can differ from one heuristic to another because the allocation dates of the previous appointment need not be identical). Thus, we simulate the heuristic over the entire time horizon, and record the $r_{(k,i)}$ and $\delta_{(k,i)}$ values for all appointments. These are used as inputs to the integer program (as perfect information known *a priori*) to determine the global minimum daily count of appointments.

In summary, we have illustrated a methodology by which to estimate the total daily appointment distribution as well as the minimum daily appointments needed in a horizon by a heterogeneous panel. Our approach incorporates a commonly observed feature in primary care practice: the presence of patient flexibility that depends on the urgency of the appointment. The use of patient flexibility by heuristics that are easily to implement when the request arises allow a practice to generate estimate the daily distribution of appointments. We also demonstrate how each heuristic can be bench-marked with mixed integer program that provides a globally optimal solution by assuming perfect information.

1.5 Quantifying Delays Specific to Each Patient Class

In this section, we look at the panel size problem the perspective of delays experienced by each patient. Unlike the previous section which used an uncapacitated model to infer distribution of daily appointments, we now assume that the provider has a strict daily limit

on the capacity; and in the absence of available capacity for a given day, patient requests are scheduled at the next available day, creating delays or, more precisely, indirect wait times. The main purpose of this section to illustrate that in a heterogeneous panel in which appointments are secured in a first-come, first-served manner, delays are concentrated among individuals with the greatest need for PCP appointments, and that these patients require reserved slots.

In the panel size literature, such appointment systems are frequently analyzed with traditional queuing approaches, for example, as $M/D/1/K$ or $M/M/1/K$ models. Such models assume a single aggregated arrival rate λ and provide outputs as the expected waiting time for the average patient. Recurring appointment behavior of individuals within a closed population (the panel) is not explicitly considered. While such aggregate results can be used for simple back-of-the-envelope analyses, they do not accurately reflect the the appointment behaviors and wait times experienced individual patients of groups of patients in the panel.

In what follows, we assume a strict limit μ on the daily PCP appointments that can be allotted. Delays arise due to unavailability of vacant appointment slots on certain days when $A_t = \mu$. As seen earlier, a patient k requests her appointment number i on day $o_{(k,i)}$ where $o_{k,i} \geq a_{k,i-1}$. When there are no vacant slots on the requested day $r_{(k,i)}$, the scheduler searches for the subsequent day with a vacant slot and allots the appointment $a_{(k,i)}$ on that day. This can be expressed as eq. (1.8), where the scheduler allocates the i^{th} appointment for patient k on the earliest day which has a vacant slot on or after day $r_{(k,i)}$.

$$\begin{aligned} a_{(k,i)} &:= \arg\min_t \{A_t | A_t < \mu, t \geq r_{(k,i)}\} \\ A_{a_{(k,i)}} &:= A_{a_{(k,i)}} + 1 \end{aligned} \tag{1.8}$$

The patient experiences an absolute delay that can be measured using eq. (1.9)

$$d_{(k,i)} := a_{(k,i)} - r_{(k,i)} \tag{1.9}$$

When the appointment allotted is for the same day same as the requested appointment the delay is zero since $a_{(k,i)} = r_{(k,i)}$.

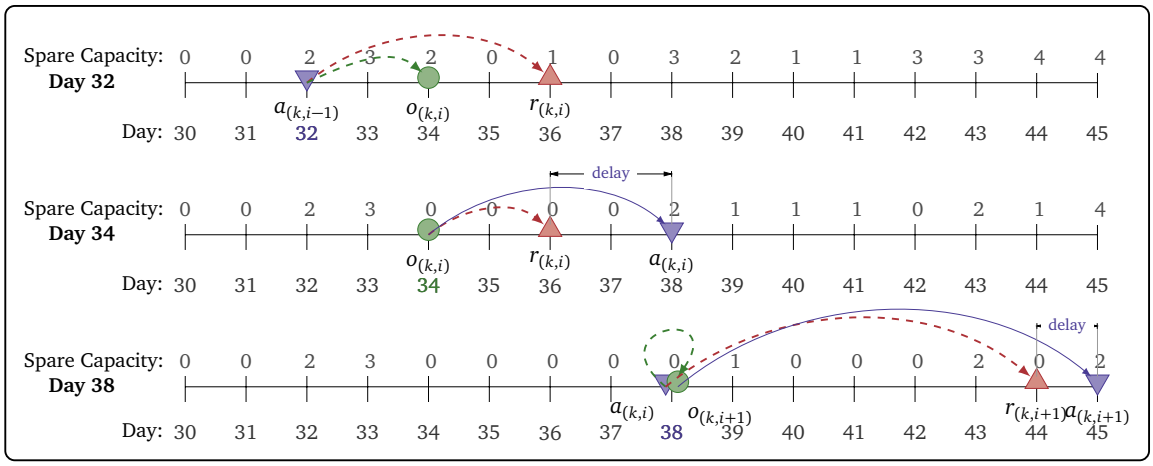


Figure 1.5: Sample illustration of allocation of appointments for requests originating on day 34 and then on day 38 for a patient k. On day 32, at the end of appointment, the patient's next request and its origin day are randomly generated as day 36 and day 34 respectively. The patient and the practice remain unaware of the request until day 34. On day 34, the patient requests for an appointment on day 36. As there are no appointment slots available, she is allotted an appointment for day 38. On day 38, after the appointment, the patients next request and its origin are randomly generated. Since the origin day is also day 38, the patient will request the next appointment for day 44 and will be allotted an appointment for day 45. The delay for appointment i is 2 days and the delay for appointment $i+1$ is 1 day.

We are also interested in the expected delay which is $\mathbb{E}\left[d_{(k,i)}\right] = \mathbb{E}\left[a_{(k,i)} - r_{(k,i)}\right] = \mathbb{E}\left[a_{(k,i)}\right] - \mathbb{E}\left[r_{(k,i)}\right]$. Since the previous appointment $a_{(k,i-1)}$ is known, we can treat it as a constant. We have the solution for $\mathbb{E}\left[r_{(k,i)}\right]$ from eq. (1.10), but we do not have the solution for $\mathbb{E}\left[a_{(k,i)}\right]$ from eq. (1.11).

$$\begin{aligned}\mathbb{E}\left[r_{(k,i)}\right] &= a_{(k,i-1)} + \mathbb{E}[X_j] \\ &= a_{(k,i-1)} + \frac{1}{p_j} \qquad (1.10)\\ \mathbb{E}\left[a_{(k,i)}\right] &= \mathbb{E}\left[\arg\min_t \{A_t | A_t < \mu, t \geq r_{(k,i)}\}\right] \qquad (1.11)\end{aligned}$$

Since an analytical solution for the expected delay is not possible, we simulate the appointment system over a sufficiently long time horizon to obtain the steady state distribution of waiting times for each class of patients.

We begin by initializing daily appointments allotted as $A_t = 0$ for each day in the simulation t. The first request and corresponding origin are determined by assuming that everyone had the previous appointment on day 0, that is $a_{(k,0)} = 0\ \forall k$. We use the eq. (1.1), eq. (1.2) to determine the appointment request $r_{(k,i)}$ and the origin day $o_{(k,i)}$ respectively. A vacant slot is available on day t when $A_t < \mu$. If the vacant slot is available, then it is allotted corresponding to the patient's request, else we search for a vacant slot on the next day. Thus the patient k is allotted their ith request for appointment using eq. (1.8) and the number of allotted slots is increased for that day.

We determine $\mathbb{E}\left[d_{(k,i)}\right] \to \overline{d_j}$, using the strong law of large numbers applied to our simulation output. The mean delay for the jth class of patients can be estimated using eq. (1.12) which is sum of delays for each appointment for each patient in class j divided by the number of appointments for each patient that class.

$$\overline{d_j} = \frac{\sum_k \sum_i d_{(k,i)} \mathbb{1}_{H_j}(k)}{\sum_k \sum_i \mathbb{1}_{H_j}(k)} \tag{1.12}$$

1.5.1 Conjecture – Delays increase with increase in p_j values

In a multi-class first-come, first-serve queueing system, for example an $M/M/1$ or $M/D/1$ system, all classes will have identical mean delay values in steady state. Differences in delays arise only when some classes are prioritized over others. In the closed population recurring appointment queueing system that we consider in this paper, appointments are also allotted on a first-come-first-serve basis. Thus, in the absence of explicit prioritization, the mean delays in principle should also be identical across the different patient classes represented by the p_j values.

However, this is not true. Indeed, we conjecture that $p_1 < p_2 < \ldots < p_J$ implies $d_1 < d_2 < \ldots < d_J$ where J represents the total number of disjoint classes in the panel H. In other words, patients in the panel who have higher need for PCP appointments will have a higher mean delay. This is because requests for a particular day in the horizon are likely to arise earlier from patients who have lower p_j values, i.e. patients have a longer interval between appointments. In a first-come, first-serve system, these requests are fulfilled early,

increasing the chance that a patient needing requesting an appointment at a short-notice (shorter interval between appointments) will not be satisfied on the requested day. Thus, a first-come, first-serve system unwittingly ends up prioritizing patients who need fewer PCP appointments.

Queuing models are not able to capture this effect is because they work at an aggregate level. The time stamps relevant to a queueing model involve the arrival time, start of service time and completion time. Delay is measured as the difference between start of service and arrival time. In our model we include the following time stamps relevant to a patient's stochastic process: previous appointment day $a_{(k,i-1)}$, origin of request $o_{(k,i)}$, request day $r_{(k,i)}$ and the day the appointment is scheduled $a_{(k,i)}$. Traditional single server queueing models keep track of $r_{(k,i)}$ as the arrival time but they are "blind" to, or have no memory of time stamps $a_{(k,i-1)}$ and $o_{(k,i)}$. This in turn implies they have no memory of the *lead time*, $r_{(k,i)} - o_{(k,i)}$ prior to the booking of an appointment. In a first-come, first-serve system, appointments requested with longer lead times are less likely to experience delays and more likely to be booked on time.

We can formalize this idea by obtaining an expression for the lead time, i.e. the expected number of days before which a patient requests an appointment using eq. (1.13).

$$\begin{aligned}
\mathbb{E}\left[r_{(k,i)} - o_{(k,i)}\right] &= \mathbb{E}\left[r_{(k,i)}\right] - \mathbb{E}\left[o_{(k,i)}\right] \\
&= a_{(k,i-1)} + \mathbb{E}\left[X_j\right] - a_{(k,i-1)} - \mathbb{E}\left[X_b X_{u_{(k,i)}}\right] \\
&= \mathbb{E}\left[X_j\right] - \mathbb{E}\left[X_b\right] \cdot \mathbb{E}\left[X_{u_{(k,i)}}\right] \\
&= \frac{1}{p_j} - p_b\left(\frac{a_{(k,i-1)} + \mathbb{E}[r_{(k,i)}] - 1}{2}\right) \\
&= \frac{1}{p_j} - \frac{p_b}{2}\left(a_{(k,i-1)} + a_{(k,i-1)} + \mathbb{E}\left[X_j\right] - 1\right) \\
&= \frac{1}{p_j} - \frac{p_b}{2}\left(2a_{(k,i-1)} + \frac{1}{p_j} - 1\right) \\
\mathbb{E}\left[r_{(k,i)} - o_{(k,i)}\right] &= \frac{1}{p_j}\left(1 - \frac{p_b}{2}\right) - p_b\left(a_{(k,i-1)} - 2\right)
\end{aligned} \tag{1.13}$$

It can be easily shown from Equation (1.13) that when we have two patients k and k' of different classes j and j' complete their previous appointments i and i' on same day, if

$p_j < p_{j'}$, then $\mathbb{E}\left[r_{(k,i)} - o_{(k,i)}\right] > \mathbb{E}\left[r_{(k',i')} - o_{(k',i')}\right]$. The patients with higher p_j will be asking the scheduler for their requested appointments earlier than patients with lower p_j. From eqs. (1.10) and (1.13), we say that patients with a lower p_j are expected to have their next appointment requests further in the future and are also expected to schedule them much earlier than the patients with higher p_j. This in turn implies that a first-come, first-serve policy will tend to benefit patients with lower p_j values and adversely impact timely access for patients with higher p_j values.

Our conjecture can be viewed as an extension of conflict between the pre-scheduled versus same-day appointments, which has been discussed before in the literature. A calendar that is packed with pre-scheduled appointments booked in advance risks timely access for same-day patients who need appointments at a short notice (typically within a few hours). Our model in this paper extends that principle to patients (typically those with multiple chronic conditions) who need recurring appointments in short intervals, whose timely access might be compromised by appointments booked well in advance by patients whose intervals between appointments are longer. To avoid this problem, reservations are needed for the short-interval patients, i.e. patients with lower p_j values. We turn to this next.

1.5.2 Slot Reservations for the Patients with the Highest Needs

Our slot reservations work in the following manner. We first determine which classes of frequently visiting patients should have access to reserve slots. We use a set V of patient classes for which $v < \mu$ appointment slots are reserved. A patient $k \in H_j \in V$ will have access to μ appointment slots and another patient $k' \in H_{j'} \notin V$ will have access to $\mu - v$ appointment slots. This concept is used when allocating slots using eq. (1.8) and is modified in eq. (1.14).

$$a_{(k,i)} := \begin{cases} \arg\min_t \{A_t | A_t < \mu, \quad t \geq r_{(k,i)}\} & \text{if } k \in H_j \in V \\ \arg\min_t \{A_t | A_t < \mu - v, \quad t \geq r_{(k,i)}\} & \text{if } k \in H_j \notin V \end{cases} \tag{1.14}$$

This method implies that patients in classes represented by V will have at least v and at the most μ slots available to them. When the demand is more than v slots, then such

patients can be allotted regular slots as everyone, if they are vacant. For other patients, the appointment slot availability remains capped at $\mu - v$. These patients cannot access the reserved slots even when there is no demand for the reserved slots. While the simulation described here aggregates the reserved slots for the classes in V, the simulation model allows additional sets of classes to have specific slots reserved for them, thus allowing higher control on the reserved slots.

1.6 Modeling Cancellations and No-Shows

We model cancellations to quantify their impact on both capacity planning as well as delays. Essentially our models (whether intended for capacity planning or delays), can switch cancellations on or off.

Recall that the i^{th} appointment of patient k is scheduled on day $a_{(k,i)}$. The patient can cancel the appointment on any day between $o_{(k,i)}$ (the day the appointment was made) and $a_{(k,i)}$. The probability that a patient will cancel the appointment depends on the time interval between when the appointment originated $o_{(k,i)}$ and the actual appointment date $a_{(k,i)}$. The longer the interval the more likely the probability of cancellation. In particular, we use the expression provided in by Linda V Green and Savin, 2008 to determine the probability that the patient will cancel the appointment:

$$\gamma(a_{(k,i)} - o_{(k,i)}) := \gamma_{\max} - (\gamma_{\max} - \gamma_0)\, e^{-(a_{(k,i)} - o_{(k,i)})/C} \tag{1.15}$$

In eq. (1.15), C is the no-show sensitivity parameter, $\gamma_0 \geq 0$ is the minimum observed no-show rate, and $\gamma_{\max} \in (\gamma_0, 1]$ is the maximum observed no-show rate. These rates can inferred for each practice. Thus, based on eq. (1.15), the cancellation probability lies somewhere between the default γ_0 and the maximum $\gamma_{\max}$. The larger the difference $a_{(k,i)} - o_{(k,i)}$ the more likely it is lie closer to $\gamma_{\max}$.

Equation (1.15) yield a probability that is used to generate a Bernoulli random variable X_γ. When $X_\gamma = 0$, the appointment is not cancelled, and when $X_\gamma = 1$ the appointment is cancelled. In the latter case, a second random variable X_Δ from a triangular distribution is used to determine the precise day between $o_{(k,i)}$ and $a_{(k,i)}$ on which the patient chooses to

cancel: $X_\Delta \sim \text{Tri}\left[o_{(k,i)}+1, a_{(k,i)}, a_{(k,i)}\right]$. This gives a random day between the day after the origin day and day of the appointment, with the mode taken as the day of the appointment. This distribution reflects the reality that patients may realize they are not going to be able to make it for the appointment as the days progress towards the appointment day. Cancellations that happen on the day of the appointment, $a_{(k,i)}$, are treated as no-shows and the appointment slot goes unused. For cancellations done earlier than the day of appointment, the vacated slot is made available for new appointment requests, if any. When a patient cancels an appointment, her new appointment day and origin day are generated as if she had been seen for an appointment on the cancellation day.

More formally we have:

$$\begin{aligned} c_{(k,i)} &:= X_\gamma \text{round}\left(X_\Delta\right) \\ &\text{where } X_\gamma \sim \text{Bern}\left(\gamma\left(a_{(k,i)} - o_{(k,i)}\right)\right) \\ &\text{and } X_\Delta \sim \text{Tri}\left[o_{(k,i)}+1, a_{(k,i)}, a_{(k,i)}\right]. \end{aligned} \tag{1.16}$$

The round$(\cdot)$ operator is used to convert the continuous values to discrete values. If $c_{(k,i)} = 0$ then the appointment will not be canceled. We have a no-show when $c_{(k,i)} = a_{(k,i)}$. When $c_{(k,i)} \neq 0$, we will cancel the appointment on day $c_{(k,i)}$ and determine a new ith appointment request and origin in a way similar to eq. (1.1) and eq. (1.2) by assuming a pseudo appointment $a'_{(k,i)}$ has taken place on the day of cancellation. This is shown in eq. (1.17).

$$\begin{aligned} a'_{(k,i)} &:= c_{(k,i)} \\ r_{(k,i)} &:= a'_{(k,i)} + X_j \\ o_{(k,i)} &:= a'_{(k,i)} + X_b X_{u_{(k,i)}} \end{aligned} \tag{1.17}$$

The allotted appointments on that day is also reduced, as shown in eq. (1.18).

$$A_{c_{(k,i)}} := A_{c_{(k,i)}} - 1 \tag{1.18}$$

The parameters used for modeling cancellation are taken from Linda V Green and Savin, 2008 and are summarized in table 1.2

Table 1.2: Cancellation parameters

Parameter	Value
γ_0	0.01
$\gamma_{\max}$	0.31
C	50

1.7 Experimental Setup

Now that we have described our two models—one for capacity planning, the other for quantifying delays—as well our method for modeling cancellations which applies to both, we are now in a position to our experimental setup. Table 1.3 summarizes the parameters of the distributions used in the simulation of the heterogeneous panel. Table 1.4 shows parameters for capacity planning while table 1.5 shows the parameters used in the simulation for delay estimation.

In both cases, we use the panel sizes of 1800, 2000 and 2200, since they fall in the range of commonly observed panel sizes in the US. The case-mix of each panel reflects the PCP visit behavior shown in table 1.1. Thus, whatever the panel size, we have 20 classes of patients based on their expected annual visits.

For the delay model, we use a daily capacity of $\mu \in \{\lceil\lambda\rceil, \lceil\lambda - 2\rceil, \lceil\lambda - 4\rceil\}$, where $\lambda = \sum_j n_j p_j$ is the daily arrival rate for a panel of size $N = \sum_j n_j$. Note that in closed

Table 1.3: Heterogeneous panel parameters

Random Variable	Distribution	Parameter
X_j	Geometric	$p(j)$ from table 1.1
X_b	Bernoulli	$p_b = 0.5$
X_u	Discrete Uniform	Unif $\left\{0, r_{(k,i)} - a_{(k,i-1)} - 1\right\}$
X_γ	Bernoulli	$\gamma\left(a_{(k,i)} - o_{(k,i)}\right)$
X_Δ	Triangular	Tri $\left[o_{(k,i)} + 1, a_{(k,i)}, a_{(k,i)}\right]$

Table 1.4: Capacity Planning parameters

Parameter		Values	Combinations
Panel Size	$N \in$	{1800, 2000, 2200}	3
Cancellation Policy	$\in$	{No, Yes}	2
Simulation Days	$=$	1250 (5 years)	1
Simulation Repetitions	$=$	100	100
TOTAL SIMULATIONS FOR EACH HEURISTIC			600

Table 1.5: Simulation parameters for Delay estimation

Parameter		Values	Combinations
Panel Size	$N \in$	{1800, 2000, 2200}	3
Capacity	$\mu \in$	$\{\lceil\lambda\rceil, \lceil\lambda - 2\rceil, \lceil\lambda - 4\rceil\}$	3
Reservation Slots	$v \in$	$\{0, 1\}$	2
Reservation Class	$V =$	{18, 19, 20}	1
Cancellation Policy	$\in$	{No, Yes}	2
Simulation Days	$\in$	2500 (10 years)	1
Simulation Repetitions	$=$	10	10
TOTAL SIMULATIONS			360

population queuing models can be stable (i.e. reach steady state) even when the arrival rate λ exceeds the service rate μ.

Finally, to illustrate the benefit of reserved slots, we reserve a single slot each day ($v = 1$) for patients in classes 18, 19 and 20—that is three patient classes that have the greatest annual visits. As a result, for patients in these classes there are μ slots available each day, while for all remaining classes, there are $\mu - 1$ slots.

1.8 Results

1.8.1 Capacity Planning

We first start with results relevant to capacity planning. Figure 1.6 shows the histograms of daily scheduled appointments for three panel sizes (1800, 2000 and 2200) and the three different heuristics (First Minimum, Last Minimum, and Uniform Random), assuming patient flexibility. For illustration purposes, over each histogram we superimpose the histogram of the no-flexibility case. In the no-flexibility case, appointments are scheduled on the requested

day and there are no decisions to be made. Therefore the no-flexibility histogram is identical in all cases. The mean, standard deviation, 20th and 80th percentiles are shown for each pair of overlapping histograms and in each of the nine figures in the panel. We note the following key conclusions from the histograms:

- The First Minimum and Last Minimum heuristics lead to tighter, less variable histograms compared to the no-flexibility case. The first minimum heuristic in particular leads to the least variable histograms. For example, under the first minimum heuristic when the panel size is 2000, the standard deviation of the daily slot distribution is 1.78 while daily slot distribution under the no-flexibility case has a standard deviation of 4.34. A comparison of the 20th and 80th percentiles further confirms the differences in variability between the two cases.

- The tighter histograms that the First Minimum heuristic produces imply less day to day variability for the PCP. Less variability implies that the PCP has fewer extremes of under and over utilization. Such extremes occur commonly in the no-flexibility case. The last minimum heuristic also performs well in this regard, but the standard deviations are slightly higher. These results illustrate that the natural flexibility that patients have around non-urgent appointments can be intelligently utilized via the first and last minimum heuristics to reduce day to day variability.

- In contrast to the above, if the scheduler decides to allocate a patient's request on a random day in the flexibility window around each appointment, the daily slot histograms are virtually *identical* to the no-flexibility case. Thus, if the natural flexibility around non-urgent PCP appointments is not utilized optimally, the PCP can experience more days where the utilization is low (excessive idle time) as days where the daily appointment demand is high (excessive overtime).

- The mean daily slots under the three different heuristics show subtle differences. The mean daily slots under the Last Minimum heuristic is slightly smaller because the scheduler first identifies days in the flexibility window that have the smallest number of slots booked, and in the case of a tie always chooses the last of these minimum

slot days. This results in slightly longer intervals between appointments. When this policy is followed consistently, the longer inter-appointment intervals lead to a slight reduction in the mean daily slots booked. This is the reason why the histograms for the Last Minimum heuristic exhibits a slight left shift compared to the no-flexibility histogram.

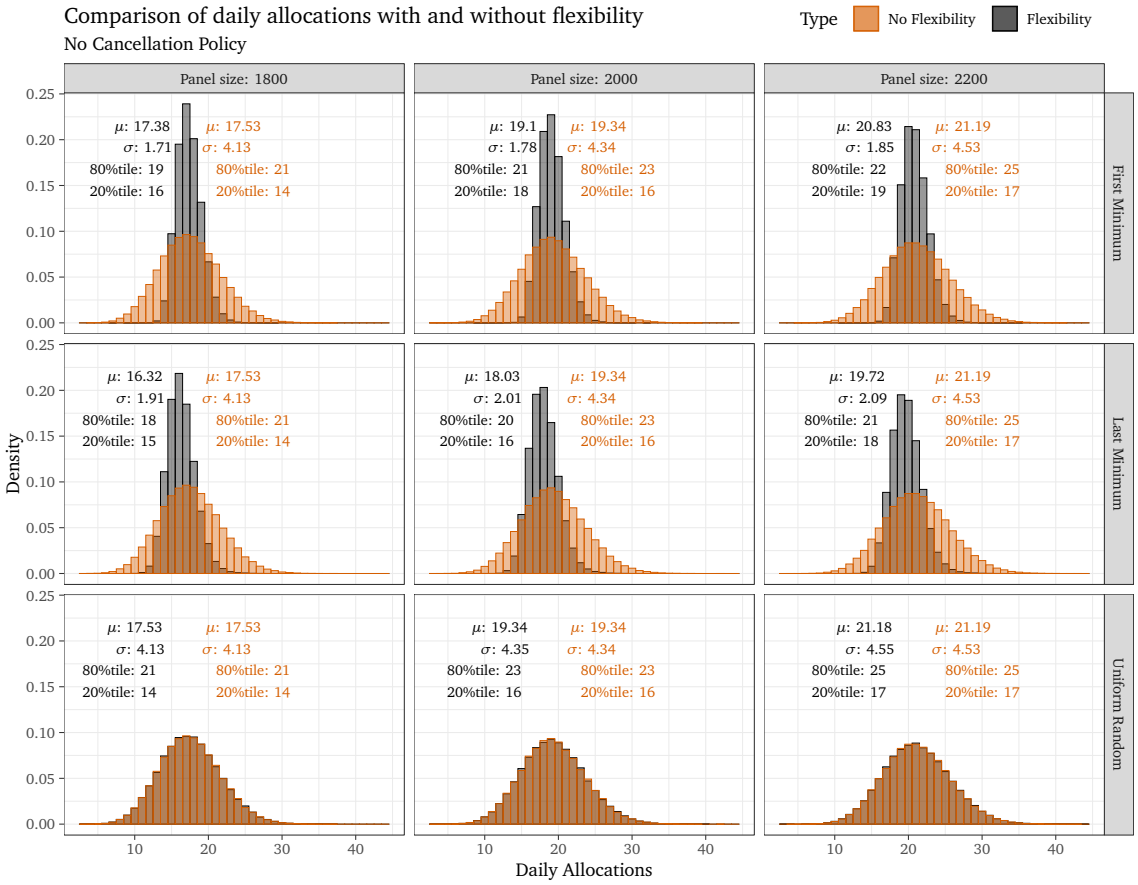


Figure 1.6: Comparison of the daily appointment distributions for the three heuristics and the base case of no-flexibility under different panel sizes. Note the significant decrease in variability under the First Minimum and Last Minimum heuristics compared to no-flexibility and Uniform Random allocation.

Next, we quantify how the heuristics compare to the globally optimal integer program. We borrow the phrase "Value of Perfect Information" from stochastic programming, to demonstrate our findings. Since the optimal integer program described in section 1.4.2 requires us to know all the requested appointments and the flexibility windows *a priori*, we can consider perfect information is available for determining the optimal solution. Recall

that in the integer program we are interested in minimizing the maximum value of daily slots booked in the 1250-day horizon. In contrast, the heuristics operate with partial information, and are used repeatedly as the schedule evolves. Each time an appointment is booked, the scheduler minimizes the maximum slots in the flexibility window of that appointment by using the First or Last Minimum heuristics. In the case of Uniform Random, the scheduler picks an arbitrary day in the window. Thus the heuristics work 'locally' or 'myopically' until the appointments are booked in the 1250-day horizon. For each heuristic hur, the maximum number of slots in the horizon (μ_i^{hur}) in replication i can be compared with the globally optimal integer program value, (μ_i^{opt}). Specifically, we can estimate the EVPI by averaging the difference in the maximum daily slots between heuristic hur and the globally optimal integer program across the 100 replications:

$$\text{EVPI} = \frac{1}{100}\sum_{i=1}^{100}(\mu_i^{hur} - \mu_i^{opt}) \tag{1.19}$$

Table 1.6 shows the EVPI results both each heuristic for three panel sizes, with and without cancellation. On average, the First Minimum heuristic has a maximum daily slot value that between 2-3 slots higher than the integer program when no cancellations are included while the Last Minimum heuristic performs slightly worse. In the cancellation case, the Last Minimum shows shows slightly better performance compared to the First Minimum and is within 1-2 slots of the integer program optimal value. In contrast, the Uniform Random heuristic has a maximum daily slots value that is between 9-12 slots higher than the integer program under no cancellations; and between 8-10 slots higher when cancellations are included. These differences are directly attributable to the less variable daily distributions produced by the First and Last Minimum heuristics and to the highly variable distribution produced by the Uniform Random heuristic. Higher variability leads to higher maximum values of daily slot booked in the horizon. These results demonstrate that the First and Last Minimum heuristics perform reasonably well compared to the integer program despite having only partial information.

Cancellation Policy	Panel Size	Expected Value of Perfect Information		
		First Minimum	Last Minimum	Uniform Random
No	1800	2.21	2.67	9.79
	2000	2.06	2.73	10.44
	2200	2.56	3.66	11.32
Yes	1800	1.91	1.55	8.71
	2000	1.54	1.24	8.36
	2200	2.24	1.66	9.16

Table 1.6: The Value of perfect information in terms of appointment slots is determined using mean of differences of Simulated Capacity and Optimal Capacity from eq. (1.19)

1.8.2 Delays

Figure 1.7 shows how the mean delay varies as a function of the lead time $r_{(k,i)} - o_{(k,i)}$, i.e. the number of days in advance the appointment was requested. The results are for a panel size of 2000, under three different capacities, $\mu \in \{\lceil\lambda\rceil, \lceil\lambda - 2\rceil, \lceil\lambda - 4\rceil\}$, with and without cancellations. The results reveal how delays are high for those patients who need appointments with short lead times, this trend gets more pronounced as the daily PCP capacity drops. Cancellations cut down the mean delays by more than 50% under all three capacity settings.

When $\mu = \lceil\lambda\rceil$ for panel size of 2000, we have $\lambda = 19.3$ and $\mu = 20$ and yields a utilization of 0.96. Such a utilization is considered quite high in queueing theory. Yet if patients request appointments with a less time of 8 days or more, they are likely to experience minimal indirect delays. When $\mu = \lceil\lambda - 2\rceil$, appointments with a lead time of longer than 12 days have virtually no wait time; if cancellations are present, appointments with lead time greater than 8 days experience minimal waits. Thus the $\mu = \lceil\lambda - 2\rceil$ case with cancellations is very similar to the $\mu = \lceil\lambda\rceil$ without cancellations. These results suggest that while panel sizes where μ is close to even slightly less than λ are feasible in practice; however, to ensure timely access for short-lead time appointments, some overtime would be necessary.

The $\mu = \lceil\lambda - 2\rceil$ and $\mu = \lceil\lambda - 4\rceil$ cases can also be interpreted as appointment systems in which physician-patient consultation times (i.e. appointment durations) are longer than the $\mu = \lceil\lambda\rceil$ case. Thus, we can conclude from fig. 1.7 that an increase in delays due an

increase in the average physician-patient consultation time is mitigated by the presence of cancellations. This is because cancellations create gaps in the schedule which benefit patients with short lead time requests, thereby reducing their indirect delays. The lowest lead times are typically observed in patients with the greatest frequency of annual visits, i.e. the highest p_j values. Thus these patients experience the highest mean delays.

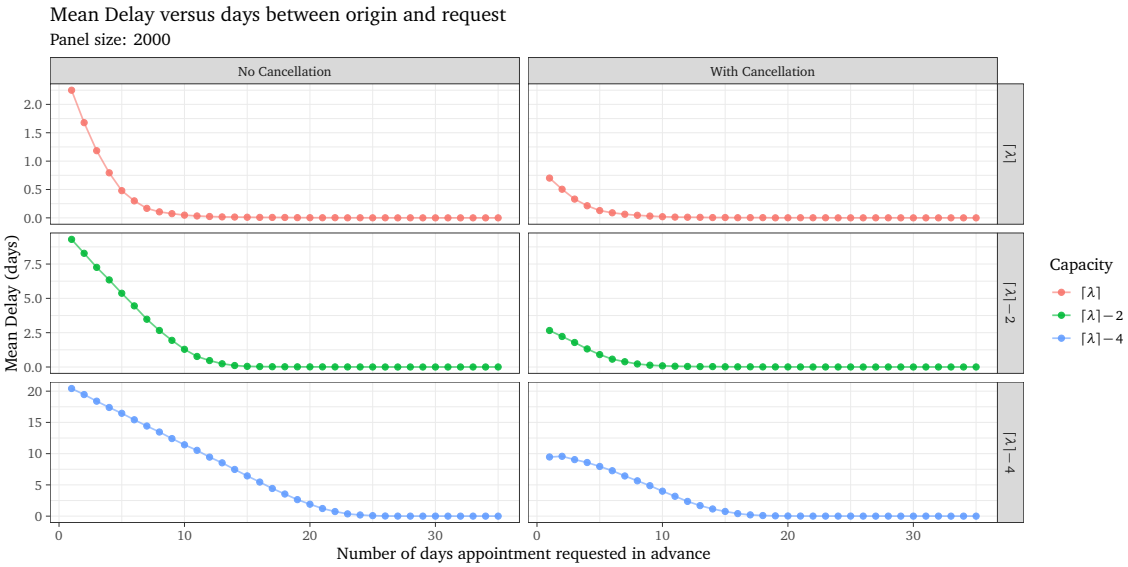


Figure 1.7: The delay is measured against the number of days before the appointment requested day that the appointment was scheduled. The X-axis is a measure of $r_{(k,i)} - o_{(k,i)}$. When patients request for an appointment in the near future, they may face more delays than patients who request for appointment in the future. This is for a panel of 2000 patients with no-flexibility policy.

Figure 1.8 shows the mean delays (Y-axis) for each of the 20 patient classes (X-axis) in our simulation for three different panel sizes (1800, 2000 and 2200), with and without cancellation; with and without reservation, and under three different daily capacity values ($\mu \in \{\lceil\lambda\rceil, \lceil\lambda - 2\rceil, \lceil\lambda - 4\rceil\}$). Each figure also shows the mean delay across all patient classes (i.e. across entire panel) under the no reservation case and with reservations. Note that mean delays for the panel are shown both in writing and with horizontal lines. We can make the following inferences from the figure:

- The mean delay for all patient classes is less than 1 day in all three panel sizes (1800, 2000 and 2200) when no cancellations are present and when $\mu \in \{\lceil\lambda\rceil\}$ (top row figures).

This suggests that even when arrival rate is very close to the demand, patients can experience reasonably low indirect delays. As a comparison, the mean wait time per patient in an $M/D/1$ model for panel size 2000 ($\lambda = 19.36, \mu = 20$) is 0.73 days, while our simulation estimates a mean wait time across all patient classes in the panel as 0.33 days (shown by the horizontal line).

- The results also show that the mean delay for the panel is an inadequate measure since it hides the significant differences in delays experienced by the various patient classes. In particular, mean delays increase with the rise in patient class number (recall that the patient class number based on the expected annual visits) when there are no reservations. The curves show slight oscillations towards the end, since the number of data points to estimate delays for patients in classes 15 and higher can be small, despite 10 replications of the 10-year simulation run. However, the basic rising trend is quite clear from the figure. Thus we can conclude that in a capacitated queueing system with first-come, first-serve appointment reservations, patients with greater needs who need more frequent visits with their primary care physician will experience longer delays. This empirically verifies our conjecture from section 1.5 under a wide range of experimental parameters.

- The presence of cancellations decreases the mean delays in all classes, since it creates empty slots in the near future which in turn benefits patients who need appointments at a short notice. Additionally, as expected, the magnitude of the mean delays increases significantly with a reduction in the daily capacity. The impact of these increases is felt by patients who need more frequent visits.

- Reserving slots for higher patient classes (18, 19 and 20) decreases the mean delay for those classes significantly. However, this comes at the expense of increased mean delays for all other patient classes. In particular, the higher patient classes among those without access to the reserved slots experience the greatest increases in comparison to the no reservation case. Since the reservation benefits a small minority of patients, the mean delay for the panel as a whole (i.e. mean across all patient classes) rises with reservations.

Further perspective on the impact of capacity constraints on a heterogeneous closed loop queueing system can be obtained from Table 1.7. We use panel size of 2000 in this table for illustration purposes. The expected annual visits for each patient class under infinite capacity (the unconstrained or uncapacitated case: $\mu = \infty$) is contrasted with expected annual visits under $\mu \in \{20, 18, 16\}$. In the unconstrained case, the expected annual visits are more or less equal to what we expect for the respective class. For example, in the unconstrained case expected annual visits in patient class 13 is 13.144. However, in the constrained cases without cancellation, we see that the expected annual visits for the patient classes start to decrease in relation to the unconstrained case. The higher the patient class, the greater the reduction. For example, patients in class 15 have an expected annual visit value of 8.9 when $\mu = 16$ while in the ideal, unconstrained case, it is 15.021. This reduction of about 6 annual visits is because the delays are higher for class 15 patients, resulting in longer time between appointments, and therefore fewer annual visits. Such missed visits often manifest in loss of continuity and increased expenses as patients see other providers in urgent care, and emergency/inpatient care. All of these pose higher chance of medical complications and burdens for patients with high needs. The table also shows that cancellations minimize the discrepancy between the unconstrained and reduced capacity cases: patients of class 15 experience 10.72 visits under $\mu = 16$ when cancellations are present. Thus, while delays are a valid outcome measure, it is also important to look into the consequences of fewer primary care visits.

1.9 Conclusion and Implications for Practice

In summary, we have introduced a granular patient-level stochastic process for scheduling appointments with a primary care physician who manages a panel of patients. Our modeling framework introduces key factors noticed in practice such as heterogeneity in appointment request rates, recurring appointments, and cancellations. We use the framework to study two types of questions: (1) capacity planning for PCPs assuming patients have flexibility in appointment dates, as a function of lead time to appointment; and (2) quantifying the differences in delays for various patient classes, and the impact of reservations to alleviate these differences. In our computational experiments, we parameterize heterogeneity and

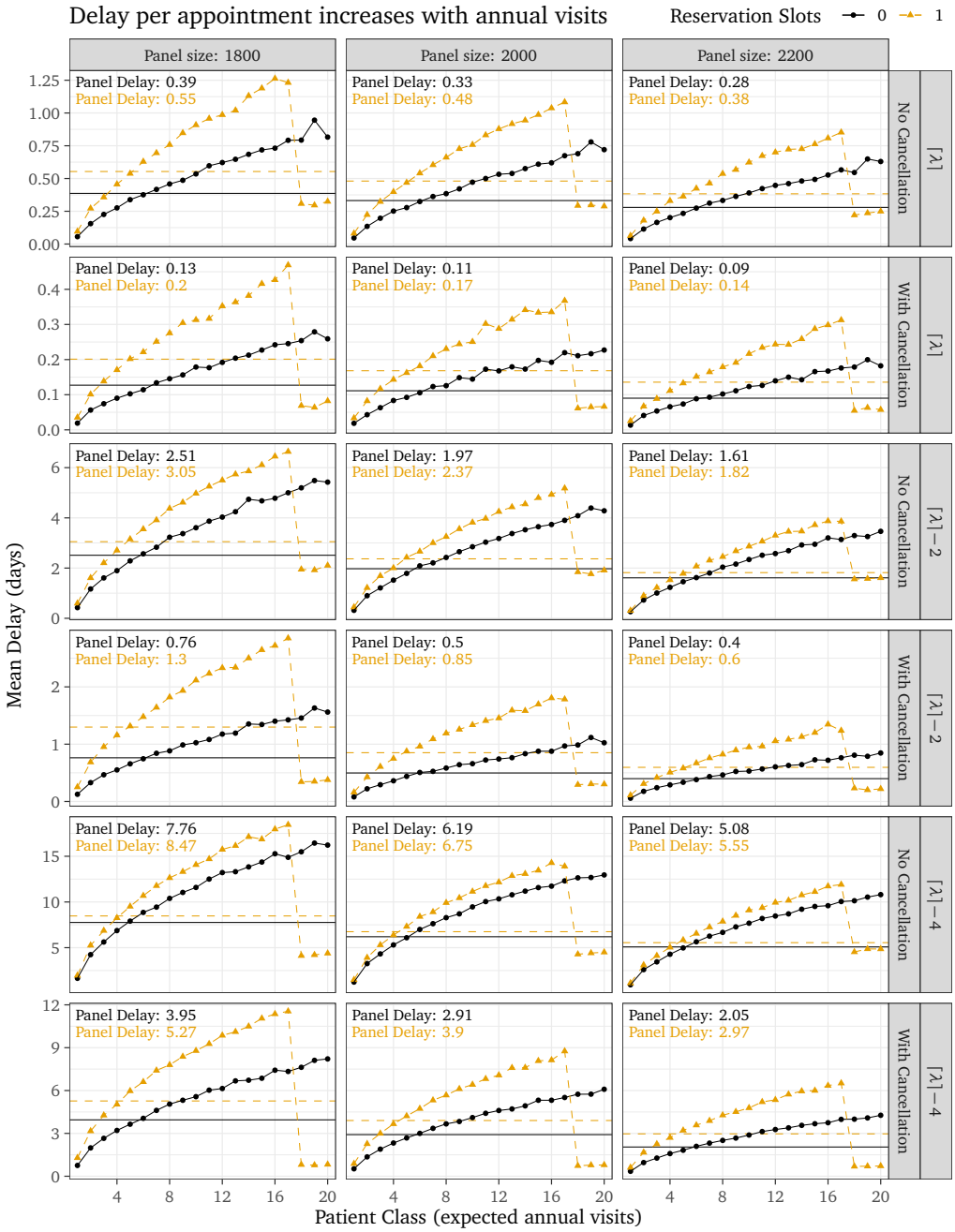


Figure 1.8: Mean delay increases with the patient class for all scenarios. We can note the sharp reduction in delay when reservation policy is applied for patient classes 18, 19, and 20. This mean delay is over all appointments as experienced on the day of the origin of the appointment. To allow stability of the simulation, only those appointments that originate after 2 years of the start of the simulation and before 2 years of the end of simulation have been considered. The mean delay for the full panel is shown as horizontal lines to compare it with delay for different patient classes.

Table 1.7: Mean annual visits for the patient classes, under different physician capacity limits for a panel of 2000 patients.

Class	Expected Annual Visits	Without Cancellation				With Cancellation			
		Capacity μ				Capacity μ			
		∞	$\lceil\lambda\rceil = 20$	$\lceil\lambda\rceil - 2 = 18$	$\lceil\lambda\rceil - 4 = 16$	∞	$\lceil\lambda\rceil = 20$	$\lceil\lambda\rceil - 2 = 18$	$\lceil\lambda\rceil - 4 = 16$
1	0.5	0.508	0.506	0.504	0.503	0.421	0.419	0.420	0.420
2	2	1.994	1.995	1.983	1.950	1.755	1.762	1.761	1.735
3	3	3.010	2.995	2.954	2.846	2.714	2.695	2.693	2.614
4	4	4.009	3.983	3.889	3.688	3.652	3.657	3.626	3.492
5	5	4.996	5.016	4.821	4.444	4.599	4.583	4.561	4.341
6	6	6.030	5.979	5.722	5.158	5.616	5.593	5.492	5.161
7	7	7.051	6.913	6.527	5.787	6.600	6.505	6.406	5.959
8	8	8.003	7.927	7.418	6.355	7.519	7.459	7.357	6.712
9	9	8.897	8.722	8.169	6.890	8.479	8.573	8.253	7.349
10	10	9.960	9.848	9.009	7.281	9.487	9.452	9.228	8.087
11	11	11.018	10.837	9.651	7.661	10.589	10.420	10.130	8.660
12	12	11.881	11.801	10.437	8.077	11.607	11.437	11.018	9.257
13	13	13.144	12.689	11.137	8.360	12.484	12.560	11.975	9.758
14	14	14.323	13.980	11.609	8.568	13.603	13.160	12.685	10.173
15	15	15.021	14.294	12.413	8.904	14.369	14.390	13.567	10.720
16	16	16.063	15.193	12.800	9.083	15.293	15.063	14.613	11.203
17	17	17.226	16.258	13.462	9.416	16.394	16.056	15.264	11.682
18	18	17.853	17.463	13.950	9.457	17.143	17.020	15.910	12.097
19	19	19.470	18.060	14.250	9.580	18.150	18.660	16.210	11.890
20	20	20.165	18.975	14.946	9.799	19.365	19.083	17.681	12.714
2000	19.346	19.408	19.144	18.003	16.020	17.796	17.707	17.365	15.877
Panel	λ	Mean Daily Arrivals							

recurring visits based on primary care visit patterns available in the nationally representative Medical Expenditure Panel Survey.

Our study reveals a number of insights for primary care practice. On the capacity planning side, we demonstrate that the use of heuristics such as First Minimum and Last Minimum when patients have flexibility in their appointment day reduces day to day variability from the provider's perspective, thus minimizing both idle time and overtime simultaneously. Furthermore, the heuristics compare well to an integer program that assumes prior knowledge and creates a globally optimal schedule. One important practical benefit of the heuristics is that a scheduler can easily implement them when the patient's call arrives by looking at the physician's calendar. The reduction in variation in daily workload by intelligent appointment allocation can be summarized as daily workload balancing by considering flexibility. This will help the provider reduce both—the overtime and idle time. The nature of flexibility is such that same-week visits have no flexibility. These immediate visits emulate advanced access and are implicitly part of the appointment system. Our

model also provides a distribution of the daily workload. We can use this distribution to design the optimal capacity for a nationally representative panel with flexibility by using a newsvendor model that balances idle time and overtime.

On the patient delay side, we demonstrate that while appointments are scheduled on a first-come, first-serve basis, they end up inadvertently penalizing patients who are sicker and need more frequent visits. Traditional analytical queuing models (M/M/1/K and M/D/1/K) used in the literature miss this effect entirely because the stochastic processes underlying them do not consider the lead time to appointments—that is, the time between the day the patient desires an appointment (appointment origin day) and the day the patient makes the call to the practice (appointment origin day). While we have considered one mitigation method by reserving appointments, service providers can use overtime for such patients that need immediate visits. Such overtime can help increase the capacity on-demand, and improve access for patients with immediate needs.

Our study has limitations which provide pathways for future investigation. While we use the geometric distribution to choose the next appointment request, the simulation can work with any distribution. We assume a nationally representative panel (United States) based on primary care visit rates, but if the healthcare needs of a local population are better known, the local population's distribution can be used. Additionally, sampling methods like the bootstrap methods are apt in such simulations thus allowing better modeling of regional / local populations than relying on the national level statistics.

The flexibility criteria used in capacity planning is currently based on our best judgment, since there are no studies that describe patient flexibility or tolerance for appointment delay. Future research could try to elicit the true nature of flexibility through data collected by the scheduler or online systems. We have assumed that appointments with short lead time have less flexibility. This assumption is appropriate but needs verification in practice.

Our model assumes each patient requests an appointment slot of the same size—typically a 20-min primary care slot. In practice, certain patients, particularly those with complex needs, might require two consecutive slots (i.e. a longer 40-min appointment). While we don not model this behavior in this study, it can be easily included in our modeling framework. A further extension to this could be allowing the stochastic nature of actual appointment

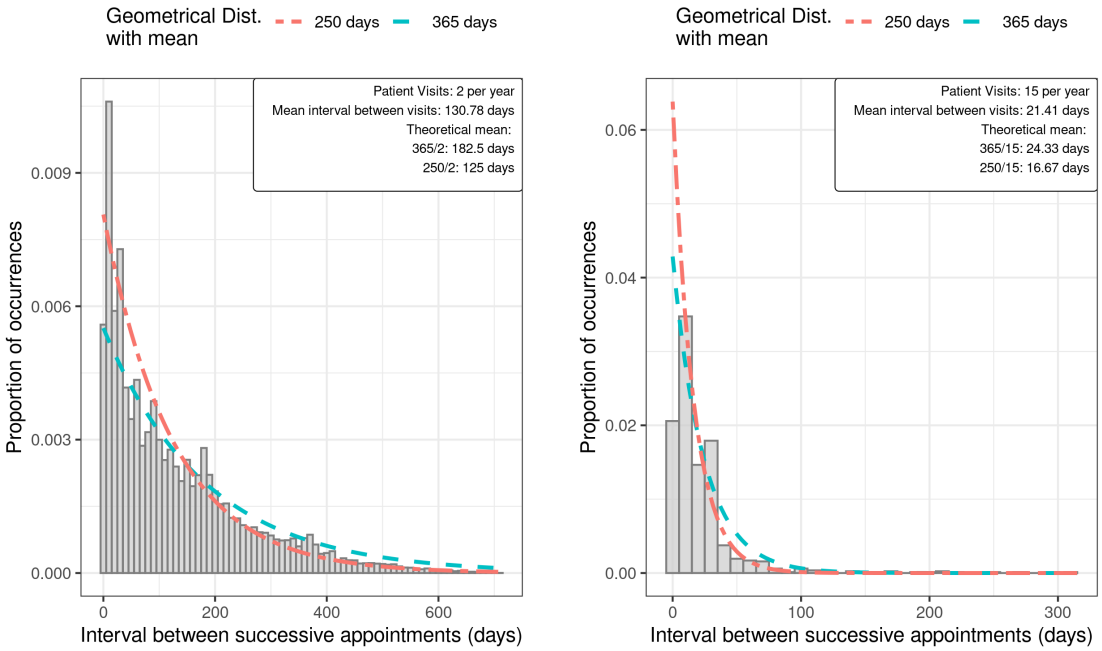
duration, which allows more precise estimation of the direct wait time, physician idle time and overtime.

Finally, while we have shown the impact of delay on patients of different classes, we have not been able to directly provide evidence on the implications of such delays on the health outcomes. Retrospective analysis of clinical and EHR data from large health systems may help quantify these implications.

1.10 APPENDIX Geometric distribution for appointment requests

In our patient-level stochastic process for recurring appointments, we assume that the time between current appointment and the next appointment request day follows a geometric distribution, with a parameter p_j. In fig. 1.9, we show histograms of time between successive PCP appointments for two patient classes: patients who visit twice a year, and patients who visit 15 times a year. This data from 2-year patient histories of PCP visits available in the Medical Expenditure Panel Survey (MEPS). Figure 1.9 also illustrates the fit of geometric distribution assuming 250 workdays in a year (primary care offices tend work 5 days per week, and do not work on holidays) and 365 workdays in a year. We use a bin width of one week (7 days) in the histograms. Figure 1.9a represents patients who visit twice a year suggests that geometric distribution is a reasonable fit, while it appears less so for patients who visit 15 times a year as seen from fig. 1.9b.

It is important, however, to note that the data on time between PCP visits includes delays, while in our stochastic process models the time between the current PCP appointment and desired/requested day of the next appointment, and therefore does not include delays. The MEPS data does not provide information on when the appointment was desired and when it was actually scheduled; if delays occurred, they are included in the interval between PCP appointments. This could explain why the height of the first bin (i.e. intervals of less than a week between appointments) is shorter than than the second (intervals longer than a week but less than two weeks): it is harder to get next appointments on a short notice, due to capacity constraints, which results in longer realized intervals between appointments.



(a) Patients with two visits per year

(b) Patients with fifteen visits per year

Figure 1.9: Histogram of interval in days between successive primary care appointments.

For model simplicity and for the purposes of this paper, we assume a geometric distribution for all patient classes in our panel. However, our stochastic process can accommodate any distribution for time between the next appointment request and current appointment.

In summary, the time interval between patient's successive appointments follows a Geometric distribution with the probability parameter p_j. Thus p_j as the probability for requesting an appointment on any day for each patient $k \in H_j$.

CHAPTER 2

A DISCRETE TIME MARKOV CHAIN FOR MODELING MULTI-CLASS PATIENT APPOINTMENT SCHEDULING SYSTEM

2.1 Introduction

As mentioned earlier in chapter 1 the patient panel for a primary care provider is not homogeneous. Addressing the heterogeneity in the panel allows us to see the system behavior for the various patient classes. One of the behaviors that we have seen in chapter 1 resulting from a heterogeneous panel was the increase in delay for each class. The appointment scheduling system, at its most stripped down version, consists of (i) an appointment calendar broken down to slot level to record slots allotted for each resource, and (ii) a scheduling method to allocate appointments when the requested slot in unavailable. The appointment calendar is a record and the scheduling method is the process. The resource in context of this dissertation is the physician. The resources may be extended by including the examination-room, the equipment, supporting staff etcetera. The appointment scheduling system is not restricted to healthcare. It can be used for various professions and services including appointments for personal grooming, travel booking, and entertainment. Any more discussion on other applicable areas of appointment scheduling systems would be simply digressing from the chapter.

2.1.1 Appointment Calendar

The appointment calendar for a single physician can be imagined as a sequence of slots lined up one after another as shown in fig. 2.1. Chapter 1 did not have any limit to the calendar horizon. Here we restrict the calendar horizon to an arbitrary size more than the panel size. This restriction is introduced for tractability as we shall see again in section 2.2.2. In fig. 2.1 the calendar horizon is 32 slots. All slots are assumed to have equal time duration. All patients are assumed to respect the allotted appointments. There is no

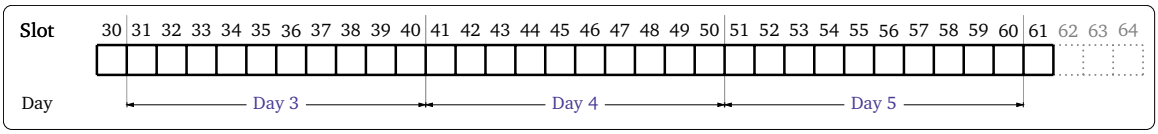


Figure 2.1: A sample sequence of appointment slots in an appointment calendar is shown. Each day is assumed to have ten slots for this illustration. The slots may be allotted to patients or may be vacant.

tardiness. No appointment will overshoot it's time to encroach on the subsequent slot. As the day progresses, there is a "current" slot that keeps moving as the appointment gets over. All appointments before such a current slot are in the past. They do not affect the appointments of the future.

2.1.2 Appointment process

At the end of the current appointment, the patient makes a request for her next appointment, based on her needs, to the scheduler. The scheduler allocates an appointment to the patient based on predefined rules. This patient visit the doctor again on her allotted appointment slot. The subsequent appointment becomes the current appointment.

2.1.3 Appointment Scheduling as a Markov Chain

Instead of labeling the slots with absolute indices, we can label them relative to the current appointment slot. As each appointment is completed, and we move to the next appointment there is progression in the appointment calendar with relative index as shown in fig. 2.2. The current appointment slot has index 0. The arbitrary calendar horizon is 32. The appointment calendar shifts to the future, but the relative indices remain. At the end of each appointment, the current appointment goes in the past and the subsequent appointment becomes the current appointment. At the same time, a new vacant appointment slot, labeled with index 31, is introduced. The appointment calendar, thus, changes at discrete time steps which is the time interval of each slot. And the calendar changes depending on the slots already allotted to patients and any newly allotted slot to the patient who has just finished her appointment. The change in the appointment calendar only depend on its current state. The past states of the appointment calendar do not matter.

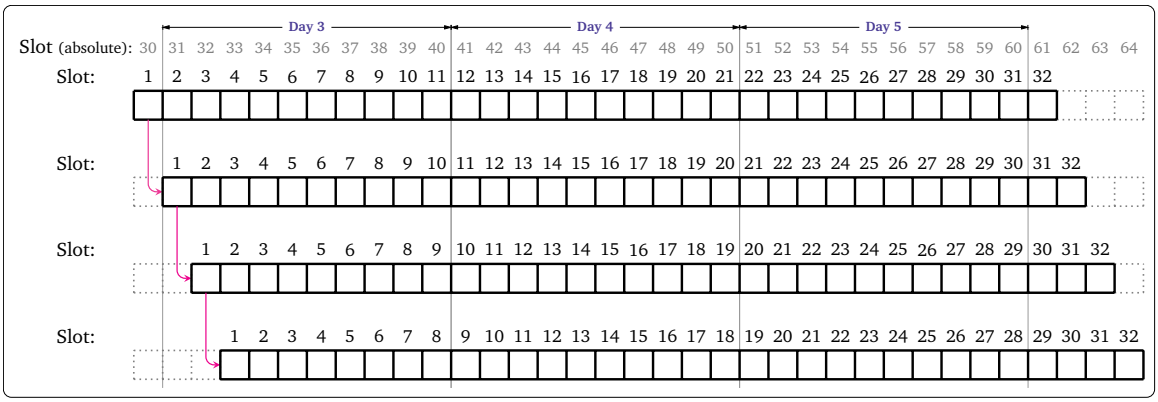


Figure 2.2: A sample sequence of appointment slots in an appointment calendar is shown. Each day is assumed to have ten slots for this illustration. The slots may be allotted to patients or may be vacant.

This behavior of the appointment calendar resembles the memory-less property of a Markov chain. We can model the appointment calendar indexed with relative labels as the state of the Markov chain. The state changes at discrete steps equal to the slot duration. When the current appointment slot is occupied, the transition in state of the appointment calendar is stochastic, since the request for next appointment is random and the allocation depends on the vacant slots. When the current appointment slot is not occupied, the appointment calendar simply shifts each slot to the lower index.

We shall see the Markov chain representation of the appointment scheduling system further in section 2.2 in order to analyze its properties.

2.2 Methodology

We consider a panel of n patients associated with a primary care physician. This panel is represented as a set H. We partition the panel such that each partition H_j represents the set of patients with similar healthcare. The number of partitions is J. We consider patients with similar healthcare require similar number of visits to their healthcare provider. The number of patients in class H_j is $n_j = |H_j|$, $\sum_j n_j = |H| = N$ and $H_j \cap H_{j'} = \emptyset$ for all $j \neq j'$. Each patient $k \in H_j$ has the same probability of requesting an appointment slot given as p_j. The vector of the probability parameters is $p = (p_1, p_2, \ldots, p_J)$ and the vector of the number of patients is $n = (n_1, n_2, \ldots, n_J)$.

In modeling the probability of a patient requesting an appointment slot, we assume that, in a primary care setting, successive requests for appointments are independent of each other.

A patient k requests her appointment for a slot $r_{(k,i)}$ at the end of the previous appointment $a_{(k,i-1)}$. If that slot has already been allotted to some other patient, then the appointment is allotted on a subsequent available slot $a_{(k,i)}$. The patient experiences an absolute delay that can be measured as $d_{(k,i)} = a_{(k,i)} - r_{(k,i)}$. When the appointment allotted is for the same slot same as the requested appointment the delay is zero since $a_{(k,i)} = r_{(k,i)}$.

The request for the next appointment is a random variable X_j having the geometric distribution for first success with the probability parameter p_j.

$$r_{(k,i)} = a_{(k,i-1)} + X_j \quad \text{where } X_j \sim \text{Geo}(p_j). \tag{2.1}$$

The allotted appointment slot $a_{(k,i)}$ is the earliest vacant slot including and after the corresponding requested appointment slot $r_{(k,i)}$. The rationale for using the geometric distribution is the same as provided in section 1.10

2.2.1 Equivalence Between the Probabilities of Visit per Slot and Visit per Day

Consider the similarity between the description of the model in this chapter with the model in chapter 1. Let $\widehat{p}_j$ be the probability of requesting a slot on any day for a patient in class j. The value of $\widehat{p}_j$ is calculated using the law of large numbers as $\widehat{p}_j = j/250$, assuming 250 annual working days. Let λ_s represent the random variable for the number of appointment requests for slot s and let $\widehat{\lambda}_d$ represent the random variable for number of appointment requests for day d. We get the expected number of appointment requests per slot as $\mathbb{E}[\lambda_s] = \sum_j n_j p_j$. Similarly, we get the number of appointment requests per day as $\mathbb{E}\left[\widehat{\lambda}_d\right] = \sum_j n_j \widehat{p}_j$. If the capacity in terms of slots per day is μ, and all slots in a day have equal preference for patients, we show how the probability parameter at a slot level may be derived from the probability parameter at a day level from eq. (2.2).

$$p_j = \frac{1}{\mu}\widehat{p}_j = \frac{j}{250\mu} \tag{2.2}$$

The expected demand for appointments are related by eq. (2.3).

$$\begin{aligned}\mathbb{E}\left[\lambda_s\right] &= \sum_j n_j p_j = \sum_j n_j \frac{1}{\mu}\widehat{p}_j = \frac{1}{\mu}\sum_j n_j \widehat{p}_j \\ \mathbb{E}\left[\lambda_s\right] &= \frac{1}{\mu}\mathbb{E}\left[\widehat{\lambda}_d\right]\end{aligned} \tag{2.3}$$

Similar to what we have seen in table 1.7 where the expected arrivals exceed capacity when $\mathbb{E}\left[\widehat{\lambda}_d\right] > \mu$, here we have expected demand exceeding supply when $\mathbb{E}\left[\lambda_s\right] > 1$.

2.2.2 Markov Chain Behavior

Continuing from section 2.1.3, we can model the appointment system as a Markov chain. The appointment calendar represents the state of the system. A new appointment allotted to a patient's request is dependent on the current state of appointment calendar. The previous states of the appointment calendar do not matter which allow for the memory-less behavior of the appointment calendar. This calendar is a sequence of slots, with the first slot representing the current appointment. We can represent the appointment calendar as a Markov chain transitioning from one state of the appointment calendar to another state. Each patient in a panel is assigned an appointment slots on the calendar. The appointment calendar horizon T is limited to an arbitrary length satisfying $N < T < \infty$, to allow tractability for analysis.

The appointment calendar itself is a vector of size T with each element representing the slots. An element representing the patient k allotted to that slot has value k while a vacant slot is represented by a zero. The sequence of random variables $\widehat{Y}_t$ representing the appointment calendar during the absolute slot t is a Markov chain. The first element of the vector is the current slot t and the last element is the slot $t + T - 1$ in the future. The number of possible states is

$$\frac{T!}{(T-N)!}.$$

The state changes from $\widehat{Y}_t$ to $\widehat{Y}_{t+1}$ when the appointment slot labeled with the absolute index t is over and the next slot with absolute index $t + 1$ starts.

As an example we try to identify the number of states when the panel size is $N = 6$ and horizon $T = 10$. This makes the number of vacant slots is $T - N = 4$. The number of possible states is $10!/4! = 151,200$.

We can reduce the state space by representing each patient $k \in H_j$ by its class j. The interchange of patients to their classes in the appointment calendar is possible because each class is considered homogeneous within itself. This appointment calendar may be represented with the random variable Y_t. The current appointment slot of Y_t is the absolute slot indexed by t. This is also the first element in the vector.

The number of possible states now reduces to

$$\frac{T!}{\left((T-N)! \prod_j n_j!\right)}.$$

For example, our panel $N = 6$, is split as three classes with $3, 2, 1$ patients in each class and horizon $T = 10$. The number of possible states is $10!/\left(4!\,3!\,2!\,1!\right) = 12,600$. The number of states has reduced by 91.6%.

Let S represent the set of all possible states and S_j represent set of all states with current appointment allotted to a patient of class j. We also have S_0 represent the set of all states with current appointment slot vacant. Here, $S_0 \subset S$ and $S_j \subset S$. The state changes from $Y_t = \mathbf{s}$ to $Y_{t+1} = \mathbf{s}'$ at the end of the slot t and vectors $\mathbf{s}, \mathbf{s}' \in S$ are two possible states of the appointment calendar.

2.2.3 Transition Probability

The patient at the current slot is the key to determine the transition probabilities. When the current slot is unoccupied, there is no patient to request a new appointment. The appointment calendar vector "left-shifts" and the newly introduced last slot in the appointment calendar stays unoccupied. There are no other transitions possible. This left-shift operation on the vector is defined using an operator `left_shift`$(\cdot)$. When the current slot is occupied, the appointment calendar vector "left-shifts" and the patient can be allotted only one of those slots that were previously vacant $T - N$ slots plus the last newly introduced slot. The number of possible slots that can be allotted to the patient is

$U = T - N + 1$. We use the index u to denote the possible allotted slot, that is $1 \leq u \leq U$ and $a_1 < a_2 < \cdots < a_{(T-N)} < a_{(T-N+1)}$. When the allotted slot is not the last slot, that is $a_u < T$, the patient would have requested for an appointment either for the vacant slot $t + a_u$, or for any of the occupied slots in $\{t + a_u - 1, t + a_u - 2, \ldots, t + a_{(u-1)} + 1\}$. For the requested slot $t + r_u$ associated to the allotted slot $t + a_u$, we use $a_0 = 0$, to get $a_{(u-1)} + 1 \leq r_u \leq a_u$. For the last allotted slot $a_U = T$, it is possible that the requested slot $t + r_U$ can be any slot in the future even beyond the calendar horizon making $r_U \geq a_{(U-1)} + 1$. In the above context of the allotted slot a_u, for a patient of class j, the transition probability when we move from the state of appointment calendar $\mathbf{s}$ to another state $\mathbf{s}'$ is the probability of requesting any of the slots r_u If the requested appointment after the current appointment for a patient of class j is a random variable $X_j = r_u$, we get elements of the transition probability matrix $\mathbf{P}$ of using:

$$p_{\mathbf{ss}'} = \begin{cases} \Pr\left(a_{(u-1)} < X_j \leq a_u\right) & \text{if } \mathbf{s} \in S_j, a_u < T, \mathbf{s}' = \texttt{left\_shift}(\mathbf{s}) + j\mathbf{e}_u \\ \Pr\left(X_j > a_{(u-1)}\right) & \text{if } \mathbf{s} \in S_j, a_u = T, \mathbf{s}' = \texttt{left\_shift}(\mathbf{s}) + j\mathbf{e}_u \\ 1 & \text{if } \mathbf{s} \in S_0, \mathbf{s}' = \texttt{left\_shift}(\mathbf{s}) \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Here $\mathbf{e}_u$ is the standard basis vector indicating position of the allotted appointment a_u. The $\texttt{left\_shift}(\cdot)$ operator shifts the vector by one element to the left, with the last element introduced as zero. As long as the probability parameter p_j associated with X_j is in the open interval $(0, 1)$, the Markov chain is ergodic. We can derive the stationary probability distribution vector Π_S associated with the transition probability matrix $\mathbf{P}$. Each element $\pi_{\mathbf{s}}$ of the vector corresponds to the steady state probability of state $\mathbf{s}$.

2.2.4 Delay

Now, we can determine the delay experienced by the patient by using the transitions between the states of the appointment calendar. We again consider the case of $\mathbf{s} \in S_j$ for transition from state $\mathbf{s}$ to $\mathbf{s}'$ associated with a_u. The expected delay for such transitions when $p_{\mathbf{ss}'} \neq 0$ is the conditional delay $\mathbb{E}[d(Y_t = \mathbf{s})|Y_{t+1} = \mathbf{s}']$

$$d_{\mathbf{ss}'} = \mathbb{E}\left[d\left(Y_t = \mathbf{s}\right)|Y_{t+1} = \mathbf{s}'\right] = \frac{\sum_{a_{(u-1)}<i\leq a_u}\left(a_u - i\right)\Pr\left(X_j = i\right)}{\sum_{a_{(u-1)}<i\leq a_u}\Pr\left(X_j = i\right)} = \frac{\sum_{a_{(u-1)}<i\leq a_u}\left(a_u - i\right)\Pr\left(X_j = i\right)}{p_{\mathbf{ss}'}}$$

Thus we can generate the expected delay associated to the transitions as a matrix $\mathbf{D}$ from its elements:

$$d_{\mathbf{ss}'} = \begin{cases} \dfrac{\displaystyle\sum_{a_{(u-1)}<i\leq a_u}\left(a_u - i\right)\Pr(X_j = i)}{p_{\mathbf{ss}'}} & \text{if } \mathbf{s} \in S_j \text{ and } p_{\mathbf{ss}'} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

The expected delay $d_{\mathbf{s}}$ of a particular state $\mathbf{s}$ of the appointment calendar is the sum of the delays of transition to each of the states $\mathbf{s}' \in S$ times the transition probability $p_{\mathbf{ss}'}$.

$$d_{\mathbf{s}} = \mathbb{E}\left[d\left(Y_t = \mathbf{s}\right)\right] = \sum_{\mathbf{s}'\in S} d_{\mathbf{ss}'} \cdot p_{\mathbf{ss}'} = \sum_{u=1}^{T-n+1}\left(\sum_{a_{(u-1)}<i\leq a_u}\left(a_u - i\right)\Pr(X_j = i)\right)$$

It is clear from this equation that the expected delay from a state is only concerned with the (1) probability parameter p_j associated with X_j and (2) position of the vacant slots. So two different states $\mathbf{s}, \overline{\mathbf{s}} \in S_j$ that have equivalent vacant slots will have the same expected delay. From here we can get the delay associated with a class j by enumerating all the possible vacant slots that can be requested. The expected delay associated with class j is

$$d_j = \mathbb{E}\left[d|(Y_t \in S_j)\right] = \frac{\sum_{\mathbf{s}\in S_j}\pi_{\mathbf{s}} d_{\mathbf{s}}}{\sum_{\mathbf{s}\in S_j}\pi_{\mathbf{s}}}$$

Here, the values of different $\pi_{\mathbf{s}}$ are dependent on the probability parameters and number of patients for all classes $(p_j, n_j)\forall j$. The summation over all the states in S_j effectively determines all possible combinations of available slots. When all the available slots for the next appointment are stacked as early as possible then $a_1 = 1, a_2 = 2, \ldots, a_{T-n} = T - n,$ and $a_{T-n+1} = T$. Similarly, when all the available slots for the next appointment are stacked as late as possible then $a_1 = n, a_2 = n+1, \ldots, a_{T-n} = T-1,$ and $a_{T-n+1} = T$. We can enumerate vectors of availability of next slot as a set A, where each vector

$\mathbf{a} = [a_1\, a_2\, \ldots\, a_{T-n}\, a_{T-n+1}] \in A$. Since many states in S can be associated to each of the next availability vector $\mathbf{a}$, we can construct a logical membership matrix $\mathbf{A}_j$ with rows representing each availability vector $\mathbf{a} \in A$ and the columns representing each state $\mathbf{s} \in S_j$. The elements of this matrix $\mathbf{A}_j$ with row $\mathbf{a}$ and column $\mathbf{s}$ are 1 when $\mathbf{s} \in S_j$ is associated to $\mathbf{a}$, and 0 otherwise. We can now determine the steady state probability of availability $\Pi_{A_j} = \mathbf{A}_j \times \Pi_S$. Let $\pi_{\mathbf{a}_j}$ represent the element in Π_{A_j} associated with the availability vector $\mathbf{a}$. From the geometric distribution, $\Pr(X_j = i) = (1-p_j)^{i-1}p_j$. Thus, we can expand the expression for d_j as:

$$d_j = \frac{\displaystyle\sum_{a_1=1}^{n}\sum_{a_2=a_1+1}^{n+1}\cdots\sum_{a_{T-n}=a_{T-n-1}+1}^{T-1}\sum_{a_{T-n+1}=T}^{T}\left(\pi_{\mathbf{a}_j}\sum_{u=1}^{T-n+1}\sum_{a_{(u-1)}<i\leq a_u}(a_u-i)(1-p_j)^{(i-1)}p_j\right)}{\displaystyle\sum_{a_1=1}^{n}\sum_{a_2=a_1+1}^{n+1}\cdots\sum_{a_{T-n}=a_{T-n-1}+1}^{T-1}\sum_{a_{T-n+1}=T}^{T}\pi_{\mathbf{a}_j}} \tag{2.5}$$

We can now define delay as a function $d : (0,1)^J \to \mathbb{R}^J$ that maps the probability vector p to the delay vector with the parameters $\mathbf{n}$ and T. The jth element of the vector d is d_j from eq. (2.5) that is $d = (d_j)$.

2.3 Results

To prove that delay of patient in class j is more than delay of patient in class j' we need to prove that the function d is monotone. In order to this we need to show that for a two class panel, when $p_1 < p_2$ we get $d_1 < d_2$. Additionally, for a three class panel, when $p_1 < p_2 < p_3$ we should expect $d_1 < d_2 < d_3$ for monotonicity.

2.3.1 Monotone Mapping

The concept of monotone mapping is described in Defn 12.1 Rockafellar and Wets, 1998, Chapter 12. This method says that a mapping is monotone when

$$(\mathbf{d}_a - \mathbf{d}_b)\cdot(\mathbf{p}_a - \mathbf{p}_b) \geq 0 \forall \mathbf{p}_a, \mathbf{p}_b \text{ and } \mathbf{d}_a = d(\mathbf{p}_a), \mathbf{d}_b = d(\mathbf{p}_b). \tag{2.6}$$

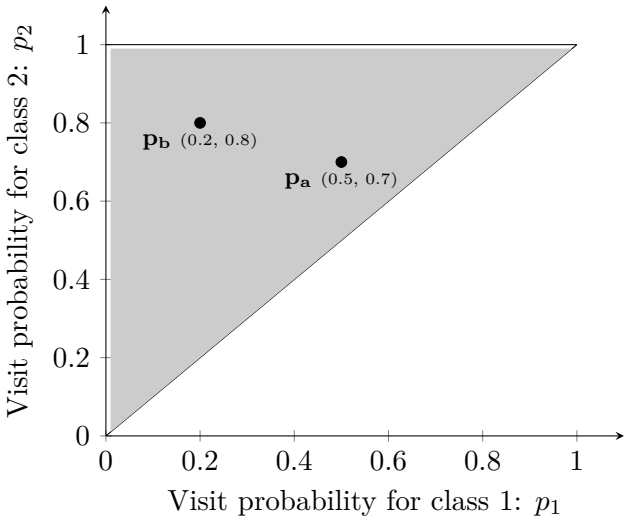


Figure 2.3: Feasible space for exploring monotonic behavior.

Figure 2.3 give a visualization of the domain for probabilities we for a two-class panel. Any point in the shaded area satisfies $0 < p_1 < p_2 < 1$. The two sample points $\mathbf{p_a}$ and $\mathbf{p_b}$ give an understanding of the feasible space.

The algebraic analysis of eq. (2.5) using the condition in eq. (2.6) is intractable because of complexity. The intractability leads us to explore numerical analysis as we shall see in section 2.3.2.

2.3.2 Numerical Analysis

We use the function defined in eq. (2.5) to generate delays using computations. We begin with exploring a 10 appointment slots with 6 patients as seenin table 2.1. The 6 patients are split in two classes with all possible combinations. We initially start with two arbitrary probabilities of 0.2 and 0.5 for each class and get the delay. Then we use an extremely small difference in the request probabilities between two classes as 0.499 and 0.5. We find that the comparison of delays to be consistent.

Next, we try a three class panel with various combinations of the 6 patient panel in the 10 appointment slots as seen in table 2.2. For every selected combination, with different probabilities, we consistently find the results that point to monotone behavior.

T	n_1	n_2	N	p_1	p_2	d_1	d_2	Is $d_1 < d_2$?
10	5	1	6	0.2	0.5	1.88039917	2.97243372	☑
10	4	2	6	0.2	0.5	2.00419889	3.28110949	☑
10	3	3	6	0.2	0.5	2.11328127	3.52388108	☑
10	2	4	6	0.2	0.5	2.20852708	3.71512407	☑
10	1	5	6	0.2	0.5	0.14803428	0.85195390	☑
10	5	1	6	0.499	0.5	3.99807873	4.00155115	☑
10	4	2	6	0.499	0.5	3.99816068	4.00163317	☑
10	3	3	6	0.499	0.5	3.99824261	4.00171516	☑
10	2	4	6	0.499	0.5	3.99832452	4.00179713	☑
10	1	5	6	0.499	0.5	3.99840641	4.00187907	☑

Table 2.1: Sample delay for two patients for horizon $T = 6$ and panel size $N = 6$ always has $d_1 < d_2$ when $p_1 < p_2$.

T	n_1	n_2	n_3	N	p_1	p_2	p_3	d_1	d_2	d_3	Is $d_1 < d_2 < d_3$?
10	1	1	4	6	0.499	0.5	0.501	3.99871886	4.00221177	4.00567377	☑
10	1	2	3	6	0.499	0.5	0.501	3.99863628	4.0021343	4.00559232	☑
10	1	3	2	6	0.499	0.5	0.501	3.99855362	4.00204864	4.00551047	☑
10	1	4	1	6	0.499	0.5	0.501	3.99846997	4.00196605	4.0054278	☑
10	2	1	3	6	0.499	0.5	0.501	3.99856121	4.00205626	4.00551402	☑
10	2	2	2	6	0.499	0.5	0.501	**Out of Memory**			☐
10	2	3	1	6	0.499	0.5	0.501	3.99839453	4.00188704	4.0053501	☑
10	3	1	2	6	0.499	0.5	0.501	3.9984016	4.0018895	4.005354	☑
10	3	2	1	6	0.499	0.5	0.501	3.99831866	4.00180762	4.00527205	☑
10	4	1	1	6	0.499	0.5	0.501	3.99824233	4.00172793	4.00519367	☑
10	1	1	3	5	0.499	0.5	0.501	2.99970856	3.00272972	3.00574434	☑
10	1	4	1	6	0.4	0.5	0.6	3.59364291	3.99232154	4.29491263	☑
10	4	1	1	6	0.4	0.5	0.6	3.55185799	3.94582715	4.24581152	☑
10	1	1	4	6	0.4	0.5	0.6	3.60646184	4.00659329	4.3099646	☑
10	1	2	3	6	0.4	0.5	0.6	3.6022577	4.00192422	4.3050541	☑
10	2	1	3	6	0.4	0.5	0.6	3.58914499	3.98747717	4.2899402	☑
10	2	3	1	6	0.4	0.5	0.6	3.58018182	3.97743229	4.27927329	☑
10	1	4	1	6	0.04	0.05	0.06	0.32873367	0.395557	0.45709815	☑
10	1	4	1	6	0.004	0.005	0.006	0.03541344	0.04397158	0.0524387	☑

Table 2.2: Sample delay for three patients for horizon 10 has $d_1 < d_2 < d_3$ when $p_1 < p_2 < p_3$. Note that one sample gave an out-of-memory error for which we could not conclude any outcome.

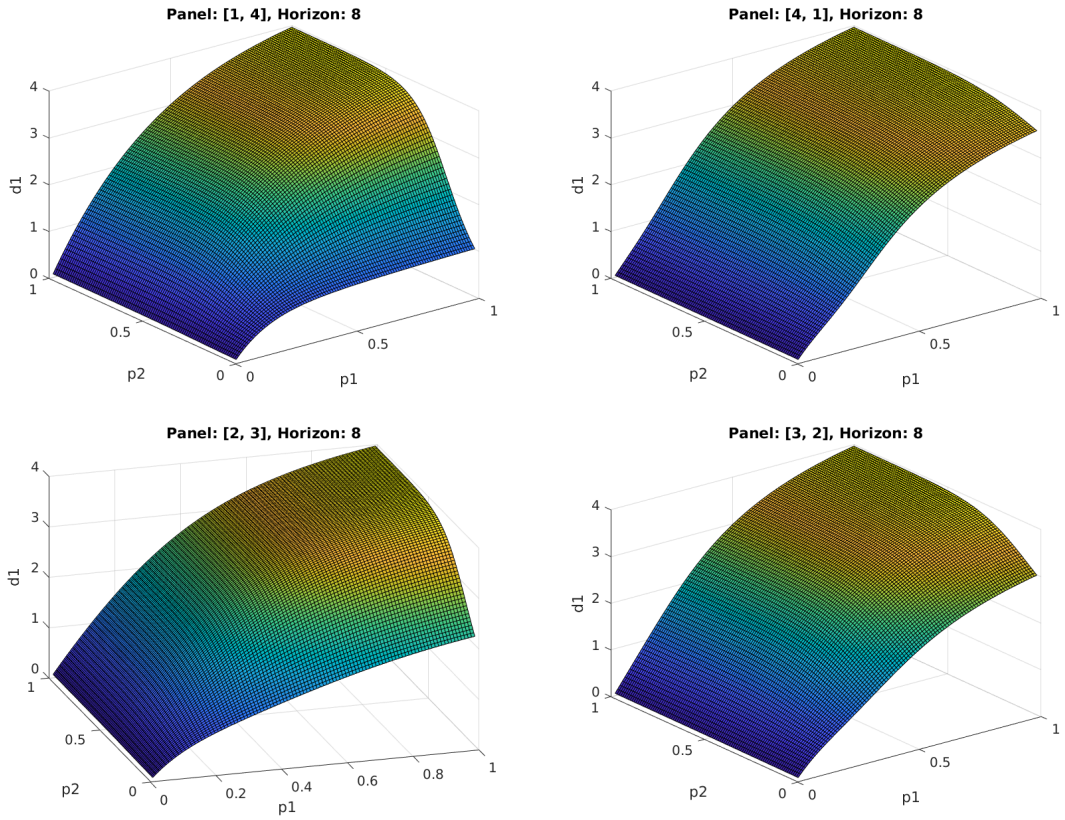


Figure 2.4: Surface plot for a panel of $N = 5$ patients and horizon $T = 8$ at various values of $\mathbf{p}$ show that the delay d_1 is monotonic. The symmetry in the definition of d_j would show similar monotone behavior for d_2.

With these tabulated results, we investigated the monotone behavior for the delay of one class d_1 in a two class panel. We sampled 10,000 instances of the vector $\mathbf{p}$ and determined the delay d_1 for each sample. Figure 2.4 shows the surface plot for this delay d_1 for different combinations of patients in each class for a panel size of 5 and horizon of 8. These samples support the idea that delay for one class has a monotone behavior on its own.

2.4 Conclusion & Discussion

We are unable to prove mathematically that the delay follows a monotone mapping because of the intractability of the delay function. The complexity of the delay function restricts the horizon size for the appointment slots and the panel size for numerical analysis. Within the computationally feasible size of the parameters, we have consistently show that our hypothesis of lower delay for low probability of requesting a slot and higher delay

for higher probability of requesting an appointment slot. The results from the simulation from chapter 1 and the numerical analysis above give more evidence towards inequity in delay for people with more health care needs.

CHAPTER 3

CAPACITY PLANNING FOR A SPECIALTY NETWORK OF OUTPATIENT HEALTH CARE PROVIDERS

3.1 Introduction

In the previous chapters, we have looked at how a panel of patients seeks appointments with a primary care provider (PCP) and have modeled relevant capacity planning and delay concerns. Now we consider that a panel of patients seeks appointments in *a network of specialty care providers* (including primary care). Our motivation comes from the need for health systems to plan for capacities not just for primary care, but for a range of other specialties. Such capacity planning is vital for large health systems such as the Veterans Affairs and Kaiser Permanente. Patient access in a specialty network often exists in the form of independent outpatient clinics that serve a region. Administrators and public health officials may need to address capacity issues for under-served patients.

In fig. 3.1, we can see the outpatient visits for three different patients over a two-year period. These figures are based on data for patients obtained from the Medical Expenditure Panel Survey. The patient labeled as 'A' goes first to ophthalmology, then to primary care twice and then twice to ophthalmology. Thus, we observe transitions between specialty types as well as intervals between visits which can vary. The duration of the appointment is minuscule when compared with the time-interval—small enough to be considered a point process. The patient transitions from one specialty to another with some revisits. The visit referrals to a different specialty or revisits to the same specialty may be from the advice of the primary care physician or from the specialty provider. A single patient may have their own unique temporal visit signature, but when every patient in the US is observed, broader patterns in the visits may emerge. While many patients may request multiple follow-up appointments with their provider for managing their chronic conditions, new symptoms and outcomes can necessitate changes in future appointment schedules, which may be influenced

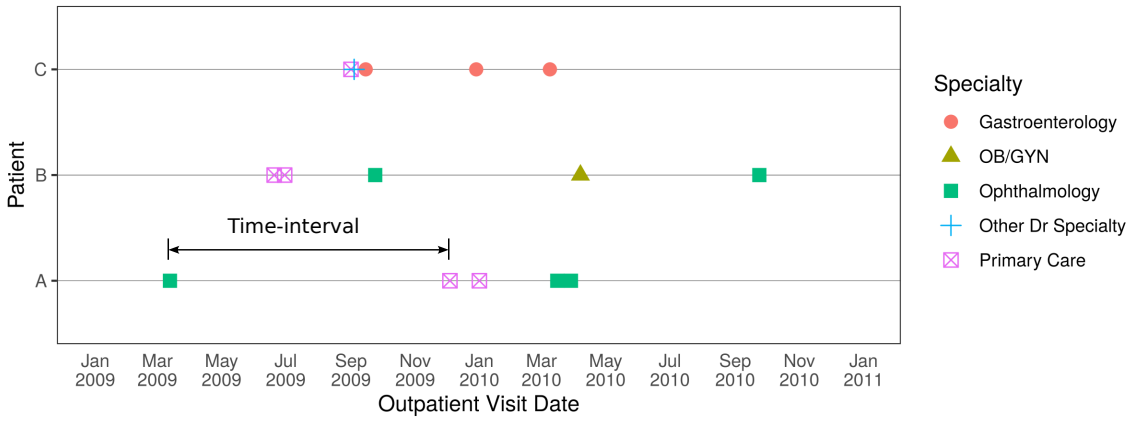


Figure 3.1: Outpatient visits for three patients, over a two-year period, 2009-2010. Data from the Medical Expenditure Panel Survey.

by the most recent healthcare visit. In the grand scheme of things, when we try to build patterns, these individual behaviors become their intrinsic constituent.

The role of the primary care provider as the point of first-contact remains important (Grumbach et al., 1999). Referrals to specialty care can add to patients' and primary care's burden of care-coordination (O'Malley & Cunningham, 2009; Patel et al., 2018). Further, the primary care providers remain unaware if the patient has acted on the referral by making appointments and showing up, nor do they always receive information from the specialty provider to help them make better decisions (Mehrotra, Forrest, & Lin, 2011). Barnett, Song, and Landon, 2012 have shown an increase of referrals from 4.8% in 1999 to 9.3% in 2009. The number of visits by Medicare beneficiaries to specialty care has increased at much higher rate than the number of visits to primary care(Barnett, Bitton, Souza, & Landon, 2021). Another study (Ganguli et al., 2020) uses data from a commercial insurer to show the number of visits to primary care has reduced by 24.2% from 2008 to 2016. Not all specialist visits require referrals from primary care. The emergence of the PPO & EPO plans for commercially available medical insurance in the US has made specialty visits accessible to patients without gatekeepers.

To the best of our knowledge, despite the increase in referrals, the sequences in which patients consult with primary care and specialists and the time-intervals between appointments have not been quantified. From the point-of-view of an observer looking at patients moving

from one node to another in a network of providers, a closed Jackson queuing network (Jackson, 2004) may provide insights in the waiting time, bottleneck service providers, and the "flow-rate" at each node. However, such closed queuing networks are difficult to characterize analytically.

Thus, there is a need for an adequate modeling framework that captures both patient heterogeneity and the pattern of referral and revisit transitions and time-intervals is observed in patient-level longitudinal datasets. Patients that use the specialty network, may face delays with a provider due to capacity bottlenecks, which may affect any other downstream visits they may have in the future. Therefore, we choose an approach that assumes infinite capacity for each specialty to analyze the impact of differences in lead time and to reverse-engineer capacity needs. In queuing theory, this approach is known as offered-load analysis (Whitt, 2013) and is well suited for capacity estimation.

Analytical queuing models become even more challenging to analyze when multiple classes of patients (heterogeneity) need to be modeled. Furthermore, appointment systems are not traditional *first-come, first-served* queues. As discussed in earlier chapters, queueing models do not consider the appointment lead time—that is, the time between desired day of the appointment and the day the request was made. Different patient classes have different appointment lead times which in turn leads to differences in delays between patient classes. However, a multi-class Jackson queueing model that uses first-come, first-served queueing discipline and which ignores lead time considerations results in identical waiting times for all patient classes.

Patients' visits to specialties come with their own lead-time for making appointment requests. This appointment lead-time may be affected by the sense of urgency that the patient and/or the referring physician may have. One can argue that patients who request appointments at very short lead-time may have an urgent need for care to treat a new condition or symptom. When providers cannot service requests, we may see a surge in visits to emergency rooms, urgent care and unplanned inpatient hospitalization. The delay in access to appointments specialty care can vary, depending on the urgency of the visit and the mismatch between the supply and demand for appointments. The rate at which appointments get filled can give some insight to the which patient subgroups or which referrals are expected

to experience delays. Since patients' specialty care visits are not always motivated by formal referrals, the visit time-interval between different pairs of specialty care will vary.

Buttorff, Ruder, and Bauman, 2017 shows that patients with 5+ chronic conditions are only 12% of the population, but they consume 41% of all healthcare expenditure, including ambulatory, inpatient and emergency care. On the other hand, 40% of the population has no chronic conditions and consume 10% of the healthcare expenditure. Patients with chronic conditions have different healthcare needs than other patients, and their burden of care-coordination is much higher. Modeling the average patient may obscure these patients with multiple chronic conditions. We make a case for considering patient subgroups differentiated by their healthcare needs.

We propose the use of Markov Renewal Process (MRP) (Çinlar, 1975) to model the patient transitions between specialties and time-interval between successive appointments. One of the contributions of this work is the parameterization of an MRP from longitudinal data to provide a nationally representative model of outpatient referrals and visits in the specialty network. Our use of a MRP is inspired from R. Hilton, Zheng, Fitzpatrick, and Serban, 2018.

We answer the following questions with the MRP model:

1. How do specialty transitions and time-intervals differ between different patient subgroups?
2. How does the expected fill-rate for each specialty change over time? How does this expected fill-rate vary by patient subgroup and referral network?
3. What is the distribution of the appointment requests for each specialty in a given time-period. How does this distribution change as the appointment allocation schemes utilize patient flexibility of day of appointment?

Our contributions from this work are (1) framework for modeling MRP for outpatient visits, for a nationally representative population, (2) analysis of lead-time between appointment requests by patient subgroups, (3) insight into fill-rate by various patient subgroups for each specialty, (4) analysis of aggregate daily appointment requests capacity and improvement in capacity with scheduling policy.

The rest of the chapter is organized as follows. In section 3.2, we provide the literature landscape and identify how our work fits in relation to other research. In section 3.3, we

give the rationale for the Markovian property in appointment transitions from one specialty to another, and model it as a MRP for patient subgroups. We then give the analytical expressions for fill-rate for patient subgroups by the referrals. We simulate the MRP of the specialty network referrals to discover the distribution of daily appointment requests and assign appointments to reduce the variance. In section 3.4 we analyze the differences in MRP and the fill-rate for different patient subgroups. We quantify the distribution of appointment requests by specialty and the allocation using heuristics. In section 3.5 we give some insights on the results and implications for future work.

3.2 Literature Review

We now summarize our work in relation to the broader literature. Appointment scheduling is a very well studied topic. Gupta and Denton, 2008 gives an overview on the appointment scheduling opportunities in healthcare for primary care, specialty clinics and elective surgery. They give guidelines for the broad framework in which appointment scheduling models can be explored. Youn, Geismar, and Pinedo, 2022 gives a literature survey on planning and scheduling in healthcare. Their literature overview covers capacity planning for hospitals, outpatient and other networks, and appointment scheduling for different modeling approaches and constraints, including recurring visits. Majority of the papers focus on a single provider or a group of providers in a single specialty. Yu, Kulkarni, and Deshpande, 2020 formulate an MDP for patients that need a series of appointments that are recurring but random for a specialty provider. They compare different scheduling polices that are tractable over MDP. Many papers also consider multiple-steps in the visit. Such multi-stage and multistep scheduling considers the simultaneous scheduling of appointments. Alvarez-Oh, Balasubramanian, Koker, and Muriel, 2018 provide analysis for a two-step stochastic scheduling problem of the visit to the nurse followed by visit to the doctor using integer linear programming. Berg, Erdogan, Lobo, and Pendleton, 2020 propose a constrained optimization model to reduce the variance of the number of doctors scheduled at each hour of the working day over the planning horizon, in a specialty clinic setting.

Patients being referred to other specialties may have urgent or non-urgent need for appointments. Deglise-Hawkinson, Helm, Huschka, Kaufman, and Van Oyen, 2018 provide

an optimization model that ensures delay within a predefined limit for urgent appointments, while having a slight increase in delay for non-urgent appointments. They use approximation measures to linearize the queuing model and use linear optimization that also reduces overtime.

Clinical research trials may require participants on a recurring appointment schedules are aligned to the resource availability. Deglise-Hawkinson, Kaufman, Roessler, and Van Oyen, 2020 uses the linear approximation model to determine such a schedule. Among other things, they aim to reduce the nurse overtime. Lee and Zenios, 2009 examine end-stage kidney disease that needs recurring appointments to come up with the optimal overbooking policy. Since a few patients may need unplanned inpatient care and others may get discharged from such inpatient care, the demand for appointments is not certain, which makes overbooking helpful. Yu et al., 2020 optimally allocates recurring series of appointments that can be scheduled in advance for treatments similar to chemotherapy, kidney dialysis. They use Markov decision process to provide optimal scheduling policies that balances revenue and costs including staffing, overtime, overbooking and delay. Marynissen and Demeulemeester, 2019 provides a literature review on multi-appointment scheduling problems in hospitals.

Outpatient providers can be represented as a network using their corresponding patient referrals. An, O'Malley, Rockmore, and Stock, 2018 infer referrals using patient visit and treatment records. The authors determine various network characteristics related to patient referrals and compare them across different states. These network characteristics are then used over a regression model to determine the relationship between the network measures and health care measures. Patients also use a network of alternatives to inpatient care like home care, assisted living, chronic care etc. Mohammadi Bidhandi, Patrick, Noghani, and Varshoei, 2019 uses a queueing network method to capacity planning for such facilities. The use simulated annealing to determine the optimal capacities with performance guarantees. Helm and Van Oyen, 2014 look beyond the hospital bed capacity to provide optimal scheduling for elective hospital admissions that use a network of facilities in the hospitals—surgery room, ICU, testing etc. The transitions between the facilities is stochastic, and resource constraints force delay due to blocking. They derive analytical expressions for the number of arrivals and use optimization models to determine the best scheduling based on such arrivals.

The use of Markov Renewal Process in our paper draws inspiration from the R. P. Hilton, Zheng, and Serban, 2018. The authors extract and MRP model of patient transitions to different provider types based on the heterogeneity of the health condition of pediatric asthma patients. They cluster the patients based on behavior and healthcare utilization patterns from longitudinal event sequences extracted from a large patient-level database. This clustering is done to find the reasons of variance in utilization of healthcare. The authors compare the cluster characteristics over six regions in US.

Though the literature frequently uses the recurring appointments and specialty network referrals, we do not have a complete picture on the nature of the referrals based on when the appointment requests are made. Most of the literature does not differentiate between patient types of subgroups for analysis. Though a few studies look at the urgent and non-urgent patient, the optimal scheduling policies fail to consider differences in patient health and behavior. Research that attempts to determine the optimal capacity looks at a single specialty or facility than the holistic view of all specialties for all the people in the region. We address these gaps in literature. We use referral and next visit transitions to determine the appointment fill-rates by lead-time. This is analyzed for patient heterogeneity based on health condition by the number of comorbodities. We use simulation to determine the distribution of demand for each type of specialty per 100,000 people. We also determine the capacity needed when appointments are allocated using the patient flexibility based on lead-time.

3.3 Methodology

Consider a person who is part of a regional community. The person may request appointments and visit different health care providers of different specialties as and when needed. We represent the random variable X_n as the n^{th} event that the person has visited a health care provider. At the end of the visit, the person schedules the next event X_{n+1} for a health care provider visit from the state space S of all the specialty providers. It could be the same specialty, or it could be for a different specialty.

We associate the probability of event X_n as $\Pr(X_n = i)$. When:

$$\Pr\left(X_{n+1} = j | X_n = i, X_{n-1}, X_{n-2}, \ldots, X_0\right) = \Pr\left(X_{n+1} = j | X_n = i\right) = p_{ij},$$

the sequence of events hold the memory-less Markovian property. For many office-based and outpatient visits, the prognosis of health conditions in that visit is provided by the health care provider which determines the next step to be taken. For chronic conditions, preventive check-ups and unanticipated flareups may give an indication of the next visit to the specific type of provider specialty. For people without any chronic conditions, the sequence of visits may be based on the preventive check-ups, steady rate of incidence of infectious diseases, or the random nature of individual accidents and injuries that can be treated with outpatient visits.

Though the Markovian property may hold, the sequence of visits cannot be represented as a discrete time Markov chain (DTMC). The next visit does not occur at a fixed time-interval after the previous visit. Instead, the next visit will occur at a random time. We represent this time with the random variable T_{n+1} when the person goes for the next visit X_{n+1}. The time-interval between the two events X_n and X_{n+1} is $T_{n+1} - T_n$.

3.3.1 Markov Renewal Process

A stochastic process $\{(X_n, T_n) | n \geq 0\}$ with state space S is called Markov Renewal Process (MRP) when:

$$\begin{aligned}\Pr(X_{n+1} = j, T_{n+1} - T_n \leq t | (X_n = i, T_n), (X_{n-1}, T_{n-1}), \ldots (X_0, T_0)) \\ = \Pr(T_{n+1} - T_n \leq t, X_{n+1} = j | X_n) \quad \forall n \geq 0, t \geq 0, i, j \in S\end{aligned}$$

In an MRP, the renewal process can change the state within the state space S at each increment and this change of state retains the Markovian property of depending only on the previous state instead of the entire history. One key difference between the MRP and the continuous time Markov chain (CTMC) is that the CTMC requires a change in the state that is $X_{n+1} = j | X_n = i \iff i \neq j$. Further we see that the mean arrival rate in a CTMC also decides the probability distribution. In contrast, the MRP allows the state to

change to the same state. Also, we are able to decouple the transition probability from the time-interval with MRP. Thus we cannot use either DTMC nor CTMC

Within each MRP we have an embedded DTMC where we strictly consider the state changes and ignore the renewal process. This allows us to partially analyze the MRP using the properties of the DTMC. The limiting distribution of a DTMC is used in generating the limiting distribution of an MRP. Let v_i be the limiting distribution for state i for the embedded DTMC. Let τ_i be the mean sojourn time for state i for the MRP. The limiting distribution of the MRP is the ratio of the time the state of the system is i to the sum of all the times in each state as given in eq. (3.1)

$$\pi_i = \frac{v_i \tau_i}{\sum_k v_k \tau_k} \tag{3.1}$$

Instead of the distribution of time spent in each state, we will find the distribution of the time spent in each state transition. That is, the distribution of time spent in a particular state with the transition to a specific next state. The limiting distribution of the transition from i to j, π_{ij}, is analogous to the limiting distribution of time spent in state i given by π_i. We start from the limiting distribution of state i of the embedded Markov chain v_i. Since the transition probability of moving to state j is p_{ij}, we will have $v_i p_{ij}$ as the fraction of transitions from i to j. The mean time spent in state i when moving to state j is τ_{ij}. So the time spent in the transition to j as a fraction of all transitions from i is $v_i p_{ij} \tau_{ij}$. The limiting distribution of this transition from i to j is the ratio of the time spent in this transition to the sum of time spent in all transitions.

$$\pi_{ij} = \frac{v_i p_{ij} \tau_{ij}}{\sum_k \sum_l v_k p_{kl} \tau_{kl}} \tag{3.2}$$

The relation of π_{ij} with π_i can be explained as follows. We can consider the sojourn time of state i as the mean of the time spent in each of the transitions as seen in eq. (3.3).

$$\tau_i = \sum_j p_{ij} \tau_{ij}. \tag{3.3}$$

Rewriting eq. (3.1) with eq. (3.3), we get eq. (3.4).

$$\pi_i = \frac{v_i \sum_j p_{ij}\tau_{ij}}{\sum_k v_k \sum_l p_{kl}\tau_{kl}} \tag{3.4}$$
$$= \frac{\sum_j v_i p_{ij}\tau_{ij}}{\sum_k \sum_l v_k p_{kl}\tau_{kl}}$$
$$\pi_i = \sum_j \left(\frac{v_i p_{ij}\tau_{ij}}{\sum_k \sum_l v_k p_{kl}\tau_{kl}} \right) \tag{3.5}$$

We can rewrite eq. (3.5) using eq. (3.2) to get eq. (3.6).

$$\pi_i = \sum_j \pi_{ij} \tag{3.6}$$

In this work, we are more concerned with the transitions from every state i to a specific state j as we shall see later. This requires us to know $\pi_{ij}\ \forall i$ to help us find $\sum_i \pi_{ij}$.

It could be argued that state space of the MRP could be the transitions themselves. For an original state space of size n, we have n^2 different original transitions. If we define the new state space made with original transitions, we would have n^2 new states and n^4 new transitions. This increases the complexity by having to define multiple improbable transitions in the new MRP. Example, transition probabilities from state ij to kl is zero when $j \neq k$. The use of transitions of the original state space reduces the complexity.

3.3.2 Modeling health care visits as a Markov Renewal Process

The Medical Expenditure Panel Survey (MEPS) provides longitudinal data on outpatient and office-based visits at the day level with the provider specialty associated with the visit. In MEPS, the visits are called as events, and also include visits over telephone in addition to in-person visits. The IPUMS MEPS combines multi-year MEPS survey to provide participants' static information into a single consistent database by using standard variable names and standard coding across different surveys (Blewett, Drew, Griffin, & Williams, 2019). We use both these datasets since IPUMS sources data from MEPS and is expected to be consistent with it when joining data. As the survey can change each year, we identify the relevant survey years using some filters that we describe next. The day of visit is available from the survey year 1999 to the survey year 2012. Survey year 2013 onward only

provides the month and removes the day-level granularity. The provider specialty associated with the visit is available from survey year 2002 till most recently released data (survey year 2020 at the time of writing). The survey questions related to chronic conditions are available from survey year 2007 till most recently released data. The relevant survey years is the intersection of the above years for which relevant data is available—survey year 2007 to survey year 2012. Since we are interested in the longitudinal data, the survey panels that start in survey year 2007 and end in survey year 2012 are panels 12, 13, 14, 15, and 16. Along with the outpatient and office-based visits, MEPS also provides medical-conditions data for each survey participant and the longitudinal weights for the survey panels. We do not consider visits to the emergency department, in-patient, and nurses and technicians, that is available in MEPS, as the rationale of the Markovian property gets difficult.

A flowchart on generating the parameters of the Markov renewal process to model health care visits is given in fig. 3.2.

The individual event files and longitudinal data files are downloaded from MEPS and the medical-conditions information taken from IPUMS. The relevant features from the `csv` files are retained and stacked over multiple survey panels to get data consistent across multiple survey periods. Within the event files, the visit dates are generated. Missing dates are cleaned by uniform random dates.

MEPS intends to represent all civilian non-institutionalized people in US. It assigns appropriate *weights* to the survey participants equivalent to their representation of the national population. The sum of all the weights corresponds to the national population estimate for that time period. The longitudinal files provide these weights to the participants to correctly represent the national population in the two-year longitudinal survey. Since the national population estimates change every year, using the weights as-is over multiple periods can add bias. We normalize the national population for multiple time periods to reduce bias. We convert the weights from whole numbers to fraction of the total. Each survey participant then has a fractional representation of the national population, thus allowing data over multiple years to be aggregated. Lumley, 2004 describes how stratified samples in complex surveys can be handed in `R` language.

The medical-conditions are a list of chronic conditions that the survey participants have been diagnosed anytime. We determine the comorbidity count for each survey participant from the following conditions:

1. Heart Condition
2. Coronary Heart Disease
3. Heart Attack
4. Stroke
5. High Cholesterol
6. Emphysema
7. Diabetes
8. Cancer
9. Arthritis
10. Asthma
11. ADHD
12. Hypertension
13. Angina Pectoris

Since people are generally healthy, not everyone in the survey can be expected to have an outpatient visit. To represent the entire population in the MRP, we add two `HOME` events—one on the day prior to the start of the panel survey, and the other on the day after the end of the survey. We use database joins to merge the three data-tables—longitudinal fractional weights, the comorbidity count, and the visits. The `HOME` events ensures that data remains consistent and nationally representative after full outer database joins. This is because every individual in the survey will have a `HOME` event, even if they did not have any outpatient visit. We assign a patient class based on the comorbidity count for each survey participant. The sequence of visits for each participant is determined. All sequences start and end at `HOME` with a two-year horizon. This allows the embedded Markov chain to be irreducible, since every sequence that ends with the `HOME` state can be regenerated as start of a new sequence. The time-interval between successive events is computed to reflect five weekdays per week instead of seven calendar days per week. This manipulation is essential to remove the effect of longer time-intervals because of weekends. The visit to a particular specialty represents

Comorbidity Count	Distribution	Class	Distribution
0	49.783%	A	70.54%
1	20.753%		
2	11.222%	B	18.72%
3	7.499%		
4	4.527%	C	10.74%
5	2.759%		
6	1.646%		
7	0.970%		
8	0.488%		
9	0.238%		
10	0.097%		
11	0.015%		
12	0.002%		

Table 3.1: Distribution of population by comorbidity count and formulation of classes.

the state of the participant. The transition probabilities of moving from every state to every other state is estimated using law of large numbers. Separate transition probabilities are computed for each patient class. The mean time-interval between every pair of states is calculated with similar grouping by the patient class. The fractional weights are used in each of these computations, so that each survey participant is represented appropriately.

We thus generate separate MRP parameters for each class of patient. We expect each patient class to be homogeneous within itself and will have heterogeneity when compared to other patient classes. Though the appointment bookings are at the same set of specialty providers, the MRP for each class doesn't interact with any other MRP.

The table 3.1 shows the distribution of the population by the comorbidity count. We can see the trend of most people are healthy and fewer people have more comorbidity count. We have split the population roughly by 70-20-10 percent in three classes A, B and C. The class A has most healthy people while the class C has people with comorbidity count of 4 or more.

The MEPS data set includes 35 different medical specialties. We combined all the events labeled as "Don't Know" / "Not Ascertained" / "Refused" in the "Other Dr Specialty". These weighted events made up 10% of all the events. Further, we combined all the least

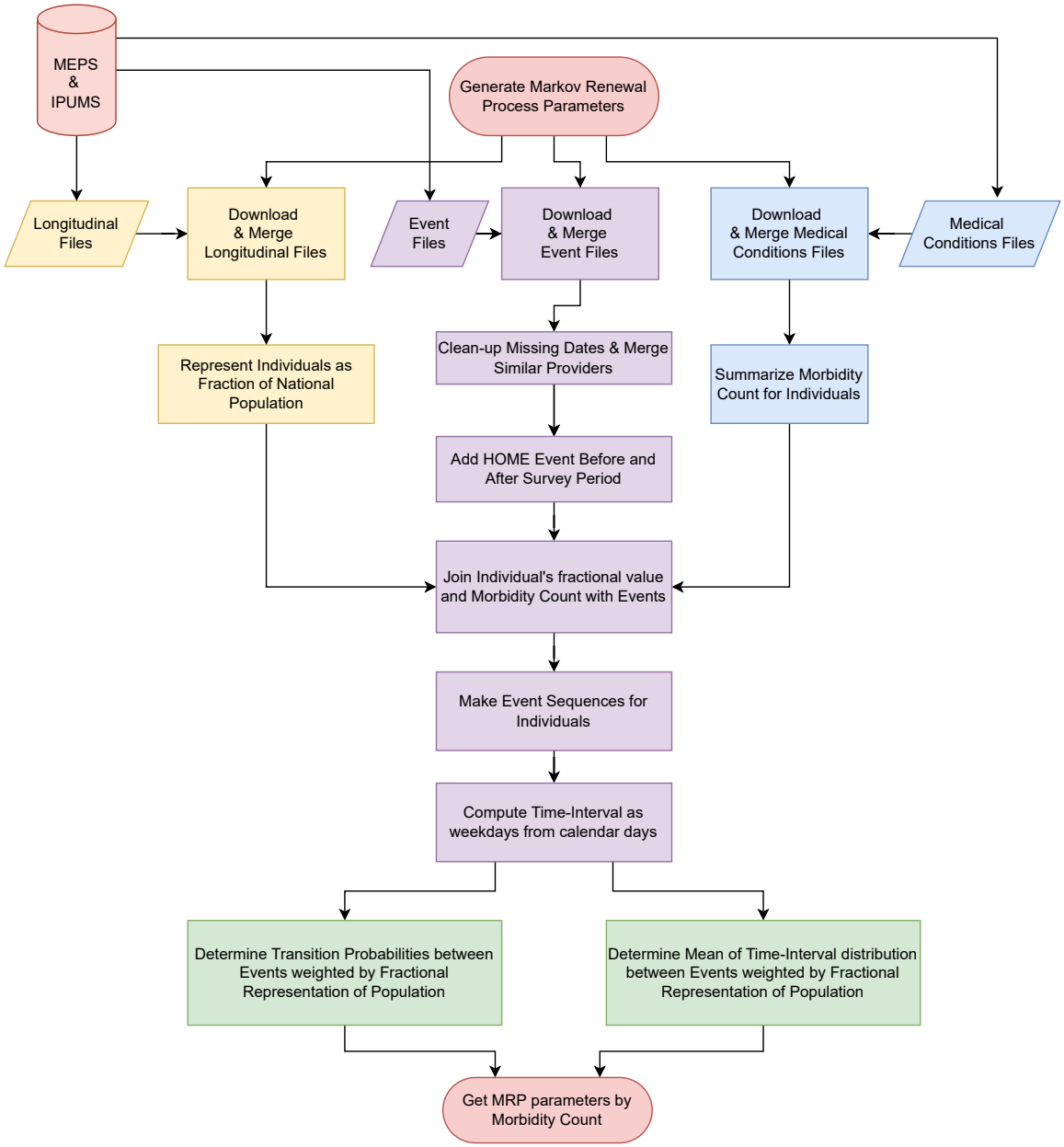


Figure 3.2: Flowchart showing approach to extract Markov Renewal Process parameters from the Medical Expenditure Panel Survey (MEPS) using the office-based visits and outpatient visits, longitudinal weights and medical conditions files.

occurring 6% of the events to "Other Dr Specialty", thus making it nearly 16% of the events. We combined the "Family Practice", "General Practice", "Internal Medicine" and "Pediatrician" as "Primary Care" to become the largest specialty by events of nearly 43%. We have only considered the events where the participant had a visit or a call to a doctor of medicine and ignored events related to nurses and technicians. Thus, the number of specialties considered is 16 with the addition of the `HOME` state making the number of states as 17. Table 3.2 shows the distribution of events considered in the model.

There are $17 \times 17 = 289$ different state transitions. In order to determine the parametric continuous probability distribution for time-interval between states, we examine the histogram. Since it is not possible to examine the histogram of all 289 states, we randomly sampled two state transitions from the top ten most frequently occurring transitions. Their histogram is shown in fig. 3.3. We determined the weighted mean of the time-intervals and found the exponential distribution aligns with the observed data.

The memory-less property of the exponential distribution helps in analysis as we shall see.

3.3.3 Fill-rate for Daily Appointment Requests for a Specialty by Lead-Time for a Homogeneous Population

In an MRP, the limiting distribution of a specialty gives us the expected fraction of the population in a state at anytime. The state of the system is the specialty visit that has been completed. So if a patient leaves for another state, another patient is expected to arrive in that state to replace the previous person. In our model, we assume that patients will request for the next specialty appointment immediately after the previous appointment. We shall try to get insight in the pattern by which appointments are requested for a particular day.

If p is the probability of requesting for an appointment for a person, $(1 - p)$ is the probability of not requesting the appointment by that person. When we consider n people, the distribution of the number of people requesting an appointment follows a binomial distribution. The expected number of people requesting for the appointment is np.

The distribution of time-interval between state transitions is exponential. The time-interval allows us to fix the requested date of appointment and determine the probability of

Specialty in Data	Specialty in Model	Distribution
Family Practice General Practice Internal Medicine Pediatrician	Primary Care	42.85%
OB/GYN	OB/GYN	6.54%
Ophthalmology	Ophthalmology	5.87%
Orthopedics	Orthopedics	5.17%
Psychiatry	Psychiatry	4.35%
Cardiology	Cardiology	3.36%
Dermatology	Dermatology	2.96%
Oncology	Oncology	2.42%
Neurology	Neurology	1.79%
Gastroenterology	Gastroenterology	1.78%
Otorhinolaryngology	Otorhinolaryngology	1.67%
Urology	Urology	1.62%
General Surgery	General Surgery	1.30%
Immunology	Immunology	1.19%
Nephrology	Nephrology	1.16%
Other Dr Specialty Endocrinology Pulmonary Rheumatology Physical Medicine/Rehab Radiology Osteopathy Plastic Surgery Hematology Anesthesiology Geriatrics Proctology Thoracic Surgery Hospital Residence Pathology Nuclear Medicine	Other Dr Specialty	15.96%
Nurse / Technician Home	[Ignore] [Add]	
TOTAL		100%

Table 3.2: Office-based and Outpatient Medical Specialty events as available in data and as used in the model.

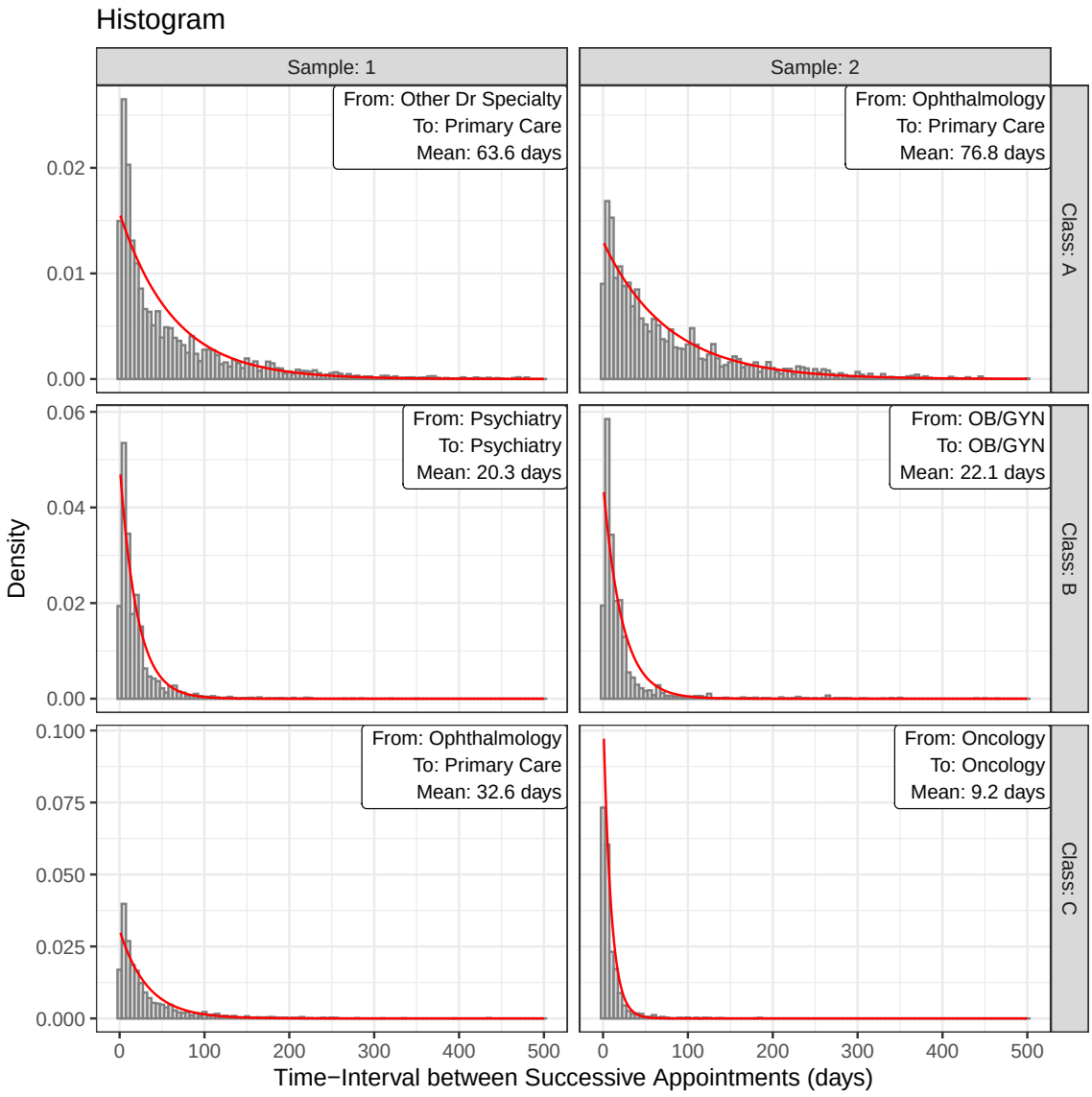


Figure 3.3: Histograms are shown for two sample state transitions for each of the three classes. The red line shows the parameterized exponential distribution function.

requesting an appointment at a time t days before the event. Consider $X_n = i, X_{n+1} = j$ and $T_{n+1} - T_n = t$. We define $\Pr(T_{n+1} - T_n = t | X_{n+1} = j, X_n = i) = f_{ij}(t)$. This is the probability distribution function for the time-interval for the transition from i to j. The corresponding probability density function is $\Pr(T_{n+1} - T_n \leq t | X_{n+1} = j, X_n = i) = F_{ij}(t)$. With population N, the number of people moving from state i to state j is $N\pi_{ij}$. Since π_i is the limiting distribution, the number of people expected to be in state i at any time stays the same, and the number of people expected to transition also remains same. The expected number of people going out of state i will equal the expected number of people coming in state i. From eqs. (3.2) and (3.6), we know π_{ij} is also constant.

Thus, using the binomial distribution, we get the expected number of people requesting appointment at a time t days before the event as $N\pi_{ij}f_{ij}(t)$. We can derive the expected number of people requesting appointment in specialty j at least t days before the event by eq. (3.7)

$$\begin{aligned} \widehat{g_{ij}}(t) &= \int_t^\infty N\pi_{ij} f_{ij}(u)du \\ &= N\pi_{ij} \int_t^\infty f_{ij}(u)du \end{aligned} \tag{3.7}$$

$$\widehat{g_{ij}}(t) = N\pi_{ij}(1 - F_{ij}(t)) \tag{3.8}$$

If we consider all the patients from every state i requesting appointments for a specific state j at least t days before, we get

$$\widehat{g_j}(t) = \sum_i \widehat{g_{ij}}(t) \tag{3.9}$$

using eqs. (3.8) and (3.9) we get eq. (3.10) as shown:

$$\begin{aligned} \widehat{g_j}(t) &= \sum_i N\pi_{ij}(1 - F_{ij}(t)) \\ \widehat{g_j}(t) &= N \sum_i \pi_{ij}(1 - F_{ij}(t)) \end{aligned} \tag{3.10}$$

We know the value of π_{ij} from eq. (3.2). The random variable for the time-interval follows exponential distribution with a mean time-interval of τ_{ij}. Thus, eq. (3.10) becomes

$$\widehat{g_j}(t) = N \sum_i \frac{v_i p_{ij} \tau_{ij}}{\sum_k \sum_j v_k p_{kj} \tau_{kj}} (1 - (1 - e^{-t/\tau_{ij}}))$$
$$\widehat{g_j}(t) = N \frac{\sum_i v_i p_{ij} \tau_{ij} e^{-t/\tau_{ij}}}{\sum_k \sum_j v_k p_{kj} \tau_{kj}} \tag{3.11}$$

We can arrive at a per-capita based appointment requests as from eq. (3.11)

$$g_j(t) = \frac{\widehat{g_j}(t)}{N} = \frac{\sum_i v_i p_{ij} \tau_{ij} e^{-t/\tau_{ij}}}{\sum_k \sum_j v_k p_{kj} \tau_{kj}} \tag{3.12}$$

The eq. (3.12) is of the form

$$g_j(t) = \sum_i a_i e^{-b_i t} \text{ where } 0 \leq a_i \leq 1, b_i \geq 0, t \geq 0. \tag{3.13}$$

The cumulative number of appointment requests expected at the time of the appointment (that it, when $t = 0$) is given in eq. (3.14). So,

$$\frac{\widehat{g_j}(0)}{N} = g_j(0) = \frac{\sum_i v_i p_{ij} \tau_{ij} e^{0/\tau_{ij}}}{\sum_k \sum_l v_k p_{kl} \tau_{kl}}$$
$$g_j(0) = \frac{\sum_i v_i p_{ij} \tau_{ij}}{\sum_k \sum_l v_k p_{kl} \tau_{kl}} \tag{3.14}$$

The $g_j(t)$ expression in eq. (3.11) is for a homogeneous population where one single MRP represents the community. We can extend our analysis for a heterogeneous population where a separate MRP for each class models the behavior of that class.

3.3.4 Fill-rate for Daily Appointment Requests for a Specialty by Lead-Time for a Heterogeneous Population

If the population can be stratified as mutually exclusive classes, we can represent each class with a separate MRP. The MRP for each class is independent of the other class. They do not interact with each other. The number of appointments booked at each specialty is the sum of the appointments booked by each class. Let the various classes by denoted by subscript c.

The distribution of population a class can be denoted as d_c where $\sum_c d_c = 1$. Heterogeneous class notations can be extended from the previous homogeneous class notations.

- The transition probability from specialty i to j for class c can be represented as p_{cij}.
- The mean time-interval between specialties (which is also the parameter of the exponential distribution) is given by τ_{cij}.
- The sojourn time in specialty i is given by τ_{ci}.
- The steady state distribution for the embedded Markov chain is v_{ci}.
- The limiting distribution for the MRP in specialty i is π_{ci}, the components of which, are π_{cij}.
- The expected number of people in specialty i requesting appointment at least t days before, for specialty j is $g_{cij}(t)$.
- The expected number of people requesting appointment at least t days before, for specialty j is $g_{cj}(t)$.
- The expected number of people requesting appointment at least t days before, for specialty j from all classes is $g_j(t)$.

We then get the appointments requested in class c for specialty j before t days is given by eq. (3.15).

$$\begin{aligned} \widehat{g_{cj}}(t) &= N d_c \frac{\sum_i v_{ci} p_{cij} \tau_{cij} e^{-t/\tau_{cij}}}{\sum_k \sum_l v_{ck} p_{ckl} \tau_{ckl}} \\ g_{cj}(t) = \frac{\widehat{g_{cj}}(t)}{N} &= d_c \frac{\sum_i v_{ci} p_{cij} \tau_{cij} e^{-t/\tau_{cij}}}{\sum_k \sum_l v_{ck} p_{ckl} \tau_{ckl}} \end{aligned} \tag{3.15}$$

The total appointments requested at least t days before, for specialty j, from every class c and every specialty i is given by eq. (3.16).

$$\begin{aligned} g_j(t) &= \sum_c g_{cj}(t) \\ g_j(t) &= \sum_c d_c \frac{\sum_i v_{ci} p_{cij} \tau_{cij} e^{-t/\tau_{cij}}}{\sum_k \sum_l v_{ck} p_{ckl} \tau_{ckl}} \end{aligned} \tag{3.16}$$

3.3.5 Simulating the Specialty Network Referrals for Patient Visits

We simulate the specialty network referral for patient visits to understand the distribution of appointment requests for each patient class and specialty. We use the MRP parameters of each class in the simulation.

To initialize the simulation, we first stratify the population to represent the class using table 3.1. The individuals in each class are assigned specialties such that their distribution resembles the limiting distribution of the corresponding MRP. This initial assignment of the simulation— assumes that every individual has completed the appointment on day 0 and the request for the next visit has to be randomly generated.

The next specialty visit, though random, is selected based on the current visit. Though not always required, referrals are frequently given to patients to consult a specialty doctor. Since the population is nationally representative, there is a composition of visits with referrals and visits without referrals. To request the next visit, we randomly select the next specialty using a probability distribution given by the transition probabilities. Using the next specialty we generate a random time-interval from the exponential distribution with the parameter as the mean time-interval between the current specialty and the next specialty. This time-interval gives the realization of the next appointment request day.

We consider a regional population where individuals are part of a set H. This set is partitioned to different classes each represented by the subscript c. So:

$$H_c \cap H_{c'} = \emptyset, \quad \forall c, c' : c \neq c',$$
$$\bigcup_c H_c = H.$$

The set is partitioned to resemble the distribution of the population in that class, that is $|H_c| \propto d_c$. An individual person in class c is $q_c \in H_c$. We know the parameters of the MRP of each class. The transition probability for moving from specialty i to specialty j for class c is p_{cij}. The mean time-interval of moving from specialty i to specialty j for class c is τ_{cij}. The n^{th} visit of patient q_c is the random variable $X_{(q_c,n)}$, and the day of the visit with random variable $T_{(q_{c,n})}$. We generate the next specialty $X_{(q_c,n+1)}$ using random sampling from a probability distribution based on the transition probabilities

of $\Pr\left(X_{(q_c,n+1)} = j | X_{(q_c,n)} = i\right) = p_{cij}$. We generate a random time-interval with this parameterized exponential distribution to get $t_{cij} \sim \text{Exp}\left[1/\tau_{\text{cij}}\right]$.

It is important to note that here we are merely generating requests for appointments. We need to allot the requested appointments, which we shall see below. The realizations of the random variables $X_{(q_c,n)}$ and $T_{(q_c,n)}$ are $s_{(q_c,n)}$ and $r_{(q_c,n)}$ respectively. The corresponding appointment is allotted on day $a_{(q_c,n)}$.

We initialize the simulation by first generating the steady state distribution of $\pi_{(c,i)} \forall (c,i)$ using eq. (3.1), related to each set H_c. Each individual in class c is then assigned an appointment at day 0 for a specialty $s_{(q_c,0)}$ so that the distribution of individuals in different specialties resembles the steady state distribution of the MRP for that class.

$$r_{(q_c,0)} := 0 \tag{3.17}$$

$$a_{(q_c,0)} := 0 \tag{3.18}$$

The simulation progresses by generating the random sequences of the embedded Markov chain and the corresponding renewal process.

$$s_{(q_c,n)} := X_{(q_c,n)} \text{ and} \tag{3.19}$$

$$r_{(q_c,n)} := a_{(q_c,n-1)} + \left\lceil \left(T_{(q_c,n)} - T_{(q_c,n-1)} | X_{(q_c,n)}, X_{(q_c,n-1)}\right)\right\rceil \tag{3.20}$$

The time-interval is a continuous distribution. Since we are only concerned with the day of the appointment and not the specific time of the day, we use the ceiling operator to round up the random time-interval to a positive integer. We allocate the appointment for the same day it has been requested and denote it using

$$a_{(q_c,n)} := r_{(q_c,n)} \tag{3.21}$$

At the end of the simulation we determine the number of appointments requested on each day for all $n > 0$. This is the same as all appointments allotted on each day.

3.3.6 Heuristics to improve allocation of appointments

We have explored the allocation of appointments using heuristics based on patient flexibility in section 1.4.1. We use the "First Minimum" heuristic to allocate appointments in the simulation. We assume each patient has flexibility of appointment allocation based on the lead-time of request similar to eq. (1.5). Each appointment request has some flexibility $\delta_{(q_c,n)}$ within which the appointment allocation may be done. The lead-time of the appointment request is the number of days difference in the previous appointment and the request for next appointment, which is $r_{(q_c,n)} - a_{(q_c,n-1)}$. We convert the lead-time to weeks using $\left\lfloor \frac{r_{(q_c,n)} - a_{(q_c,n-1)} - 1}{5} \right\rfloor$. Here we define the flexibility using eq. (3.22), where each week of lead-time contributes to one day of flexibility. This flexibility is capped to 7 days. Appointments requested at shorter lead-time tend to have less flexibility because of the urgency of the visit at short notice, and also because patients' own schedules are firmed up for the near-future as compared to the far-future.

$$\delta_{(q_c,n)} := \min\left(\left\lfloor \frac{r_{(q_c,n)} - a_{(q_c,n-1)} - 1}{5} \right\rfloor, 7\right) \tag{3.22}$$

The constraints on the appointment allocation is similar to eq. (1.4) as given in

$$\begin{aligned} r_{(q_c,n)} - \delta_{(q_c,n)} \leq a_{(q_c,n)} \leq r_{(q_c,n)} + \delta_{(q_c,n)} \\ l_{(q_c,n)} \leq a_{(q_c,n)} \leq u_{(q_c,n)} \end{aligned} \tag{3.23}$$

The first minimum heuristic searches the day on which the least number of appointments have been allotted within the limits of $\left[l_{(q_c,n)}, u_{(q_c,n)}\right]$. The earliest day is chosen as a tie-breaker for the appointment allocation.

The simulation uses two patient flexibility parameters. When there is no flexibility we have $\delta_{(q_c,n)} = 0$, so eq. (3.23) reduces to eq. (3.21). The flexibility allows us to use the heuristic with the constraints of eq. (3.23). The simulation is run for a population of $|H| = 100,000$ individuals for ten years (2500 days). We run 50 different simulations by controlling the random number seed.

We use Python with Numpy and Pandas(McKinney, 2010; van der Walt, Colbert, & Varoquaux, 2011) to run the simulation. The simulations are run in parallel over multiple processor cores using GNU Parallel(Tange, 2018).

3.4 Results

3.4.1 Markov Renewal Process Parameters

The parameters for the MRP with seventeen states are generated from the MEPS data for a nationally representative population. We show the MRP parameters for a sample of five specialties in fig. 3.4, split by the patient-class. We represent the specialties as nodes, and the transitions as edges in the graph. The edges are labeled with the transition parameters. The first label shows the transition probability as expressed in percentage. The second label shows the mean time-interval of the exponential distribution between the specialties.

The juxtaposition of the MRP parameters can help us compare the patient-classes. For instance, we look at the first row related to transitions from Primary Care to other specialties. The mean time-interval for nearly all transitions reduces as we go from low morbidity count to higher morbidity count. Transition mean time-interval Primary Care to Cardiology may be an exception for class B. The probability of going from Primary Care back to Primary Care reduces slightly as the comorbidity count increases. This is because patients with more comorbidities are more likely to be referred to other specialties. The differences observed in the mean time-interval between successive visits implies that some patient subgroups will have more appointments on shorter notice than others. We now examine this property in the next section.

3.4.2 Fill-rate Analysis of Appointment Requests Over Time

Since the daily appointment requests for a particular specialty is a linear combination of the independent requests from the referring specialties for every class, we can analyze each set of request and their relation to the overall requests. The expected appointment requests from different specialties and classes as a fraction of the total appointment requests over time is somewhat analogous to the order fill-rate concept in supply chain. Since some specialties will have very small number of referrals as compared to others, we can examine the fraction

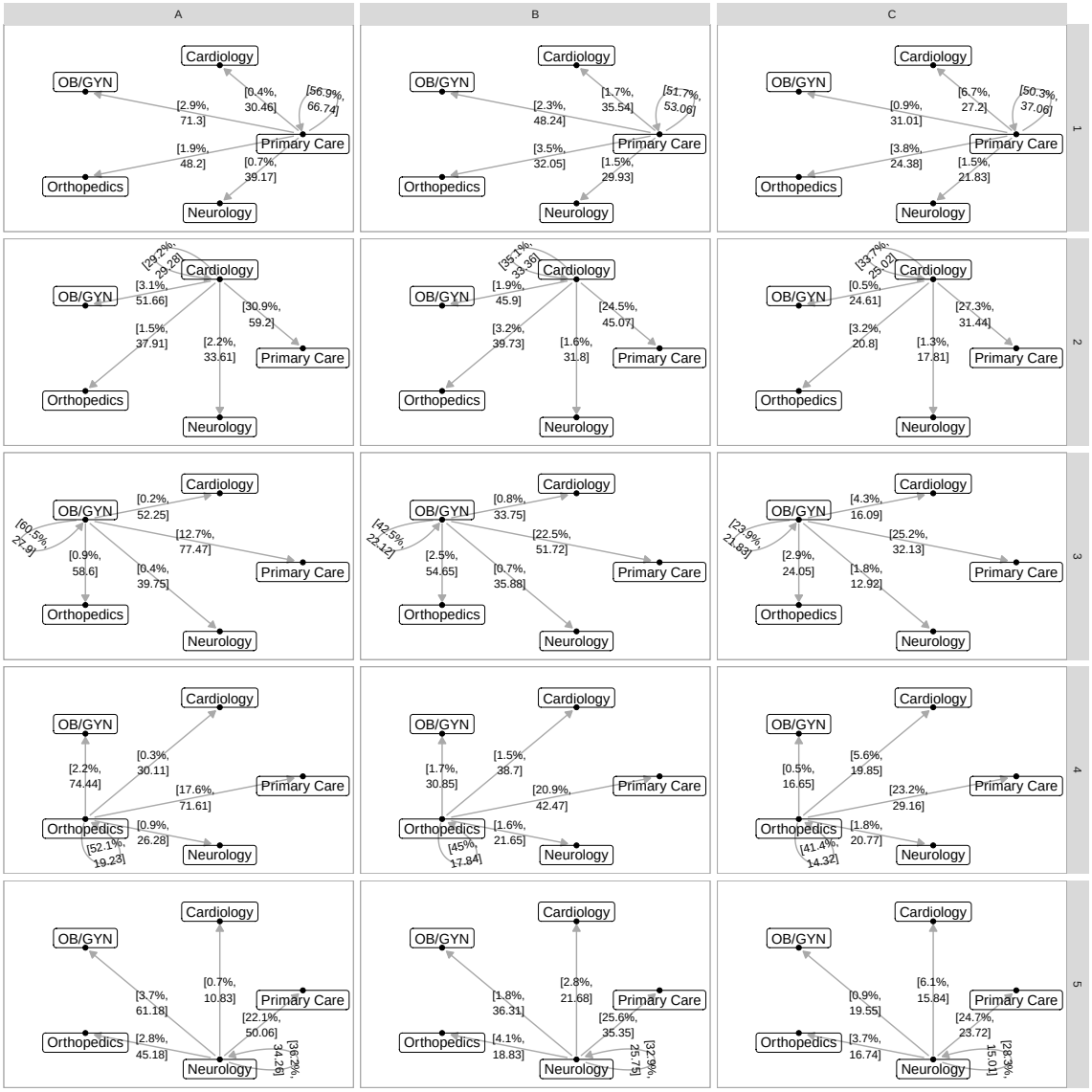


Figure 3.4: Plot shows MRP parameters sampled for five specialties. The nodes of the graph represent the specialty and the edges show the transition probability expressed as percentage along with the mean time-interval between specialty visits. Each column in the figure describes a different patient class. Derived from the Medical Expenditure Panel Survey.

of the appointment requests coming from each specialty over time, by normalizing the said requests. For the sake of depth of analysis, we shall focus on one specialty—Cardiology—to examine details.

Figure 3.5 shows the pattern of appointment requests coming from a specialty network for a homogeneous panel. The appointment requests from each specialty referral are normalized to reach 1 on the day of the appointment. The plot marks the seventh day before the appointment as a vertical intercept. Here we expect 0.79 of all the appointment requests, which implies that 21% of the appointment requests for Cardiology are made in the last seven days. All specialties do not have referrals at the same rate. Patients coming from directly from `HOME`, without any relevant specialty referrals have 92% of all of their appointment requests made before seven days, while only 67% of all the referrals from Neurology have been made in that period. This imbalance in slower rate of appointment requests from Neurology may be look insignificant if the appointments referred from Neurology are a very small fraction of all the Cardiology appointments. But for the patients referred from Neurology, a third of all appointment requests to Cardiology are expected in the last seven days. The problem here is that in case of a capacity limit, most of the appointment requests from `HOME`, would have been allotted, while many of the appointment requests from Neurology will not be allotted.

We can analyze the fill-rate further, by looking at the specialty referrals based on the comorbidity count from fig. 3.6. The patients with different comorbidity counts being referred from OB/GYN to Cardiology have a wide range in the fraction of appointment requests made before seven days. The patients in class A and class B have 13% and 19% of the expected appointment requests in the last seven days respectively, while for class C 35% of the appointment requests are expected in the same period. Similarly, let us look at referrals from Orthopedics. Class A and Class B have 21% and 17% of the appointment requests will be made in the last seven days during which 30% of appointment requests for class C will be made. The referrals from Neurology for Class A are expected to have 48% of appointment requests in the last seven days. This is higher than the class C (36% pending) and class B (28% pending). This particular specialty referral shows that it is possible for

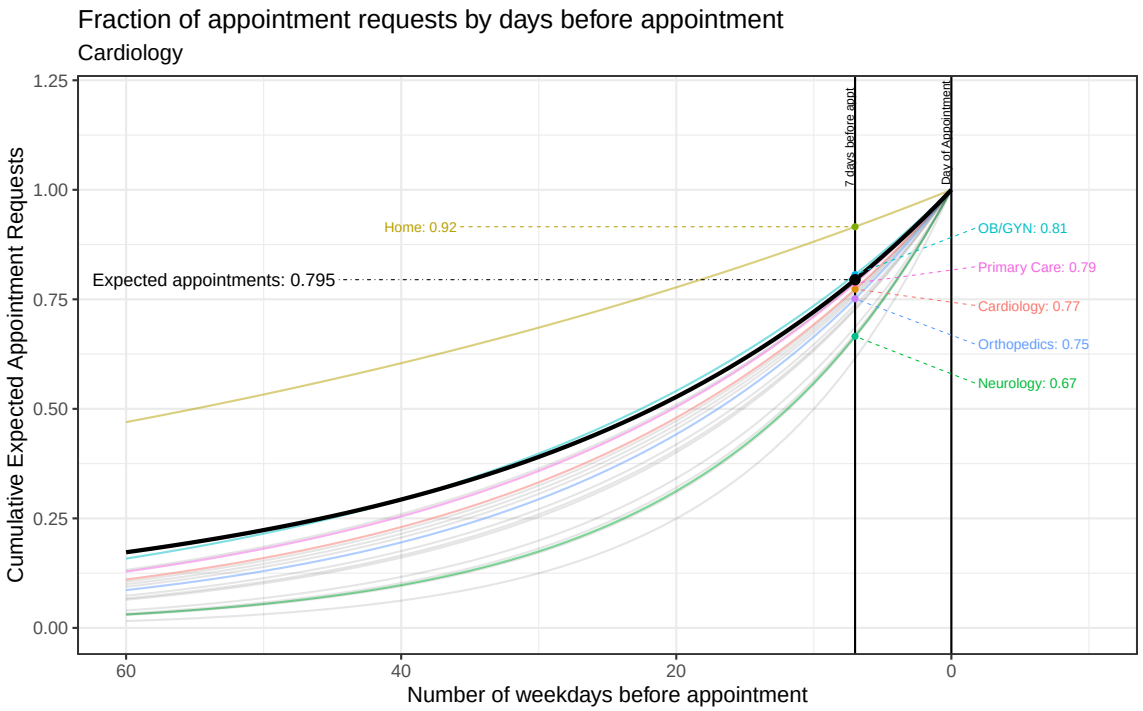


Figure 3.5: The plot shows the normalized appointment requests for Cardiology from patients coming from different specialties by weekdays before the appointment for a homogeneous population. The requests are made from different incoming specialties of which some are highlighted by color, including `HOME`. We can see the fraction of requests which have been completed by some incoming specialties is much lower than others.

the patients with no or low comorbidity count to still make appointment requests at a very short notice.

We take a step back again at the big picture to see the pending appointment requests for Cardiology only by the comorbidity count in fig. 3.7. The class A and class B patients will have only 15% and 17% of their appointment requests in the last seven days during which class C will have 23% of their appointment requests. The implication is that despite being a 10% of the population, the class C patients which have comorbidity count of 4+ will have most of their requests at shorter notice compared to patients in class A & B. If the capacity is filled up by class A & B patients, then class C patients will face more delay leading to inequity in access to healthcare.

We can see the other specialties in table 3.3. The first three columns show the fraction of expected appointment requests in the last seven days for normalized by each specialty for

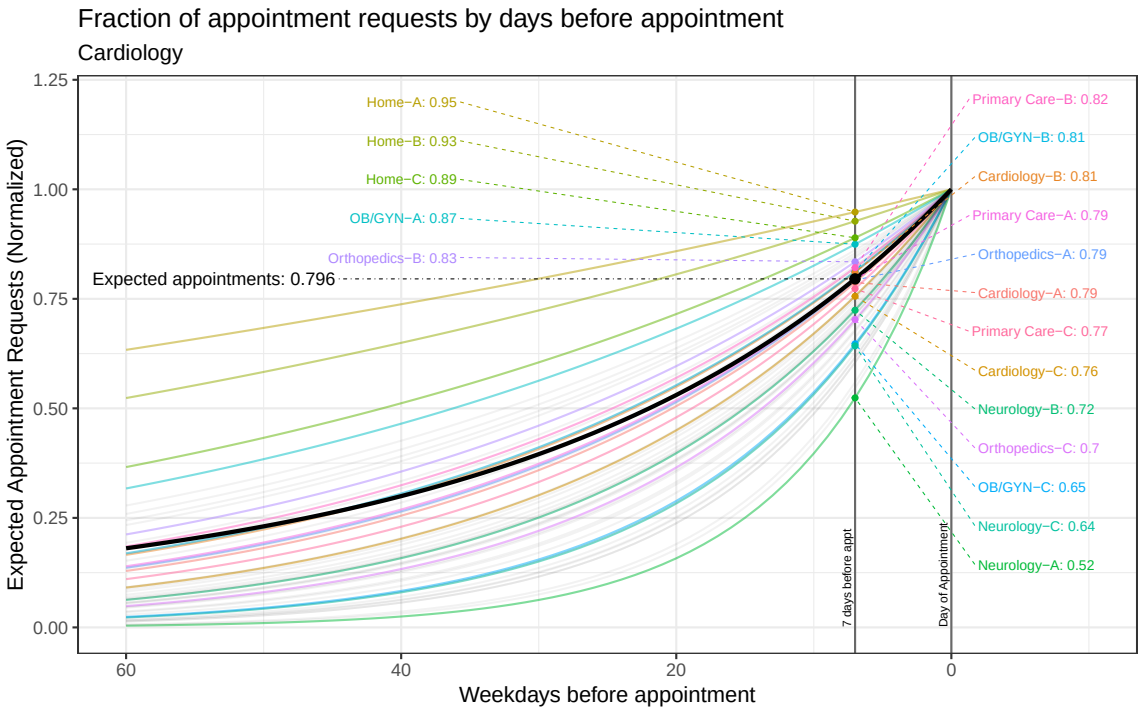


Figure 3.6: The normalized appointment requests for Cardiology by weekdays before the appointment is shown. The appointment requests are split by various incoming specialties and the morbidity class. The highlighted specialties show how significant the differences in fraction of requests completed.

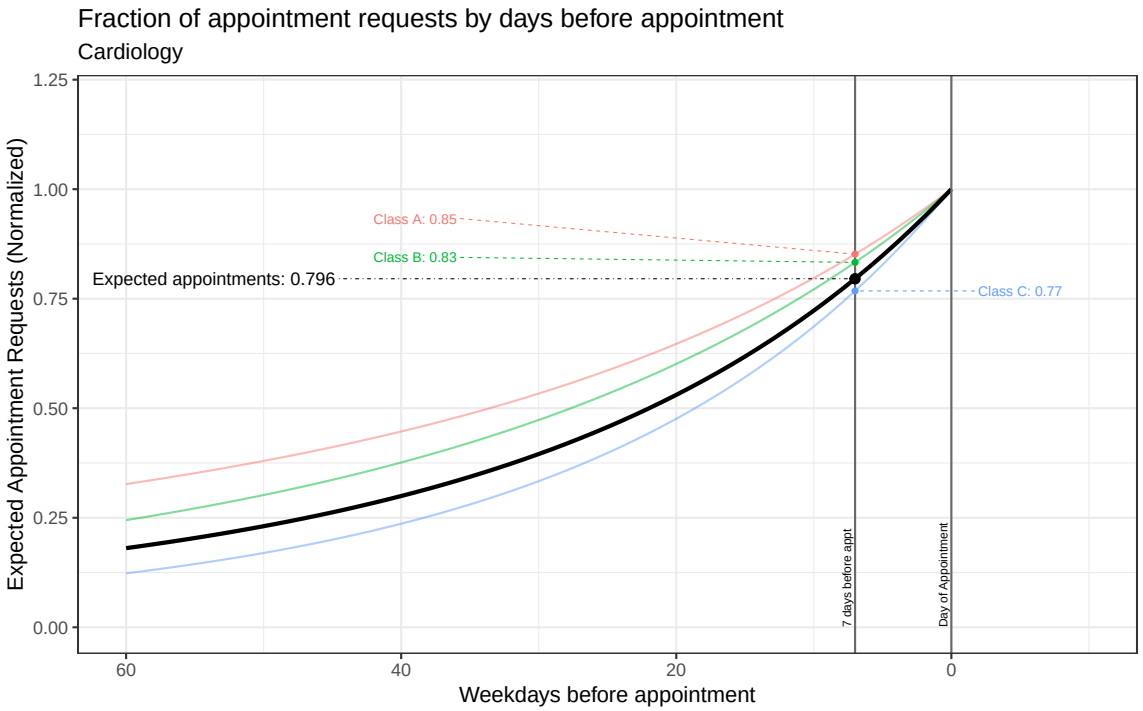


Figure 3.7: The normalized appointment requests for Cardiology by weekdays before the appointment is shown. The appointment requests are split by various classes, for all incoming specialties grouped together. We can see how the patients in class C which have most number of morbidities have lower fraction of requests made 7 days before the appointment. This implies a larger fraction of the appointment requests will be done within 7 days of the appointment than for patients with lower number of car.

that class. Each column is calculated using below expression with $t = 7$ days:

$$1 - \frac{\sum_i v_{ci} p_{cij} \tau_{cij} e^{-t/\tau_{cij}}}{\frac{\left(\sum_i v_{ci} p_{cij} \tau_{cij}\right)}{\left(\sum_k \sum_l v_{ck} p_{ckl} \tau_{ckl}\right)} \left(\sum_k \sum_l v_{ck} p_{ckl} \tau_{ckl}\right)} = 1 - \frac{\sum_i v_{ci} p_{cij} \tau_{cij} e^{-t/\tau_{cij}}}{\sum_i v_{ci} p_{cij} \tau_{cij}}.$$

The column with heading "Overall" is calculated with $t = 7$ days using:

$$\sum_c d_c \left(1 - \frac{\sum_i v_{ci} p_{cij} \tau_{cij} e^{-t/\tau_{cij}}}{\sum_k \sum_l v_{ck} p_{ckl} \tau_{ckl}}\right)$$

The last three columns show the overall fraction of appointment requests expected in the last seven days for each specialty when split by the different classes. Each of these columns is calculated using:

$$d_c \left(1 - \frac{\sum_i v_{ci} p_{cij} \tau_{cij} e^{-t/\tau_{cij}}}{\sum_k \sum_l v_{ck} p_{ckl} \tau_{ckl}}\right)$$

One example is for OB/GYN. Class C patients have 22.8% of their appointment requests in the last seven days as compared to class A and B. But they will have only 0.6% of appointment requests of all the OB/GYN appointment requests.

3.4.3 Appointment distribution and effect of scheduling heuristics

Using the simulations, we determine the daily number of appointments requested for each specialty, using the no-flexibility parameter as shown in fig. 3.8. We remove the first 500 days and last 500 days to ensure steady state. The number of daily appointments per 100,000 individuals represents aggregations. We are not reflecting on the number of physicians since each specialty will have a different number of mean appointments per day. If we assume the number of appointments per physician, we will have the lower bound of the physicians needed, as the variability in appointment requests needs to be considered. This is why we retain the capacity as the number of daily appointments.

We can again consider Cardiology for discussion. The mean of the daily appointments is 26.44 and eightieth percentile is at 31 per 100,000 individuals. This indicates that 80% of all daily visits can be fulfilled by having a daily appointment capacity of 31. The coefficient of variation is 21.1%, Similarly for Neurology, the mean of the daily appointments required

Normalized for the Specialty-Class			Requests in last seven days of appointment		Split by Specialty				
A	B	C	Specialty	Overall =	A	+	B	+	C
0.148	0.167	0.232	Cardiology	0.205	0.018		0.046		0.141
0.105	0.170	0.259	Dermatology	0.145	0.063		0.044		0.038
0.150	0.205	0.295	Gastroenterology	0.193	0.077		0.063		0.053
0.141	0.225	0.287	General Surgery	0.193	0.073		0.062		0.058
0.028	0.051	0.079	Home	0.032	0.023		0.006		0.003
0.200	0.287	0.339	Immunology	0.238	0.128		0.074		0.036
0.176	0.336	0.414	Nephrology	0.359	0.021		0.116		0.222
0.155	0.214	0.297	Neurology	0.205	0.071		0.070		0.064
0.126	0.167	0.228	OB/GYN	0.134	0.104		0.024		0.006
0.241	0.274	0.364	Oncology	0.299	0.058		0.107		0.134
0.088	0.162	0.248	Ophthalmology	0.134	0.050		0.042		0.042
0.144	0.222	0.295	Orthopedics	0.195	0.075		0.065		0.055
0.147	0.230	0.318	Other Dr Specialty	0.205	0.076		0.063		0.066
0.169	0.238	0.291	Otorhinolaryngology	0.202	0.106		0.056		0.040
0.078	0.117	0.174	Primary Care	0.098	0.051		0.025		0.022
0.202	0.259	0.322	Psychiatry	0.232	0.121		0.076		0.035
0.127	0.209	0.274	Urology	0.195	0.052		0.065		0.078

Table 3.3: The first three columns show the fraction of all appointment requests as expected within seven days before the appointment by class. We can see the trend in all specialties where the class with most number of comorbidity also has the highest fraction of appointment requests made within last seven days. In these seven days before the appointment, the patients with higher comorbidity count will request disproportionately more visits than other classes. The last three columns show the fraction of the requests to the specialty that are expected from each of the three classes within seven days before the appointment.

is at 14.94, while the eightieth percentile is at 18 per 100,000 individuals. The coefficient of variation is 25%.

Higher variation results in some days being idle, while other days, there are appointment requests.

To reduce the fluctuations in the appointments allotted, we look at the results when an appointment scheduling heuristic is applied, as seen in fig. 3.9. For Cardiology, we see the mean has reduced to 25.39 appointments and the eightieth percentile is at 28 appointments per day. The coefficient of variation is 12.9%. Similarly, for Neurology, we see the mean of the daily appointments is 14.48 and the eightieth percentile is at 16 appointments per day.

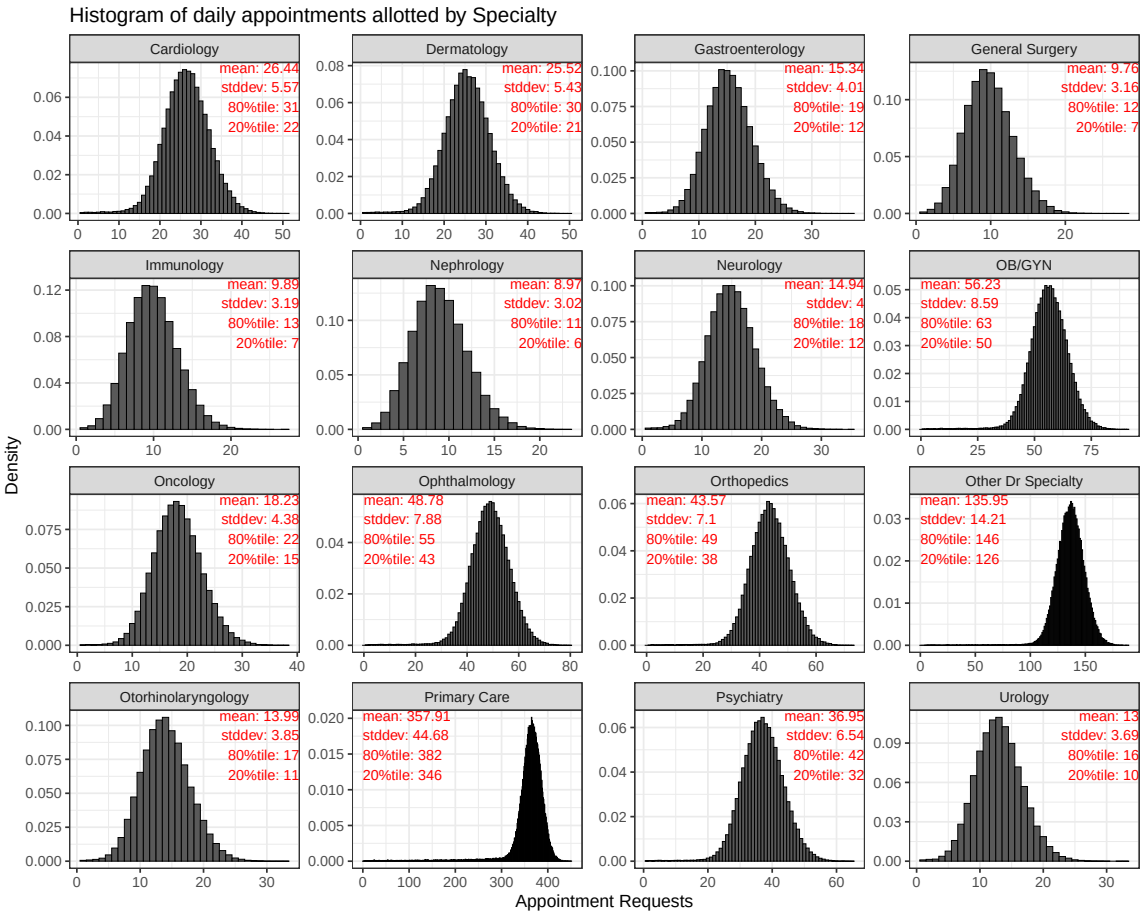


Figure 3.8: Distribution of daily appointment requests and allocation for nationally representative population of 100,000 individuals in a specialty network.

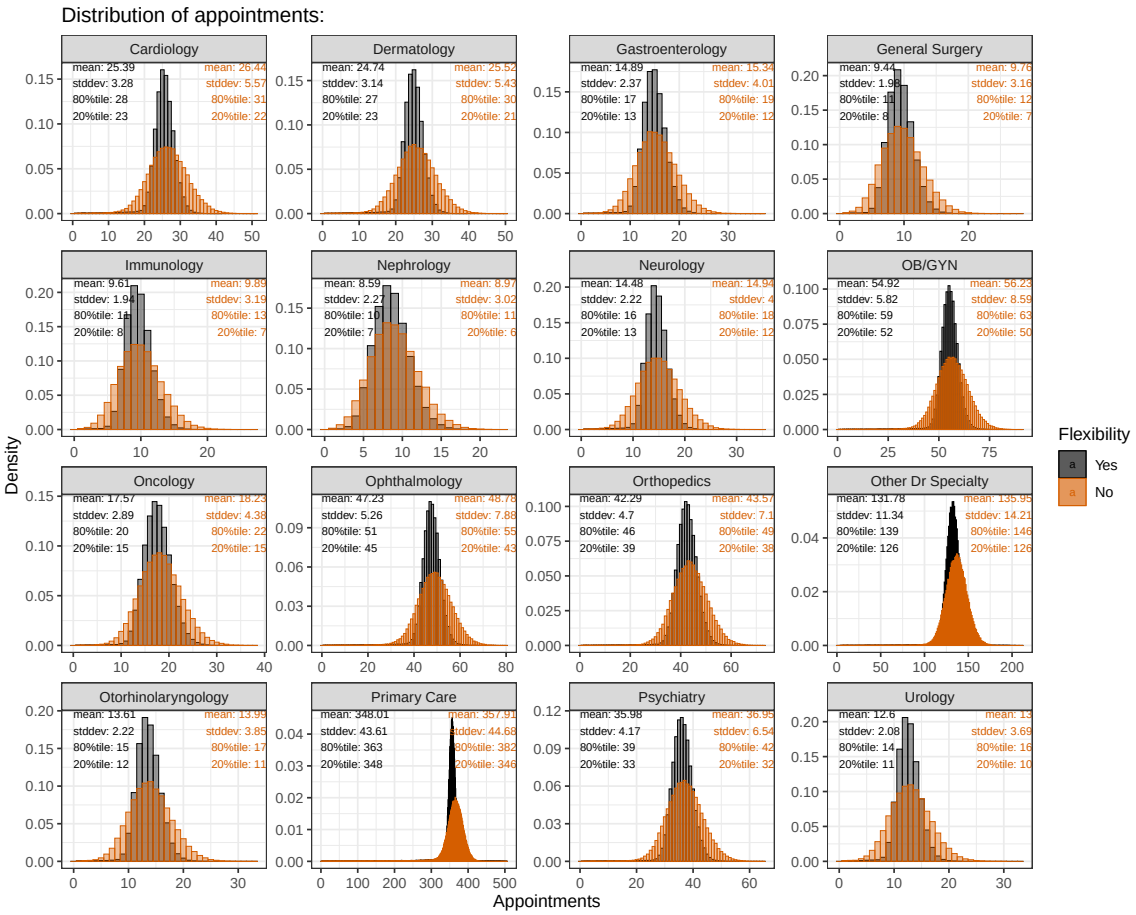


Figure 3.9: Comparison of daily appointment requests and allocation for nationally representative population of 100,000 individuals in a specialty network with the "First Minimum" scheduling heuristic applied.

The coefficient of variation is 15.3%. The reduction in coefficient of variation is nearly 10 percentage points in both cases.

3.5 Conclusion

We are able to analyze the behavior of the entire network of specialty providers simultaneously instead of focusing on a specific specialty. This can help for planning regional health systems at a more granular level than aggregating annual demand.

Our data-driven construction of an MRP ensures that it is nationally representative. We can apply this MRP to regional populations when we don't have accurate estimates of patient visits for such regions. Additionally, we can redefine the states of the MRP to

focus on certain specialties that require detailed analysis, while grouping together other specialties. We can also reshape the patient classes to focus on a subgroup of patients relative to everyone else, to better analyze their needs.

Instead of looking at the average patient, we use heterogeneous models to ensure that the patients with the most healthcare needs, yet extremely low in proportion of the population, are not overlooked.

The use of MRP for modeling outpatient sequences of visits to specialty network captures both—time-interval between visits and transitions to other specialties including the specialty already visited. We get additional insight in the referral patterns, which can determine how many patients from different specialties and different classes will make requests for appointment at short notice. This insight can be used for deciding on overtime or reservations to accommodate the patients with most healthcare needs. While our MRP model only considers requests, an appointment scheduling model can use the expected arrival patterns to improve allocation. It can help medical providers to predict and plan their overtime in advance to manage variations. Using heuristics in appointment scheduling, we show that variation in demand can be leveled with the help of relatively healthy patients with longer appointment lead-times. Our simulation follows the patient-level visits, and is flexible to add additional constraints including appointment cancellations, no-shows, urgent visits. The simulation also allows the distribution of the lead time for appointment to be arbitrary if needed.

This approach of modeling will also be useful for specialty referral decisions. One specialty referring to another specialty notices their patients will need more last minute appointments that others, they can request the specialty to have some kind of *reservation* or priority access for such referrals.

Finally, we feel this model can give planners the big-picture view for all specialties while highlighting the needs of all patient classes.

CHAPTER 4

CONCLUSION

4.1 Summary of work and findings

In this dissertation, we have focused on models for appointment scheduling in primary care and specialty networks. Specifically, we track appointment seeking behavior of individual patients in our models using stochastic processes with unique features that have not been considered in the literature. We also provide insights on the allocation of slots and capacity needs that can satisfactorily address appointment needs. Our models are tested with data extracted from a nationally representative surveys.

Patient heterogeneity is a key feature in our models. We have used techniques to stratify the patient population based on their healthcare needs. Market research experts examine customers' and consumers' needs by market segmentation. This segmentation helps both consumers and providers work with product and service offerings tailored to their specific needs. While such a similar segmentation based approach may informally exist in appointment scheduling, we have been able to highlight the need of analyzing appointment scheduling by looking at different patient subgroups rather than examining the average patient. Models that use the average patient are more analytically tractable, yet they mask important differences between patient subgroups that become more evident in our analysis in both the primary and specialty care setting. In both these settings, we have traced these differences in access for patient subgroups to the appointment lead time.

In chapter 1, we have developed a model for the appointment scheduling in primary care for a patient panel that has recurring appointments. We were able to provide analysis on two fronts: (i) near-optimal heuristics for allocating appointments based on patients' flexibility, and (ii) inequity in delay experienced by patients having more healthcare needs. We have shown that using the "First Minimum" or the "Last Minimum" heuristics for appointment scheduling results in reduced variance of daily appointments allotted and is closer to the

optimal appointment schedule. We have also shown that the appointment delay experienced by primary care patients with more healthcare needs is much higher than patients with lower healthcare need. This delay can be reduced by introducing reservations for critical patient subgroups at the expense of increase in delay for other less critical patient subgroups.

In chapter 2, we have modeled the primary care appointment request and allocation as a discrete time Markov chain. We have derived the expression for the expected delay faced by patient subgroups based on their probability of appointment request. Although we cannot prove mathematically that higher probability of visit results in higher delay because of intractability, we have been able to demonstrate it numerically.

In chapter 3, we have extended our appointment scheduling from a patient panel for primary care to a population that uses a specialty network of outpatient providers. We have extracted the Markov Renewal Process (MRP) parameters for outpatient specialty referrals for various patient subgroups from longitudinal and nationally representative patient level data. Using the limiting distribution of the MRP, we have used a novel method to derive specialty provider's fill-rate of appointment requests by lead-time for all the referring specialties and patient subgroups. We have shown that the patient subgroups that need most access to healthcare also require more appointments at a shorter notice than other patient subgroups. Using simulation, we have determined the distribution of daily appointment requests for all the specialties for a nationally representative population of 100,000 people. This distribution gives an estimate of the number of appointments to be serviced. It implies a lower bound of the appointment capacity required for every specialty. We have used the "First Minimum" heuristics for allocation based on patient flexibility to further reduce the variance in the daily appointment allocations.

4.2 Discussions & Future Work

The future work possible from here can be looked at from three aspects (i) empirical data analysis, (ii) clinical and outcome based research, and (iii) optimal policy analysis.

The effect of delay is well known: delay in access to primary care frequently leads to urgent care or unplanned emergency room visits (Cheung et al., 2012) and access within seven days after hospital discharge reduces the rate of readmission (Wiest, Yang, Wilson,

& Dravid, 2019). Since we now have mathematical evidence of the inequity in access to healthcare faced by patients with more healthcare needs, we should explore the empirical evidence for the differences in delay.

To improve the typical appointment scheduling process for primary care and outpatient specialty providers, we need to collect and analyze data related to every aspect of the scheduling process. This involves recording patient profile, their preferences for appointment request, the discussions and negotiations that occur and the final appointment allocations. This should also involve data related to appointment cancellations, rescheduling, no-shows and future appointments. Although there are many surveys and studies analyzing patient no-shows for primary care and for outpatient specialty provider visits, there is limited knowledge on the health outcome or fallout from no-shows, cancellations and delays.

In our appointment scheduling heuristics we expect the daily workload leveling to help reduce staff overtime and physician burnout. The distribution of the appointments allotted still leaves room to explore newsvendor models for the optimal balance between overtime and idle-time. The optimal online appointment allocation policy can be explored using Markov decision processes.

Specialty providers can use segmented fill-rates for appointment requests for capacity planning for their referring specialties. As an example, if an ophthalmologist visit is frequently recommended to patients with diabetes mellitus, the ophthalmologist can make arrangements with most frequently referring specialty providers to make such visits easy.

While we have analyzed the outpatient specialty network, we can additionally add emergency visits and inpatient visits to have integrated analysis for public health planners. We can further incorporate the population's progression by age and risks to make the analysis more complete, at the cost of computational complexity.

The Markov renewal process may be an under-utilized tool in the Operations Research practitioner's toolkit. We can use it further for modeling disease progression, machine breakdown and maintenance, spare part usage, among other things.

Thus, we conclude this dissertation with the hope that our contributions may lead to improvement in appointment scheduling and capacity planning by considering the diversity in patient needs and the alignment of physician capacity to such needs.

BIBLIOGRAPHY

Ahmadi-Javid, A., Jalali, Z., & Klassen, K. J. (2017, April 1). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, *258*(1), 3–34. doi:10.1016/j.ejor.2016.06.064

Alvarez-Oh, H.-J., Balasubramanian, H., Koker, E., & Muriel, A. (2018, September 1). Stochastic Appointment Scheduling in a Team Primary Care Practice with Two Flexible Nurses and Two Dedicated Providers. *Service Science*, *10*(3), 241–260. doi:10.1287/serv.2018.0219

An, C., O'Malley, A. J., Rockmore, D. N., & Stock, C. D. (2018, February 28). Analysis of the U.S. patient referral network: Analyze US Patient Referral Network and its Relationship to Healthcare. *Statistics in Medicine*, *37*(5), 847–866. doi:10.1002/sim.7565

Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., & Stahl, J. (2010, October 1). Improving Clinical Access and Continuity through Physician Panel Redesign. *Journal of General Internal Medicine*, *25*(10), 1109–1115. doi:10.1007/s11606-010-1417-7

Barnett, M. L., Bitton, A., Souza, J., & Landon, B. E. (2021, December). Trends in Outpatient Care for Medicare Beneficiaries and Implications for Primary Care, 2000–19. *Annals of internal medicine*, *174*(12), 1658–1665. doi:10.7326/M21-1523. pmid: 34724406

Barnett, M. L., Song, Z., & Landon, B. E. (2012, January 23). Trends in Physician Referrals in the United States, 1999-2009. *Archives of Internal Medicine*, *172*(2), 163. doi:10.1001/archinternmed.2011.722

Bavafa, H., Savin, S., & Terwiesch, C. (2019, June 13). *Redesigning Primary Care Delivery: Customized Office Revisit Intervals and E-Visits* (SSRN Scholarly Paper No. ID 2363685). Social Science Research Network. Rochester, NY. Retrieved July 25, 2019, from https://papers.ssrn.com/abstract=2363685

Berg, B. P., Erdogan, S. A., Lobo, J. M., & Pendleton, K. (2020, July 1). A Method for Balancing Provider Schedules in Outpatient Specialty Clinics. *MDM Policy & Practice*, *5*(2), 2381468320963063. doi:10.1177/2381468320963063

Blewett, L. A., Drew, J. A. R., Griffin, R., & Williams, K. C. (2019). *IPUMS Health Surveys: Medical Expenditure Panel Survey: Version 1.1.* Version 1.1. doi:10.18128/D071.V1.1

Boersma, P., Black, L. I., & Ward, B. W. (2020, September 17). Prevalence of Multiple Chronic Conditions Among US Adults, 2018. *Preventing Chronic Disease*, *17*, 200130. doi:10.5888/pcd17.200130

Buttorff, C., Ruder, T., & Bauman, M. (2017). *Multiple chronic conditions in the United States*. Santa Monica, CA: RAND Corporation.

Cayirli, T., & Veral, E. (2003). Outpatient Scheduling in Health Care: A Review of Literature. *Production and Operations Management*, *12*(4), 519–549. doi:10.1111/j.1937-5956.2003.tb00218.x

Cheung, P. T., Wiler, J. L., Lowe, R. A., & Ginde, A. A. (2012, July). National study of barriers to timely primary care and emergency department utilization among Medicaid beneficiaries. *Annals of Emergency Medicine*, *60*(1), 4–10.e2. doi:10.1016/j.annemergmed.2012.01.035. pmid: 22418570

Çinlar, E. (1975, March 1). Exceptional Paper—Markov Renewal Theory: A Survey. *Management Science*, *21*(7), 727–752. doi:10.1287/mnsc.21.7.727

Cook, L. L., Golonka, R. P., Cook, C. M., Walker, R. L., Faris, P., Spenceley, S., ... Oddie, S. (2020, October). Association between continuity and access in primary care: A retrospective cohort study. *CMAJ Open*, *8*(4), E722–E730. doi:10.9778/cmajo.20200014

Deglise-Hawkinson, J., Helm, J. E., Huschka, T., Kaufman, D. L., & Van Oyen, M. P. (2018, December). A Capacity Allocation Planning Model for Integrated Care and Access Management. *Production and Operations Management*, *27*(12), 2270–2290. doi:10.1111/poms.12941

Deglise-Hawkinson, J., Kaufman, D. L., Roessler, B., & Van Oyen, M. P. (2020, August 2). Access planning and resource coordination for clinical research operations. *IISE Transactions*, *52*(8), 832–849. doi:10.1080/24725854.2019.1675202

Emily M. Mitchell. (2019, February 19). *Concentration of Health Expenditures and Selected Characteristics of High Spenders, U.S. Civilian Noninstitutionalized Population, 2016* (No. Statistical Brief 521). Agency for Healthcare Research and Quality. Rockville (MD). Retrieved from https://meps.ahrq.gov/mepsweb/data_files/publications/st521/stat521.shtml

Emily M. Mitchell. (2020, February 20). *Concentration of Healthcare Expenditures and Selected Characteristics of High Spenders, U.S. Civilian Noninstitutionalized Population, 2017* (No. Statistical Brief 528). Agency for Healthcare Research and Quality. Rockville (MD). Retrieved from https://meps.ahrq.gov/mepsweb/data_files/publications/st528/stat528.shtml

Emily M. Mitchell. (2021, January 15). *Concentration of Healthcare Expenditures and Selected Characteristics of High Spenders, U.S. Civilian Noninstitutionalized Population, 2018* (No. Statistical Brief 533). Agency for Healthcare Research and Quality. Rockville (MD). Retrieved from https://meps.ahrq.gov/mepsweb/data_files/publications/st533/stat533.shtml

Ganguli, I., Shi, Z., Orav, E. J., Rao, A., Ray, K. N., & Mehrotra, A. (2020, February 18). Declining Use of Primary Care Among Commercially Insured Adults in the United

States, 2008-2016. *Annals of Internal Medicine*, *172*(4), 240–247. doi:10.7326/M19-1834. pmid: 32016285

Green, L. V. [Linda V.]. (2008, September 1). Using Operations Research to Reduce Delays for Healthcare. In *State-of-the-Art Decision-Making Tools in the Information-Intensive Age* (Vols. 0, pp. 1–16). INFORMS TutORials in Operations Research. doi:10.1287/educ.1080.0049

Green, L. V. [Linda V], & Savin, S. (2008). Reducing delays for medical appointments: A queueing approach. *Operations Research*, *56*(6), 1526–1538. doi:10.1287/opre.1080.0575

Green, L. V. [Linda V], Savin, S., & Murray, M. (2007). Providing timely access to care: What is the right patient panel size? *The Joint Commission Journal on Quality and Patient Safety*, *33*(4), 211–218. doi:10.1016/S1553-7250(07)33025-0

Grumbach, K., Selby, J. V., Damberg, C., Bindman, A. B., Charles Quesenberry, J., Truman, A., & Uratsu, C. (1999, July 21). Resolving the Gatekeeper Conundrum: What Patients Value in Primary Care and Referrals to Specialists. *JAMA*, *282*(3), 261–266. doi:10.1001/jama.282.3.261

Gupta, D., & Denton, B. (2008, July 21). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, *40*(9), 800–819. doi:10.1080/07408170802165880

Gupta, D., & Wang, L. (2008, June 1). Revenue Management for a Primary-Care Clinic in the Presence of Patient Choice. *Operations Research*, *56*(3), 576–592. doi:10.1287/opre.1080.0542

Harrington, W., Rubin, P. A., & Bai, L. (2021, December 1). An optimization approach to panel size management. *Operations Research for Health Care*, *31*, 100313. doi:10.1016/j.orhc.2021.100313

Helm, J. E., & Van Oyen, M. P. (2014, December). Design and Optimization Methods for Elective Hospital Admissions. *Operations Research*, *62*(6), 1265–1282. doi:10.1287/opre.2014.1317

Hilton, R. P., Zheng, Y., & Serban, N. (2018, January 2). Modeling Heterogeneity in Healthcare Utilization Using Massive Medical Claims Data. *Journal of the American Statistical Association*, *113*(521), 111–121. doi:10.1080/01621459.2017.1330203. pmid: 30294054

Hilton, R., Zheng, Y., Fitzpatrick, A., & Serban, N. (2018, January 1). Uncovering Longitudinal Health Care Behaviors for Millions of Medicaid Enrollees: A Multistate Comparison of Pediatric Asthma Utilization. *Medical Decision Making*, *38*(1), 107–119. doi:10.1177/0272989X17731753

Jackson, J. R. (2004). Jobshop-Like Queueing Systems. *Management Science*, *50*(12), 1796–1802. doi:10.1287/mnsc.1040.0268. JSTOR: 30046149

Lee, D. K. K., & Zenios, S. A. (2009, August). Optimal Capacity Overbooking for the Regular Treatment of Chronic Conditions. *Operations Research*, *57*(4), 852–865. doi:10.1287/opre.1080.0666

Liu, N., & Ziya, S. (2014). Panel Size and Overbooking Decisions for Appointment-Based Services under Patient No-Shows. *Production and Operations Management*, *23*(12), 2209–2223. doi:10.1111/poms.12200

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, *9*(8). doi:10.18637/jss.v009.i08

Macinko, J., Starfield, B., & Shi, L. (2007, January 1). Quantifying the Health Benefits of Primary Care Physician Supply in the United States. *International Journal of Health Services*, *37*(1), 111–126. doi:10.2190/3431-G6T7-37M8-P224

Marynissen, J., & Demeulemeester, E. (2019, January). Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research*, *272*(2), 407–419. doi:10.1016/j.ejor.2018.03.001

McKinney, W. (2010). Data Structures for Statistical Computing in Python. (pp. 56–61). Python in Science Conference. doi:10.25080/Majora-92bf1922-00a

McQueenie, R., Ellis, D. A., McConnachie, A., Wilson, P., & Williamson, A. E. (2019, January 11). Morbidity, mortality and missed appointments in healthcare: A national retrospective data linkage study. *BMC Medicine*, *17*(1), 2. doi:10.1186/s12916-018-1234-0

Mehrotra, A., Forrest, C. B., & Lin, C. Y. (2011, March). Dropping the Baton: Specialty Referrals in the United States: Specialty Referrals in the United States. *Milbank Quarterly*, *89*(1), 39–68. doi:10.1111/j.1468-0009.2011.00619.x

Mohammadi Bidhandi, H., Patrick, J., Noghani, P., & Varshoei, P. (2019, May). Capacity planning for a network of community health services. *European Journal of Operational Research*, *275*(1), 266–279. doi:10.1016/j.ejor.2018.11.008

Murray, M., Davies, M., & Boushon, B. (2007, April). Panel size: How many patients can one doctor manage? *Family Practice Management*, *14*(4), 44–51. pmid: 17458336

O'Malley, A. S., & Cunningham, P. J. (2009, February 1). Patient Experiences with Coordination of Care: The Benefit of Continuity and Primary Care Physician as Referral Source. *Journal of General Internal Medicine*, *24*(2), 170–177. doi:10.1007/s11606-008-0885-5

Office of Disease Prevention and Health Promotion. (n.d.-a). Healthy People 2020 Framework. Retrieved from https://www.healthypeople.gov/sites/default/files/HP2020Framework.pdf

Office of Disease Prevention and Health Promotion. (n.d.-b). Healthy People 2030 Framework. Retrieved March 3, 2021, from https://health.gov/healthypeople/about/healthy-people-2030-framework

Office of Disease Prevention and Health Promotion. (n.d.-c). Reduce the Proportion of People Who Can't Get Medical Care When They Need It — AHS-04 - Healthy People 2030. Retrieved March 3, 2021, from https://health.gov/healthypeople/objectives-and-data/browse-objectives/health-care-access-and-quality/reduce-proportion-people-who-cant-get-medical-care-when-they-need-it-ahs-04

Ozen, A., & Balasubramanian, H. (2013). The impact of case mix on timely access to appointments in a primary care group practice. *Health care management science*, *16*(2), 101–118. doi:10.1007/s10729-012-9214-y

Patel, M. P., Schettini, P., O'Leary, C. P., Bosworth, H. B., Anderson, J. B., & Shah, K. P. (2018, May). Closing the Referral Loop: An Analysis of Primary Care Referrals to Specialists in a Large Health System. *Journal of General Internal Medicine*, *33*(5), 715–721. doi:10.1007/s11606-018-4392-z

Rockafellar, R. T., & Wets, R. J.-B. (1998). *Variational Analysis*. Grundlehren Der Mathematischen Wissenschaften. doi:10.1007/978-3-642-02431-3

Rossi, M. C., & Balasubramanian, H. (2018, July 1). Panel Size, Office Visits, and Care Coordination Events: A New Workload Estimation Methodology Based on Patient Longitudinal Event Histories. *MDM Policy & Practice*, *3*(2), 2381468318787188. doi:10.1177/2381468318787188

Rust, G., Ye, J., Baltrus, P., Daniels, E., Adesunloye, B., & Fryer, G. E. (2008, August 11). Practical Barriers to Timely Primary Care Access: Impact on Adult Use of Emergency Department Services. *Archives of Internal Medicine*, *168*(15), 1705–1710. doi:10.1001/archinte.168.15.1705

Starfield, B., Shi, L., & Macinko, J. (2005, September 1). Contribution of Primary Care to Health Systems and Health. *The Milbank Quarterly*, *83*(3), 457–502. doi:10.1111/j.1468-0009.2005.00409.x

Tange, O. (2018, April 27). *GNU Parallel 2018*. doi:10.5281/zenodo.1146014

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011, March). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, *13*(2), 22–30. doi:10.1109/MCSE.2011.37

Vanberkel, P. T., Litvak, N., Puterman, M. L., & Tyldesley, S. (2018, December 1). Queuing network models for panel sizing in oncology. *Queueing Systems*, *90*(3), 291–306. doi:10.1007/s11134-018-9571-4

Wang, J., & Fung, R. Y. K. (2015, January 1). Adaptive dynamic programming algorithms for sequential appointment scheduling with patient preferences. *Artificial Intelligence in Medicine*, *63*(1), 33–40. doi:10.1016/j.artmed.2014.12.002

Wang, W.-Y., & Gupta, D. (2011, June 8). Adaptive Appointment Systems with Patient Preferences. *Manufacturing & Service Operations Management*, *13*(3), 373–389. doi:10.1287/msom.1110.0332

Whitt, W. (2013, May). **OM Forum** —Offered Load Analysis for Staffing. *Manufacturing & Service Operations Management*, *15*(2), 166–169. doi:10.1287/msom.1120.0428

Wiest, D., Yang, Q., Wilson, C., & Dravid, N. (2019, January 4). Outcomes of a Citywide Campaign to Reduce Medicaid Hospital Readmissions With Connection to Primary Care Within 7 Days of Hospital Discharge. *JAMA Network Open*, *2*(1), e187369–e187369. doi:10.1001/jamanetworkopen.2018.7369

Youn, S., Geismar, H. N., & Pinedo, M. (2022, December). Planning and scheduling in healthcare for better care coordination: Current understanding, trending topics, and future opportunities. *Production and Operations Management*, *31*(12), 4407–4423. doi:10.1111/poms.13867

Yu, S., Kulkarni, V. G., & Deshpande, V. (2020, February). Appointment Scheduling for a Health Care Facility with Series Patients. *Production and Operations Management*, *29*(2), 388–409. doi:10.1111/poms.13117

Zacharias, C., & Armony, M. (2016, September 12). Joint Panel Sizing and Appointment Scheduling in Outpatient Care. *Management Science*, *63*(11), 3978–3997. doi:10.1287/mnsc.2016.2532

Zander, A. (2017). Modeling indirect waiting times with an M/D/1/K/N queue. In *Proceedings of the second KSS research workshop : Karlsruhe, germany, february 2016. Ed.: P. Hottum* (Vol. 69, pp. 110–119). KIT Scientific Working Papers. KIT, Karlsruhe.

Zander, A., Nickel, S., & Vanberkel, P. (2021, January 30). Managing the intake of new patients into a physician panel over time. *European Journal of Operational Research*. doi:10.1016/j.ejor.2021.01.035