Capacity Management in Health Care Delivery Systems

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Wen-Ya Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Adviser: Dr. Diwakar Gupta

August, 2012

Acknowledgements

I would like to express my deepest gratitude to my adviser, Dr. Diwakar Gupta, for his guidance. Without his tireless feedback and encouragement, this dissertation would not be possible. I am blessed with the privilege to stay close to him to observe, and to study and learn under his supervision. He is my role model and a true mentor who always shines a light for me to lead me forward. His dedication to my education helped me improve myself in all aspects.

I would also like to thank Dr. Sandra Potthoff for her generous inputs that helped me identify research problems that are both theoretically and practically significant, and Dr. Karen Donohue, and Dr. François Sainfort for their valuable comments to this dissertation. I am also very grateful to Dr. James Lepkowski, who inspired me to pursue a career in academia.

I thank my graduate school friends, Kannapha Amaruchkul, Yi-Su Chen, Daphne Chen, Laura Lee, Wen-Tai Hsu, Chia-Ning Wang, and David Zepeda, for great conversations that ranged from mathematical models to social issues. These different perspectives broadened the scope of my thinking.

Finally, I thank my husband Hao-Wei Chen for his full support and being a great colleague who is happy to hear me talking about my research at any time. His company made my doctoral program an exceptionally rewarding process.

Dedication

To Hao-Wei, and my parents Nyron Wang and Yu-Shan Lai.

Abstract

This dissertation contains three capacity management problems in health care delivery systems. In particular, Chapter 2 evaluates a panel design problem regarding how clinics may wish to best allocate a pool of heterogeneous patients (i.e. non-acute and acute patients) into physician panels. The analytical results show that neither specialization (i.e. each panel contains patients that are as homogeneous as possible) or equal assignment (i.e. identical panels with same types of patient mix) is a dominant patient allocation strategy. The results also show that equal assignment strategy works better when acute demand is relatively low or high as compared to the capacity, and specialization works better when acute demand is moderate. This chapter serves to highlight the impact of patient composition on the performance of a clinic profile. Chapter 3 investigates how clinics may learn and utilize patients' preference information through an existing web-based interface in appointment booking decisions. Analytical results leading to a partial characterization of an optimal booking policy are presented. Examples show that heuristic decision rules, based on this characterization, perform well and reveal insights about trade-offs among a variety of performance metrics such as expected revenue, patient-PCP match rate, number of patients served, and capacity spoilage rate. Chapter 4 focuses on identifying observable predictors of nurse absenteeism and incorporates these factors into staffing decisions. The analysis highlights the importance of paying attention to unit-level factors and absentee-rate heterogeneity among individual nurses. The data-based investigation confirms that nurses' absence history is a good predictor of their future absences. This result is used as the nurse absenteeism assumption in the model-based investigation that evaluates how to assign nurses to identical nursing units when nurses' absentee rates are heterogeneous. We propose and test several easy-to-use heuristics to identify near optimal staffing strategies for inpatient units.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Capacity management in health care delivery systems is a complex problem because it involves multiple stakeholders whose interests may not always align. Broadly speaking, health care services can be categorized into outpatient and inpatient services (see, Figure 1.1). The major capacity management challenges differ by the nature of the service such as service time variability, demand patterns, and the criticality of stake holders' preferences (Gupta and Denton 2008). We start the discussion with outpatient services' stakeholders and capacity management challenges.

Figure 1.1: Outpatient versus Inpatient Services

Outpatient services generally involves booking appointments for fixed-length time slots that are booked for randomly arriving requests from patients. These patients may

have appointment preferences regarding physician and/or time of day, and some may require urgent medical attention. Physicians who work in outpatient settings may establish their work patterns with the clinic, which typically do not change frequently. Many physicians specialize (i.e. serving only one type of patients) in serving sub populations of patients (e.g. geriatric) and also have appointment preferences. Clinics care about the financial performance of the services and patients' satisfaction regarding not only the quality of medical services but also appointment booking experiences.

In outpatient environments, the main capacity management challenges include (1) how to set up the clinic profile (e.g. clinic capacity, physician work hours, appointment length per slot, physicians' work patterns, and panel sizes), and (2) how to book randomly arriving patients into available appointment slots (Gupta and Wang 2011). We frame the outpatient problems discussed in this dissertation within this categorization.

The clinic profile setup problem occurs at the long-term planning stage because clinic capacity strategy depends not only on demand trends, but also on factors such as the health care policy and population changes. In addition, physicians may have their preferred work patterns that cannot be arbitrarily changed by the clinic. In this dissertation, we discuss one capacity management problem that falls in this category in Chapter 2 — *panel design*. A panel is a group of patients who choose the same physician as their preferred care provider (PCP). The panel design problems can occur on a regular basis in many clinics when physicians retire or residents (especially in teaching hospitals) relocate, or when patients' insurance status changes. In these scenarios, patients may be re-assigned to different physician panels, and the question regarding how clinics should allocate different types of patients into different physician panels arises. In this dissertation, we use a stylized model to capture the panel design problem for two identical physicians in Chapter 2. We assume two types of patients (urgent and acute) who have different rates of requesting an appointment. We evaluate patient allocation rules that can be applied by a clinic to improve its overall efficiency.

Once a clinic sorts its patients into different physician panels, the day-to-day capacity management challenges fall in the appointment scheduling category. The clinic needs to book sequentially arriving patients based on the clinic profile that has been set up at the previous planning stage. Many studies have focused on appointment scheduling

problems (see, Chapter 3 for details). However, little attention has been paid to modeling the impact of patients' appointment choices on the performance of an appointment system. In Chapter 3 of this dissertation, we focus on an adaptive appointment system design that learns patients' preference information and takes into account that information to dynamically make appointment booking decisions upon receiving patients' appointment requests. We show that patients' preferences information can be collected through existing web appointment interfaces. In addition, clinics' performance in terms of expected revenue, patient and PCP match, and clinic utilization can be improved by taking into account patients' preferences in booking decisions.

We switch from outpatient to inpatient services in Chapter 4. In inpatient environments, patients' lengths-of-stay (LOS) are random. Primary service providers are registered nurses (RNs) whose work schedules need to be determined a few weeks in advance and unplanned nurse absences may occur. Nurse managers' staffing performance is typically evaluated by the difference between realized and benchmark nursing hours per patient day (NHPPD). Hence, the major challenge here also lies in matching random demand and supply. In addition, some union rules may result in limited flexibility in short-term staffing level adjustments. Therefore, it is also important to reduce frequent staffing adjustments. Some studies focus on smoothing out the demand to the extent possible through surgical or operating room schedules. Nevertheless, demand uncertainty cannot be completely eliminated as patients may come through emergency department and patients' health conditions may change unexpectedly.

In order to make informed RN staffing decisions, there is a need to forecast/estimate nursing needs and assign available nurses to different shifts and nursing units accordingly. Many industrial engineering (IE) and operations research (OR) studies have focused on nurse scheduling problems that aim to minimize staffing costs while taking into account union rules and nurses' scheduling preferences; see Lim et al. (2011) for a recent review. However, these studies generally assume that demand is known (either deterministic or stochastic) and ignore the uncertain in supply (i.e. unplanned nurse absences). In this dissertation, we discuss the challenges of forecasting nurse requirements in Chapter 4. Then in the main body of Chapter 4, we evaluate actionable factors that contribute to nurse absenteeism. We show that nurses' absentee rates may depend on group-level factors such as unit culture and shift start/end time. Furthermore, nurses'

absentee rates differ by nurse and nurses' attendance history is a good predictor of their absentee rates over a reasonable period of time into the future. Based on this result, we model the staffing problem of assigning heterogeneous nurses to multiple nursing units. We show that hospitals may achieve substantial savings through easy-to-use heuristics that account for individual nurses' absentee rates.

Each topic in this dissertation typically contains three main elements: (1) understanding and identifying research problems with exploratory data analysis/hypothesis testing (if data are available) and literature analysis (2) formulating and analyzing models to suggest operational strategies, and (3) evaluating the proposed model/strategy's performance with analytical and/or confirmatory data analysis.

The remaining of this dissertation is organized as follows. Chapters 2 describes the panel design problem and Chapter 3 evaluates the impact of patients' preferences on appointment systems' performance. Chapter 4 investigates inpatient units' nurse assignment problems when nurses may be absent. Chapter 5 concludes the dissertation.

Two papers have been completed based on the work reported in this thesis. One paper based on Chapter 3 has appeared in the *Manufacturing & Services Operations Management* journal. One paper based on Chapter 4 is currently under review.

Chapter 2

Physician Panel Design

2.1 Introduction

Many health systems ask patients to designate a service provider who specialized in primary care or internal medicine as their preferred care provider (PCP). Patients who choose the same PCP form a panel. Clinics prefer patients to book appointments with their PCPs to the extent possible to ensure continuity of care and improve clinic revenue (O'Hare and Corlett 2004). Therefore, it is important to pay attention to the balance between physicians' capacity and the demand generated from their panels. In this dissertation, we refer the problem of determining the patient composition (e.g. panel size, age distribution, and health status) for physicians' panels as the panel design problem.

Several studies have addressed the problem of determining panel sizes in the primary care setting. When a physician's panel becomes too large, not only the provider who has a large panel would experience backlogs, but also other providers who may end up serving urgent and semi-urgent demand for the former provider. This also results in patient-PCP mismatches for providers whose capacity is utilized by the overflow patients from the congested panel. Panel sizes are often used to benchmark providers' workload. For example, Murray et al. (2007) provide a capacity balance equation that matches average demand (panel size $\times$ visits per patient per year) and available capacity (available provider slots per day $\times$ provider days per year). Green et al. (2007) focus on the impact of panel size on the level of overflow frequency.

However, workload is affected not only by the size of a physician's panel, but also other factors such as age distribution and health status of the patients in the panel. Different types of patients may have different frequency of visits and their needs also vary. Therefore, it is difficult to compare the load of heterogeneous panels purely based on their sizes. Gupta and Wang (2011) address this issue using a fluid model that assumes two classes of patients with different visit rates (acute care and follow-up patients), and models the choice of follow-up inter-visit times in panel size decisions for a single physician.

In many health systems, clinics can encourage patients to join different physicians' panels or recommend physicians to new patients. When physicians retire or relocate, their patients are re-assigned to different panels. If patients' insurance status changes, then their appointment request patterns may also change. Therefore, a clinic may need to rebalance the workload among different physicians by adjusting physician panel sizes and patient composition.

Nevertheless, it is not clear how a clinic may design its multiple physician panels for a given pool of heterogeneous patients. At one end of the spectrum, the clinic may wish to use a specialization strategy that allows a physician to specialize in serving a particular type of patients. At the other end, a clinic may evenly distributed each type of patients among different panels to ensure that the workload is the same across physicians. It is not clear which of these two strategies is superior, or whether there are other patient allocation rules that allow the clinic to achieve better performance.

Ideally, a clinic may want to manage its capacity by taking into account the setup of clinic profiles (i.e. panel composition, physician capacity, etc) and the design of appointment systems simultaneously. However, due to the complexity of the problem, the solution to the capacity management problem becomes analytically intractable. For this reason, many clinics use a two-step process – set up the clinic profile first, then determine appointment booking rules. We follow the two-step process convention in this dissertation.

In this chapter, we focus on a stylized problem of assigning two types of patients (non-acute and acute) to two identical physicians. This analysis is used to provide insights concerning how clinics may design physician panels to accommodate the demand for urgent (i.e. acute, walk-in, same-day) and non-urgent (i.e. non-acute, advance-book)

patients. Note that the urgency of an appointment is generally self-defined by the patients. We use the word *acute* interchangeably with urgent, walk-in, and same-day patients throughout the dissertation.

We present the problem formulation and analyze different panel design strategies in Section 2.2. We conclude with findings and insights in Section 2.3. In the panel design problem, we do not account for patients' preference regarding different appointment slots. Then in the Chapter 3, we assume that each physician's panel composition is given and focus on accommodating patients' preferences while taking into account urgent and non-urgent requests.

2.2 Formulation and Analysis

In this section, we formulate the panel design problem when there are two identical physicians each with capacity κ per session and two types of patients:

- Type 1: Non-acute patients who request appointments in a non-urgent manner. This group of patients generally schedule visits for reasons such as chronic conditions or follow-ups. We assume that type-1 patients' appointment requests arrive prior to the appointment day and view these patients as advance-book patients. The average revenue and cost for serving and turning away a non-acute patient's request are r_1 and π, respectively. There are a total of n type-1 patients.

- Type 2: Acute patients. This group of patients are generally healthy, but may occasionally need urgent attention for reasons such as flu or sport injury. We assume that type-2 patients' appointment needs are urgent and they book appointments on the same-day that they need medical attention. The average revenue and cost for serving and turning away an acute patient's request are r_2 and c, respectively. There are a total of m type-2 patients.

The revenue parameters (i.e. r_1 and r_2) may be estimated through hospital's financial data and appointment records. The cost parameters (i.e. π and c) may be interpreted as functions of physicians' overtime costs or patients' wait costs. These parameters may be estimated as relative costs as compared to r_1 and r_2 (e.g. Robinson and Chen 2010b), or as the cost of patient waiting (e.g. Yabroff et al. 2005, Russell 2009). Note that we

define the revenue and cost parameters from a clinic's perspective. Therefore, prior to the realization of a service, it is reasonable for the clinics to use expected revenue (resp. expected cost) to evaluate the benefit (resp. penalty) for serving (resp. not serving) a patient.

The expected revenues for the two types of patients (i.e. r_1 and r_2) are not necessarily different, however, the expected costs for turning away different types of patients (i.e. π and c) could be very different. The cost of turning away a non-urgent patient may be negligibly small so long as the turned-away patient is able to book an appointment within a short time from his/her desirable appointment day. However, the cost of not serving an urgent/same-day/walk-in patient could be substantial for several reasons. First, an unserved urgent patient may go to a different clinic that can provide care to him/her right away, which results in lost revenue for the clinic and potentially a higher cost for the health insurer. This is because many clinics are part of a health system and the cost for serving a patient within the system's clinic is lower for the insurer. Second, if the unserved urgent patient goes to emergency department of the same health system, then the cost of caring for the patient is still higher because providing care in the emergency department is more expensive. Furthermore, if many urgent patients cannot get medical attention in a timely manner, then there may be a negative effect on the clinic's reputation. For the above reasons, we assume that $(r_2 + c) \geq (r_1 + \pi)$ because turning away an acute patient is generally less desirable than letting an non-acute patient wait.

We take the clinic's perspective in formulating the panel design problem where the objective is to find a patient allocation between two identical physicians such that the expected total reward is maximized. To keep our model tractable, we assume that each patient belongs to one type although in reality some non-acute patient may also occasionally need urgent care. Although this assumption does not allow a patient to generate different types of appointment requests, it still allows different patient mix to result in different aggregate urgent and non-urgent demand distributions.

Let $X_1(\alpha_i)$ and $X_2(\beta_i)$ be the demand for type-1 and type-2 patients, where α_i, and β_i are the proportion of type-i patients who are assigned to panel i. We assume the following demand properties.

- $X_1(\alpha_i) = n\alpha_i Y_1$ and $X_1(\alpha_i + \alpha_j) = n\alpha_i Y_1 + n\alpha_j Y_1 = nY_1$ (because $\alpha_i + \alpha_j = 1$);

- $X_2(\beta_i) = m\beta_i Y_2$ and $X_2(\beta_i + \beta_j) = m\beta_i Y_2 + m\beta_j Y_2 = mY_2$ (because $\beta_i + \beta_j = 1$);

where $Y_1 \in [0,1]$ and $Y_2 \in [0,1]$ respectively are the uncertainty in type-1 and type-2 demand. Let $f(y_1)$ and $g(y_2)$ respectively be the probability density functions for Y_1 and Y_2. Let $F(y_1)$ and $G(y_2)$ respectively be the cumulative density function for Y_1 and Y_2. The distribution function for $X_1(\alpha_i)$ is $P(X_1(\alpha_i) \le a) = P(n\alpha_i Y_1 \le a) = P(Y_1 \le a/(n\alpha_i)) = F(a/(n\alpha_i))$, and $P(X_2(\beta_i) \le a) = G(a/(\beta_i m))$. In addition, $P(X_1(\alpha_i) > a) = \bar{F}(a/(n\alpha_i))$ and $P(X_2(\beta_i) > a) = \bar{G}(a/(\beta_i m))$.

Let κ_i and $\kappa - \kappa_i$ respectively be panel-i's capacity allocated to type-1 and type-2 patients. Note that we assume type-1 patients' requests arrive before type-2 patients' requests. Therefore, the unused type-1 capacity may be utilized by type-2 patients. If $\kappa_i = \kappa$, then the clinic has no capacity control over who will book the appointments and the appointment slots are based on a first-come-first serve policy. If $\kappa_i = 0$, then the clinic does not allow any advance-book appointments. Let $\Pi = \Pi(\alpha_1, \beta_1, \kappa_1) + \Pi(\alpha_2, \beta_2, \kappa_2)$ be the total expected revenue for the two physicians, where the revenue function for physician i is

$$\begin{aligned}
\Pi(\alpha_i, \beta_i, \kappa_i) &= E[r_1 \min\{X_1(\alpha_i), \kappa_i\} + r_2 \min\{X_2(\beta_i), \kappa - \kappa_i + (\kappa_i - X_1(\alpha_i))^+\} \\
&\quad -\pi(X_1(\alpha_i) - \kappa_i)^+ - c(X_2(\beta_i) - (\kappa - \kappa_i + (\kappa_i - X_1(\alpha_i))^+))^+] \qquad (2.1) \\
&= r_1 E(X_1(\alpha_i)) + r_2 E(X_2(\beta_i)) - (r_1 + \pi) E(X_1(\alpha_i) - \kappa_i)^+ \\
&\quad -(r_2 + c) E(X_2(\beta_i) - (\kappa - \kappa_i))^+ P(X_1(\alpha_i) > \kappa_i) \\
&\quad -(r_2 + c) \int_0^{\kappa_i/n\alpha_i} E(X_2(\beta_i) - \kappa + x_1)^+ f(y_1) dy_1 \\
&= r_1 n\alpha_i E(Y_1) + r_2 m\beta E(Y_2) - (r_1 + \pi) \int_{\frac{\kappa_i}{\alpha_i n}}^1 (\alpha_i n y_1 - \kappa_i) f(y_1) dy_1 \\
&\quad -(r_2 + c) \int_{\frac{\kappa - \kappa_i}{\beta_i m}}^1 (\beta_i m y_2 - \kappa + \kappa_i) g(y_2) dy_2 \bar{F}\Big(\frac{\kappa_i}{\alpha_i n}\Big) \\
&\quad -(r_2 + c) \int_0^{\frac{\kappa_i}{\alpha_i n}} \int_{\frac{\kappa - \alpha_i n y_1}{\beta_i m}}^1 (\beta_i m y_2 - \kappa + \alpha_i n y_1) g(y_2) f(y_1) dy_2 dy_1,
\end{aligned}$$

where in equation (2.1) the first (resp. second) term is the expected revenue from type-1 (resp. type-2) demand, and the third (resp. fourth) term is the expected penalty from unserved type-1 (resp. type-2) demand.

Next, we show that for any given α_i and β_i, there is an optimal capacity level reserved for non-acute patients that maximizes the expected revenue for physician i.

$$\begin{aligned}\frac{\partial \Pi(\alpha_i,\beta_i,\kappa_i)}{\partial \kappa_i} &= (r_1+\pi)P(X_1(\alpha_i)>\kappa_i)\\ &\quad -(r_2+c)P(X_2(\beta_i)>\kappa-\kappa_i)P(X_1(\alpha_i)>\kappa_i)\\ &= [(r_1+\pi)-(r_2+c)P(X_2(\beta_i)>\kappa-\kappa_i)]P(X_1(\alpha_i)>\kappa_i)\\ &= \Big[(r_1+\pi)-(r_2+c)P\Big(Y_2>\frac{\kappa-\kappa_i}{\beta_i m}\Big)\Big]P\Big(Y_1>\frac{\kappa_i}{\alpha_i n}\Big).\end{aligned}$$

Note that $P(X_1(\alpha_i)>\kappa_i)\geq 0$ for any κ_i, and $P(X_2(\beta_i)>\kappa-\kappa_i)$ increases in κ_i. Therefore, if $(r_1+\pi)-(r_2+c)P(X_2(\beta_i)>\kappa)\leq 0$, then any $\kappa_i>0$ will decrease the revenue, and the optimal capacity reserved for non-acute patient is 0. If $(r_1+\pi)-(r_2+c)P(X_2(\beta_i)>\kappa)>0$, then the revenue function increases in κ_i for any κ_i less than or equal to $\kappa_i^*=\arg\max\{\kappa_i:(r_1+\pi)-(r_2+c)P(X_2(\beta_i)>\kappa-\kappa_i)=0|0\leq\kappa_i\leq\kappa\}=\arg\max\{\kappa_i:\bar{G}\Big(\frac{\kappa-\kappa_i}{\beta_i m}\Big)=\frac{r_1+\pi}{r_2+c}|0\leq\kappa_i\leq\kappa\}$, and the revenue function decreases in κ_i for any κ_i greater than κ_i^*. Let $0\leq z^*\leq 1$ be a number such that $(r_1+\pi)=(r_2+c)\bar{G}(z^*)$. This will be useful in deriving the properties of the optimal assignments in the ensuing discussion. Then the optimal capacity reserved for type-1 patients for panel i is $k_i^*=\kappa-\beta_i m z$, where

$$z=\begin{cases}\frac{\kappa}{\beta_i m} & \text{if } (r_1+\pi)<(r_2+c)\bar{G}\Big(\frac{\kappa}{\beta_i m}\Big);\\ z^* & \text{otherwise.}\end{cases} \tag{2.2}$$

This booking-limit type capacity reservation for type-2 patients is reasonable because of the linear cost induced by any additional acute or non-acute patient who is not served. Note that the best capacity reserved for acute (respectively non-acute) patients only depends on the amount of type-2 patients allotted to the panel. If we re-arrange equation (2.2), it can be shown that if $\beta_i\geq\kappa/(mz^*)$, then physician-i will reserve all capacity for type-2 patients. If $0<\beta_i\leq\kappa/(mz^*)$, then the capacity reserved for type-1 (resp. type-2) patients is $\kappa-\beta_i m z^*$ (resp. $\beta_i m z^*$).

In the ensuing analysis, we drop the subscript of panel index in patient allocation notation and replace α_1 by α and β_1 by β for simplicity. Recall that we are considering two physician panels in which we assign α type-1 and β type-2 patients to one panel, and $(1-\alpha)$ type-1 and $(1-\beta)$ type-2 patients to the other panel. Next, we evaluate the relationship between β and the revenue function $\Pi=\Pi(\alpha,\beta,\kappa_1^*)+\Pi(1-\alpha,1-\beta,\kappa_2^*)$.

There are four scenarios to be considered based on the capacity reserved for type-1 and type-2 patients in each panel:

Case 1: If $\beta \geq \kappa/(mz^*)$ and $(1-\beta) \geq \kappa/(mz^*)$, then $\kappa_1^* = 0$ and $\kappa_2^* = 0$. This case only exist when $\kappa/(mz^*) \leq 0.5$ because $\kappa/mz^* \leq \beta \leq 1 - \kappa/(mz^*)$.

Case 2a: If $\beta \geq \kappa/(mz^*)$ and $(1-\beta) \leq \kappa/(mz^*)$, then $\kappa_1^* = 0$ and $\kappa_2^* = \kappa - (1-\beta)mz^*$. This case exists for $0 \leq \kappa/(mz^*) \leq 1$.

Case 2b: If $\beta \leq \kappa/(mz^*)$ and $(1-\beta) \geq \kappa/(mz^*)$, then $\kappa_1^* = \kappa - \beta mz^*$ and $\kappa_2^* = 0$. This is a case that is identical to case 2a due to symmetry. This case exists for $0 \leq \kappa/(mz^*) \leq 1$.

Case 3: If $\beta \leq \kappa/(mz^*)$ and $(1-\beta) \leq \kappa/(mz^*)$, then $\kappa_1^* = \kappa - \beta mz^*$ and $\kappa_2^* = \kappa - (1-\beta)mz^*$. This case only exists when $\kappa/(mz^*) \geq 0.5$. Note that when $\kappa/(mz^*) = 0.5$, case 3 is identical to case 1.

For each case, we replace κ_i with κ_i^* in the revenue function to obtain the optimal β. Then we use the optimal β to derive the optimal α. Note that because we assume identical physicians, cases 2a and 2b are symmetric and only one of them needs to be considered.

Let α^* and β^* be the optimal allocation of type-1 and type-2 patients, and κ_i^* be the optimal capacity reserved for type-1 patients for physician i. Upon evaluating the three scenarios, we obtain the following results depending on the value of $\kappa/(mz^*)$.

Corollary 2.2.1 *If $\kappa/(mz^*) \leq \beta \leq 1 - \kappa/(mz^*)$, then the optimal patient allocation is to assign α arbitrarily, and $\beta^* = 0.5$ with the $\kappa_1^* = \kappa_2^* = 0$. Note that this scenario only exists when $\kappa/(mz^*) \leq 0.5$.*

Corollary 2.2.1 suggests that if the clinic's capacity is relatively scarce as compared to type-2 demand, and the clinic needs to choose an allocation of type-2 patients such that $\kappa/(mz^*) \leq \beta \leq 1 - \kappa/(mz^*)$, then the clinic will assign equal amount of type-2 patients to the two panels. This result comes from the analysis of case 1. A proof of Corollary 2.2.1 is included in Appendix A.1.

Corollary 2.2.2 *If* $\beta > \kappa/(mz^*)$ *and* $(1-\beta) \leq \kappa/(mz^*)$, *then* $\kappa_1^* = 0$, $\kappa_2^* = \kappa - (1-\beta)mz^*$, $\alpha^* = 0$ *and* $\beta^* = \max\{\kappa/(mz^*), 1-\kappa/(mz^*)\}$.

This corollary is obtained upon analyzing case 2. A proof can be found in Appendix A.2.

Corollary 2.2.3 *If* $\beta \leq \kappa/(mz^*)$ *and* $(1-\beta) \leq \kappa/(mz^*)$, *then* $\kappa_1^* = \kappa - \beta mz^*$ *and* $\kappa_2 = \kappa - (1-\beta)mz^*$ *and* $\alpha^* = 0.5$ *and* $\beta^* = 0.5$ *is a local optimal solution.*

The derivation of Corollary 2.2.3 is included in Appendix A.3.

Corollary 2.2.4 *If* $\kappa/(mz^*) = 1$, *then* $\alpha^* = 0$, $\beta^* = 1$, $\kappa_1 = 0$, *and* $\kappa_2 = \kappa$.

Corollary 2.2.4 says that when a single physician's capacity is relatively adequate as compared to type-2 demand, then the clinic would prefer to specialize and let each physician only deal with one type of patients. Note that cases 1 and 3 are identical when $\kappa/(mz^*) = 1$. See Appendix A.4 for the Proof of Corollary 2.2.4.

Upon comparing the three types of patient allocation, the choice of the optimal solution can be evaluated according to different ranges of the value of $\kappa/(mz^*)$. When $0 < \kappa/(mz^*) \leq 0.5$, we only need to compare the performance of cases 1 and 2. When $0.5 < \kappa/(mz^*) \leq 1$, we only need to compare the performance of cases 2 and 3. When $\kappa/(mz^*) > 1$, only case 3 has a valid range for β. The results of the comparison are summarized below.

Table 2.1: Patient Allocation Strategy by $\kappa/(mz^*)$

	$0 < \kappa/(mz^*) \leq 0.5$	$0.5 < \kappa/(mz^*) \leq 1$	$\kappa/(mz^*) > 1$
Strategy:	(a)	(c) if $\kappa/(mz^*) \leq \kappa/(\tilde{m}z^*)$ (b) if $\kappa/(mz^*) > \kappa/(\tilde{m}z^*)$	(c)

(a): Arbitrary-Equal Assignment Strategy; (b) Specialization Strategy (i.e. a physician only serves one type of patients); (c) Equal Assignment Strategy.

As shown in Table 2.1, if $\kappa/(mz^*) > 1$, then the clinic will adopt the equal assignment strategy to ensure that each physician's panel composition is the same. When $\kappa/(mz^*) = 1$, the clinic will choose to specialize the physicians. When $0.5 < \frac{\kappa}{mz^*} < 1$,

then the clinic needs to choose between either equal assignment or some degree of specialization. In particular, for any random variable Y_2, there is a cut-off point $\tilde{m}$ such that when $\kappa/(mz^*) \leq \kappa/(\tilde{m}z^*)$ equal assignment is better (case 2), and when $\kappa/(mz^*) > \kappa/(\tilde{m}z^*)$, specialization is better. When $0 < \kappa/(mz^*) \leq 0.5$ then the clinic will choose an equal assignment for acute patients and an arbitrary assignment for non-acute patients. Proofs of results reported in Table 2.1 are included in Appendix A.4

2.3 Conclusion

From the above analysis, we draw the following conclusions. The clinic is better-off to have identical panels (i.e. equal number of patients for each type for each panel), when a single physician's capacity is abundant (i.e. $\kappa/(mz^*) > 1$) or inadequate (i.e. $0 < \kappa/(mz^*) \leq 0.5$) relative to the total demand for acute patients. When capacity is moderately adequate ($0.5 \leq \kappa/(mz^*) \leq 1$) relative to the acute demand, the optimal strategy depends on cost parameters, capacity, and panel sizes. Specialization is a dominant strategy only for scenarios where capacity is moderately adequate.

The analysis provided in this chapter has limitations. The assumption for demand distribution did not capture pooling effects because the coefficient of variation (c.v.) is constant regardless of the number of patients assigned to a panel (i.e. the c.v. of the demand of either type is independent of n, m, α or β). Therefore, the benefit of specialization may not be fully captured. However, the analysis in this chapter is able to show that equal assignment is not always a dominant strategy without accounting for the potential benefit of pooling.

The panel design problem discussed in this chapter may also be result for patient-centered medical home model of care. In a medical home environment, patients are taken care of by a team, which typically consists of a physician, a registered nurse, and/or a nursing aid, and a technician. In this environment, each medical team can be viewed as a single service provider. Then the clinics may also wish to evaluate using the model provided above whether it is beneficial to let each team take care of one type of patients (i.e. specialization), or whether it is beneficial to let each team take care of a heterogeneous panel of patients.

An extension of the panel design problem is to allow patients to book appointments with a non-PCP. Another research avenue is to evaluate the impact on the capacity allocation decision when there are strategic non-urgent patients who identify themselves as urgent and request for urgent appointments, or when some of the non-urgent patients are willing to stand by for same-day/walk-in appointments.

This chapter serves to highlight the importance of paying attention to setting a clinic's profile. The composition of a panel affects physicians' workload and the clinic's capacity utilization. This problem is more complicated when patients may book appointments with a non-PCP and have strong physician and time of day preferences. The clinic profile setup problem is not a frequent decision. However, appointment booking decisions occur on a daily basis. We discuss the appointment booking problem assuming a given clinic profile in the next chapter.

Chapter 3

Adaptive Appointment Systems with Patient Preferences

3.1 Introduction

An outpatient appointment is a contract between a patient and a clinic by which the latter reserves a certain amount of service providers' time and physical assets for the exclusive use of the patient who holds the appointment. Patients' satisfaction with their health care clinic is affected not only by the perceived quality of medical services that they receive during their visit, but also by their appointment booking experiences. Clinic managers care about having high scores on patient satisfaction surveys because that helps them attract new patients and negotiate better rates with insurers. Because the vast majority of medical appointments are booked with physicians working in primary care clinics, we focus in this chapter on the design of primary care appointment systems. A detailed description of the primary care service environment is provided in Gupta and Denton (2008). The ensuing abbreviated description focuses on features that are central to this chapter.

Patients that belong to a health system choose both a preferred clinic and a preferred physician. The latter is commonly referred to as the PCP (preferred care provider) for the patient. The term panel is used to describe a group of patients who have chosen the same PCP. Patients usually call in advance to book an appointment. Patients' satisfaction with an appointment system when they attempt to book a non-urgent appointment

is affected by their ability to book with their doctor of choice and at a convenient time of day (Cheraghi-Sohi et al. 2008, Gerard et al. 2008). Patients also prefer a sooner rather than a later appointment so long as it meets their time and physician preferences. For urgent medical conditions, patients want quick access to a physician. Clinics plan for such appointment requests and have open slots each day that allow same-day (urgent) access.

Because appointments are booked one at a time without knowledge of the number, sequence, and service requirements of future arrivals, many clinics use a two-step process to design appointment systems, which we call *clinic profile setup* and *appointment booking* steps, respectively. Clinic profile setup refers to the common practice of dividing physicians' available time on each work day into appointment slots. All slots need not be of the same length. For example, whereas a standard slot may be appropriate for the vast majority of routine appointments, physical exams and in-office procedures may require longer slots. In the appointment-booking (second) step, the clinic profile is known and the decision concerns which available appointment slot to book for each incoming appointment request. This chapter is concerned with the second step. That is, we assume that the number of appointments and the length of each appointment slot have been determined for each physician. Clinic profile setup may take into account a whole host of factors, including physicians' willingness to work overtime, no-show rates, service time variability, and demand for physicians' slots (see LaGanga and Lawrence 2007, Robinson and Chen 2003, Denton and Gupta 2003, Ho and Lau 1992, and Weiss 1990).

What makes the appointment booking problem (the focus of this chapter) difficult is that booking preferences are different for each patient, and they change over time for the same patient. For example, some patients are willing to see any available doctor if they can have an appointment sooner whereas others prefer to wait until a slot becomes available with their PCPs. Some patients are able to visit the clinic only within a short time window because of job-related constraints or personal schedules (Jennings et al. 2005, Olowokure et al. 2006), whereas others can be quite flexible. Finally, changes in work schedule, marital status, and family size can alter a patient's booking pattern.

There is evidence showing that clinics benefit by accommodating patients' preferences. First, matching patients with their PCPs ensures continuity (quality) of care

(Doescher et al. 2004), and allows physicians to provide more value-added services to their patients, which increases clinic's revenues (O'Hare and Corlett 2004). Second, matching patients with their PCPs and offering them a convenient appointment time can decrease the number of no-shows and thereby increase operational efficiency (Barron 1980, Carlson 2002, Smith and Yawn 1994). The above arguments provide the motivation for paying attention to patients' physician and time preferences and adapting appointment booking practices as these preferences change. The purpose of this article is to develop a framework for the design of such adaptive appointment systems. We use patient-PCP match rate, advance-book failure rate, and the total number of patients served as surrogate measures for patients' satisfaction with the appointment booking system.

We assume a known clinic profile, which may include overbooking, and develop algorithms for making appointment booking decisions to maximize clinic revenue. We model each panel as a different revenue class and allow the revenue from each appointment to depend on whether the appointment is with a patient's PCP. Patients have different acceptance probabilities for each physician and time-block combination, and each patient may have several acceptable combinations when (s)he attempts to book an appointment. We also model advance-book (non urgent) and same-day (urgent) demand. Inadequate capacity to serve urgent demand results in a higher cost to the health system. If a patient's service-time class can be ascertained at the time of booking an appointment, then such information can be incorporated in the proposed system by checking that the offered appointment slot is appropriate for the services requested. However, in numerical examples presented in this chapter, the availability of such information is not assumed.

Booking decisions do not depend on each patient's individual no-show probability because such probabilities are difficult to estimate from historical data. We comment on this issue in Section 4.3 based on an analysis of data from a large health system. Thus, our approach is suitable for health systems with low no-show rates. For the problem features mentioned above, we show that certain types of information that may be retrieved from existing web-based appointment request systems can be used to estimate patients' preferences, and improve booking decisions. Our approach may be viewed as an application of the Bayesian learning approach for directly estimating

empirical distributions of patient acceptance probabilities (e.g., see Carlin and Louis 2000). Our booking algorithm is a two-step process based on a partial characterization of the optimal booking decisions.

In the remainder of this section, we compare and contrast our approach with other approaches used to design appointment systems. A detailed review of relevant operations research (OR) literature can be found in Gupta and Wang (2008). Commonly used appointment systems can be categorized into four main types: (1) traditional systems that accept any booking request so long as the requested slot is open when the booking request is made, (2) carve-out systems that reserve a certain amount of capacity for specific procedures or urgent services, (3) Advanced Access (or Open Access) systems that accommodate patients' appointment requests on the day they call, and (4) hybrid approaches that accommodate both advance-book and same-day appointments. The traditional system allows each open slot to be booked by any patient who happens to be the first person to request it. This approach usually results in large backlogs of appointments for popular physicians as well as a significant spoilage of slots (Savin 2006). Same-day requests are often deflected to urgent care clinics, emergency rooms, or double booked. Because of these shortcomings, some clinics choose a carve-out approach in which a certain amount of capacity is reserved for later-arriving patients. Once available capacity drops to the reservation level, a variety of rules are used to release this capacity for specific procedures or urgent-need patients. The urgency of each patient's needs is determined by a triage nurse. Non-urgent patients generally cannot obtain same-day appointments (Murray and Berwick 2003).

An Advanced Access system is designed to offer each patient an appointment with his/her PCP on the day (s)he calls. In many cases, the implementation of an Advanced Access system allows patients to be seen sooner and improves clinics' operational efficiency (Murray and Tantau 2000). However, physicians are typically unable to cover all appointment requests that arise each day and push some demand to future days. In addition, some patients prefer to book appointments in advance, at a time and day of their choice, rather than call on the day they wish to see a doctor (Gerard et al. 2008, Parente et al. 2005, Salisbury et al. 2007). For reasons such as these, implementations of Advanced Access systems are not always successful (Murray et al. 2003).

Clinics that implement Advanced Access systems usually adopt hybrid approaches

that allow both advance and same-day bookings. Gupta and Wang (2008) provides a model of a hybrid approach in the presence of patients' preferences upon assuming knowledge of the conditional probability that a patient belonging to physician ℓ's panel, after calling in period t and observing the state of the appointment system s, will request an appointment for slot j of physician i, for each i, j, s, t, and ℓ. The study shows that the optimal policy for a single-physician clinic is a threshold-type policy so long as patient-choice probabilities satisfy a weak condition. The authors also partially characterize the structure of an optimal policy for multiple-doctor clinics. This work provides insights into the importance of modeling patients' choices in the primary care setting. However, patient-choice probabilities are not easily obtained from appointment records and patients generally do not have complete knowledge of the system state when requesting an appointment. We address both these issues in this chapter.

In Table 3.1, we compare the study in this chapter with some recent papers in the appointment scheduling (AS) literature in terms of (1) the objectives of the study, (2) patient classification scheme, (3) key model assumptions, and (4) performance criteria that drive parameter selection. Each major attribute is further divided into sub attributes, which we describe next. Study objectives may consist of one or more of clinic profile setup (1.a), booking decisions (1.b), learning/adaptive approach for improving booking decisions (1.c), and comparison of different system designs (1.d). Furthermore, clinic profile setup may be static or dynamic, and include one or more of the following decisions: number of of appointments per slot/session/day (1.a.i), appointment intervals/start times (1.a.ii), panel sizes (1.a.iii), and sequencing groups of appointments (1.a.iv). The decisions at the appointment booking stage include whether to accept a patients' request (1.b.v), which slot to book (1.b.vi), which appointment day to book (1.b.vii), whether to reserve capacity for same-day/urgent demand (1.b.viii), and sequencing of individual appointments (1.b.ix).

Patient classification may be based on revenue/costs (2.a), patient preferences (2.b), no-show rates (2.c), service time distribution (2.d), and same-day vs. advance-book requests (2.e). Classification typically helps improve capacity allocation decisions.

Key modeling assumptions concern no-show patterns (3.a), the decision stage at

which no shows affect AS design (3.a'), service time randomness (3.b), patients' punctuality (3.c), and patients' preferences (3.d). Patterns of no shows may be homogeneous (3.a.1), patient characteristics dependent (3.a.2), lead time dependent (3.a.3), and zero no shows (3.a.4). No shows may be modeled at the clinic profile setup stage (3.a'.I) and/or appointment booking stage (3.a'.II). Performance criteria used to select AS parameters are revenue/cost (4.a), patient-PCP match (4.b), booking failure rate/utilization (4.c), patients' wait (4.d), and physicians' idle/overtime (4.e).

Studies reported in Table 3.1, except Liu et al. (2010), focus on single session/day appointment problems. Liu et al. (2010) assumes that patients have no preference for a particular appointment day, and that the clinic decides which day to book after taking into account system state and lead time dependent no-show probabilities. In the proposed adaptive appointment system, advance-book patients first pick a desired appointment date. Booking decisions are made separately for each day and depend on the combinations of physician and appointment time blocks that are deemed acceptable by patients on the chosen date. It also reserves capacity for same-day requests. The proposed approach is novel because it learns (1.c) and utilizes patients' preference information (2.b) in the booking process, and because it prioritizes patient-PCP match (4.b).

Table 3.1: Literature Analysis.

A "√" (resp. "-") indicates that the corresponding attribute is included in (resp. absent from) the study.

Study	**1. Objectives** (a)	(b)	(c)(d)	**2. Class** (a)(b)(c)(d)(e)	**3. Assptn.** (a)(a')(b)(c)(d)	**4. Criteria** (a)(b)(c)(d)(e)
Adaptive Appt. System	-	v,vi,viii	√ √	√ √ - - √	1,2 I √ √ √	√ √ √ - -
Gupta and Wang (2008)	-	v,viii	- √	√ √ - - √	4 - √ √ √	√ - √ - -
Rohleder and Klassen (2000)	-	vi	- √	- - - √ -	4 - √ √ √	√ - - - -
Liu et al. (2010)	-	vii	- √	- - √ - -	3 II √ √ -	√ - - - -
Muthuraman and Lawley (2008)	i	vi	- -	- - √ - -	2 I √ √ -	√ - - √ √
Cayirli et al. (2008)	ii	ix	- √	- - - √ -	1 I √ - -	- - - √ √
Klassen and Rohleder (1996)	-	vi	- -	- - - √ √	4 - √ √ -	- - - √ √
Robinson and Chen (2010a)	i	-	- √	- - - - √	1 I - √ -	- - - √ √
Kim and Giachetti (2006)	i	-	- -	- - - - -	1 I - - -	√ - - - -
Denton and Gupta (2003)	i,ii	-	- -	√ - - - -	4 - √ √ -	√ - - √ √
Hassin and Mendel (2008)	ii	-	- √	- - - - -	1 I √ √ -	- - - √ √
LaGanga and Lawrence (2007)	i	-	- √	- - - - -	1 I √ √ -	- - - √ √
Kaandorp and Koole (2007)	ii	-	- -	- - - - -	1 I √ √ -	- - - √ √
Robinson and Chen (2003)	ii	-	- -	- - - - -	4 - √ √ -	- - - √ √
Weiss (1990)	ii,iv	-	- -	- - - - -	4 - √ √ -	- - - √ √
Green and Savin (2008)	iii	-	- -	- - - - √	3 I - - -	- - √ √ √
Vanden Bosch and Dietz (2000)	ii,iv	-	- √	- - √ √ -	4 - √ √ -	- - - √ √
Wang (1999)	ii,iv	-	- √	- - - √ -	4 - √ √ -	- - - √ √

We further summarize our model features based on the above scheme in Table 3.2.

Table 3.2: Summary of Literature Analysis.

Objectives	**Patient Classification**
(a) Clinic profile setup	(a) Revenue/costs
(b) Booking decisions	(b) Patient preference ⋆
(c) Adaptive approach for improving booking decisions ⋆	(c) No-show rates
(d) Comparison of different designs	(d) Service time distribution
	(e) Sam-day vs. advance-book requests
Model Assumption	**Performance Criteria**
(a) No-Shows	(a) Revenue/cost
(b) Service time randomness	(b) Patient-PCP match ⋆
(c) Patients' punctuality	(c) Booking failure rate/utilization
(d) Patients' preferences ⋆	(d) Patients' wait
	(e) Physicians' idle/overtime

Underlined items are considered in this study. New features are marked with ⋆.

Because our approach considers patients' preferences and learning, discrete choice models such as probit or logit models that have been studied extensively in economics, marketing and OR literatures are also relevant. These methods usually derive choice probabilities from the assumed utility-maximizing behavior of individual decision makers. Each decision maker, upon receiving an offer of a choice set, selects one of the alternatives in the set. The individual choices are then aggregated to obtain group-level measures of choice, e.g. the probability that an arbitrary member of the group will choose a particular option in the choice set. McFadden (2001) and Train (2003) present extensive surveys of discrete choice models and Talluri and van Ryzin (2004) is an example of models involving customer-choice in revenue management. The contrast between revenue management studies and our approach can be explained in terms of the ownership of the choice set and booking decisions. In the former, the choice set is determined by the service provider and customers decide which product to purchase, whereas in our framework each patient (customer) reveals an acceptable set of slots and the clinic (service provider) decides which slot to book.

The remainder of this chapter is organized as follows. In Section 4.3, we present empirical evidence that supports the proposed model. Model formulation is presented in Section 4.5. Then, we analyze properties of optimal booking decisions and present two heuristics to help clinics make real-time booking decisions in Section 4.6. Section 3.5 contains an evaluation of the impact of patients' preferences on different performance metrics, including those that are affected by no-shows and service time variability. Section 3.6 concludes the chapter.

3.2 Analysis of a Health System's Appointment Data

We studied appointment processes of a large health system and obtained historical appointment data concerning 37 primary care clinics that operate in urban, suburban and rural areas. We analyzed this data to guide the choice of model features in Section 4.5. The data covered appointment times with a range of 13 months that were booked over 18 months. It contained 1,461,948 records pertaining to 377,284 patients. The data elements were blinded medical record number (MRN), date and time of call and appointment, blinded PCP ID and provider ID (provider was the doctor that actually saw the patient for that appointment), age category, insurance status, 5-digit zip code for each patient's address on file, and clinic location. Patient ages were divided into 5-year intervals to obtain age categories.

The data reveals that both the panel size and its age distribution are different for each physician. Although we did not have access to revenue data, publicly available data support a strong correlation between patients' age and the different types and costs of services they need (U.S. Bureau of Labor Statistics 2008). This implies that both the demand and the expected revenue generated by patients of different panels are different. To make our point, we show distributions of patients' ages, loyalty (determined by the proportion of patient-PCP matched visits among that patient's past visits in 10% increments) for three physicians' panels in our data set in Figures 3.1(a) – 3.1(c). Chi-square tests showed that the distributions of age, loyalty, and time preferences were significantly different for different panels (p-values were < 0.0005 in each case). Moreover, the number of unique MRNs within the 13-month data for the three panels were 495, 719, and 1631 respectively, which suggests that panel sizes also differ by

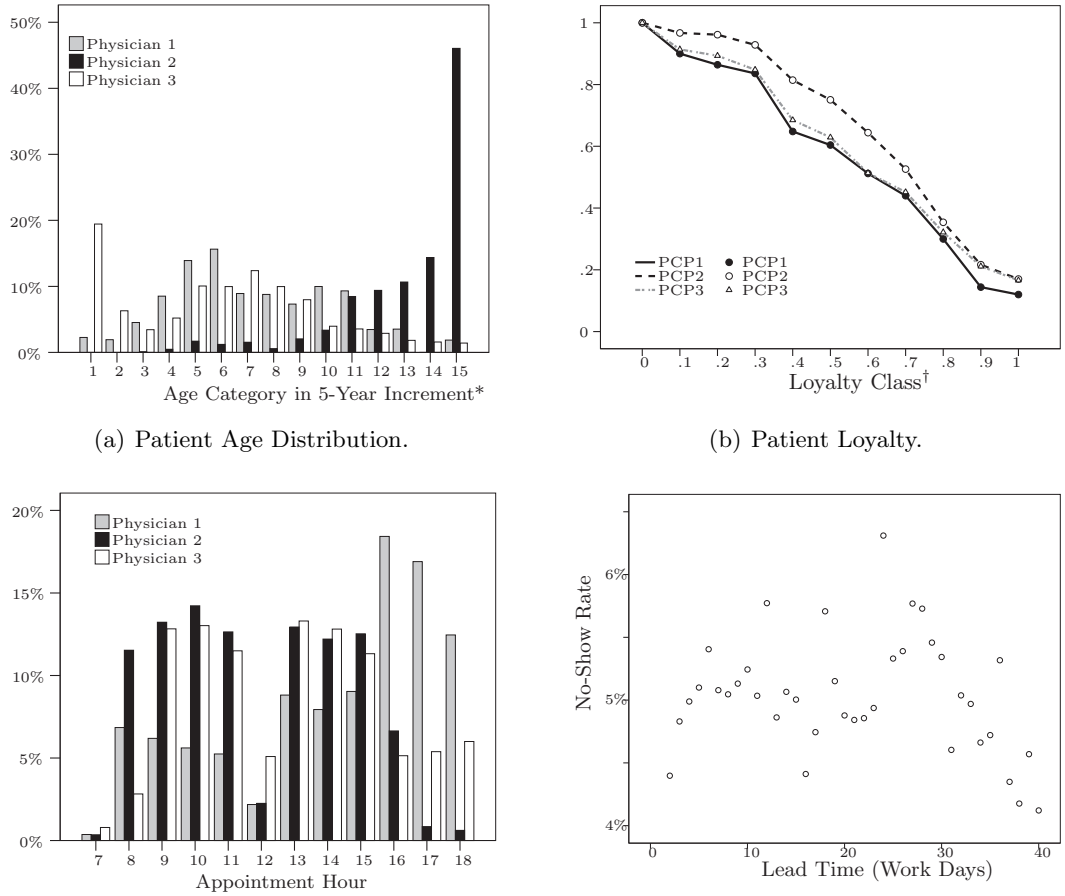
physician.

We recognize that realized appointment times may not reflect true time preferences because booking success is also affected by the availability of requested slots. For example, it is possible that Physician 2 rarely works after 4 PM, and that patients in his/her panel have adapted by accepting morning appointments. However, it is also possible that service providers respond to patients' needs. For example, families with teenagers and young adults often prefer appointment times after school hours, so as not to disrupt school attendance. Physician 1 may have chosen his/her work pattern with more availability in the afternoon in response to such demand. Irrespective of the underlying root causes, Figures 3.1(a) – 3.1(c) serve to highlight the fact that panels provide a reasonable means by which to define revenue classes and aggregate patients' preferences.

Next, we investigate the ability to predict patient-specific no-show probabilities from a data set such as ours. We first excluded canceled appointments from our data because the vast majority of the slots freed up in this fashion are rebooked. This resulted in a 1,171,950 encounters. Two factors that have been identified in previous studies are (1) history of no-shows, and (2) appointment lead time (i.e. the time between the appointment request and the appointment date). It has been suggested that patients with a history of no-shows are more likely to be a no-show and that longer appointment lead times increase the likelihood of no-shows; see Dove and Schneider (1981), Lee et al. (2005), Gallucci et al. (2005), and Whittle et al. (2008). Figure 3.1(d) shows that appointment delays are not significantly correlated with no-show rates in our data (Pearson correlation test shows no significant correlation with p-value > 0.4). A similar conclusion is also reached in Snow et al. (2009), Starkenburg et al. (1988), Irwin et al. (1981), Fosarelli et al. (1985), Neinstein (1982), and Dervin et al. (1978).

Turning to the history of no-shows, our data contained appointment times that ranged over 13 months. Therefore, we normalized the number of appointments per patient to a yearly basis and found that more than 75% of the patients in our data had fewer than 4 appointments per year, which would make it difficult to estimate individuals' no-show probabilities reliably. We believe such estimation problems could arise in many practical settings.

Finally, the overall no-show rate for the 37 clinics is 4.06% for all appointment and



(a) Patient Age Distribution.

(b) Patient Loyalty.

(c) Appointment Times.

(d) Effect of lead time on no-show rate.

Figure 3.1: Evidence from the Analysis of Data from 37 Clinics.

*In Figure 3.1(a), Group 15 includes all patients who are 70 years of age or older. †Figure 3.1(b) shows the proportion of panel patients that belonged to a higher loyalty class for patients with more than 3 encounters.

2.97% for patient-PCP matched appointments. The overall patient-PCP match rate was 45.7%. This implies that there may be a substantial opportunity to reduce no-show rates further by increasing patient-PCP match rates, which the adaptive appointment system is designed to do.

3.3 Model Formulation and Assumptions

The vast majority of large health systems operate call centers where patients call to book appointments. With the adoption of EMR (Electronic Medical Record) systems, however, many health systems are also able to provide a parallel web-based option to patients for requesting non-urgent appointments. Patients are instructed to call if their needs are urgent. Similar instructions may also apply for special appointments such as physical exams and in-office procedures that take more time and for which physicians reserve specific slots in their daily schedule. It is generally believed that web-based systems will become the primary means by which patients book non-urgent appointments in the future. Therefore, our model assumes the availability of real-time data from a web-based system. We illustrate the types of information that can be obtained from existing web interfaces in a mock-up in Figure 3.2. This mock-up is fashioned after existing systems familiar to the authors. However, it is not an exact replica of any particular system.

In the mock-up, a patient indicates a preferred appointment date and acceptable combinations of physicians and time blocks. Clinics use time blocks rather than individual time slots because patients tend to have similar acceptance rates for time slots within each half-hour or one-hour time block. Note, our formulation allows clinics to choose arbitrary block size and number of slots in each block. That is, appointment lengths may not be uniform and may depend on anticipated service-time class and no-show rates. Upon receiving a patient's request, the clinic considers any checked combination of the blocks of time and physicians to be acceptable to the patient on the chosen day of appointment request. The clinic either books an appointment in one of the combinations indicated by the patient, or responds that none of the requested combinations are available. Patients are encouraged to try a different date if their request is denied.

To increase clinics' flexibility in scheduling patients in a manner that maximizes patient-PCP match and revenue, patients are asked to provide their acceptable sets, but not rank their preferences among the acceptable combinations. If patients were asked to rank their preferences, clinics would be obligated to book appointments in the most preferred and available slots first, which would prevent them from keeping more capacity available in more popular combinations.

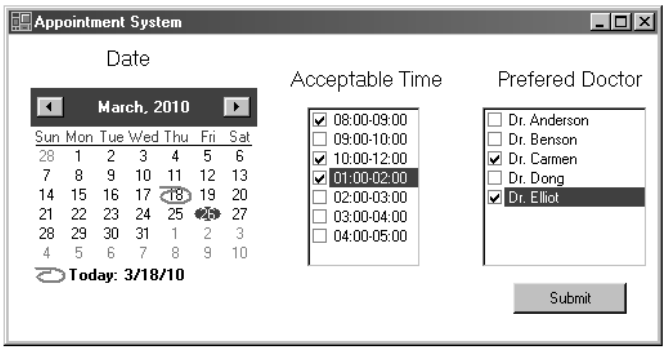


Figure 3.2: A Web-Based Patient-Clinic Interface

The proposed adaptive appointment system has two components— a component that updates estimates of acceptance probabilities, and a component that makes booking decisions after receiving patients' requests. Below, we describe each component in a separate section. Each section states model assumptions first and then presents a formulation. We show in Section 3.3.2 that for making booking decisions, clinics only need to estimate panel-level acceptance probabilities. Therefore, Section 3.3.1 deals only with panel-level probabilities. Throughout the chapter, we use m to denote the number of physicians and b to denote the number of time blocks on a work day.

3.3.1 Learning Acceptance Probabilities

Given that web-based options similar to that in Figure 3.2 are in existence today, our approach models each patient's preferences in terms of acceptance probabilities. For each physician indexed i and time-block indexed j, the probability that the k-th patient in physician ℓ's panel will find combination (i, j) acceptable is denoted by $p_{i,j}^{\ell,k}$. Furthermore, we assume that physician and time preferences are independently captured by

terms $\alpha_i^{\ell,k}$ and $\beta_j^{\ell,k}$, with $p_{i,j}^{\ell,k} = \alpha_i^{\ell,k}\beta_j^{\ell,k}$. This is consistent with the implied decomposition of physician and time preferences in Figure 3.2. From a technical viewpoint, it is possible to generalize our approach to situations where acceptance probabilities do not have the multiplicative form that we assume. However, we did not find any evidence to suggest that the multiplicative form is an unreasonable assumption.

We assume that each patient reveals his/her true acceptable set in each request (prior to receiving an appointment), and that each booking attempt is an independent draw from a patient's preference distribution. The first assumption is based on the argument that if a patient's utility from booking an appointment for a particular physician and time-block combination is higher than the utility from not booking an appointment, then the patient will include that combination in his/her acceptable set. The second assumption is based on anecdotal evidence that patients' time preferences vary by calendar day.

As shown in Section 3.3.2, booking decisions depend only on panel-level acceptance probabilities $p_{i,j}^{\ell} = \alpha_i^{\ell}\beta_j^{\ell}$, where α_i^{ℓ} and β_j^{ℓ} are the physician-i and block-j acceptance probabilities for panel-ℓ. We propose direct estimation of these probabilities; see Section B.1 of Appendix B for details. This is not the only way to estimate patients' choices. A parallel utility-based model can be constructed as well, and subsequently its parameters can be estimated. It can be shown that the strong independence of attributes assumed by clinics (as implied by Figure 3.2) leads to an equivalent model with similar estimation effort. We provide details of this alternative approach and compare it to the proposed approach in Section B.2 of Appendix B .

3.3.2 Making Appointment Booking Decisions

At the time of booking appointments, clinic profiles exist for all future work days on which appointments may be booked. The model that is used to obtain a partial characterization of optimal booking decisions also assumes that patients and physicians are punctual, patient no-show rates are negligibly small, and all same-day patients call before the start of the day. The existence of clinics with relatively low no-show rates has been documented in the literature (see, e.g. Cayirli and Veral 2003) and supported by our data (see Section 4.3). However, some clinics are also reported to have high no-show rates and the proposed adaptive appointment system design may not be the best choice

for such clinics. Clinics typically count all requests received within 24 hours before the start of a work day as same-day demand. This makes it reasonable to assume that same-day demand is realized just before the start of each work day.

Our model considers each workday's appointment booking problem separately. This is justified, in part, by assumptions in Section 3.3.1 that patients' preferences may differ by calendar day, and that patients are encouraged to try other dates if earlier booking attempts fail. Because clinic profile is assumed known, the clinic's objective function considered in this section does not include patient wait times and physician overtime, which are caused by service time variability, and choices of appointment lengths and overbooking. However, when evaluating different heuristics in Section 3.5, we also compare these metrics for different approaches.

The following information is needed to make booking decisions: (1) patients' acceptance probabilities and arrival rates at the panel level, (2) clinic's average revenue, by panel, of each PCP matched/mismatched appointment, (3) average costs of delaying an advance-book and same-day appointment, and (4) each physician's same-day demand distribution. We define the inputs to the booking decision model in Table 3.3 and explain model features below.

Table 3.3: Inputs of the Booking Decision Model

X_ℓ:	same-day demand for physician-ℓ
X:	total same-day demand; $X = \sum_{i=1}^{m} X_i$
$\kappa = (\kappa_{ij})$:	matrix of capacity of each combination (i, j) of physician and time-block combination
$s = (s_{ij})$:	matrix of number of appointments that have been booked for (i, j) combination
$\bar{\kappa}_i$:	physician i's capacity; $\bar{\kappa}_i = \sum_{j=1}^{b} \kappa_{ij}$
$\bar{\bar{\kappa}}$:	clinic's capacity; $\bar{\bar{\kappa}} = \sum_{i=1}^{m} \bar{\kappa}_i$
$\bar{s}_i$:	physician i's booked appointments; $\bar{s}_i = \sum_{j=1}^{b} s_{ij}$
$\bar{\bar{s}}$:	number of booked appointments at the clinic level; $\bar{\bar{s}} = \sum_{i=1}^{m} \bar{s}_i$
$r^{\ell}_{1,\ell}$:	average revenue of a PCP matched advance-book panel-ℓ appointment
$r^{i}_{1,\ell}$:	average revenue of a PCP mismatched advance-book panel-ℓ appointment, $i \neq \ell$
r_2:	average revenue of a PCP matched same-day appointment
r'_2:	average revenue of a PCP mismatched same-day appointment
π_t:	time-dependent average penalty induced by a failure to satisfy an advance-book request
c:	average cost of insufficient same-day capacity
τ:	number of potential advance-book appointment request epochs for a particular work day
t:	an arbitrary advance-book appointment request epoch
λ^{ℓ}_t:	the probability of having an arrival from physician ℓ's panel at epoch t
α^{ℓ}_I:	the probability that an arbitrary panel-ℓ patient's set of acceptable physicians is I
β^{ℓ}_J:	the probability that an arbitrary panel-ℓ patient's set of acceptable time blocks is J
$p^{\ell}_{I,J}$:	the probability that a panel-ℓ patient's acceptable combinations are (I, J); $p^{\ell}_{I,J} = \alpha^{\ell}_I \, \beta^{\ell}_J$

In reality, patients' true acceptance probabilities are unknown. Therefore, we propose to use the best available estimates of acceptance probabilities at each decision epoch (from the updating procedure of Section 3.3.1). This can be justified by the fact that the updating procedure is independent of booking decisions and converges quickly to the true acceptance probabilities. Unit revenues from each booked appointment satisfy the following inequalities: $r^{\ell}_{1,\ell} \geq r^{i}_{1,\ell}$ for each ℓ and $i \neq \ell$, and $r_2 \geq r'_2$; see O'Hare and Corlett (2004) for supporting evidence. Same-day visit revenue does not depend on panel index because these appointments are usually for acute symptoms for which the treatments offered are more likely to be independent of panel characteristics.

The time between the start of advance-book requests for each work day and that work day is divided into τ intervals such that the probability of obtaining more than one arrival in each interval is infinitesimally small. Time is counted backwards. Specifically, advance bookings occur from period τ to period 1 and all same-day bookings occur in period 0. Because patients who try to book appointments must have at least one acceptable combination, neither I nor J is an empty set.

In our model, the penalty for denying a patient's appointment requests n time periods before the appointment date is assumed to be different (smaller) than the penalty for doing so $(n+d)$ period before the appointment date, where $d \geq 1$. This makes sense for two reasons. First, patients who call well in advance are generally more particular about their time and physician preferences. The clinic harmonizes its booking practices with this behavior by setting $\pi_n \leq \pi_{n+d},\ \forall\ d \geq 1$. Second, this assumption leads to a fair allocation of slots as we shall show in Section 3.4.1. In particular, this means that if a physician-ℓ patient's request for a particular combination is denied in period $(n+d)$ given a particular system state, then another physician-ℓ patient's request for that combination will be denied in period n as well for the same system state.

Researchers have made a variety of attempts in recent years to estimate the cost of patient waiting (terms c and π_t in our models. For example, Yabroff et al. (2005) and Russell (2009) estimate the cost of patient waiting based on wage rates whereas Robinson and Chen (2010b) provide an observation-based method for estimating the relative cost of customer waiting time. Clinic administrators can either perform a study similar to those reported in the literature, or use the results in these articles to guide their choice of patient waiting costs.

We are now ready to set up the clinic's revenue function for the appointment booking problem. For this purpose, it helps to conceptualize the availability of different levels of information about the arriving patient. Specifically, we identify three levels of information and label them (1) the patient-level, (2) the panel-level, and (3) the clinic-level information. At the patient-level, known information includes the patient label (ℓ, k) (i.e. the fact that the arrival in period t is the k-th patient in physician ℓ's panel), the system state s, and the patient's acceptable set (I, J). In contrast, panel-level information consists of the arriving patient's panel index and the system state, whereas the clinic-level information includes only the system state.

We use notation $u_t^{\ell,k}(s)$, $u_t^{\ell}(s)$ and $u_t(s)$ to denote the maximum expected revenue from time t onwards given patient-level, panel-level, and clinic-level information, respectively. With this notation in hand, the following recursive relationship holds.

$$u_t^{\ell,k}(s) = \max_{(i,j)\in(I,J)} \left\{ r_{1,\ell}^i + u_{t-1}(s + e_{i,j}), u_{t-1}(s) - \pi_t \right\}, \tag{3.1}$$

where $e_{i,j}$ is an $m \times b$ matrix with the (i, j)-th entry equal to 1 and all other entries equal to 0. The first two terms in the curly brackets above capture the benefit of accepting the patient's request for the (i, j) physician and time-block combination, whereas the next two terms capture the benefit of rejecting the patient's requests. Equation (4.2) suggests that the clinic should accept a slot, say (i^{opt}, j^{opt}), among the arriving patient's requests (I, J) for which $r_{1,\ell}^i + u_{t-1}(s + e_{i,j}) \geq u_{t-1}(s) - \pi_t$ and the clinic's revenue is maximized. That is, $(i^{opt}, j^{opt}) \in \arg\max_{(i,j)\in(I,J)} \left\{ r_{1,\ell}^i + u_{t-1}(s + e_{ij}) : \quad r_{1,\ell}^i + u_{t-1}(s + e_{i,j}) \geq u_{t-1}(s) - \pi_t] \right\}$. Ties may be broken arbitrarily.

Using a logic similar to that behind equation (4.2), revenue functions with panel- and clinic-level information can be written as follows.

$$u_t^{\ell}(s) = \sum_{\text{all } (I,\, J)} p_{I,J}^{\ell} \max_{(i,j)\in(I,J)} \left\{ r_{1,\ell}^i + u_{t-1}(s + e_{i,j}),\ u_{t-1}(s) - \pi_t \right\}. \tag{3.2}$$

$$u_t(s) = \sum_{\ell=1}^{m} \lambda_t^{\ell} u_t^{\ell}(s) + (1 - \sum_{\ell=1}^{m} \lambda_t^{\ell}) u_{t-1}(s). \tag{3.3}$$

Upon comparing (4.2) with (3.2), we observe that the decision rule for accepting or denying a particular (i, j) request is the same for all patients in the same panel. This comes

from the fact that the arriving patient's information does not affect clinic's valuation of its benefit from saving each combination for future arrivals. Similarly, upon comparing (3.2) and (3.3), we observe that the revenue function with clinic-level information is a weighted sum of revenue functions with panel-level information.

Next, we turn to the revenue function corresponding to same-day requests, which has a different form because all same-day requests are assumed to arrive just before the start of the work day. In the model, we assume that we can optimally match them with available capacity. Therefore, it suffices to define the same-day revenue function with clinic-level information only, as shown below.

$$\begin{aligned} u_0(s) &= E\big\{r_2 \sum_{i=1}^{m} \min\{X_i, (\bar{\kappa}_i - \bar{s}_i)\} + r_2' \min\{\sum_{i=1}^{m} (\bar{\kappa}_i - \bar{s}_i - X_i)^+, \sum_{i=1}^{m} (X_i - \bar{\kappa}_i + \bar{s}_i)^+\} \\ &\quad -c(\sum_{i=1}^{m} X_i - \sum_{i=1}^{m} (\bar{\kappa}_i - \bar{s}_i))^+\big\}. \end{aligned} \tag{3.4}$$

In equation (3.4), the first term is the expected revenue from same-day patient-PCP matched visits, the second term is the expected revenue from mismatched visits, and the third term is the expected cost due to excess same-day demand.

3.4 Analysis

The formulation of the appointment booking decision problem in Section 3.3.2 has a high-dimensional state space, which precludes the use of real-time and stored solutions of the stochastic dynamic program for every system state in each period. In what follows, we show with the help of an example that there also does not exist a pattern or structure to booking decisions.

Consider a single physician clinic with 4 slots each in 4 time blocks. We omit the physician label for simplicity. The panel-level acceptance probabilities for these blocks are $\beta = (0.1, 0.2, 0.6, 1)$. Other parameters are $(r_{1,1}^1, r_2, c, \pi_t, \lambda, \mu, \tau) = (6, 6, 10, 5, 0.7, 5, 16)$, where $\pi_t = \pi$ and $\lambda_t = \lambda$ for each $t = 1, \cdots, \tau$, and μ is the arrival rate for the same-day demand, which is assumed to be Poisson distributed. The expected total demand is 16.2 whereas the total capacity is 16. Because this problem has a small state space, we are able to solve the underlying stochastic dynamic program to obtain an ordering of slots from the clinic's perspective for each system state and decision epoch. If the optimal

decision is to deny the request for time block j in every decision epoch at and after time t, then we say the system is in a no-book (NB) state for block j. In Table 3.4, the best slot to book for an arriving patient is the highest ranked available slot that is acceptable to the patient and that is not designated NB.

We use 2 cases, each with 3 examples, to illustrate how an optimal decision may depend on the remaining capacity, time preferences of future arrivals, and the acceptable set of the next appointment request (see Table 3.4). In the first example, $t = 13$ and the total remaining capacity is 13. For Case 1 (state $s = (3, 0, 0, 0)$), the clinic's first choice is to book either block 1, 2, or 3, and the second choice is to book block 4. For Case 2 (state $s = (0, 0, 0, 3)$), the rank order of available time blocks is as follows: 1 $\succ$ 2 $\succ$ 3 $\succ$ 4. That is, a patient whose acceptable set includes blocks 1 and 3 may be booked into either block 1 or 3 in Case 1, but only in block 1 in Case 2. In the second example, when $t = 8$ and the total remaining capacity is 7, block 3 (resp. block 4) is a NB block if $s = (3, 2, 0, 4)$ (resp. $s = (4, 2, 0, 3)$) and a patient whose acceptable set includes blocks 1 and 3 will be booked into block 1 in Case 1 and block 3 in Case 2. In the third example, $t = 3$, the total remaining capacity is 6, and the clinic is in a no-book state for all blocks for both cases.

Table 3.4: An Ordering of Blocks from the Clinic's Perspective.

Case 1					
t	s	block 1	block 2	block 3	block 4
13	(3,0,0,0)	1	1	1	2
8	(3,2,0,4)	1	1	NB	–
3	(3,3,3,1)	NB	NB	NB	NB
Case 2					
t	s	block 1	block 2	block 3	block 4
13	(0,0,0,3)	1	2	3	4
8	(4,2,0,3)	–	1	1	NB
3	(1,3,3,3)	NB	NB	NB	NB

These examples show that the optimal decision depends in a non-trivial fashion on the vector of remaining capacities, the index of the decision epoch, and acceptable sets. In addition, certain blocks are designated NB, which means that they are reserved for

future same-day demand. The complexity of decisions increases when there are multiple physicians. Therefore in the next section, we characterize certain properties of optimal booking decisions, which are subsequently used to construct heuristic solutions.

3.4.1 Properties of Optimal Booking Decisions

For modeling convenience, we may think of the booking decision as a two-step processes. Given that a panel-ℓ patient makes a booking request in period t with acceptable set (I, J), the clinic in the first step identifies sets of no-book states $S_t^{i,\ell}$ for each $i \in I$, i.e. states in which a panel-ℓ patient's request for an appointment with physician i is denied irrespective of J. If the current state is in the set of no-book states for all physicians in I, then the requesting patient is asked to try another date. However, if the process proceeds to the second step, then the clinic decides which of the acceptable and available appointments to book. That is, in stage two, the clinic ranks available (i, j) combinations in (I, J). It is also possible at this stage to deny a patient's request. Denials may happen either when the intersection set of patients' requested appointments and available appointments is empty or when the clinic earns a greater expected revenue by protecting specific appointments requested by the patient for future arrivals. This two-step process can be operationalized by developing procedures for identifying no-book states, and procedures for rank ordering requested appointments (from clinic's viewpoint) when the system state is not in the no-book set. We obtain partial solutions for these two tasks in Sections 3.4.1 and 3.4.1, which form the basis for the heuristics proposed in Section 3.4.2.

No-Book States

In this section, we obtain $S_t^{i,\ell}$ for $t = 1$, and for $t > 1$ we identify a set of states $\hat{S}_t^{i,\ell}$ such that $\hat{S}_t^{i,\ell} \subseteq S_t^{i,\ell}$. We also show that for $t > 2$, when $\pi_t \geq \pi_{t-1}$ (which we assume), $\hat{S}_t^{i,\ell} \subseteq \hat{S}_{t-1}^{i,\ell}$. That is, patients who call earlier encounter smaller sets of no-book states.

Consider a time-t decision epoch when the k-th panel-ℓ patient makes a booking request, and assume that there will be no more future advance-book requests after this decision epoch. Let (I, J) denote this patient's acceptable set of appointments. Then,

the clinic's decision problem is encapsulated in the following revenue function.

$$u_t^{\ell,k}(s) = \max_{(i,j)\in(I,J)} \{r_{1,\ell}^i + u_0(s+e_{i,j}), u_0(s) - \pi_t\}.$$

The above revenue function is identical to (4.2) when $t = 1$. For $t > 1$, the difference is that the right hand side contains u_0 instead of u_{t-1} because we assume no advance-book arrivals after period t. The clinic should consider booking an appointment for a panel-ℓ patient if there is at least one (i, j) combination such that $u_0(s)-u_0(s+e_{i,j}) \leq r_{1,\ell}^i+\pi_t$.

Let $F_i(\cdot)$ and $F(\cdot)$ denote the CDF of physician i's and clinic's same-day demand, respectively. Upon rearranging the terms in Equation (3.4), we obtain $u_0(s) = r_2E(X)-r_2\sum_{\ell=1}^m E(X_\ell-\bar{\kappa}_\ell+\bar{s}_\ell)^+ + r_2'\sum_{\ell=1}^m E(\bar{\kappa}_\ell-\bar{s}_\ell-X_\ell)^+ - r_2'E(\bar{\bar{\kappa}}-\bar{\bar{s}}-X)^+ - cE(X-\bar{\bar{\kappa}}+\bar{\bar{s}})^+$. Let $\bar{s}_{-i}$ be the total number of slots booked for all physicians except physician i. After a few more steps of algebra, the marginal benefit for reserving a physician-i's slot in the last period can be further simplified to

$$\begin{aligned}\Delta(\bar{s}_i, \bar{s}_{-i}) &\doteq u_0(s) - u_0(s+e_{i,j}) \\ &= r_2 + c - (r_2 - r_2')F_i(\bar{\kappa}_i - \bar{s}_i - 1) - (r_2' + c)F(\bar{\bar{\kappa}} - \bar{s}_i - \bar{s}_{-i} - 1).\end{aligned}$$

Same-day patients do not have time preferences. Therefore, the value of $\Delta(\bar{s}_i, \bar{s}_{-i})$ does not depend on which block j is being considered.

Let $\bar{a}_i$ and $\bar{a}_{-i}$ respectively be the number of available slots of physician i and the clinic not including i. Because CDF is a non-decreasing function, for any fixed value of $\bar{s}_{-i}$, $\Delta(\bar{s}_i, \bar{s}_{-i})$ increases in $\bar{s}_i$. Therefore, there exists a protection level $a_i^\ell(s) = \min\{\bar{a}_i : \Delta(\bar{\kappa}_i - \bar{a}_i, \bar{s}_{-i}) > r_{1,\ell}^i + \pi_t \text{ given } \bar{a}_i \geq 0 \text{ and fixed } \bar{s}_{-i}\}$ such that no physician-i slot should be booked for a panel-ℓ patient if $\bar{k}_i - \bar{s}_i$ is less than $a_i^\ell(s)$. Similarly, for any fixed value of $\bar{s}_i$, $\Delta(\bar{s}_i, \bar{s}_{-i})$ increases in $\bar{s}_{-i}$, which implies that there exists a protection level $a_{-i}^\ell(s) = \min\{\bar{a}_{-i} : \Delta(\bar{s}_i, \bar{\kappa}_{-i} - \bar{a}_{-i}) > r_{1,\ell}^i + \pi_t \text{ given } \bar{a}_{-i} \geq 0 \text{ and fixed } \bar{s}_i\}$ such that no physician-i slot should be booked for a panel-ℓ patient if the remaining clinic capacity, not counting physician i, is less than $a_{-i}^\ell(s)$. Similar protection levels also exist with convex cost of unmet same-day demand (see Section B.3 of Appendix B for details).

Proposition 3.4.1 *Given a panel-ℓ patient's booking request for an appointment with physician i at decision-epoch t and no more advance-book requests after t, the set of no-book states is $\hat{S}_t^{i,\ell} = \{s : \bar{\kappa}_i - \bar{s}_i \leq a_i^\ell(s)\}$.*

An immediate corollary of Proposition 3.4.1 is that $S_1^{i,\ell} = \hat{S}_1^{i,\ell}$ for each (i, ℓ) pair because after $t = 1$, there are indeed no more advance-book requests. Also, the booking decision for a type-ℓ arrival regarding a physician-i's slot depends on the current state of the clinic only through $a_i^{\ell}(s)$ and $a_{-i}^{\ell}(s)$, which leads to a two-dimensional booking profile. Gupta and Wang (2008) obtain a similar result when advance-book revenue is independent of panel index. However, in their paper, all open slots of a physician are equally valued and are made available to the arriving patient so long as the remaining capacity is higher than the protection level. In our framework the protection level serves only as an availability check in the first-step of the booking process. We refer the reader to Section B.4 of Appendix Bfor an example that identifies no-book states for a two-physician clinic.

Proposition 3.4.2 *The set of no-book states assuming no more advance-book requests is a subset of the true set of no-book states, i.e.* $\hat{S}_t^{i,\ell} \subseteq S_t^{i,\ell}$*, and if* π_t *is non-decreasing in* t*, then* $\hat{S}_t^{i,\ell} \subseteq \hat{S}_{t-1}^{i,\ell}$.

A formal proof of Proposition 3.4.2 is included in Section B.5 of Appendix B. On an intuitive level, the first part of this proposition holds because when there are no more advance-book requests, there are no competing advance-book requests for the same slot. The only demand for a slot is from same-day requests. Therefore, the protection level after making the assumption of no more advance-book requests is never greater than the true protection level when advance-book requests do occur. The second result follows from the fact that higher cost of denying a patient's request leads to lower protection levels.

Rank Order of Appointment Slots

Consider a single-physician clinic with block-j capacity κ_j and state s_j. In this section, a block is deemed available when $s_j < \kappa_j$ and the current state s is not in the set of no-book states. We analyze this simpler problem instance because in this case an advance-book patient's request is denied only when it is optimal to reserve capacity for same-day patients. This happens because each advance-book appointment results in the same revenue. This means that when there is a single physician labeled ℓ, $\hat{S}_t^{\ell} = S_t^{\ell}$ for each t. A formal argument is provided in Section B.6 of Appendix B.

The clinic faces the problem of deciding which of the requested appointments in the acceptable set J to book. We consider only those instances in which for at least one $j \in J$, $s_j < \kappa_j$. If there is at least one block $j \in J$ such that $\kappa_j - s_j > \tau - t$ and state s is not a no-book state, then it is straightforward to show that the clinic can book the patient in block j without affecting its ability to book future patients because all those patients still have a chance to book block j. Similarly, if the system is not in a no-book state and there is only one $j \in J$ such that $\kappa_j - s_j > 0$, then a slot in block j should be booked. This means that a clinic needs guidance only when $\kappa_j - s_j \leq \tau - t$ for all j, and there is more than one acceptable block with remaining capacity. We focus on such cases in the remainder of this section.

Let $\phi(s) = 1 - \prod_{j:s_j<\kappa_j}(1-\beta_j)$ be the probability that at least one time block is acceptable to an arriving patient and has remaining capacity when system state is s and consider a decision epoch after which the clinic expects at most one additional advance-book arrival. Suppose that the patient's acceptable set includes blocks j and k, both of which have at least one open slot. The clinic may then base its decision on the value of $\phi(s)$. The higher the value of $\phi(s)$, the higher the chance of satisfying a future arrival's request. The clinic may consider the relative magnitudes of $\phi(s+e_j)$ and $\phi(s+e_k)$ when deciding which block to book. When both j and k have exactly one remaining slot, it may also consider the relative magnitudes of β_j and β_k. The above informal arguments are formalized in Proposition 3.4.3; a proof of Proposition 3.4.3 can be found in Section B.7 of Appendix B.

Proposition 3.4.3 *When choosing between blocks j and k, the clinic prefers to book in block j so long as $\beta_j < \beta_k$ and $\phi(s+e_j) > \phi(s+e_k)$. Mathematically, if $\beta_j < \beta_k$ and $\phi(s+e_j) > \phi(s+e_k)$, then $u_t(s+e_j) \geq u_t(s+e_k)$ for all $t \geq 1$.*

Proposition 3.4.3 gives a partial ordering of acceptable time blocks of a single physician. It suggests that among the available and acceptable combinations, a particular block is more likely to be a clinic's top choice if it has greater remaining capacity and if assigning a slot in that block has a smaller effect on the clinic's ability to meet future demand. It is difficult to show a similar result when multiple physician's slots are being compared because of different time-preference patterns of patients belonging to different panels and because of different revenue rates. However, we use the insights from

Proposition 3.4.3 to develop a metric, $q^t_{i,j}$, to rank order available and acceptable blocks from the clinic's viewpoint. This metric is used in heuristic rules for making booking decisions (see Section 3.4.2).

We define $q^t_{i,j}$ as a measure of popularity of each (i,j) combination when $\kappa_{i,j} - s_{i,j} > 0$ in period t as follows:

$$q^t_{i,j} = \sum_{z=1}^{t-1} \lambda^i_z p^i_{i,j} / (\kappa_{i,j} - s_{i,j}). \tag{3.5}$$

The numerator of (3.5) is the expected number of times that (i,j) combination will be included in the acceptable set by panel-i patients in the remaining advance-book periods, and the denominator is the remaining capacity of the (i,j) combination. The popularity measure does not account for anticipated demand from non-panel patients because both heuristics proposed in the next section give priority to achieving high patient-PCP match.

3.4.2 Heuristic Approaches

We present two heuristics (H1 and H2) that utilize the popularity index in (3.5), and give priority to matching patients with their PCPs. In describing the heuristics below, we assume that a panel-ℓ patient has tendered an appointment request with acceptable set (I,J) and that the system state is s. The booking decisions generated by H1 and H2 are appealing on an intuitive level for two reasons. First, because $r^\ell_{1,\ell} \geq r^i_{1,\ell}$ for $i \neq \ell$, and there are a variety of other benefits of matching patients with their PCPs, it is reasonable to strive for a high patient-PCP match. Second, because any combination in (I,J) is acceptable to the patient who tendered that request, it can be beneficial to reserve slots with higher $q^t_{i,j}$ values for future patients.

H1 books an appointment so long as the intersection set of open slots and (I,J) is not empty. That is, H1 assumes that the set of no-book states is empty. It attempts to first book a patient with his/her PCP. If multiple PCP slots are open and included in J, then H1 books a slot with the smallest value of $q^t_{\ell,j}$. If none of the acceptable PCP slots are available, then H1 books the slot with the smallest value of $q^t_{i,j}$, $i \neq \ell$, among all non-PCP slots in the acceptable set.

H2 calculates $\hat{S}^{i,\ell}_t$ and only considers physicians i included in I for which $s \notin \hat{S}^{i,\ell}_t$. Upon ascertaining that $s \notin \hat{S}^{\ell,\ell}_t$, H2 attempts to first book a patient with his/her PCP. If

multiple PCP slots are open and included in J, then H2 books a slot with the smallest value of $q^t_{\ell,j}$. If none of the acceptable PCP slots are available, then H2 books the slot with the smallest value of $q^t_{i,j}$, $i \neq \ell$, among all non-PCP slots in the acceptable set for which $s \notin \hat{S}^{i,\ell}_t$. The key difference between H1 and H2 is that H1 does not protect slots for same-day demand.

3.4.3 Tests of Performance of H1 and H2

For the single physician example presented at the beginning of Section 4.6, the expected daily revenue evaluated at the beginning of the advance-book period, when the system starts empty and we use H1 and H2 to make booking decisions turns out to be 99.76% and 99.81%, respectively, of the optimal expected revenue. This suggests that the performance of H1 and H2 is reasonable in problem instances with a single physician.

However, problems with multiple physicians are not tractable and the corresponding optimal expected revenue cannot be determined exactly. Therefore, we compare the expected revenues obtained from the two heuristics to the expected maximum attainable revenue, which is an upper bound. To calculate this bound, we simulate sequences of advance-book and same-day arrivals and then use an integer program, shown in Section 3.4.3, to calculate the maximum attainable revenue for each sample path.

Maximum Attainable Revenue

Let K (resp. K_ℓ) be the set of decision epochs with an arrival from an arbitrary panel (resp. panel-ℓ). In addition, let $a^t_{i,j} = 1$ if (i, j) physician and time-block combination is acceptable to the advance-book patient who arrives in period t and $a^t_{i,j} = 0$ otherwise. Let x_ℓ denote the realized same-day demand from panel ℓ. The decision variables are $y^\ell_{i,j}$ and $o^t_{i,j}$, where $y^\ell_{i,j}$ is the number of slots that belong to the (i, j) combination and that are assigned to same-day panel-ℓ patients. Furthermore, $o^t_{i,j} = 1$ if the clinic assigns a slot of the (i, j) combination to the patient who arrives in period t, and $o^t_{i,j} = 0$ otherwise. Let $M(\ell)$ be the set of physicians excluding ℓ. Then the maximum attainable revenue of a sequence of arrivals can be obtained by solving the following

integer program.

$$\max \quad \sum_{\ell=1}^{m}\sum_{j=1}^{b}\sum_{t\in K_\ell} r^{\ell}_{1,\ell} o^{t}_{\ell,j} + \sum_{\ell=1}^{m}\sum_{i\in M(\ell)}\sum_{j=1}^{b}\sum_{t\in K_\ell} r^{i}_{1,\ell} o^{t}_{i,j} - \sum_{\ell=1}^{m}\sum_{t\in K_\ell}\pi_t\big(1-\sum_{i=1}^{m}\sum_{j=1}^{b} o^{t}_{i,j}\big)$$
$$+\sum_{\ell=1}^{m}\sum_{j=1}^{b} r_2 y^{\ell}_{\ell,j} + \sum_{\ell=1}^{m}\sum_{i\in M(\ell)}\sum_{j=1}^{b} r'_2 y^{\ell}_{i,j} - c\big(\sum_{\ell=1}^{m} x_\ell - \sum_{\ell=1}^{m}\sum_{i=1}^{m}\sum_{j=1}^{b} y^{\ell}_{i,j}\big),$$

subject to:

$$o^{t}_{i,j} \le a^{t}_{i,j} \qquad \forall\ i=1,\cdots,m;\ j=1,\cdots,b;\ t\in K,$$
$$\sum_{i=1}^{m}\sum_{j=1}^{b} o^{t}_{i,j} \le 1 \qquad \forall\ t\in K,$$
$$\sum_{t\in K} o^{t}_{i,j} + \sum_{\ell=1}^{m} y^{\ell}_{i,j} \le \kappa_{i,j} \qquad \forall\ i=1,\cdots,m;\ j=1,\cdots,b,$$
$$\sum_{i=1}^{m}\sum_{j=1}^{b} y^{\ell}_{i,j} \le x_\ell \qquad \forall\ \ell=1,\cdots,m,$$
$$o^{t}_{i,j} \in \{0,1\} \qquad \forall\ i=1,\cdots,m;\ j=1,\cdots,b;\ t\in K,\ \textit{and}$$
$$y^{\ell}_{i,j} \ge 0 \qquad \forall\ i=1,\cdots,m;\ j=1,\cdots,b;\ \ell=1,\cdots,m.$$

Using CPLEX 8.1 solver, the maximum attainable revenue for each sequence of advance-book and same-day arrivals in the examples reported in Section 3.4.3 was obtained in less than a second.

Results of Performance Tests

We tested H1 and H2 with the help of a 5-factor design of experiments. The factors were — (1) 4 clinic sizes [m = 2, 4, 6, and 8], (2) 5 clinic loads [expected demand/average capacity = 85%, 90%, 100%, 110%, and 115%], (3) 2 types of panel loads [homogeneous or heterogeneous], (4) 4 preference types [time dominant in Table 3.5, physician dominant in Table 3.6, moderate in Table 3.7, and no preferences], and (5) 2 levels of information accuracy [perfect or biased] — for a total of 320 different scenarios. We repeated the evaluation of the 320 scenarios under 2 cost structures: $c/\pi = 2$ and $c/\pi = 8$. A higher c/π ratio is appropriate for clinics that place a high priority on meeting same-day appointment requests. Results are summarized in Table 3.8 and Figure 3.3. They confirm

that H1 and H2 are robust under a variety of different clinic environments. However, before discussing the results, we first describe the experimental setup in more detail below.

Table 3.5: Time Dominant Acceptance Probabilities

ℓ	α_i^ℓ	β_1^ℓ	β_2^ℓ	β_3^ℓ	β_4^ℓ
1	1	0.2	0.4	0.6	1
2	1	1	0.5	0.5	0.5
odd $\ell \geq 3$	1	1	0.5	0.5	0.3
even $\ell \geq 4$	1	0.3	0.5	0.5	1

Table 3.6: Physician Dominant Acceptance Probabilities

ℓ	α_{odd}^ℓ	α_{even}^ℓ	β_j^ℓ
odd ℓ	1	0.3	1
even ℓ	0.3	1	1

Table 3.7: Moderate Acceptance Probabilities

ℓ	α_{odd}^ℓ	α_{even}^ℓ	β_{odd}^ℓ	β_{even}^ℓ
odd ℓ	1	0.4	1	0.4
even ℓ	0.4	1	0.4	1

A clinic may set up time blocks with different lengths and/or different number of appointment slots within a time block. For example, a clinic may divide physicians' morning sessions into three 1-hour blocks, each with two 30-minute slots, and afternoon sessions into two 2-hour blocks, each with three 20-minute appointments. On any given day, a particular physician's slots in each block may vary on account of staff meetings, training, variable work schedules, and differences in the number of work-in/overbook slots. To capture this variability, we assume that clinics have on average 5 slots within each of 4 daily blocks for each physician, but the actual number of slots within each block for each physician is independently sampled from a Uniform [4,6] distribution.

Each physician's same-day demand is assumed to be independent and Poisson distributed with mean 6 (30% of the average capacity). Different levels of clinic load are simulated by choosing $\tau = 0.7\times$ (20 slots/physician) $\times$ m $\times$ (clinic load)$/\sum_{\ell=1}^m \lambda_t^\ell$, where $\sum_{\ell=1}^m \lambda_t^\ell = 0.1$. In the homogenous panel load scenario, the arrival probability

for each panel equals $\sum_{\ell=1}^{m} \lambda_t^\ell/m$, whereas in the heterogeneous panel load scenario, $\lambda_t^1 = 0.8\sum_{\ell=1}^{m} \lambda_t^\ell/m$, $\lambda_t^2 = 1.2\sum_{\ell=1}^{m} \lambda_t^\ell/m$, and $\lambda_t^{\ell'} = 0.8\sum_{\ell=1}^{m} \lambda_t^\ell/m$ for $\ell' \geq 3$. We also varied the clinic load by keeping τ fixed and changing each decision epoch's arrival rate. The performance of H1 and H2 was similar to what we report in Table 3.8 below. Therefore those results are not presented in the interest of brevity.

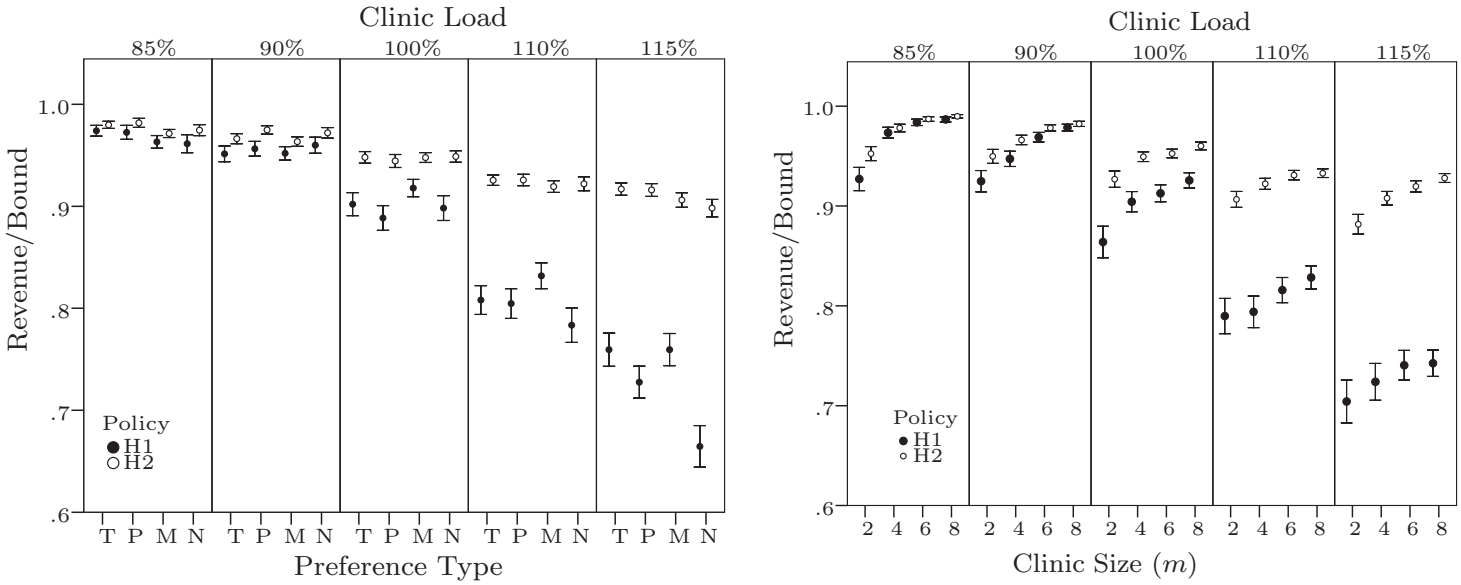
We assume that information bias results in inaccurate estimates of β_j^ℓ. Let $\beta_{j(d)}^\ell$ be panel-ℓ's acceptance probability for time-block j, where (d) indicates that block j has the d-th highest probability among the b time blocks for panel ℓ. In the biased case, the clinic's estimate is assumed to be sufficiently inaccurate that it reverses the ordering for each panel's time-block acceptance probabilities. That is, the clinic uses $\hat{\beta}_{j(d)}^\ell = \beta_{(b-d+1)}^\ell$ when making booking decisions. For example, the clinic would use $\hat{\beta}^1 = (1, 0.6, 0.4, 0.2)$ as the clinic's biased estimates for β^1 in Table 3.5.

To focus attention on the impact of patient preferences and to not confound this effect with the effect of different revenue classes, we assumed that all panels had the same expected revenue. In particular, $(r_{1,\ell}^\ell, \pi, r_2, r_2') = (100, 35, 100, 85)$, $r_{1,\ell}^i = 85$ for $i \neq \ell$, and $\pi_t = \pi$ for all t. We generated 50 sample paths for each scenario, and tracked the performance of H1 and H2 by average relative revenue (as compared to the bound discussed in Section 4.3.1), average patient-PCP match rate, average advance-book failure rate as a percentage of non-urgent requests for a particular appointment date, and average spoilage rate as a percentage of slots unused.

We first compared H1's and H2's performance with accurate and biased acceptance probabilities for each sample path. Neither H1's nor H2's average performance is affected much by using inaccurate acceptance probabilities — relative revenue on average increased by 0.53% for H1 and decreased by 0.49% for H2; all other metrics were on average affected less than 1.8% and 0.7% for H1 and H2, respectively. Note that the improvement in H1's performance is due to the higher advance-book failure rate induced by biased estimates of acceptance probabilities, which increases availability of slots for same-day demand.

Table 3.8: Aggregate Performance

		Clinic	Rel. Rev.		PCP Match		Adv-Bk Failure		Spoilage	
c/π		Load	Mean	SD	Mean	SD	Mean	SD	Mean	SD
2	H1	85%	0.992	(0.017)	0.958	(0.040)	0.007	(0.018)	0.123	(0.088)
		90%	0.989	(0.021)	0.952	(0.041)	0.009	(0.022)	0.094	(0.083)
		100%	0.981	(0.026)	0.943	(0.042)	0.013	(0.027)	0.047	(0.068)
		110%	0.965	(0.033)	0.942	(0.041)	0.022	(0.033)	0.022	(0.055)
		115%	0.958	(0.036)	0.943	(0.040)	0.027	(0.038)	0.013	(0.048)
	H2	85%	0.992	(0.017)	0.958	(0.040)	0.010	(0.024)	0.124	(0.088)
		90%	0.989	(0.020)	0.952	(0.041)	0.016	(0.035)	0.095	(0.082)
		100%	0.983	(0.024)	0.942	(0.042)	0.037	(0.054)	0.050	(0.068)
		110%	0.974	(0.028)	0.940	(0.039)	0.083	(0.077)	0.028	(0.057)
		115%	0.971	(0.028)	0.939	(0.039)	0.108	(0.083)	0.021	(0.051)
8	H1	85%	0.968	(0.100)	0.957	(0.040)	0.008	(0.020)	0.119	(0.085)
		90%	0.955	(0.106)	0.950	(0.042)	0.010	(0.024)	0.093	(0.084)
		100%	0.902	(0.161)	0.943	(0.041)	0.014	(0.026)	0.051	(0.071)
		110%	0.807	(0.212)	0.941	(0.040)	0.022	(0.034)	0.021	(0.053)
		115%	0.728	(0.250)	0.944	(0.038)	0.027	(0.037)	0.014	(0.047)
	H2	85%	0.977	(0.063)	0.957	(0.040)	0.016	(0.037)	0.121	(0.084)
		90%	0.969	(0.069)	0.949	(0.042)	0.022	(0.044)	0.095	(0.083)
		100%	0.947	(0.081)	0.941	(0.041)	0.049	(0.063)	0.058	(0.071)
		110%	0.923	(0.085)	0.936	(0.040)	0.101	(0.083)	0.035	(0.059)
		115%	0.909	(0.102)	0.936	(0.039)	0.135	(0.090)	0.031	(0.055)



(a) By Preference Type and Clinic Load

(b) By Clinic Load and Clinic Size

Figure 3.3: Average Relative Revenue and 95% Confidence Intervals When $c/\pi = 8$.
T=Time Dominant. P=Physician Dominant. M=Moderate. N=No Preferences. HO: homogeneous. HE: heterogeneous.

Table 3.8 reports average performance measures sorted by c/π ratio and clinic load. For each combination, the reported performance metrics are aggregated over all possible scenarios of panel load, preference type, and information accuracy. Both H1 and H2 have high average Patient-PCP match rates (94.7% and 94.5%), low average advance-book failure rates (1.6% and 5.7%), and low average spoilage rates (6.0% and 6.6%). The relative revenue performance of H1 and H2 is respectable with low variability when $c/\pi = 2$ — the relative revenue is on average 97.7% and 98.2% of the bound for H1 and H2, respectively. When $c/\pi = 8$, the relative revenue performance is worse (on average 87.2% and 94.5% of the bound for H1 and H2, respectively), but H2 performs better. This suggests that when a health system has other options for taking care of urgent requests (e.g. urgent clinics), its cost of turning away same-day requests is smaller (low c/π) and it may be justified in using H1, which is much simpler to implement.

Next, we report more detailed results in Figure 3.3 for the case when $c/\pi = 8$. As seen in Figure 3.3(a), for each clinic load, the relative revenue performance of H1 and H2 is robust across preference types. Similar results were also observed for panel loads, which are not reported here in the interest of brevity. Note that the means and confidence intervals are based on all sample paths generated across different clinic environments conditioned on the levels of factors presented in each sub graph. The relative performance of H1 and H2 deteriorates when clinic load exceeds capacity (Figure 3.3(a)), but improves as the clinic size increases (Figure 3.3(b)). The latter happens because each panel's same-day demand is assumed to be independent and clinics with more physicians benefit from pooling available capacity to take care of same-day demand. The size effect may disappear when same-day demand patterns are correlated across physician panels.

3.5 Insights

In this section, we first compare the performance of H1 and H2 to a straw policy that does not utilize patients' preference information when making booking decisions. The straw policy attempts to book each arriving patient with the earliest available and acceptable patient-PCP matched slot. If none of matched acceptable slots is available, the straw policy then books an appointment in the earliest available non-PCP slot,

paying no attention to remaining capacity and time preferences. Next, we compare H1 (or H2) to itself when using true acceptance probabilities and naive acceptance probabilities. The purpose of this comparison is to tease out the value of information if a clinic decides to adopt either H1 or H2 booking heuristic. Finally, we evaluate the effect of low-levels of no-show rates and service time variability by comparing H1, H2, and the straw policy.

All examples of this section use the following common parameters : $(r^{\ell}_{1,\ell}, \pi_t, r_2, r_2^{'}, c) = (100, 35, 100, 85, 280)$, $r^{i}_{1,\ell} = 85$ for $i \neq \ell$, Poisson same-day demand with $E(X_\ell) = 6 = 30\%$ of each physician's capacity of 20 appointments per day. If desired, each panel's advance-book arrival rate can be varied to realize different workloads for different physicians. Total advance-book periods equal $(0.7 \times \text{clinic capacity})/(\sum_{\ell=1}^{m} \lambda_t^{\ell})$, which ensures expected clinic demand equals clinic capacity. We report results when $\sum_{\ell=1}^{m} \lambda_t^{\ell} = 0.1$. Each experimental set up is simulated for 200 sample paths and all booking strategies are evaluated for the same sample paths.

The first set of comparisons consider a 6-PCP and 4-time-block clinic whose patients always show up and find all physicians acceptable, but these patients have the following time preferences: $\beta^{\ell} = (\bar{\beta}, \bar{\beta}, 1, 1)$ for $\ell = 1, 2$; $\beta^{\ell} = (1, \bar{\beta}, \bar{\beta}, 1)$ for $\ell = 3, 4$; $\beta^{\ell} = (1, 1, \bar{\beta}, \bar{\beta})$ for $\ell = 5, 6$. Each physician's clinic profile has more slots in blocks that are more acceptable to their panel patients (i.e. $\kappa_{\ell,j} = 6$ if $\beta_j^{\ell} = 1$, and $\kappa_{\ell,j} = 4$ otherwise). We vary $\bar{\beta}$ from 0.2 to 0.8 in 0.1 increments and study two arrival patterns: (1) constant arrival rates: $\lambda_t^{\ell} = 0.1/m$ for all $\ell = 1, \cdots, 6$ and $t = 1, \cdots, \tau$, and (2) varying arrival rates: when $t \leq (1/3)\tau$, $\lambda_t^k = 3\lambda_t^i$ for $k = 1, 2$ and $i = 3, 4, 5, 6$; when $(1/3)\tau < t \leq (2/3)\tau$, $\lambda_t^k = 3\lambda_t^i$ for $k = 3, 4$ and $i = 1, 2, 5, 6$; when $t > (2/3)\tau$, $\lambda_t^k = 3\lambda_t^i$ for $k = 5, 6$ and $i = 1, 2, 3, 4$.

H1 and H2 on average result in about 1% and 8% higher revenue as compared to the straw policy regardless of the value of $\bar{\beta}$ and the arrival pattern. We report only the aggregate results in Table 3.9. H1 achieves a higher PCP match, relative to the straw policy, by reserving more popular slots for future advance-book arrivals. In contrast, H2 with higher spoilage and advance-book failure rates achieves a high PCP match rate for different reasons. By reserving slots for same-day patients, it allows more of those patients to have an appointment with their PCPs. H2 has much higher advance-book failure rate and slightly smaller number of patients served because some advance-book

requests are denied when no-book states are reached. However, overall revenue is higher because it is costlier to turn away same-day patients.

Table 3.9: Performance of H1 and H2 Compared to the Straw Policy

Policy	PCP match	Adv-Bk Failure Rate	Spoilage Rate	# Scheduled/ Served	Rev. Improv. compared to Straw
Straw	92.88%	0.00%	3.41%	115.9	–
H1	94.34%	0.04%	3.50%	115.8	1%
H2	94.06%	0.43%	4.17%	115.0	8%

Next, we evaluate the performance of H1 and H2 with two levels of information: (1) true acceptance probabilities, and (2) naive acceptance probabilities. The latter assumes that every physician and time combination is acceptable to every patient. For each heuristic, we calculate the value of preference information by comparing that heuristic's average daily revenue to itself when the clinic uses true versus naive acceptance probabilities as inputs. We also monitor changes in patient-PCP match rates, advance-book failure rates, and number of patients served. In the results reported here, the clinic has 8 full-time physicians, 4 time blocks, and 5 slots per block. Arrival pattern is time homogenous, but expected demand rates can vary by panel resulting in imbalanced workload across physicians.

We use a full factorial design of three factors, each with 2 levels. For a panel with strong (resp. weak) time preferences, we allow one block to be always acceptable and the remaining blocks to be accepted with probability 0.3 (resp. 0.7). For a panel with strong (resp. weak) physician preferences, the PCP is always acceptable and each non-PCP is acceptable with probability 0.3 (resp. 0.7). For a panel with adequate (resp. inadequate) capacity, we let the expected demand to be 66.7% (resp. 133.3%) of the capacity. These combinations lead to 8 stylized panel types.

Accurate preference information on average increases daily revenue by \$20.35 and \$93.15 for H1 and H2, respectively. H2 reduces advance-book failure rate by 1.3% and serves on average 1.16 more patients per day when using true acceptance probabilities as inputs. That is, our example clinic would be able to serve on average 423.4 more patients per year by updating patients' acceptance probabilities and using H2. We also studied a different scenario (results not reported for brevity) in which physician workloads were balanced and found that in such cases, knowledge of accurate acceptance probabilities

does not significantly affect average daily revenues, patient-PCP match rates, advance-book failure rates, and the number of patients served (paired sample tests were not significant in all comparisons). This suggests that clinics whose physicians' workloads are imbalanced are more likely to benefit from accurate preference information when using H1 or H2 booking schemes. Imbalanced workloads are a common occurrence in practice.

In the last set of examples, we evaluate the effect of no-shows and service-time variability, assuming punctual physicians and patients, i.i.d. service times and equal-length appointment slots. We test two levels of average no-show rates: 5% and 10% with each patient's no-show probability drawn independently from Beta(0.05,0.95) and Beta(0.1,0.9) distribution, respectively. To test the impact of service time variability, we sample 5 distributions (see Table 3.10) that have the same mean but different co-efficients of variation (0.33, 0.58, 0.58, 0.71, and 1 respectively). These distributions cover the range of service time variability observed in empirical studies (0.3 – 0.85); see Cayirli and Veral (2003). Finally, $\bar{\beta} = 0.5$ and $\lambda_t^\ell = 0.1/m$ for each ℓ and t. We report performance comparisons in terms of paired sample t statistics for the average difference in revenue, average patient wait and average physician overtime between H1 (or H2) and the straw policy in Table 3.10.

H1 and H2 on average have a statistically higher revenue than the straw policy at each level of no-shows. The difference in average patient wait and physician overtime time between H1 (or H2) and the straw policy is statistically insignificant in most cases. However, when the difference is significant, H1 and H2 perform better. If the length of the appointment time slot is 30 minutes and average service time is 27.3 minutes, then patients' average wait ranges from 6.8 to 45 minutes while physicians' average overtime ranges from 6 to 40 minutes across these 10 scenarios. These results show that it is reasonable to use H1 and H2 when a clinic's no-show probability is not too high ($\leq 10\%$) and the service time variability is not more extreme than the variability observed in empirical studies.

Table 3.10: Performance comparison in terms of t statistics

No-show rate	Service time distribution	Revenue		PCP match		Avg wait		Avg OT	
		H1	H2	H1	H2	H1	H2	H1	H2
5%	Unif[0.4,1.6]	6.5*	8.1*	14.6*	13.5*	-1.8	-1.7	-1.6	-1.9
	Unif[0,2]	5.9*	8.2*	14.4*	12.4*	-1.0	-1.3	-0.7	-1.9
	Gamma(3,1/3)	7.5*	8.4*	14.7*	11.1*	0.5	-0.9	-0.8	-2.7*
	Gamma(2,0.5)	5.7*	7.9*	15.4*	12.7*	-0.8	-1.2	-0.3	-0.9
	Exp(1)	6.9*	8.8*	14.8*	12.5*	-1.1	-1.8	-0.3	-0.5
10%	Unif[0.4,1.6]	8.3*	3.9*	15.1*	11.7*	0.4	-0.5	-1.9	-3.4*
	Unif[0,2]	7.6*	7.8*	15.8*	12.5*	-0.1	-1.0	-1.6	-3.8
	Gamma(3,1/3)	6.7*	9.2*	15.1*	13.3*	-0.6	-0.1	-1.2	-1.9
	Gamma(2,0.5)	6.7*	7.9*	15.2*	11.9*	-0.1	-0.2	-0.8	-3.2*
	Exp(1)	7.3*	8.9*	13.8*	10.4*	1.7	1.4	0.4	0.3

Degree of freedom = 199 under paired sample tests.
Asterisk means significant at 0.05 level.

3.6 Concluding Remarks

This chapter presents a framework for using appointment request data to update patients' preferences and to subsequently use this information to improve clinics' revenues, serve more patients, and increase patient-PCP match rates. This approach can be implemented by utilizing data that can be retrieved from existing web-based appointment request systems. However, it may not be suitable for clinics with high no-show rates that cannot be controlled by the use of a reminder system and patient education, or by better matching patients' preferences with available slots. Such clinics may benefit from using approaches that explicitly consider no shows when making booking decisions.

Our model is limited because it considers each workday's booking problem separately. A clinic may benefit from knowing all acceptable dates and the physician and time-block combinations that are acceptable to each arriving patient on each date before making a booking decision. However, that will make the booking process tedious for the patients and the state space of the appointment system will become unmanageable because acceptable dates may span an arbitrarily large period of time. It is perhaps for this reason that common web-based booking request systems accept requests for one day at a time.

The contribution of this chapter is to incorporate patients' preferences in modeling a design of appointment systems. The model provides one way to capture patients' preference information using existing web-based interface, and utilize preference information in real-time booking decisions. The high dimensional state space in the appointment booking problem results from multiple physician and multiple time blocks, which makes it not only unrealistic to calculate the optimal booking decision in real time, but also to prepare an off-line look-up booking policy table based on an optimal booking policy.

There are several future research directions. For example, clinics may further benefit from obtaining additional preference information such as the rank ordering of different physician/time combinations in addition to the combinations that are acceptable to the patients. It would be worthwhile to investigate whether the proposed booking heuristics and preference capturing procedures will remain robust when some patients do not truthfully reveal their acceptable sets. Alternatively, future studies may explore whether the clinic can achieve better performance by showing patients a set of available appointment choices up front rather than asking the patients to reveal their acceptable set.

Patient-centered service models have attracted much attention in recent health policy literature. For example, a medical home model is a one-stop model that matches each patient with a team of providers based on the patient's needs. This team monitors patient's health status and coordinates appointments for acute, chronic, and preventive services. Similarly, many health systems allow patients to see several service providers in a day or within a short period of time so that out-of-town patients do not need to travel to the service facility multiple times. Both models require matching patients' needs and preferences to multiple providers' availability. An interesting avenue of future research along the lines presented in this chapter is the development of a model-based design of an adaptive appointment system for integrated medical services.

Chapter 4

Nurse Absenteeism and Staffing Strategies for Inpatient Units

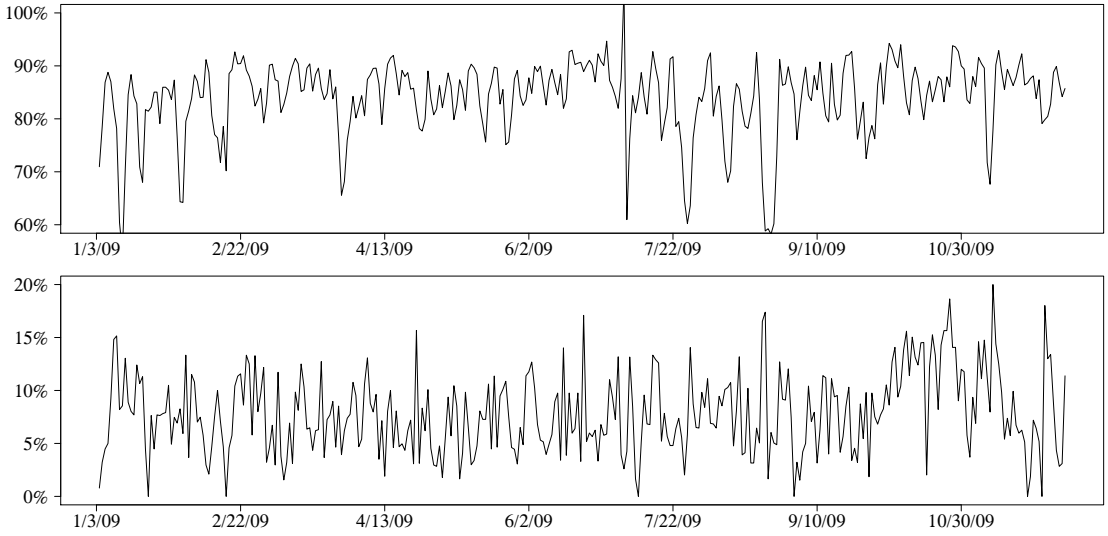
4.1 Introduction

We discussed some outpatient clinics' capacity management challenges in Chapters 2 and 3. Now we turn to capacity management challenges faced by inpatient units where the primary care providers are registered nurses (RNs). This problem is complicated because nurses may be absent and patients' arrivals and lengths of stay are random.

Inpatient units are often organized by nursing skills required to provide care. A typical classification of inpatient units includes the following tiers: intensive care (ICU), step-down, and medical/surgical. Multiple units may exist within a tier, each with a somewhat different specialization. For example, different step-down units may focus on cardiac, neurological, and general patient populations. Frequently, each nurse is matched to a home unit in which skill requirements are consistent with his or her training and experience. Nurses' work schedules are fixed several weeks in advance, and once finalized, schedules may not be changed unless nurses agree to such changes. A finalized staffing schedule is also subject to random changes due to unplanned nurse absences. These facts complicate a nurse manager's job of scheduling nurses to match varying demand and supply. To illustrate these points, we provide in Figure 4.1 a time-series plot of percent of occupied beds (i.e. average census divided by bed capacity), and

percent of absent nurse shifts (unplanned) from three step-down units of an urban 466-bed community hospital between Jan 3rd, 2009 and Dec 4th, 2009. Note that patient census and nurse absentee rate vary significantly from one day to the next, which makes staffing decisions challenging.

Figure 4.1: Daily Percent Occupied Beds (Top) and Absent Nurse Shifts (Bottom).



In this chapter, we focus on staffing decisions at the stage in which the total number of nurses, their skill levels, and work patterns are known and the problem is that of making unit/shift assignments. This problem is complicated by uncertainty in future shifts' requirements and available nursing care hours. The former is caused by randomness in census and other contributors of nurses' workload and the latter by unplanned absences. Therefore, for determining the nurse assignment, one challenge is to forecast nurse requirements and another challenge is to cope with nurses' unplanned absences. In this chapter, we mainly focus on nurses' unplanned absences. The challenges of forecasting nurse requirement is discussed briefly in Section 4.2. Next, we discuss the background of nurse staffing with the presence of unplanned absences that is central to this chapter.

We studied data from two hospitals and found that the average absentee rates among registered nurses were 8.3% and 7.7% (see Section 4.3 for details). We also obtained

statistics from Veterans Administration (VA) Nursing Outcomes Database for a period of 24-week period between September 2011 and February 2012. The average unplanned[1] absentee rate across all hospitals in the VA Health Care System was 6.4%. These statistics are significantly higher than absentee rates among healthcare practitioner/technical occupations and all occupations in the United States, which happen to be 3.7% and 3% respectively (Bureau of Labor Statistics 2011), underscoring the importance of considering absenteeism in staffing models. Nurse absenteeism exacerbates the problem of inadequate staffing, which affects quality of care, patients' safety and length of stay, nurses' job satisfaction, and hospitals' financial performance (e.g. Unruh 2008, Aiken et al. 2002, Needleman et al. 2002, Cho et al. 2003, Lang et al. 2004, and Kane et al. 2007).

In addition to nursing, absenteeism related labor costs are high in a whole host of labor intensive environments such as fast food restaurants, automobile manufacturing and assembly, and call center operations. Many companies rely on a pool of workers to substitute for absent employees, however, the replacements may not be as skilled or as efficient as those who are absent. Among fast food restaurants, failure to match capacity and demand can significantly reduce or delay customer service. Ordones (2000) reports that in a typical McDonald's restaurant, every six-second delay at the drive-thru leads to one percent lost sales. Connelly (2003) reports that absenteeism adversely affects safety, quality, delivery, cost, and morale in the automobile industry. Absenteeism makes call-center staffing decision difficult because absence affects the ability to meet requirements for different operator skills (Aksin et al. 2007). Although this chapter is motivated by high absenteeism among nurses, the mathematical models developed here are likely to be useful in other applications as well.

Nurses' absence from work may be either planned or unplanned. Planned absences, such as scheduled vacations, continuing education classes, and training are easier to cope with because a nurse manager has advance knowledge of potential staff shortages created by such absences. In contrast, unplanned absences are costly and may compromise patient safety as well as quality of care because well qualified replacements can be expensive and difficult to find at a short notice. For these reasons, our focus in this chapter is on unplanned absences.

[1] Includes sick leaves and leaves without pay, both of which are often unplanned.

We begin by describing two examples that highlight the interaction between staffing strategies and a nurse manager's model of absenteeism. In the first example, a nurse manager becomes aware that there will be a staff shortage in a particular future shift. Such assessments are based on inpatients' health status, nurses' time-off requests arising after their schedules are finalized, more than anticipated admissions, and fewer discharges. Nurse managers like to recruit part-time nurses to work extra shifts to satisfy excess demand. This is cheaper and less stressful than finding overtime or agency nurses at a short notice (May et al. 2006). In hospitals with unionized workforce, nurse managers need to announce the opportunity to pick up extra shifts to all nurses who are qualified for these shifts, and union rules may dictate the order in which extra shift requests must be granted; e.g. one of the hospitals whose data we use to shine light on possible predictors of absenteeism is required to prioritize such requests by seniority. Consequently, nurse managers have little control over who may be selected to work extra shifts.

Suppose that the projected excess demand equals 5 RN shifts and the absentee rate among nurses available for extra shift assignments is either 5% or 15%, with an average absentee rate of 10%. Cost per shift of an extra nurse shift is r and the cost of an overtime/agency nurse shift is $r' = 1.5r$. If the nurse manager assumes independent homogeneous absentee rate of 10%, then he or she will recruit 6 extra-shift nurses because that minimizes the expected total cost (sum of under- and over-staffing cost) of $1.5r \sum_{q=1}^{n} (5-q)^+ P(Q=q) + r \sum_{q=1}^{n} (q-5)^+ P(Q=q)$. In this expression, Q is a binomial random variable representing the number of nurses who show up for work among the n scheduled nurses and the parameter n is the decision variable. However, if the 6 nurses selected for the extra shifts happen to all have absentee rate of 5%, then the optimal number of nurse shifts is 5 and the expected cost with 6 nurses will be twice that of the cost with 5 nurses. The example highlights the importance of considering heterogeneous absentee rates when making staffing decisions.

In the second example, a nurse manager needs to assign nurses to inpatient units either to rebalance workload or in response to reorganization of beds caused by changes in either patient volumes or flow patterns. Nurse assignment decisions are not frequent because nurses receive orientation and training specific to their home unit. Nurses generally do not like to float to other units due to the potential negative impact on

patient safety (especially when patient requirements are different), and the stress of working in an unfamiliar physical environment, using unfamiliar equipment, and having unfamiliar coworkers (Ferlise and Baggot 2009). In fact, some contract rules prevent nurse managers from floating or temporarily reassigning nurses to work in another unit (California Nurse Association and National Nurses Organizing Committee 2012). Therefore, we do not consider the possibility of floating nurses to level workload based on realized demand.

Suppose ten nurses need to be assigned to work in a particular shift in two inpatient units with independent discrete uniform demand distribution between 0 and 8, and a mean of 4 nurses per shift. Their absentee rates are (0, 0, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.2, 0.2), respectively, and the average absentee rate is 10% . How should the nurse manager assign these nurses? There are a number of different staffing strategies that would result in the same average absentee rate but different costs. For example, the manager may assign five nurses with absentee rates (0, 0.1, 0.1, 0.1, 0.2) to each unit (Strategy 1). Alternatively, he or she may assign nurses whose absentee rates are (0.1, 0.1, 0.1, 0.1, 0.1) to Unit 1 and those whose absentee rates are (0, 0, 0.1, 0.2, 0.2) to Unit 2 (Strategy 2). Unit 1's average shortage is smaller under Strategy 1 (0.693 versus 0.699) whereas Unit 2's average shortage is smaller under Strategy 2 (0.688 versus 0.693). Strategy 1 is better in the aggregate with a lower total shortage cost. We will later show that Strategy 1 is an optimal assignment. However, the problem of choosing an optimal staffing strategy with heterogeneous absentee rates is in general a nontrivial combinatorial problem.

The two examples serve to highlight the importance of identifying key variables that affect staffing decisions and after embedding them in appropriate optimization models, finding easy-to-implement methodologies for assisting nurse managers in making staffing decisions. This chapter focuses on both these aspects. We analyze data from two hospitals to build predictive models that explain nurse absenteeism as a function of observable unit-level and nurse-specific variables. (These variables were selected after consulting with nurse managers at both hospitals.) We also propose and analyze a model for making nurse assignment decisions.

Statistical models are used in this chapter to ascertain whether nurse absence patterns are better explained by a model that assumes nurses are homogeneous decision

makers or a model that assumes a different absence pattern for each nurse. Within the first model, each nurse's time-off decision is hypothesized to be a realization from a common underlying distribution, which is parameterized by unit-level factors such as unit index (which captures unit culture, manager effectiveness and long-term workload), shift time, short-term anticipated workload, and interactions among these factors. The second model assumes that absentee rates are not homogeneous and tests the hypothesis that nurses' past absence records can be used to predict their absences in the near future.

A factorial design with all two-way interactions was used to analyze the first model. We found that unit index had a significant effect on how nurses as a group responded to the anticipated workload, but that there did not exist a consistent relationship between workload and nurses' absenteeism after controlling for other factors. This is important because it means that absenteeism need not be treated as an endogenous variable in staffing optimization models for inpatient units. The second model utilized nurse-specific data. We found that each nurse's absentee rate was relatively stable over the period of time for which data were obtained and that nurses' history of absence from an earlier period was a good predictor of their absentee rates in a future period. The latter is consistent with findings of other papers (see Davey et al. 2009 for a systematic review).

From the statistical models, we conclude that nurse managers need to account for heterogeneous attendance history when making staffing plans. We then propose mathematical models to determine nurse assignments that minimize the total expected shortage cost for multiple nursing units, given a cohort of nurses who may have different absentee rates. The demand for each nursing unit is assumed independent. We propose three different models, each of which may be useful in a different setting. These are: (1) deterministic, (2) random aggregate, and (3) nurse-specific Bernoulli. In the first model, the number of absences is a deterministic function of the number of nurses assigned to a unit. In the second model, nurses are grouped into subsets and each subset (or nurse type) has a different absentee rate distribution. In the third model, each nurse is different and the overall absentee rate distribution of a unit is the convolution of absentee rate distributions of particular nurses assigned to that unit.

The models are used to show that greater variability in demand and attendance patterns increases a hospital's costs. Moreover, when inpatient units face identical

demand, the hospital minimizes costs by choosing the same staffing plan for each unit. However, from the perspective of each unit, for the same overall aggregate absentee rate, its cost is minimized by choosing a more heterogeneous cohort of nurses. This suggests a staffing strategy that maximizes heterogeneity within a unit but creates uniform plans across units. We also establish that the hospital's objective function is supermodular. Because greedy heuristics generally work well when objective function is supermodular, we explore greedy and two other heuristics for solving the staffing-plan optimization problem. Two of the three heuristics are shown to work extremely well in numerical experiments. These experiments suggest that hospitals can reduce staffing costs by utilizing historical attendance data and relatively easy-to-use heuristic approaches for staff assignment.

The nurse staffing problem studied in this chapter is related to literature in operations management (OM) and health services research (HSR) disciplines. The staffing problem is similar to the random yield problem studied in the OM literature in the sense that the realized staffing level (equivalently, the yield of good items produced) may be lower than the planned staffing level (production lot size) due to nurses' show uncertainty (random yield). However, existing random yield models do not fully capture the features of the nurse staffing problem. For example, random yield models generally characterize yield uncertainty in one of the following ways: (1) For any given lot size n, the yield $Q(n)$ is a binomial process with a yield rate p. (2) The yield is a product of the lot size and a random yield rate (i.e. $Q(n) = n \cdot \xi$, where n is the lot size, and ξ is the random yield rate). (3) The production process is in control for a period of time followed by a period when it is out of control (e.g. yield Q may have a geometric distribution). (4) Yield is a result of having random capacity (i.e. $Q = \min\{n, C\}$, where C is the random capacity that captures the unreliability of the equipment). (5) the distribution of yield is known (i.e. $p(q|n)$ is the probability of q good units given a lot size n). The first two types of models may be reasonable for the nurse staffing problem when nurses' absentee rates are independent and homogeneous. However, random yields models have not focused on the types of issues that arise in the nurse assignment problem – e.g. different show rates for different nurses.

Green et al. (2011) formulate a model with endogenous yield rates for the nurse staffing problem. The authors use data from one emergency department (ED) of a

single hospital and observe that nurses' anticipated workload (measured by the ratio of staffing level in a shift and the long-term average census) is positively correlated with their absentee rate. Green et al.'s model of the staffing problem minimizes the expected staffing cost $w_r n + w_e \sum_{q=0}^{n} E(X-q)^+ p(q; n, \gamma(n))$, where n is the staffing level, w_r and w_e are respectively the regular and extra/overtime wage rates, X is the nursing demand, and $p(q; n, \gamma(n))$ is the probability that q nurses show up for work given that the staffing level is n and $\gamma(n)$ is the probability that a scheduled nurse will be absent from work given that the staffing level is n. The absentee rate $\gamma(n)$ is assumed homogeneous among all nurses who are scheduled to work for a shift.

The results from our analysis are significantly different from those reported in Green et al. (2011). In particular, short-term workload is not correlated with absenteeism in our data. The differences arise because of the fundamental differences in the type of data and problem scenarios modeled, which we explain next. First, inpatient units and EDs face different demand patterns and patients' length-of-stay with patients staying significantly longer in inpatient units.[2] Second, it may be argued that EDs present a particularly stressful work environment for nurses and therefore ED nurses may react differently to workload variation than nurses who work in inpatient units. Third, we use data from multiple units and two hospitals, which allows us to quantify the effects due to unit index and the interaction between unit index and shift index, whereas Green et al. (2011) examine data from a single ED of a single hospital.

Much of the OM literature dealing with nurse staffing has focused on developing nurse schedules to minimize costs while satisfying nurses' work preferences; see Lim et al. (2011) for a recent review. These works are not closely related to this chapter. There are numerous papers that are motivated by applications outside healthcare domain that take into account staff absenteeism; see e.g. Hur et al. (2004), Whitt (2006) and Blumenfeld and Inman (2009). However, to our best knowledge, they do not focus on identifying predictors of absenteeism and assigning personnel with heterogeneous absentee rates.

[2] According to surveys done in 2006 and 2010, average length of stay in emergency departments (delay between entering emergency and being admitted or discharged) was 3.7 hours and 4.1 hours respectively (Ken 2006, Anonymous 2010). In contrast, the average length of inpatient stay in short-stay hospitals was 4.8 days according to 2007 data (Table 99, part 3 in National Center for Health Statistics 2011).

The HSR literature attempts to explain why nurses take unplanned time off; see Davey et al. (2009) for a systematic review. This literature concludes that causes of absenteeism vary among different groups of nurses in the same hospital, and fluctuate over time (Johnson et al. 2003), and that nurse absences are associated with organization norms, nurses' personal characteristics, chronic work overload and burn out. Our results are consistent with this literature with the difference that HSR papers do not deal with staffing optimization.

The remaining of this chapter is organized as follows. In Section 4.2, we discuss the challenges in forecasting nurse requirements. Then we describe the data and institutional background in Section 4.3 before the statistical investigation in Section 4.4. The staffing problem is formulated in Section 4.5, and the results are presented in Section 4.6.

4.2 Challenges in Forecasting Nursing Requirements

Nurse staffing decisions affect quality of care, patients' safety, nurses' job satisfaction, and hospitals' financial performance (Unruh 2008, Aiken et al. 2002, Needleman et al. 2002, Cho et al. 2003, Lang et al. 2004, Kane et al. 2007). These decisions fall into three time-based hierarchies (Abernathy et al. 1973, Brusco et al. 1993). Long-term staffing decisions, with a typical planning horizon of 1 year, focus on choosing the number of nurses of each skill type and required staffing levels in terms of full time equivalents (FTEs). Medium-term staffing decisions concern the development of a schedule of days-off and shift assignments for each nurse in each unit. Nurses' work schedules are usually set for several weeks at a time and posted a few weeks before the start of each planning period. Short-term staffing decisions increase/decrease staffing levels by using overtime or agency nurses, asking nurses to exercise benefit time, or transferring personnel from one unit to another to match staffing levels and realized nursing needs. Short-term decisions are usually made before the start of each shift.

In this chapter, we focus on the long-term staffing stage in which the total number of nurses, their skill levels, and work patterns are known and the problem is that of assigning nurses of the same skill type into nursing units.

For determining the number of nurses to assign to an inpatient unit, a key challenge

is how to forecast nurse requirements. The data that are available to the hospitals for forecasting purposes typically include hourly patient census, hourly number of admissions, discharges, and transfers (ADT). Hospitals generally assume an average mix of patients and use some formula (based on expert opinion) to convert census and ADT information into nurse requirement (usually in nursing hours required for each activity). This type of workload conversion generally forms the basis for determining nurse requirements.

A variety of time-series (Shumway and Stoffer 2006) techniques are developed for forecasting. Several previous studies (Côté and Tucker 2001, Kao and Tung 1980, Wood 1976, Earnest et al. 2005) discuss the use of such methods for forecasting nursing needs at different levels of resource aggregation and time scales. For applying time series methods to forecasting nurse requirements, Wang et al. (2009) use a time-series model that utilizes moving averages of the realized nursing requirements to improve the accuracy of 2-4 week look-ahead forecast for medium-term planning, and use the forecast to evaluate the trade-offs between accuracy of demand forecasting and scheduling flexibility. This type of forecasting models assume the historical data is complete and produce point estimates rather than an estimation of the nursing requirement distribution.

In reality, over- and under-staffing costs are unequal and target levels need to balance capacity excess and shortage costs. Target staffing levels can be estimated ex ante only upon knowing the demand distribution and overage/underage cost parameters. However, forecasting techniques recommended in the health services literature do not obtain the distribution of nursing needs. Furthermore, demand information is usually only partially observed due to *data censoring*, which is often overlooked. For example, inpatient census data might be censored at times when a unit's staffed bed capacity was reached and no additional patients were admitted to that unit even if it had empty beds. Therefore, the observed demand might be smaller than the true demand.

We use an example to highlight the impact of data censoring. Suppose that a 20-bed unit experiences a uniformly distributed demand for beds (between 12 and 20) in each shift, but that this information is not known to the hospital. The hospital has been staffing historically at 80% bed capacity (16 beds). Note that 16 is also the unit's mean and median demand. In the historical data, the hospital did not keep track of instances when additional patients could not be placed because bed census had reached

16. Suppose it decides to staff at mean observed demand level for future shifts. How will that decision affect staffing costs when per hour costs of staffing shortage and overage are equal?

The observed census at time t equals $O_t = \min\{N_t, s_t\}$, where N_t is the demand (sampled from a Uniform distribution over [12, 20]), and s_t is the staffed bed capacity at time t (fixed at 16). It is straightforward to calculate that the average observed demand $\bar{O}_t = 14\frac{8}{9}$, which is significantly smaller than the true mean demand 16. Scheduling 15 nurses per shift costs on average 5% more than the minimum cost, even though this approach employs fewer staff.

There are several methods for estimating demand distributions in the presence of censored data (Klein and Moeschberger 2005). However, these methods have not been tested for nurse staffing applications. Some recent inventory management studies use Kaplan-Meier (K-M) estimator (Kaplan and Meier 1958) to correct for censored demand data. These models iterate between collecting new data to update demand distribution, correcting for censoring, and then using the most recent corrected demand distribution to determine order quantity. The latter serves as the censoring variable for the next demand observation. The performance of these models improves as the number of observations increases (Huh and Rusmevichientong 2009, Huh et al. 2009). However, these models require well-defined censoring variable (i.e. whether the demand is censored or not should be clear cut), and the censoring variable needs to be independent of the demand.

The demand censoring events are not always observable for hospitals' inpatient units for several reasons. For example, nurse requirements could be subjective. Sometimes even though patient census and other nominal and observable demand information such as admissions, discharges, and transfers (ADT) may exceed the desirable nurse-to-patient ratios, nurses may work harder to cover the excess nursing requirements. In addition, nurse managers may make short-term staffing adjustments to react to demand variations. Because hospitals typically do not track censoring events and the censoring variable (e.g. realized staffing level) is not always independent of the nurse requirements, modification in the K-M methods for estimating the demand distribution is necessary if it is applied for demand estimating/forecasting purpose.

For example, Gupta et al. (2011) use data from multiple telemetry nursing units in

a hospital to evaluate whether the staffing decisions can be improved by applying a modified K-M estimator to improve the performance of staffing targets. The authors assume that if the realized staffing level is lower than the realized demand's desirable staffing level calculated based on the desirable nurse-to-patient ratio, then censoring occurs. Upon applying this methodology to data from telemetry units of that hospital, it was found that the hospital can realize significant savings by using this approach.

This chapter does not focus on forecasting nursing requirements. The discussion in this section serves to highlight the challenges of estimating nurse staffing requirements, which are assumed known in the remainder of this chapter. We will assume that a hospital has already obtained reasonable estimates of the demand distribution, hired a certain number of nurses, and determined their work patterns based on their full-time equivalents (FTEs). With the assumption that demand distribution is known and nurse availability is fixed, in the remaining of this chapter, we focus on the capacity allocation challenge of assigning available nurses to different nursing units when nurses may be absent.

4.3 Institutional Background

We studied de-identified census and absentee records from two hospitals located in a large metropolitan area. Census and absentee data from Hospital 1 were for the period January 3, 2009 through December 4, 2009, whereas Hospital 2's data were for the period September 1, 2008 through August 31, 2009. Basic information about these hospitals from fiscal year 2009 is summarized in Table 4.1. The differences between the maximum and the minimum patient census were 52.7% and 49.5% of the average census for Hospitals 1 and 2, respectively, indicating that the overall variability in nursing demand was high. Patients' average lengths of stay were 4.0 and 4.9 days and registered nurse (RN) salary accounted for 17.9% and 15.5% of total operating expenses of the two hospitals. This indicates that a reduction in nurse staffing cost could result in significant savings in operating expenses.

Hospital 1 had five shift types. There were three 8-hour shifts designated Day, Evening, and Night shifts, which operated from 7 AM to 3 PM, from 3 PM to 11 PM, and from 11 PM to 7 AM, respectively. There were also two 12-hour shifts, which

Table 4.1: Fiscal Year 2009 Statistics for the Two Hospitals.

	Hospital 1	Hospital 2
Size/Volume		
Available beds	466	284
Maximum Daily Census	359	242
Minimum Daily Census	205	149
Average Daily Census	292	188
Admissions	30,748	14,851
Patient Days	114,591	68,924
Admissions through ER	14,205	6,019
Acute Care Admissions	26,751	13,694
Acute Patient Days	106,625	66,842
Average Acute Care Length of Stay (days)	4.0	4.9
Number of RN FTEs	855.9	415.4
Expenses/Income (in millions)		
RN salary expenses	$73.2	$36.6
Total Operating Revenue	$431.2	$231,4
Total Operating Expenses	$408.5	$235.5
Total Operating Income	$22.7	$-4.1

Available beds = number of beds immediately available for use.

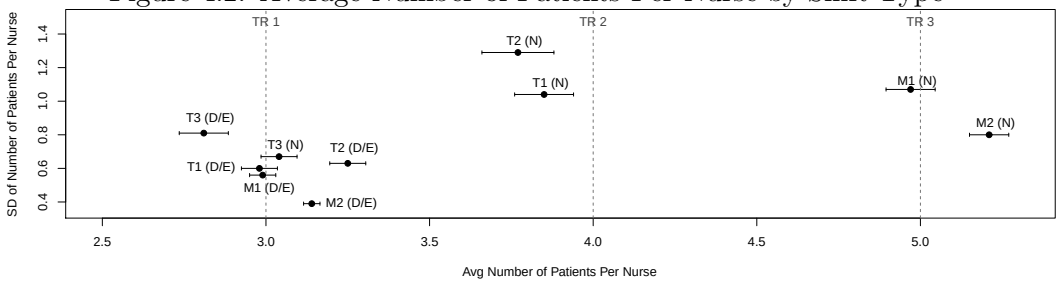
RN = registered nurse. FTE = full-time equivalent.

were designated Day-12 and Night-12 shifts. These operated from 7 AM to 7 PM, and 7 PM to 7 AM, respectively. Hospital 2 had only three shift types, namely the 8-hour Day, Evening, and Night shifts. Hospital 1's data pertained to three step-down (telemetry) units labeled T1, T2, and T3 with 22, 22, and 24 beds, and Hospital 2's data pertained to two medical/surgical units labeled M1 and M2 with 32 and 31 beds. The common data elements were hourly census, admissions, discharges, and transfers (ADT), planned/realized staffing levels, and the count of absentees for each shift. Hospital 1's data also contained individual nurses' attendance history. The two health systems' data were analyzed independently because (1) the data pertained to different time periods, (2) the target nurse-to-patient ratios were different for the two types of inpatient units, and (3) the two hospitals used different staffing strategies.

Hospital 1's target nurse-to-patient ratios for telemetry units were 1:3 for Day and Evening shifts during week days and 1:4 for Night and weekend shifts. Hospital 2's target nurse-to-patient ratios for medical/surgical units were 1:4 for Day and Evening shifts and 1:5 for Night shifts. Hospital 1's planned staffing levels were based on the mode of the midnight census in the previous planning period. Nurse managers would further tweak the staffing levels up or down to account for holidays and to meet nurses' planned-time-off requests and shift preferences. Hospital 2's medical/surgical units had fixed staffing levels based on the long-run average patient census by day of week and shift. In both cases, staff planning was done in 4-week increments and planned staffing levels were posted 2-weeks in advance of the first day of each 4-week plan. Consistent with the fact that average lengths of stay in these hospitals were between 4 and 5 days, staffing levels were not based on a projection of short-term demand forecast. When the number of patients exceeded the target nurse-to-patient ratios, nurse managers attempted to increase staffing by utilizing extra-time or overtime shifts, or calling in agency nurses. Similarly, when census was less than anticipated, nurses were assigned to indirect patient care tasks or education activities, or else asked to take voluntary time off. These efforts were not always successful and realized nurse-to-patient ratios often differed from the target ratios. For example, Hospital 1's unit T3 on average staffed lower than the target ratios, whereas Hospital 2's unit M2 on average staffed higher than the target ratios during weekends; see Figure 4.2.

In Figure 4.2, each dot's horizontal and vertical coordinates are the mean and the

Figure 4.2: Average Number of Patients Per Nurse by Shift Type



Note: D = Day. E = Evening. N = Night. TR 1 = Target ratio for T1, T2, and T3 in D and E shifts during weekdays. TR 2 = Target ratio for T1, T2, and T3 in N shifts or weekends; TR 2 is also the target ratio for M1 and M2 in D and E Shifts. TR 3 = Target ratio for M1 and M2 in N shifts.

standard deviation of the realized nurse-to-patient ratio for the corresponding unit and shift type. The bars show the 95% confidence intervals for the mean realized nurse-to-patient ratios. The higher the altitude of the dots, the more variable the nurse-to-patient ratio for that unit-shift combination. A numerical summary of these statistics is included in Table 4.2.

Table 4.2: Number of Patients per Nurse by Shift Type.

Shift Type	TR	Unit	Mean	SD	95% CI
Weekday	3	T1	2.98	0.60	(2.92, 3.03)
Day &	3	T2	3.25	0.63	(3.20, 3,31)
Evening shifts	3	T3	2.81	0.81	(2.73, 2.88)
Weekends	4	T1	3.85	1.04	(3.76, 3.94)
and	4	T2	3.77	1.29	(3.66, 3.88)
Night shifts	4	T3	3.04	0.67	(2.98, 3.09)
Day &	4	M1	2.99	0.56	(2.95, 3.03)
Evening Shifts	4	M2	3.14	0.39	(3.12, 3.17)
Night	5	M1	4.97	1.07	(4.90, 5.05)
Shifts	5	M2	5.21	0.80	(5.15, 5.27)

TR = target number of patients per nurse

SD = standard deviation. CI = confidence interval.

The absentee rate for Hospital 1's three inpatient units varied from 3.4% (T1, Sunday, Day Shift) to 18.3% (T1, Saturday, Night Shift) depending on unit, shift time, and

day of week. Among the 154 nurses who worked in the three units of Hospital 1, the average absentee rate was 10.78% with a standard deviation of 9%. The first and third quartile were 4.8% and 14.9% respectively. Similarly, the absentee rate for Hospital 2's two inpatient units varied from 2.99% (M1, Wednesday, Evening Shift) to 12.98% (M2, Tuesday, Evening Shift) depending on unit, shift time, and day of week. The 95% confidence intervals for absentee rates do not overlap across all units and shift types (Figures 4.3 and 4.4) but overlap across all days of week (see Table 4.3). These statistics suggest that absentee rates were significantly different by unit and shift, but not by the day of week. We also compared absentee rates for regular days and holidays, and fair-weather days and storm days. At 5% significance level, holidays had a lower average absentee rate than non holidays for Hospital 2 and storm days had a higher average absentee rate for Hospital 1. A numerical summary of absentee rates is included in Table 4.3.

Figure 4.3: Absentee Rate by Unit

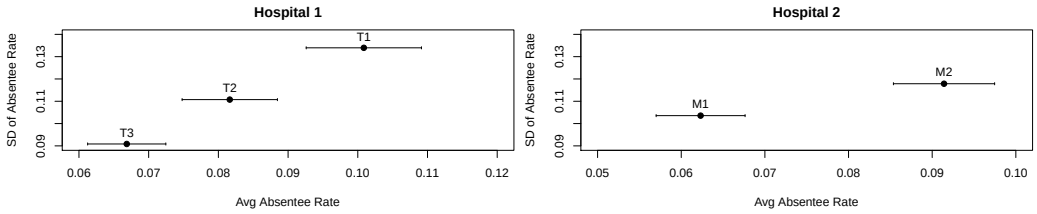


Figure 4.4: Absentee Rate by Shift Type

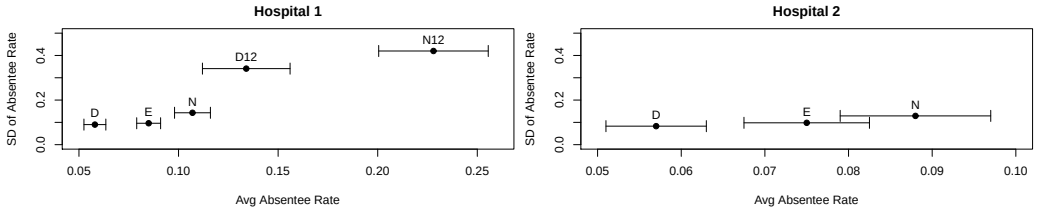


Table 4.3: Absentee Rate by Unit, Day of Week, Shift, Holidays, and Storm Days.

Hospital 1

Unit	Mean	SD	95% CI
T1	0.101	0.134	(0.093, 0.109)
T2	0.082	0.111	(0.075, 0.089)
T3	0.067	0.091	(0.061, 0.073)
DoW	Mean	SD	95% CI
Sun	0.077	0.118	(0.066, 0.089)
Mon	0.093	0.124	(0.081, 0.104)
Tue	0.077	0.103	(0.068, 0.087)
Wed	0.074	0.100	(0.065, 0.083)
Thu	0.086	0.111	(0.075, 0.096)
Fri	0.088	0.116	(0.077, 0.099)
Sat	0.086	0.123	(0.074, 0.098)
Shift	Mean	SD	95% CI
Day	0.058	0.090	(0.053, 0.064)
Evening	0.085	0.096	(0.079, 0.091)
Night	0.107	0.143	(0.098, 0.116)
Day 12-hour	0.134	0.341	(0.112, 0.156)
Night 12-hour	0.228	0.420	(0.201, 0.256)
Holiday[a]	Mean	SD	95% CI
Non holiday	0.083	0.113	(0.079, 0.087)
Holidays	0.084	0.138	(0.054, 0.114)
Storm Days[b]	Mean	SD	95% CI
No	0.082	0.114	(0.078, 0.087)
Yes	0.129	0.149	(0.088, 0.171)

Hospital 2

Unit	Mean	SD	95% CI
M1	0.062	0.104	(0.057, 0.068)
M2	0.091	0.118	(0.085, 0.098)
–	–	–	–
DoW	Mean	SD	95% CI
Sun	0.076	0.112	(0.065, 0.087)
Mon	0.082	0.120	(0.071, 0.094)
Tue	0.084	0.120	(0.072, 0.095)
Wed	0.072	0.108	(0.062, 0.083)
Thu	0.077	0.112	(0.066, 0.087)
Fri	0.073	0.102	(0.063, 0.083)
Sat	0.075	0.107	(0.064, 0.085)
Shift	Mean	SD	95% CI
Day	0.057	0.083	(0.051, 0.063)
Evening	0.075	0.098	(0.067, 0.082)
Night	0.088	0.129	(0.079, 0.097)
–	–	–	–
–	–	–	–
Holidays[a]	Mean	SD	95% CI
Non holiday	0.078	0.112	(0.074, 0.082)
Holiday	0.051	0.101	(0.032, 0.070)
Storm Days[b]	Mean	SD	95% CI
No	0.077	0.112	(0.073, 0.081)
Yes	0.10	0.126	(0.052, 0.143)

[a]Holidays include US federal holidays and the day before Thanksgiving and Christmas.

[b]Storm days were 2/26/09, 5/5/09, 8/2/09, 8/8/09, 10/12/09, 12/8/09, and 12/23/09 according to National Climatic Data Center (2012).

SD = Standard Deviation. CI = Confidence Interval.

In addition to the two hospitals whose data is analyzed in this chapter, the authors have interacted with nurse managers at numerous other urban hospitals. Staffing practices do vary from one hospital to another. However, the variation in practices prevalent at these two hospitals are representative of many other hospitals. That is, models based on such practices should be useful for other hospitals as well.

4.4 Statistical Models & Results

Our staffing objective is to minimize the expected shortage cost for any given level of nurse availability. We use $\boldsymbol{a}$ to denote a nurse assignment, where the (i, j)-th element is 1 if nurse i is assigned to work in unit j, and 0 otherwise. Therefore, the nurse assignment $\boldsymbol{a}$ needs to account for not only demand uncertainty, but also uncertainty concerning the number of nurses who show up for work in each unit. Let $\boldsymbol{Q}(\boldsymbol{a})$ denote the vector of the random number of nurses who show up for work for each unit under assignment $\boldsymbol{a}$. Consequently, it is important to evaluate the predictors of the function of $\boldsymbol{Q}(\boldsymbol{a})$. We next present two models to evaluate different predictors of nurse absenteeism. The choice of potential predictors was based on interactions with nurse managers and findings in previous studies.

4.4.1 Unit-Effects Model

In the first model, nurses' absentee rate for a particular shift is assumed to depend both on factors that are relatively stable and factors that vary. Factors in the former category include long-term average demand and staffing levels, unit culture, and desirability of certain shift start times. These factors are represented by fixed effects for unit, day of week, and shift. The factors that vary within our data are census levels and nurse availability, which is represented by the short-term anticipated workload w_t. We used three different versions of w_t in our analysis: (1) $w_t^{(1)} = n_t/E[C_t]$ and (2) $w_t^{(2)} = \sum_{i=1}^{m}(c_{t-m}/m)(1/n_t)$, and (3) $w_t^{(3)} = \sum_{i=1}^{m}(c_{t-m}/m)$, where n_t is the planned staffing level, c_t is the start-of-shift census for shift t, and $E[C_t]$ is the long-run expected census. Put differently, $w_t^{(1)}$ equals the anticipated nurse-to-patient ratio; $w_t^{(2)}$ equals the m-period moving average of estimated number of patients per nurse; and $w_t^{(3)}$ equals the m-period moving average census. The choice of $w_t^{(1)}$ is appropriate for units with stable

nursing demand, $w_t^{(2)}$ for units in which both census and staffing levels vary from shift to shift, and $w_t^{(3)}$ for units that have constant staffing levels (such as in Hospital 2).

Given unit index $i \in \{1, \cdots, u\}$, shift type $j \in \{1, \cdots, v\}$, and day-of-week $k \in \{1, \cdots, 7\}$, a logistic regression model was used to estimate p_t, the probability that a nurse will be absent in shift t. The parameter u equals three and two for Hospitals 1 and 2, respectively. The number of shift types v equals three for both hospitals where each shift index represents a different shift start time. Hospital 1's Day-12 shifts were classified as Day shifts because these shifts' start time was the same as Day shifts and 2/3 of the shift time overlapped with Day shifts. Although the shift start-time of Hospital 1's Night-12 shifts did not synchronize with any of the 8-hour shifts, we indexed these shifts as Night shifts because 2/3 of the shift time overlaps with Night shifts. Day-12 and Night-12 shifts account for a small fraction of the total number of nurse shifts (5.6% in each case). The models were evaluated with data that included in one instance and excluded in another the 12-hour shifts, and the conclusion did not change. However, the results reported in this section include these 12-hour shifts. Note that index t represents an unique shift in the data. Each shift t is mapped to exactly one (i, j, k) triplet. However, each (i, j, k) may map to several shifts with different shift indices. For example, different Monday Day shifts in the unit indexed 1 are associated with the same triplet $(1, 1, 1)$, but each such shift is assigned a different index t.

A full factorial model for estimating p_t is

$$\begin{aligned} log(\frac{p_t}{1-p_t}) &= \mu + \sum_{i=2}^{u} \beta_i U_i + \sum_{j=2}^{v} \alpha_j S_j + \sum_{k=2}^{7} \xi_k D_k + \rho H_t + \lambda Y_t + \gamma w_t \\ &\quad + \sum_{i=2}^{u}\sum_{j=2}^{v} \eta_{i,j}(U_i * S_j) + \sum_{i=2}^{u}\sum_{k=2}^{7} \vartheta_{i,k}(U_i * D_k) + \sum_{i=2}^{u} \iota_i(U_i * w_t) \\ &\quad + \sum_{j=2}^{v}\sum_{k=2}^{7} \varsigma_{j,k}(S_j * D_k) + \sum_{j=2}^{v} \phi_j(S_j * w_t) + \sum_{k=2}^{7} \nu_j(D_k * w_t) \\ &\quad + (\text{the remaining higher-order interaction terms}), \end{aligned} \tag{4.1}$$

where U_i, S_j, and D_k are indicator variables. In particular, $U_i = 1$ if the nurse under evaluation worked in unit i and $U_i = 0$ otherwise. Similarly, $S_j = 1$ (respectively $D_k = 1$) if the nurse was scheduled to work on a type-j shift (respectively day k of the week). H_t and Y_t are also indicator variables that are set equal to 1 if shift t

occurred either on a holiday or on a bad-weather day. Notation and assumptions are also summarized in Table 4.4. We use $(a*b)$ to denote the interaction term of a and b. In the ensuing analysis, all two-way interactions are included in the initial model but higher-order interaction terms are omitted. This is done because higher order interaction terms do not have a practical interpretation (see Faraway 2006 for supporting arguments).

The explanatory variables in (4.1) capture the systematic variation in nurses' absentee rates due to unit, shift time, day of week, and their interactions. Because long-term workload is included in these factors, we did not include that as a separate predictor. We also did not include week- or month-of-year effect because of data limitations[3] . The unit, shift, and day-of-week with the smallest indices are used as the benchmark group in the above model.

Table 4.4: Unit-Effects Model Notation and Assumptions

Covariate	Description	Coefficient
p_t	absentee rate for a shift t	none
U_i	indicator variable for unit i.	β_i
S_j	indicator variable for shift type j	α_j
D_k	indicator variable for day k of the week	ξ_k
H_t	indicator variable for holiday shifts	ρ
Y_t	indicator variable for storm-day shifts	λ
w_t	short-term anticipated workload for shift t	γ
$(U_i * S_j)$	unit and shift interaction	$\eta_{i,j}$
$(U_i * D_k)$	unit and day of week interaction	$\vartheta_{i,k}$
$(U_i * w_t)$	unit and workload interaction	ι_i
$(S_j * D_k)$	shift and day of week interaction	$\varsigma_{j,k}$
$(S_j * w_t)$	shift and workload interaction	ϕ_j
$(D_k * w_t)$	day of week and workload interaction	ν_j

Assumptions:

1. Independent and homogeneous nurses.

2. A nurse's attendance decision for a particular shift is independent of his/her decisions for other shifts.

[3] With approximately 1 year of data, observations of higher/lower absentee rate in certain weeks are not informative about future absentee rates in those weeks. Also, when week of year was included as an explanatory variable, this resulted in some covariate classes with too few observations. For example, there were only 3 nurses who were scheduled to work during week 2 (the week of 1/4/09 – 1/10/09) Monday Night shift in Unit 1 of Hospital 1.

We used stepwise variable selection processes to leave out insignificant explanatory factors by comparing nested models' deviances via Chi-square tests. Only factors that significantly improved model fits at the 5% significance level were retained in the model. A summary of our results with $m = 6$[4] is reported in Tables 4.5 and 4.6. To highlight that the impacts of these factors differ by unit and shift, we followed the prevailing norms for reporting such results and kept covariates of the same index group (unit index, shift index, or unit-shift index) in the summary so long as at least one in the group was statistically significant. We found that Unit and Shift effect were significant for both datasets (e.g. the estimated odds of being absent for T3's Day (resp. Night) shift was 27% lower (resp. 35% higher) than that for T1's Day shift), but bad weather effect was not. However, holiday effect was significant for Hospital 2's data. Two of three workload measures produced consistent results for both hospitals' data – neither $w_t^{(1)}$ nor $w_t^{(2)}$ were statistically significant. Anticipated workload was significant only for Hospital 2's data when $w_t = w_t^{(3)}$, and the coefficient was positive for one unit and negative for another. In particular, for every additional patient assigned to the unit, the odds of a nurse being absent would decrease by 2.5% for M1 and increase 1% for M2.. We also tried a variant of our model in which w_t was replaced by the m-shift realized nurse-to-patient ratios[5] for both datasets and the conclusion that there was not a consistent relationship between short-term anticipated workload and nurses' absenteeism remained intact. The absence of consistent relationship makes it difficult to incorporate short-term workload related absenteeism in staffing decisions.

Upon further examination, we found that the logistic regression model did not fit the data well and the goodness of fit test rejected the null hypothesis that the model was a good fit[6] – the two models in Tables 4.5 and 4.6 respectively resulted in a p-value of 0.029 and 0.011. This happened because of the large residual deviances of the unit-level model. The lack of fit may be caused by a variety of reasons. For example, it is possible

[4] We tested $w_t^{(2)}$ and $w_t^{(3)}$ with $m = 1, 2, \cdots, 12$. The results were consistent in the sign and significance of the parameters estimated and there was little variation in the estimated values of coefficients.

[5] Note that this assumes that nurses have advance knowledge of how many nurse shifts will be short relative to n_t after taking into account absences as well as management action to restore staffing levels in shift t.

[6] Nurses who are scheduled to work in the same shift have the same values of the covariates. Therefore, the data can be analyzed with these covariate classes as the grouped data. Because almost all shifts scheduled more than five nurses, the deviance is asymptotically Chi-square distributed and χ^2-statistic may be used to evaluate the goodness of fit.

Table 4.5: Hospital 1 Summary

Coefficients	Estimate	SE	Wald Test p-value
(Intercept)	-2.76	0.09	< 0.001
T2	0.159	0.12	0.182
T3	-0.317	0.12	0.010
Evening	0.499	0.11	< 0.001
Night	1.044	0.11	< 0.001
T2*Evening	-0.150	0.16	0.333
T3*Evening	0.184	0.16	0.236
T2*Night	-0.759	0.16	< 0.001
T3*Night	-0.430	0.16	0.009

Benchmark unit = T1; Benchmark shift = Day.
Null deviance: 3345.5 on 3023 degrees of freedom.
Residual deviance: 3163.3 on 3015 degrees of freedom.
Goodness of fit test: p-value = 0.029 (see footnote 6).

that unit-level factors/covariates do not adequately explain nurses' absentee rates, or that overdispersion[7] occurred due to non-constant probability within a covariate class (e.g. population heterogeneity). It is also possible that the unit, shift, and day of week patterns are confounded with individual nurses' work patterns – some high absentee rate nurses may have a fixed work pattern that contributed to the high absentee rates for some shifts. Some nurses also changed their work patterns during the data collection period, which might result in large deviance if nurse-specific effects were strong.

One of the critical assumptions underlying the generalized linear model (GLM) in Equation (4.1) is the statistical independence of observations; i.e. all observations, regardless of whether the attendance records belong to the same or different nurses, are assumed independent. This assumption may not hold when there is a natural clustering of the data and possible intra-cluster correlation. If there is a positive correlation among observations, the variance of the coefficients will be underestimated upon ignoring the covariances and inferences for these coefficients may be inaccurate.

[7] Overdispersion means that the variability around the model's fitted value is higher than what is consistent with the formulated model.

Table 4.6: Hospital 2 Summary ($w_t = w_t^{(3)}$)

Coefficients	Estimate	SE	Wald Test p-value
(Intercept)	-2.471	0.28	< 0.001
M2	-1.004	0.44	0.023
Evening	0.150	0.12	0.223
Night	0.644	0.11	< 0.001
Holiday	-0.413	0.18	0.019
w_t	-0.025	0.01	0.017
M2*Evening	0.208	0.15	0.176
M2*Night	-0.306	0.14	0.027
M2*w_t	0.035	0.01	0.006

Benchmark unit = M1; Benchmark shift = Day.
Null deviance: 3232.8 on 2907 degrees of freedom.
Residual deviance: 3076.1 on 2899 degrees of freedom.
Goodness of fit test: p-value = 0.011 (see footnote 6).

Hospital 1's data are comprised of multiple observations of each nurse's outcome (response) variable (whether a nurse was absent or present for a scheduled shift) and a set of unit-level covariates (e.g. unit index, shift time, day of week, workload, etc). Each nurse can be viewed as a cluster with correlation among the observations (attendance outcomes) of a nurse. Therefore, we also evaluated whether unit, shift, and day of week effects still exist after we accounted for individual nurses' effect. We used generalized estimating equations (GEE) to fit a repeated measure logistic regression model with Hospital 1's data. GEE, first introduced in Zeger and Liang (1986), is a method for analyzing correlated data. Although there are other statistical methods that may be used to evaluate repeated measures data (e.g. generalized linear mixed models, hierarchical generalized linear models), we chose GEE approach because this method has a track record of being useful in many applications and it has been implemented in several statistical packages (e.g. SAS, SPSS, R).

We applied the GEE approach to estimate the effect of unit-level factors (see Equation (4.1)). We estimated associations between outcomes of the same nurse ℓ, and

assumed that outcomes of different nurses were independent (i.e. nurses were still assumed to be independent decision makers). The number of observations differed across nurses resulting in unbalanced data. We had observations for 160 nurses. The average number of observations per nurse was 107.44, and the standard deviation of the number of shifts scheduled among these nurses was 52.23.

Table 4.7: Hospital 1 GEE Model Summary.

Coefficients	Estimate	SE	Wald Test p-value
(Intercept)	2.455	0.185	< 0.0005
T2	-0.015	0.205	0.942
T3	0.243	0.215	0.259
Evening	-0.543	0.136	< 0.0005
Night	-0.736	0.193	< 0.0005
Mon	-0.095	0.124	0.443
Tue	0.158	0.1232	0.199
Wed	0.254	0.1202	0.035
Thu	0.071	0.1055	0.501
Fri	0.127	0.1197	0.29
Sat	-0.126	0.0956	0.188

Benchmark unit = T1; Benchmark shift = Day;
Benchmark day of week = Sunday.

The results in Table 4.7 show that shift effect and Wednesday's day of week effect were significant while accounting for individual nurses' effect. However, unit effect was no longer significant. This observation was different from the model in which we assumed independent and homogeneous nurses. The differences in results from the GLM and GEE models suggest that it is not reasonable to ignore differences among nurses. Therefore, we next investigate a nurse-effects model and its implications for staffing decisions.

4.4.2 Nurse-Effects Model

We divided Hospital 1's staffing data into two periods – before and after June 30, 2009. There were 146 nurses who worked for more than 10 shifts in both periods.

Paired sample t-test showed that the average absentee rate did not change across these two periods, which indicates an overall stable absentee rate. Among these nurses, we calculated the absentee rate prior to June 30, 2009 for each nurse. The mean and median absentee rates among those nurses were 11.0% and 7.6%, respectively, and the standard deviation was 12.0%. For both time periods, we identified nurses whose absentee rates were higher (resp. lower) than 7.6% during the period and categorized these nurses as type-1 (resp. type-2) nurses for that period. The Phi coefficient was 0.43 with the two-by-two nurse classification for the two periods, which indicated a positive association between nurse types – nurses who were categorized as a particular type in period 1 were more likely to be categorized as the same type in period 2. For each shift, we model the impact of nurse-effects via the percent of type-1 nurses scheduled for that shift. We used the data between July 1st and December 4th, 2009 to evaluate the impact of having different proportions of period 1 type-1 nurses scheduled for a shift. We fitted the following model:

$$\begin{aligned} log(\frac{p_t}{1-p_t}) &= \mu + \sum_{i=2}^{u} \beta_i U_i + \sum_{j=2}^{v} \alpha_j S_j + \sum_{k=2}^{7} \xi_k D_k + \gamma w_t + \nu z_t \\ &\quad + (two-way\, interaction\, terms), \end{aligned} \tag{4.2}$$

where p_t is the group absentee rate for a given unit, shift, and day of week, w_t (measured by $w_t^{(1)}$, $w_t^{(2)}$, or $w_t^{(3)}$) is the workload, and $(100 \times z_t)\%$ of the scheduled nurses are type 1. Coefficients γ and ν in Equation (4.2) capture the workload and nurse-effects in each shift, whereas β_i, α_j, and ξ_i captures the effect of unit i, shift j, and day of week k relative to the benchmark group.

The response variable (nurses' attendance outcomes) were no longer assumed binomial because nurses were assumed to have heterogeneous absentee rates. Therefore, we fitted the regression model using quasi-binomial responses. We used F-tests to evaluate the impact of leaving each factor out of the model, and the results showed that the model could be simplified by excluding w_t for all three versions of w_t. The results, shown in Table 4.8, suggest that individual nurses' attendance history could be used to explain future shifts' absentee rates.

Table 4.8: Nurse-Effects Model Summary

	Estimate	Std. Error	Wald z value	Wald Test p-value
(Intercept)	-2.79933	0.23806	-11.759	$<$ 2e-16
T2	-0.02670	0.19205	-0.139	0.889436
T3	0.11006	0.18103	0.608	0.543324
Evening	0.26054	0.17216	1.513	0.130427
Night	0.72032	0.20591	3.498	0.000484
z_t	1.11769	0.23110	4.836	1.48e-06
$w_t^{(1)}$	-0.46005	0.51143	-0.900	0.368533
T2*Evening	0.26639	0.23416	1.138	0.255471
T3*Evening	-0.03931	0.22963	-0.171	0.864089
T2*Night	-0.65075	0.27097	-2.402	0.016466
T3*Night	-0.71374	0.25606	-2.787	0.005390

Benchmark unit = T1; Benchmark shift = Day.

Estimated dispersion parameter: 1.042.

Null deviance: 1583.8 on 1323 degrees of freedom.

Residual deviance: 1460.0 on 1314 degrees of freedom.

The take away from Section 4.4 is that a model that assumes nurses are homogeneous and their decisions independent across shifts does not fit our data well. Moreover, short-term workload either does not explain shift absentee rate or its effect is both positive and negative (which makes it non-actionable for nurse managers). In contrast, if nurses are assumed to be heterogenous but consistent decision makers, then nurse-effects explain shift absentee rates reasonably well.

In the next section, we model the nurse staffing problem with heterogeneous absentee rates to gain insights into the nature of optimal assignments. In Section 4.6, we develop heuristics to solve the assignment problem.

4.5 Model Formulation and Analysis

Suppose there are u inpatient units that require nurses with a particular skill set, n nurses with this skill set are available, and these nurses can be divided into m types based on their absentee rates. In particular, nurses that belong to the same type have the same probability of being absent in an arbitrary shift. The objective is to minimize expected

total cost. We consider three models: (1) deterministic, (2) random aggregate, and (3) nurse-specific Bernoulli. These models are denoted by letters d, r and b, respectively and additional notation needed for model formulation is presented in Table 4.9.

Table 4.9: Notation

Indices		
j	=	nurse index, $j = 1, \cdots, n$
t	=	nurse type index, $t = 1, \cdots, m$; $m \leq n$
i	=	unit index, $i = 1, \cdots, u$
Parameters		
$\boldsymbol{p}$	=	$(p_1, \cdots, p_m)$, absent probabilities by nurse type
$\boldsymbol{n}$	=	$(n_1, \cdots, n_m)$, number of nurses by type
$\boldsymbol{X}$	=	$(X_1, \cdots, X_u)$, (random) nursing needs vector
$c_0(\cdot)$	=	an increasing convex shortage cost function
Decision Variables		
$\boldsymbol{a}^{(i)}$	=	$(a_1^{(i)}, \cdots, a_m^{(i)})$, number of nurses assigned to unit i by type
$\boldsymbol{a}$	=	$(\boldsymbol{a}^{(1)}, \cdots, \boldsymbol{a}^{(u)})$, staffing plan
$\mathcal{A}$	=	set of all possible assignments, $\mathcal{A} \in \times_{i=1}^{u}([0, n_1] \times [0, n_2] \cdots \times [0, n_m])$
Calculated Quantities		
$\phi_i(\boldsymbol{a})$	=	the number of absent nurses in unit i given assignment $\boldsymbol{a}$
$\phi_i^{[k]}(\boldsymbol{a})$	=	$\phi_i(\boldsymbol{a})$ under absentee model k, $k \in \{d, r, b\}$
$Q_i(\boldsymbol{a})$	=	$\sum_{t=1}^{m} a_t^{(i)} - \phi_i(\boldsymbol{a})$ = number of nurses who show up in unit i
$\pi_i(\boldsymbol{a})$	=	total expected shortage cost for Unit i when staffing plan $\boldsymbol{a}$ is used
$\pi(\boldsymbol{a})$	=	total expected shortage cost when staffing plan $\boldsymbol{a}$ is used $= \sum_{i=1}^{u} \pi_i(\boldsymbol{a})$

The first model assumes that there is a deterministic mapping from $\boldsymbol{a}^{(i)}$ to the number of absentees, i.e. $\phi_i(\boldsymbol{a}) = \phi_i^{[d]}(\boldsymbol{a}) \in [0, \sum_{t=1}^{m} a_t^{(i)}]$. A deterministic model is appropriate when nurse absences are predictable. It also serves as a benchmark mean value approximation of the underlying stochastic optimization problem (Birge and Louveaux 2011). In the random aggregate model, nurse absence is viewed as a group characteristic such that there is a type-specific random absence uncertainty ξ_t for each type-t nurse that is independent of $a_t^{(i)}$, and $\phi_i^{[r]}(\boldsymbol{a}, \boldsymbol{\xi}) = \sum_{t=1}^{m} g_t(a_t^{(i)}, \xi_t)$. Note that this is similar to the multiplicative random yield model studied extensively in the OM literature

(see e.g. Yano and Lee 1995). The third model considers each nurse as an independent decision maker with a no-show probability p_t. Note that in this case $m = n$ and $n_t = 1$ because each nurse is a type. The number of nurses who are absent equals $\phi_i(\boldsymbol{a}) = \phi_i^{[b]}(\boldsymbol{a}) = \sum_{t=1}^{n} a_t^{(i)} B(p_t)$, where $B(p)$ is a Bernoulli random variable with parameter p, and each $a_t^{(i)}$ is either 0 or 1.

The staffing-plan optimization problem is now formulated as follows.

$$\min_{\boldsymbol{a}} \pi(\boldsymbol{a}) = \sum_{i=1}^{u} E\left(c_0(X_i - Q_i(\boldsymbol{a}))^+\right), \text{ subject to } \sum_{i=1}^{u} a_t^{(i)} \leq n_t \text{ and } \boldsymbol{a} \in \mathcal{A}. \quad (4.3)$$

In what follows, we index nurse types such that $p_1 \geq p_2 \cdots \geq p_m$ and evaluate the effect of demand variability and absentee rate variability on the performance of an assignment. For this purpose, we use concepts from the theory of stochastic orders and majorization. These concepts are described briefly next. Further details can be found in Shaked and Shanthikumar (2007) and Marshall et al. (2011). Given $\boldsymbol{X}$ and $\boldsymbol{X'}$ nursing requirement vectors, $\boldsymbol{X} \leq_{icx} \boldsymbol{X'}$ in a component-wise manner if $E[g(X_i) \leq E[g(X_i')]$ for every increasing convex function g for which the expectations exist and every $i = 1, \cdots, u$. Similarly, given vectors $\boldsymbol{p}$ and $\boldsymbol{p'}$, where the components of these vectors are indexed such that $p_1 \geq p_2 \cdots \geq p_m$, $p_1' \geq p_2' \cdots \geq p_m'$, and $\sum_{t=1}^{m} p_t = \sum_{t=1}^{m} p_t'$, we say that vector $\boldsymbol{p}$ is majorized by $\boldsymbol{p'}$, written $\boldsymbol{p} \leq_M \boldsymbol{p'}$ if $\sum_{t=1}^{\ell} p_t \leq \sum_{t=1}^{\ell} p_t'$ for every $\ell \leq m$. Let $\{Z(\theta), \theta \in \Theta\}$ be a family of random variables with survival functions $\bar{F}_\theta(z) = P(Z(\theta) > z), \theta \in \Theta$. The family $\{Z(\theta), \theta \in \Theta\}$ is said to be stochastically increasing and linear in the sense of usual stochastic order, denoted SIL(st), if $E[g(Z(\theta)]$ is increasing linear for all increasing functions g. It is easy to verify that the family $\{B(p), p \in (0,1)\}$ is SIL(st).

We use asterisk notation to denote optimal quantities. In particular, $\boldsymbol{a}_\alpha^*$ denotes an optimal assignment of nurses when the problem is characterized by the problem parameter α. For example, suppose $\boldsymbol{X}$ and $\boldsymbol{X'}$ denote two different nurse requirement vectors. Then, $\boldsymbol{a}_{\boldsymbol{X}}^*$ and $\boldsymbol{a}_{\boldsymbol{X'}}^*$ are used to denote optimal staffing plans with nurse requirements $\boldsymbol{X}$ and $\boldsymbol{X'}$. With these notation in hand, we carry out certain stochastic comparisons and obtain the following results.

Proposition 4.5.1 *If $\boldsymbol{X} \leq_{icx} \boldsymbol{X'}$ in a component-wise sense, then $\pi(\boldsymbol{a}_{\boldsymbol{X}}^*) \leq \pi(\boldsymbol{a}_{\boldsymbol{X'}}^*)$.*

Proposition 4.5.1 states that larger and more variable demand leads to greater expected shortage costs. This statement applies to all three models. The proof of Proposition 4.5.1 is included in Appendix C. The result in Proposition 4.5.1 is intuitive because for a fixed level of available staff, shortage costs increase when demand is larger and more uncertain.

Proposition 4.5.2 *Let $\phi_i(\boldsymbol{a}) = \phi_i^{[d]}(\boldsymbol{a})$ be a deterministic mapping from $\boldsymbol{a}^{(i)}$ to $[0, \sum_{t=1}^m a_t^{(i)}]$. If $\{X_i\}$ are independent and identically distributed (i.i.d.), then an optimal staffing plan is realized upon making $Q_i(\boldsymbol{a}^*)$ equal for all $i = 1, \cdots, m$.*

Proposition 4.5.2 states that if absenteeism can be predicted reasonably well and inpatient units have i.i.d. demand, then the nurse manager should assign nurses such that each unit has the same realized staffing level. A proof of Proposition 4.5.2 is also presented in Appendix C.

For the random aggregate absence model, we assume that the number of absentees upon scheduling $a_t^{(i)}$ type-t nurse in unit i is determined by a function $g_t(a_t^{(i)}, \xi_t)$, where ξ_t represents the uncertainty regarding the number of type-t absentees. In particular, we assume $\phi_i^{[r]}(\boldsymbol{a}, \boldsymbol{\xi}) = \sum_{t=1}^m g_t(a_t^{(i)}, \xi_t) = \sum_{t=1}^m [g_{t,1}(a_t^{(i)}) + g_{t,2}(a_t^{(i)})\xi_t]$, and for each ξ_t, $g_t(a, \xi_t) \leq a$, $0 \leq g_t'(a, \xi_t) \leq 1$, and $g_t''(a, \xi_t) \leq 0$. In the following analysis, we assume that shortage is linear in the number of shifts short, i.e. $c_o(x) = c_o \cdot x$ for every x for ease of exposition.

Let $\psi_i(\{a_t^{(i)}\}) = x_i - \sum_{t=1}^m a_t^{(i)} + \sum_{t=1}^m g_t(a_t^{(i)}, \xi_t)$ denote the difference between demand and availability of nurses given realizations x_i and ξ_t. Then staffing problem can be written as follows:

$$\min \sum_{i=1}^u E\left(c_0 \cdot (X_i - Q_i(\boldsymbol{a}))^+\right) = \sum_{i=1}^u E\left(c_0 \cdot (\psi_i(\{a_t^{(i)}\}))^+\right), \tag{4.4}$$

subject to:

$$\sum_{i=1}^u a_t^{(i)} = n_t, \qquad \text{for each } t = 1, \cdots, m. \tag{4.5}$$

Each function $\psi_i(\{a_t^{(i)}\})$ is decreasing concave in each $a_t^{(i)}$. Furthermore, because $\psi_i(\{a_t^{(i)}\})$ consists of functions that are separable in $a_t^{(i)}$, it is also jointly concave in $a_t^{(i)}$. The function $(\cdot)^+$ is increasing convex in its argument. The composition of an increasing

function and a decreasing function is a decreasing function. From above, we infer that $(\psi_i(\{a_t^{(i)}\}))^+$ is decreasing in $a_t(i)$. Furthermore, realizations x_i and ξ_t are independent of the choice of $a_t^{(i)}$. Therefore, $E\left(c_0 \cdot (\psi_i(\{a_t^{(i)}\}))^+\right)$ is also decreasing in $a_t^{(i)}$. Finally, $\sum_{i=1}^{u} E\left(c_0 \cdot (\psi_i(\{a_t^{(i)}\}))^+\right)$ is decreasing in $a_t^{(i)}$ because the sum of decreasing functions is a decreasing function.

Because the problem is to minimize a decreasing function under linear constraints, the first-order necessary Karush-Kuhn-Tucker (KKT) conditions imply that there exist γ_k's that are unrestricted in sign and

$$c_0 E\left(\frac{\partial(\psi_i(\{a_t^{(i)}\}))^+}{\partial a_k^{(i)}}\right) + \gamma_k = 0, \qquad \text{for each } k \text{ and each } i. \tag{4.6}$$

The derivative of $\psi_i(\{a_t^{(i)}\}))^+$ with respect to $a_k^{(i)}$ is zero if $x_i \leq \sum_{t=1}^{m} a_t^{(i)} - \sum_{t=1}^{m} g_t(a_t^{(i)}, \xi_t)$ and $-1+g_k'(a_k^{(i)}, \xi_t)$ otherwise. Therefore, Equation (4.6) can be rewritten as follows:

$$\gamma_k = c_0 \cdot E_{X_i}\left[E_{\xi_t}\left(1 - g_k'(a_k^{(i)}, \xi_t) \mid X_i + \sum_{t=1}^{m} g_t(a_t^{(i)}, \xi_t) > \sum_{t=1}^{m} a_t^{(i)}\right)\right]. \tag{4.7}$$

Equation (4.7) can be interpreted as follows. The right-hand side of Equation (4.7) is the expected rate of decrease in unit i's shortage cost as a function of the staffing level of type-k nurses when unit i experiences nurse shortage. Note that the left-hand side does not depend on unit index i. This means that under an optimal allocation, the expected rate of decrease in each unit's shortage cost as a function of the staffing level of each nurse type when that unit experiences nurse shortage should be the same. When X_i's are i.i.d., one way to achieve this equality is to set $a_t^{(i)} = a_t^{(j)}$ for each pair (i, j), and $\sum_{t=1}^{m} a_t^{(i)} = n_t$ for every t. That is, to staff such that each unit has the same number of type-t nurses, for each t. This may not be the only optimal solution, but it is a straightfroward allocation that achieves optimality. We found such an allocation to be optimal for the second example we considered in the Introduction section.

A special case of the random aggregate model arises when $g_{t,1}(a_t^{(i)}) = 0$ and $g_{t,2}(a_t^{(i)}) = a_t^{(i)}$. In this case, for a given $\boldsymbol{a}$, we show in Proposition 4.5.3 that a nurse manager would prefer a cohort of nurses with a stochastically smaller absentee rate (i.e. nurses with smaller mean and/or variance of aggregate absentee rate). A proof of Proposition 4.5.3 is presented in Appendix C.

Proposition 4.5.3 *Let* $\phi_i(\boldsymbol{a}) = \sum_{t=1}^{m} a_t^{(i)} \xi_t$, *where* $\xi_i \in [0,1]$ *is the random absentee rate for type-t nurses and* $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_m)$. *For each fixed* $\boldsymbol{a}$, *if* $\boldsymbol{\xi} \leq_{icx} \boldsymbol{\xi'}$ *in a component-wise manner, then* $\pi(\boldsymbol{a}^*_{\xi}) \leq \pi(\boldsymbol{a}^*_{\xi'})$.

So far we have shown that a nurse manager prefers smaller and less variable absentee rates for each nurse type from which it needs to choose a certain number of nurses. Next, we address a different question. Given a cohort of heterogeneous nurses, which nurses should a nurse manager choose to realize less variable absentee rate for his or her unit? To avoid situations where the manager of each unit would want only nurses that never take unplanned time off, we fix the total aggregate absentee rate that any choice of nurses must satisfy for that unit. We use the concept of majorization to answer this question. We define $\boldsymbol{p}^{(i)}$ to be the absentee probabilities of nurses assigned to unit i. That is, components of $\boldsymbol{p}^{(i)}$ contain information about only those nurses that are assigned to unit i. Furthermore, let $\pi_i(\boldsymbol{p}^{(i)})$ denote the expected shortage cost incurred in unit i with no-show probability vector $\boldsymbol{p}^{(i)}$. Then, with individual Bernoulli no-show model, we can prove that a nurse manager would prefer a more variable mix of absentee rates.

Proposition 4.5.4 *If* $\boldsymbol{p}^{(i)} \leq_M \boldsymbol{p'}^{(i)}$, *then* $\pi_i(\boldsymbol{p'}^{(i)}) \leq \pi_i(\boldsymbol{p}^{(i)})$.

A proof of Proposition 4.5.4 is included in Appendix C. Proposition 4.5.4 shows that for fixed overall absentee rate and number of nurses, the nurse manager would prefer to utilize a more heterogeneous cohort of nurses. However, this may not be the best overall strategy when costs across different units need to be balanced. It is the difficulty of balancing staffing across units while maximizing heterogeneity within a unit that makes it difficult to identify an optimal assignment strategy.

Next, we show that the results in Propositions 4.5.3 and 4.5.4 are related. If we view each nurse type as a unit, then Proposition 4.5.4 establishes that greater heterogeneity within a cohort of nurses leads to smaller variability in attendance pattern of that cohort taken together, which makes it more desirable according to Proposition 4.5.3. A proof of Corollary 4.5.5, which establishes this correspondence between Propositions 4.5.3 and 4.5.4, is presented in Appendix C.

Corollary 4.5.5 *Let* $\bar{B}_i(\boldsymbol{p'}) = \frac{\sum_{t=1}^{n} a_t^{(i)} B(p_t)}{\sum_{t=1}^{n} a_t^{(i)}}$. *If* $\boldsymbol{p} \leq_M \boldsymbol{p'}$, *then* $\bar{B}_i(\boldsymbol{p'}) \leq_{icx} \bar{B}_i(\boldsymbol{p})$.

The structural results in Propositions 4.5.2-4.5.4 and Corollary 4.5.5 show that for inpatient units with identical demand patterns, hospitals' costs are lower when (i) nurse assignments are as heterogenous as possible within a unit, but (2) uniform across units. Section 4.6 contains heuristics that attempt to achieve assignments that are consistent with these principles.

Propositions 4.5.2 and 4.5.4 can be used to explain observations in the second example of Section 4.1. We considered two staffing strategies. Both were consistent with Proposition 4.5.2 and maintained the same expected number of nurses who show up in each unit. The absentee rate vector (0.1, 0.1, 0.1, 0.1, 0.1) is majorized by the absentee rate vector (0, 0.1, 0.1, 0.1, 0.2), so the Unit 1's performance is worse under Strategy 2. However, the absentee rate vector (0, 0.1, 0.1, 0.1, 0.2) is majorized by the absentee rate vector (0, 0, 0.1, 0.2, 0.2), so Unit 2's performance is worse under Strategy 1. Finally, Strategy 1 emerges as the overall best choice because the two units are identical and shortage costs are increasing convex. Note that the two strategies assign equal number of nurses to each unit.

4.6 Heuristics and Performance Comparisons

Solving (4.3) with nurse-specific Bernoulli representation of absenteeism is a combinatorially hard problem. Therefore, we propose three heuristics in this section, which can be used to obtain assignment of nurses to inpatient units. We also test these heuristics in numerical experiments. In this section, we assume that the nurse manager has divided nurses into a manageable number of classes and for the purpose of making assignment decisions, nurses belonging to the same group are treated as having identical no-show probability. In particular, this means that n_t could be greater than 1 and therefore $a_t^{(i)} \leq n_t$ could be greater than 1 as well. Also, $\phi^{[b]}(\boldsymbol{a}) = \sum_{t=1}^{n} \sum_{j=1}^{a_t^{(i)}} B_j(p_t)$, where $B_j(p_t)$ are i.i.d. Bernoulli random variables with parameter p_t.

First, we show that the objective function in (4.3) is supermodular, which helps to motivate the heuristics in the sequel. We define $\delta_t^{(i)}(\boldsymbol{a}) = \pi(\boldsymbol{a}) - \pi(\boldsymbol{a} + e_{ti})$, where $\boldsymbol{e}_{ti}$ is a $m \times u$ matrix with the (t, i)-th component equal to 1 and the remaining components equal to 0, as the incremental benefit of adding a type-t nurse to unit i. Then, it can be shown that

- $\delta_t^{(i)}(\boldsymbol{a}) \geq 0 \ \forall \ t = 1, \cdots, m$, and $\forall \ i = 1, \cdots, u$. In addition, when no-show probabilities are ordered such that $p_1 \geq p_2 \cdots \geq p_m$, $\delta_t^{(i)}(\boldsymbol{a}) \leq \delta_{t'}^{(i)}(\boldsymbol{a})$ for $t \leq t'$.
- $\delta_t^{(i)}(\boldsymbol{a}) - \delta_t^{(i)}(\boldsymbol{a} + \boldsymbol{e}_{t'i}) \geq 0 \ \forall \ t' = 1, \cdots, m$ and $\forall \ i = 1, \cdots, u$. Note that $\delta_t^{(i)}(\boldsymbol{a}) - \delta_t^{(i)}(\boldsymbol{a} + \boldsymbol{e}_{t'i})$ is the difference in incremental benefits of adding a type-t nurse to unit i under two situations: one in which this unit had $a_{t'}^{(i)}$ type-t' nurses and another in which the number of type-t' nurses was $a_{t'}^{(i)} + 1$.

These results have a straightforward intuitive explanation. The first bullet confirms that it is better to add a nurse with a lower absentee rate. The second bullet says that the benefit (reduction in cost) of adding one more type-t nurse to a particular unit diminishes in the number of type-t' nurses in the unit when the staffing levels of the remaining groups are held constant (note that t' is an arbitrary type, which includes t). Together, these observations imply that the objective function in (4.3) is supermodular (Topkis 1998).

The reason why supermodularity is relevant is that a greedy heuristic has been shown to work well when the objective function is monotone supermodular (Topkis 1998). Therefore, the nurse manager may wish to sort nurses by increasing absentee rates, and assign them to different units sequentially to maximize the marginal benefit from each assignment (i.e. in a greedy fashion) until all nurses are exhausted. For the problem of adjusting staffing levels, the nurse manager can accept extra shift volunteers in the sequence dictated by union rules until the cost of adding the next volunteer is higher than the expected benefit. We call this strategy the greedy assignment. It can be argued that when nurses have identical no-show probabilities, the greedy strategy results in an optimal assignment. We omit the details in the interest of brevity. We also propose two other assignment strategies, as described below.

$\boldsymbol{H}_1$: Greedy Assignment: If there is no pre-determined assignment sequence among a group of nurses, assign one nurse at a time to a unit that generates the highest expected marginal benefit. If there is a pre-determined sequence and each assignment incurs a cost, assign nurses according to the sequence until the expected marginal benefit is at least as large as the cost.

$\boldsymbol{H}_2$: Arbitrary Assignment: This strategy randomly assigns each nurse to the two units

while ensuring that the total number assigned to each unit is proportional to the expected demand for that unit.

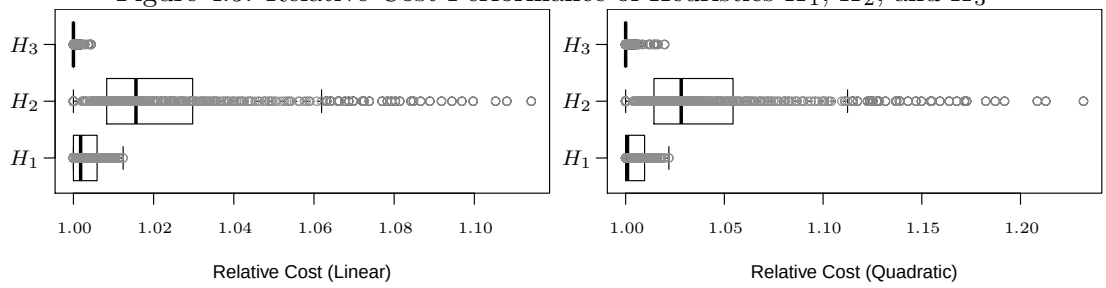
$\boldsymbol{H}_3$: Balanced Assignment: This strategy searches for an assignment that minimizes the difference in the expected demand-supply ratios (i.e. $E(X_i)/E(Q_i(\boldsymbol{a}))$ across units. When there are multiple assignments that result in identical expected staffing levels, any one of the balanced assignments is picked at random.

All ensuing comparisons are performed with the assumption that there are two units with independent and identically distributed nursing requirements and that nurses may be grouped into two types with absentee rates p_1 and p_2. In that case, each unit's staffing level under H_2 will be either $\lceil (n_1+n_2)/2 \rceil$ or $\lfloor (n_1+n_2)/2 \rfloor$ and H_3 will minimize the absolute difference between $(a_1^{(1)}(1-p_1)+a_2^{(1)}(1-p_2))$ and $(a_1^{(2)}(1-p_1)+a_2^{(2)}(1-p_2))$.

In computational experiments, we fixed the total number of nurses to be $n = n_1 + n_2 = 15$, and varied n_1 from 0 to 15. When $n_1 = 0$ or $n_1 = 15$, nurses have homogeneous absentee rates. We also varied p_1 from 0 to 0.2 in 0.05 increments. The show probability for type-2 nurses were set as $(1-p_2) = \theta(1-p_1)$, where θ was varied from 0.5 to 0.9 in 0.1 increments. Nurse requirements were assumed to be Poisson distributed and independent across units with rate $\lambda = (n_1(1-p_1)+n_2(1-p_2))/2$, ensuring that overall mean requirements and supply were matched. This experimental design resulted in 400 scenarios. For each heuristic, we compared its expected shortage cost relative to the optimal cost upon assuming two shortage cost functions: (1) linear, i.e. $c_o(x_i-q)^+ = x_i - q$ if $x_i \geq q$ and 0 otherwise, or (2) quadratic, i.e. $c_o(x_i-q)^+ = (x_i-q)^2$ if $x_i \geq q$ and 0 otherwise. The optimal assignment and associated minimum expected shortage cost were obtained through an exhaustive search over all possible assignments.

The performance for the heuristics in terms of relative expected costs are shown as box plots in Figure 4.5 with each grey dot representing an outcome. H_2 is dominated by other solution approaches. Both H_1 and H_3 perform quite well. There is statistically no difference in average performance of H_1 and H_3, but H_3 performs better than H_1 in more problem instances (see Table 4.10 for pairwise comparisons and detailed summary statistics for numerical experiments). However, H_3 requires knowledge of all available nurses up front, whereas H_1 can be used when we must assign one nurse at a time in a pre-determined sequence (as happens in the first example in Section 4.1). Thus, nurse

Figure 4.5: Relative Cost Performance of Heuristics H_1, H_2, and H_3



managers may choose between H_1 and H_3 depending on the problem scenarios that they face. H_2 can also be viewed as a strategy that treats a heterogeneous cohort of nurses as a homogeneous cohort of nurses in the staffing decision. The numerical study shows that by taking into account nurses' heterogeneous absentee rates (H_1 and H_3), the nurse manager can reduce shortage costs by 2% to 4% depending on the cost structure. This result is promising because nurse staffing costs are a significant fraction of a hospital's overall operating expenses.

Table 4.10: Performance Comparisons

Linear Cost function

Cost Relative to the Optimal Cost

	H_1	H_2	H_3
Avg	100.03%	102.24%	100.01%
SD	0.34%	2.18%	0.05%

Pairwise Comparisons*

	H_1	H_2	H_3
H_1	–	0%	59.8%
H_2	87.5%	–	87.5%
H_3	2.3%	0%	–

Quadratic Cost function

Cost Relative to the Optimal Cost

	H_1	H_2	H_3
Avg	100.50%	104.15%	100.05%
SD	0.58%	4.19%	0.21%

Pairwise Comparisons*

	H_1	H_2	H_3
H_1	–	0%	52.75%
H_2	87.5%	–	87.50%
H_3	9.5%	0%	–

*The left (resp. right) panel shows the performance comparisons under a linear (resp. quadratic) cost function. Within each panel, the upper table reports the mean and standard deviation of the ratio of the cost associated with each heuristic and the optimal cost, expressed in percent. Each cell in the lower table summarizes the percent of scenarios in which the column strategy performed better than the row strategy.

In summary, this chapter shows that nurse managers may use each nurse's attendance history to predict his or her likelihood of being absent in a future shift. This information can be utilized within easy-to-implement staffing heuristics, e.g. heuristics labeled H_1 and H_3, to reduce staffing costs. The use of this approach does not require significant effort on part of nurse managers.

The contribution of this chapter lies in (1) developing detailed analyses of data from multiple inpatient units and multiple hospitals to identify observable predictors of nurse absenteeism, and (2) establishing structural properties of optimal assignment strategies that lead to easy-to-implement heuristics for use by nurse managers. The mathematical models presented in this chapter are motivated by nurse staffing decisions, but they are not domain specific. Such models and proposed heuristics are applicable in a whole host of situations involving assignment of staff with heterogeneous attendance patterns to teams responsible for different pieces of work.

The future research directions emerging from this chapter include modeling the impact of unit and/or culture effects on nurse absenteeism and staffing decisions. For instance, if social loafing (e.g. some nurses may be more likely to be absent when more low absentee rate nurses are scheduled) may occur, whether heterogeneity in nurses' absentee rates may still benefit inpatient staffing and how should nurse managers staff their units under this circumstance. Similarly, whether peer pressure (e.g. some nurses may be less likely to be absent when they are schedule with colleagues who are rarely absent from work) would lead to a different long-term staffing strategy. This chapter also draws attention to whether there are other actionable absenteeism predictors that can be used to improve staffing decisions.

Chapter 5

Conclusion

Health care expenditure in the U.S. has reached $2.6 trillion, which is about 17.9% of the nation's Gross Domestic Product (Center for Medicare & Medicaid Services 2012). The high expenditure points to a need to focus on providing quality care by utilizing available resources in an efficient manner. A significant operations management problem in health care delivery systems lies in matching uncertain demand and supply. This includes estimating demand, re-distributing demand to the extent possible to match existing capacity, choosing staffing levels or allocating capacity in a way that can better respond to demand uncertainty, managing workforce absenteeism and incorporating unplanned absences in staffing decisions, and/or a combination of all of the above challenges. In this dissertation, we discuss several capacity management challenges. In particular, we discuss how clinics may assign heterogeneous patients to different physician panels to minimize the cost due to unserved demand while maximizing revenue, how to book randomly arriving patients into different appointment slots while accounting for anticipated demand and patients' preferences, and how to assign nurses to different nursing units based on nurses' attendance history.

A common feature for these problems is the benefit of obtaining relevant information and utilizing that information in capacity decisions. In this dissertation, we emphasize not only operations models, but also the inputs for these models. The performance of an operations model depends on whether the inputs and the assumptions to the model are representative of the reality. For example, it is important for clinics to obtain information regarding physicians' panel sizes, panel composition, and patients'

appointment preferences regarding the choice of physician and appointment time of day in order to evaluate physicians' workload. In real time appointment booking systems, it is essential to obtain patients' preference information in order to hedge against the uncertainty of appointment requests from advance-book patients and same-day urgent patients. For inpatient nurse staffing, it is beneficial to understand nurses' absentee patterns based on observable data such as nurses attendance history or types of shifts scheduled, and use that information to assign nurses to meet the staffing requirements. For this reason, we discuss the types of data that are (or may be) collected in existing systems, and how these data may be utilized to generate insights for modeling purposes.

We address a clinic profile setup problem for ambulatory clinics that serve randomly arriving patients who may wish to book appointments in advance or for the same day in Chapter 2. In particular, we investigate a panel design problem where two identical physicians need to serve two groups of patients and the clinic's goal is to determine a patient allocation scheme to maximize utilization while minimizing unserved demand. The results show that although equal patient assignment may be an intuitive choice, it does not always dominate an allocation that has some degrees of specialization. The lessons from the panel design study are as follows. Clinics's capacity allocation decisions interact with the composition of patients in each physician panel. For example, the types of patients could be very different for different physician panels and this affects the capacity allocation decisions. Therefore, clinics need to pay attention to patients' composition when allocating patients to different physician panels. This chapter serves to highlight the complexity and importance of the clinic profile setup problem.

Some future research directions based on Chapter 2 include incorporating features that allow patients to book with non-PCPs, and modeling the impact of having strategic non-urgent patients who identify themselves as urgent to obtain same-day appointment when they cannot secure a desirable appointment during the advance-book period. It is also an interesting research topic to extend the panel design concept to patient-centered medical home environments.

We also emphasize that a physician panel's patient composition will have an impact on the efficiency of an outpatient appointment booking system as discussed in Chapter 3. For any given clinic profile (and panel design), we can evaluate the impact of patients' preferences regarding physician and time of day on appointment booking decisions. The

proposed methods for estimating patients' preferences will converge within a short data collection period. The proposed heuristics that partially characterize the optimal booking policy produce respectable performance not only in terms of expected revenue, but also patient-PCP match rate, number of patients served, and capacity spoilage rate. In addition, clinics can continuously update patients' preference information upon receiving their appointment requests. A research question that emerges from this work is whether patients will react to a clinic's appointment scheduling rules. If so, how many patients will change their behavior or booking preferences, and whether the adaptive appointment system will respond to these changes accordingly in a non-stationary environment. Future research directions that emerge from Chapter 3 point to modeling the interaction between appointment scheduling rules and patients' booking strategies, evaluating whether clinics can benefit from getting additional preference information such as each patient's rank ordering of different physician/time combinations in each booking attempt, and investigating whether clinics may achieve better performance if the appointment system adopts a different booking scheme such as offering patients a set of available appointments in the appointment process.

For inpatient services, we tackle a challenge for staffing registered nurses to different nursing units in Chapter 4. In this context, there are forecasting challenges in estimating nursing requirement distribution for inpatient units due to data availability (e.g. indicators of demand censoring) and subjective nurse requirements due to inpatients' case-mix/health conditions. In the investigation of observable nurse absenteeism predictors that can be used to account for supply uncertainty in staffing decisions, we show that there are group-level factors that may contribute to nurses' absenteeism, and nurses' absentee rates are heterogeneous and relatively stable over a short period of time for the same individuals. We show that nurse managers can utilize individual nurses' absentee rates to improve nurse assignment decision. The results in this study also highlight the importance of paying attention to modeling assumptions as different assumptions about nurse absenteeism could suggest very different operational strategies. This chapter also provides heuristics that can be applied to situations that involve assigning staff of heterogeneous absence patterns to different projects.

Findings in Chapter 4 lead to an interesting question of whether nurses will react

differently if its recommendations were to be implemented. In particular, whether heterogeneity in nurses' absentee rates leads to better or worse staffing outcomes when social loafing or peer pressure (i.e. some nurses' may become more or less likely to be absent when they are scheduled with responsible colleagues) may occur. Whether there are other unit-level and individual-level absenteeism predictors that can lead to better staffing decisions.

This dissertation also leads to several future research avenues that are not addressed in the examples studied here. For example, how can hospitals utilize fuzzy censored demand information of different nursing units to improve capacity allocation decisions? What are the conditions that are suitable for hospitals to operate with specialized nursing units? What are the conditions for hospitals to benefit from using universal beds that are suitable for patients with different types of nursing needs? How would patients react to different appointment system designs? How would physicians' workload change if the clinic applies the panel-design concept to patient-centered medical homes?

To sum up, the capacity management problems discussed in this dissertation focus on directions that may be used to improve the efficiency of a healthcare delivery system. We answer the questions at the level such that the models are constructed to improve the utilization of a given level of resources and capacity by better matching demand and supply. At a higher level, clinics may also wish to answer questions regarding what is the right capacity or amount of resources to be used to cope with demand and supply uncertainty to make hiring decisions. At a finer level, clinics may wish to answer the question regarding what is the best patient composition to suit a particular physician's specialty or work pattern, or what is an optimal schedule given some nurses' time-off requests. Although these capacity problems may all be interdependent, most of them need to be divided into smaller problems in order to be analyzed. The capacity problems analyzed in this dissertation provide insights about important factors that affect performance of capacity allocation.

References

Abernathy, W. J., Ni. Baloff, J. C. Hershey, S. Wandel. 1973. Three-stage manpower planning and scheduling model – a service-sector example. *Operations Research* **21**(3) 693–711.

Aiken, L. H., S. P. Clarke, D. M. Sloane, J. Sochalski, J. H. Siber. 2002. Hospital nurse staffing and patient mortality, nurse burnout and job dissatisfaction. *The Journal of the American Medical Association* **288**(16) 1987 – 1993.

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Anonymous. 2010. 2010 Emergency Department Pulse Report – Patient Perspectives on American Health Care. Available at: http://goo.gl/vz4eI, downloaded 1/31/2012.

Barron, W.M. 1980. Failed appointments. Who misses them, why they are missed, and what can be done. *Primary Care* **7**(4) 563–574.

Birge, J. R., F. Louveaux. 2011. *Introduction to stochastic programming*. 2nd ed. Springer New York; Berlin; Heidelberg. SpringerLink (Online service).

Blumenfeld, D. E., R. R. Inman. 2009. Impact of absenteeism on assembly line quality and throughput. *Production and Operations Managment* **18**(3) 333–343.

Brusco, M. J., J. Futch, M. J. Showalter. 1993. Nurse staff planning under conditions of a nursing shortage. *Journal of Nursing Administration* **23**(7/8) 58–64.

Bureau of Labor Statistics. 2011. Household data annual averages. *Current Population Survey* Abailable at: http://www.bls.gov/cps/cpsaat47.pdf.

California Nurse Association, National Nurses Organizing Committee. 2012. Floating according to the rules. *Nursing Practice & Patient Advocacy Alert* Available at http://goo.gl/nMqcp, downloaded 4/19/2012.

Carlin, B. P., T. A. Louis. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd ed. New York: Chapman and Hall.

Carlson, B. 2002. Same-day appointments promise increased productivity. *Managed Care* **11**(12) 43–44.

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: A review of literature. *Production & Operations Management* **12**(4) 519 – 549.

Cayirli, T., E. Veral, H. Rosen. 2008. Assessment of patient classification in appointment system design. *Production & Operations Management* **17**(3) 338 – 353.

Center for Medicare & Medicaid Services. 2012. National health expenditure data. Available at http://goo.gl/qsPph, downloaded 7/16/2012.

Cheraghi-Sohi, S., A. Hole, N. Mead, R. McDonald, D. Whalley, P. Bower, M. Roland. 2008. What patients want from primary care consultations: A discrete choice experiment to identify patients' priorities. *Annals of Family Medicine* **6**(2) 107–115.

Cho, S-H, S. Ketefian, V. H. Barkauskas, D. G. Smith. 2003. The effects of nurse staffing on adverse events, morbidity, mortality, and medical costs. *Nursing Research* **52**(2) 71 – 79.

Connelly, M. 2003. Chrysler group cracks down on absenteeism. *Automotive News* Available at http://goo.gl/PcfIi, downloaded 5/29/2012.

Côté, M. J., S. L. Tucker. 2001. Four methodologies to improve healthcare demand forecasting. *Healthcare Financial Management* **55**(5) 54–58.

Davey, M. M., G. Cummings, C. V. Newburn-Cook, E. A. Lo. 2009. Predictors of nurse absenteeism in hospitals: a systematic review. *Journal of nursing management* **17**(3) 312–30.

Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* **35**(11) 1003–1016.

Dervin, J. V., D. L. Stone, C. H. Beck. 1978. The no-show patient in the model family practice unit. *Journal of Family Practice* **7**(6) 1177–1180.

Doescher, M. P., B. G. Saver, K. Fiscella, P. Franks. 2004. Preventive care: does continuity count? *Journal of General Internal Medicine* **19**(6) 632 – 637.

Dove, H. G., K. C. Schneider. 1981. The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics. *Medical Care* **XIX**(7) 734–740.

Earnest, A., M. I. Chen, D. Ng, L. Y. Sin. 2005. Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research* **5**(36).

Faraway, J. J. 2006. *Extending the linear model with R – generalized linear, mixed effects and nonparametric regression models*. Chapman & Hall/CRC – Boca Raton, FL.

Ferlise, P., D. Baggot. 2009. Improving staff nurse satisfaction and nurse turnover: use of a closed-unit staffing model. *Journal of Nursing Administration* **39**(7/8) 318–320.

Fosarelli, P., C. DeAngelis, A. Kaszuba. 1985. Compliance with follow-up appointments generated in a pediatric emergency room. *American Journal of Preventive Medicine* **1**(3) 23–29.

Gallucci, G., W. Swartz, F. Hackerman. 2005. Brief reparts: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services* **56** 344–346.

Gerard, K., C. Salisbury, D. Street, C. Pope, H. Baxter. 2008. Is fast access to general practice all that should matter? A discrete choice experiment of patients' preferences. *Journal of Health Services Research & Policy* **13** 3–10.

Green, L., S. Savin, N. Savva. 2011. "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. Working paper. Available at http://goo.gl/G401z, downloaded 1/29/2012.

Green, L. V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* **56**(6) 1526–1538.

Green, L. V., S. Savin, M. Murray. 2007. Providing timely access to care: what is the right patient panel size? *Joint Commission Journal on Quality and Patient Safety* **33**(4) 211 – 218.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* **40** 800–819.

Gupta, D., L. Wang. 2008. Revenue management for a primary care clinic in the presence of patient choice. *Operations Research* **56**(3) 576–592.

Gupta, D., W-Y. Wang. 2011. Patient appointments in ambulatory care. R. W. Hall, ed., *Handbook of Healthcare System Scheduling: Deliverying Care When and Where It is Needed*, chap. 4. Springer, NY.

Gupta, D., W-Y. Wang, E. Tokar-Erdemir, S. Potthoff, L. P. Ettestad, S. Tredal, C Tupy. 2011. Nurse staffing for inpatient care. Working Paper.

Hassin, R., S. Mendel. 2008. Scheduling arrivals to queues: A single-server model with no-shows. *Management Science* **54**(3) 565–572.

Ho, C-J, H-S Lau. 1992. Minimizing total cost in scheduling outpatient appointments. *Management Science* **38**(12) 1750–764.

Huh, T., R Levi, P. Rusmevichientong, J. Orlin. 2009. Adaptive data-driven inventory control policies based on kaplan-meir estimator. *Working paper* .

Huh, W. T., P. Rusmevichientong. 2009. A Nonparametric Asymptotic Analysis of Inventory Planning with Censored Demand. *Mathematics of Operations Research* **34**(1) 103–123.

Hur, D., V. A. Mabert, K. M. Bretthauer. 2004. Real-time work schedule adjustment decisions: An investigation and evaluation. *Production and Operations Management* **13**(4) 322–339.

Irwin, C. E., Jr., S. G. Millstein, M.-A. B. Shafer. 1981. Appointment-keeping behavior in adolescents. *The Journal of Pediatrics* **99**(5) 799–802.

Jennings, B. M., L. A. Loan, S. L. Heiner, E. A. Hemman, K. M. Swanson. 2005. Soldiers' experiences with military health care. *Military Medicine* **170**(12) 999–1004.

Johnson, C. J., E. Croghan, J. Crawford. 2003. The problem and management of sickness absence in the NHS: considerations for nurse managers. *Journal of nursing management* **11**(5) 336–42.

Kaandorp, G. C., G. Koole. 2007. Optimal outpatient appointment scheduling. *Health Care Management Science* **10** 217–229.

Kane, R. L., T. A. Shamliyan, C. Mueller, S. Duval, T. J. Wilt. 2007. The association of registered nurse staffing levels and patient outcomes. systematic review and meta-analysis. *Medical Care* **45**(12) 1195 – 1204.

Kao, E. P. C., G. G. Tung. 1980. Forecasting demands for inpatient services in a large public health care delivery system. *Socio-Economic Planning Science* **14**(2) 97–106.

Kaplan, E. L., P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**(282) 457–481.

Ken, F. 2006. 'Patient' says it all. *USA Today* Available at http://goo.gl/w1ufF, downloaded 1/31/2012.

Kim, S., R.E. Giachetti. 2006. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transections on Systems, Man and Cybernetics–Part A: Systems and Humans* **36**(6) 1211–19.

Klassen, K. J., T. R. Rohleder. 1996. Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management* **14**(2) 83 – 101.

Klein, J., M. L. Moeschberger. 2005. *Survival Analysis*. New York: Springer-Verlag.

LaGanga, L. R., S. R. Lawrence. 2007. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences* **38**(2) 251–276.

Lang, T. A., M. Hodge, V. Olson, P. S. Romano, R. L. Kravitz. 2004. Nurse–patient ratios: A systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes. *The Journal of Nursing Administration* **34**(7-8) 326 – 337.

Lee, V., A. Earnest, M. Chen, B. Krishnan. 2005. Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. *BMC Health Services Research* **5**(1) 51.

Lim, G. J., A. Mobasher, L. Kardar, M. J. Côté. 2011. Nurse scheduling. R. W. Hall, ed., *Handbook of Healthcare System Scheduling: Deliverying Care When and Where It is Needed*, chap. 3. Springer, NY, 31 – 64.

Liu, N., S. Ziya, V. G. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management* **12**(2) 347–364.

Liyanage, L., J. G. Shanthikumar. 1992. Allocation through stochastic schur convexity and stochastic transposition increasingness. *Lecture Notes-Monograph Series* **22** pp. 253–273.

Marshall, A. W., I. Olkin, B. C. Arnold. 2011. *Inequalities: Theory of Majorization and Its Applications*. 2nd ed. Springer, NY.

May, J. H., G. J. Bazzoli, A. M. Gerland. 2006. Hospitals' responses to nurse staffing shortages. *Health Affairs* **25**(4) W316–W323.

McFadden, D. 2001. Economic choices. *American Economic Review* **91**(3) 351–378.

Müller, A., D. Stoyan. 2002. *Comparison Methods for Stochastic Models and Risks*. John Wiley & Sons: New York, NY.

Murray, M., D. M. Berwick. 2003. Advanced Access: reducing waiting and delays in primary care. *Journal of the American Medical Association* **289**(8) 1035–1040.

Murray, M., T. Bodenheimer, D. Rittenhouse, K. Grumbach. 2003. Improving timely access to primary care: Case studies of the advanced access model. *Journal of the American Medical Association* **289**(8) 1042–1046.

Murray, M., M. Davies, B. Boushon. 2007. Panel size: how many patients can one doctor manage? *Family practice management* **14**(4) 44–51.

Murray, M., C. Tantau. 2000. Same-day appointments: Exploding the access paradigm. *Family Practice Management* **7**(8) 45–50.

Muthuraman, K., M. Lawley. 2008. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transections* **40** 820–837.

National Center for Health Statistics. 2011. Health, United States, 2010: With Special Feature on Death and Dying. Available at http://www.cdc.gov/nchs/data/hus/hus10.pdf, downloaded 1/31/2012.

National Climatic Data Center. 2012. NCDC Storm Event Database. Data available at http://goo.gl/V3zWs, downloaded 1/12/2012.

Needleman, J., P. Buerhaus, S. Mattke, M. Stewart, K. Zelevinsky. 2002. Nurse-staffing levels and the quality of care in hospitals. *The New England Journal of Medicine* **346**(22) 1715 – 1722.

Neinstein, L. S. 1982. Lowering broken appointment rates at a teenage health center. *Journal of Adolescent Health Care* **3**(2) 110–113.

O'Hare, C. D., J. Corlett. 2004. The outcomes of open-access scheduling. *Family Practice Management* **11**(1) 35–38.

Olowokure, B., M. Caswell, H.V. Duggal. 2006. What women want: convenient appointment times for cervical screening tests. *European Journal of Cancer Care* **15** 489–492.

Ordones, J. 2000. An efficiency drive: Fast-food lanes, equipped with timers, get even faster. *USA Today* Available at http://goo.gl/kXVrY, downloaded 5/29/2012.

Parente, D. H., M. B. Pinto, J. C. Barber. 2005. A pre-post comparison of service operational efficiency and patient satisfaction under open access scheduling. *Health Care Management Review* **30**(3) 220–228.

Robinson, L. W., R. R. Chen. 2003. Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions* **35**(3) 295–307.

Robinson, L. W., R. R. Chen. 2010a. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management* **12** 330–346.

Robinson, L. W., R. R. Chen. 2010b. Estimating the implied value of the customer's waiting time. *Manufacturing & Service Operations Management.* Forthcoming.

Rohleder, T. R., K. J. Klassen. 2000. Using client-variance information to improve dynamic appointment scheduling performance. *Omega* **28**(3) 293 – 302.

Russell, L. B. 2009. Completing costs: Patients' time. *Medical Care* **47**(7) S89–S93.

Salisbury, C., A. A. Montgomery, L. Simons, F. Sampson, S. Edwards, H. Baxter, S. Goodall, H. Smith, V. Lattimer, D. M. Pickin. 2007. Impact of Advanced Access on access, workload, and continuity: Controlled before-and-after and simulated-patient study. *British Journal of General Practice* **57**(541) 608–614.

Savin, S. 2006. Managing patient appointments in primary care. S. Hillier, F, R. W. Hall, eds., *Patient Flow: Reducing Delay in Healthcare Delivery*, *International Series in Operations Research & Management Science*, vol. 91. Springer US, 123–150.

Shaked, M., J. G. Shanthikumar. 2007. *Stochastic orders*. New York : Springer.

Shumway, Robert H., David S. Stoffer. 2006. *Time Series Analysis and Its Applications*. New York, NY: Springer-Verlag.

Smith, C. M., B. P. Yawn. 1994. Factors associated with appointment keeping in a family practice residency clinic. *Journal of Family Practice* **38**(1) 25–29.

Snow, B. W., P. C. Cartwright, S. Everitt, M. Ekins, W. Maudsley, S. Aloi. 2009. A method to improve patient access in urological practice. *The Journal of Urology* **182**(2) 663 – 667.

Starkenburg, R. J., F. Rosner, K. Crowley. 1988. Missed appointments among patients new to a general medical clinic. *New York State Journal of Medicine* **88**(9) 437–435.

Talluri, K., G. van Ryzin. 2004. Revenue management under a general discrete choice model of consumer behavior. *Management Science* **50**(1) 15–33.

Topkis, D. 1998. *Supermodularity and complementarity*. 2nd ed. Princeton University Press, New Jersey.

Train, K. E. 2003. *Discrete choice methods with simulation*. New York: Cambridge University Press.

Unruh, L. 2008. Nurse staffing and patient, nurse, and financial outcomes. *American Journal of Nursing* **108**(1) 62–71.

U.S. Bureau of Labor Statistics. 2008. Consumer expenditures in 2006. Tech. Rep. 1010, U.S. Department of Labor.

Vanden Bosch, P. M., D. C. Dietz. 2000. Minimizing expected waiting in a medical appointment system. *IIE Transactions* **32**(9) 841 – 848.

Wang, P. P. 1999. Sequencing and scheduling n customers for a stochastic server. *European Journal of Operational Research* **119**(3) 729 – 738.

Wang, W-Y., D. Gupta, S. Potthoff. 2009. On evaluating the impact of flexibility enhancing strategies on the performance of nurse schedules. *Health Policy* **93** 188–200.

Weiss, E. N. 1990. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions* **22**(2) 143–150.

Whitt, W. 2006. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* **15**(1) 88–102.

Whittle, J., G. Schectman, N. Lu, B. Baar, M. F. Mayo-Smith. 2008. Relationship of scheduling interval to missed and cancelled clinic appointments. *Journal of Ambulatory Care Management* **31**(4) 290–302.

Wood, S. D. 1976. Forecasting patient census: commonalities in time series models. *Health Services Research* **11**(2) 158–165.

Yabroff, K. R., J. L. Warren, K. Knopf, W. W. Davis, M. L. Brown. 2005. Estimating patient time costs associated with colorectal cancer care. *Medical Care* **43**(7) 640–648.

Yano, C. A., H. L. Lee. 1995. Lot sizing with random yields: A review. *Operations Research* **43**(2) pp. 311–334.

Zeger, S. L., K-Y. Liang. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**(1) 121 – 130.

Appendix A

Proofs for Chapter 2

A.1 Proof for Corollary 2.2.1

For $\beta > \kappa/(mz^*)$ and $(1-\beta) > \kappa/(mz^*)$, i.e. $\kappa_1^* = 0$ and $\kappa_2^* = 0$ and $\kappa/(mz^*) < \beta < 1 - \kappa/(mz^*)$. This case only exists when $\kappa/(mz^*) < 0.5$.

$$\begin{aligned}\Pi &= r_1 nE(Y_1) + r_2 mE(Y_2) - (r_1+\pi)\int_0^1 ny_1 f(y_1)dy_1 \\ &\quad -(r_2+c)\int_{\frac{\kappa}{\beta m}}^1 (\beta m y_2 - \kappa)g(y_2)dy_2 \\ &\quad -(r_2+c)\int_{\frac{\kappa}{(1-\beta)m}}^1 ((1-\beta)my_2 - \kappa)g(y_2)dy_2. \end{aligned} \tag{A.1}$$

$$\frac{\partial \Pi}{\partial \beta} = (r_2+c)\int_{\frac{\kappa}{(1-\beta)m}}^{\frac{\kappa}{\beta m}} my_2 g(y_2)dy_2. \tag{A.2}$$

Note that the revenue function increases (respectively decreases) in β when $\beta < 0.5$ (respectively $\beta > 0.5$). Note that $\kappa/(mz^*) < 0.5$. Therefore, $\beta^* = 0.5$. In this case, because no capacity is reserved for type-1 patients for both panels, the optimal α is any value between 0 and 1. That is, case 1's expected revenue does not depend on α.

A.2 Proof for Corollary 2.2.2

The revenue function for cases 2a and 2b can be written as follows.

$$
\begin{aligned}
\Pi = {} & r_1 n E(Y_1) + r_2 m E(Y_2) \\
& - (r_1 + \pi) \int_0^1 \alpha n y_1 f(y_1) dy_1 \\
& - (r_2 + c) \int_{\frac{\kappa}{\beta m}}^1 (\beta m y_2 - \kappa) g(y_2) dy_2 \\
& - (r_1 + \pi) \int_{\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}}^1 ((1-\alpha)ny_1 - \kappa + (1-\beta)mz^*) f(y_1) dy_1 \\
& - (r_2 + c) \int_{z^*}^1 (1-\beta)m(y_2 - z^*) g(y_2) dy_2 \bar{F}\Big(\frac{\kappa - (1-\beta)mz^*}{(1-\alpha)n}\Big) \\
& - (r_2 + c) \int_0^{\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)ny_1}{(1-\beta)m}}^1 ((1-\beta)my_2 - \kappa + (1-\alpha)ny_1) dG(y_2) dF(y_1).
\end{aligned}
\tag{A.3}
$$

$$
\begin{aligned}
\frac{\partial \Pi}{\partial \beta} &= -(r_2+c)\int_{\frac{\kappa}{\beta m}}^{1} m y_2 g(y_2) dy_2 \\
&\quad +(r_1+\pi)\int_{\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}}^{1} mz^* f(y_1) dy_1 \\
&\quad +(r_2+c)\int_{z^*}^{1} m(y_2-z^*) g(y_2) dy_2 \bar{F}\Big(\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}\Big) \\
&\quad +(r_2+c)\int_{z^*}^{1} (1-\beta)m(y_2-z^*) g(y_2) dy_2 f\Big(\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}\Big)\frac{mz^*}{(1-\alpha)n} \\
&\quad -(r_2+c)\frac{mz^*}{(1-\alpha)n}\int_{z^*}^{1} (1-\beta)m(y_2-z^*) g(y_2) dy_2 f\Big(\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}\Big) \\
&\quad +(r_2+c)\int_{0}^{\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)ny_1}{(1-\beta)m}}^{1} m y_2 g(y_2) f(y_1) dy_2 dy_1 \\
&= -(r_2+c)\int_{\frac{\kappa}{\beta m}}^{1} m y_2 g(y_2) dy_2 \\
&\quad +(r_2+c)\int_{z^*}^{1} m y_2 g(y_2) dy_2 \bar{F}\Big(\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}\Big) \\
&\quad +(r_2+c)\int_{0}^{\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)ny_1}{(1-\beta)m}}^{1} m y_2 g(y_2) f(y_1) dy_2 dy_1.
\end{aligned}
\tag{A.4}
$$

With some algebra,

$$
\begin{aligned}
\frac{\partial^2 \Pi}{\partial \beta^2} &= -(r_2+c)\frac{\kappa^2}{\beta^3 m} g\Big(\frac{\kappa}{\beta m}\Big) \\
&\quad -(r_2+c)\int_{0}^{\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}} \frac{(\kappa-(1-\alpha)ny_1)^2}{(1-\beta)^3 m} f(y_1) dy_1 < 0.
\end{aligned}
\tag{A.5}
$$

Therefore, the optimal β is a function of α that ensures equation (A.6) equals 0. That is, $\beta^* = h(\alpha)$. Note that $\kappa_1^* = 0$, therefore, $\alpha^* = 0$ because no capacity is reserved for type-1 patients for the first panel. We replace α by 0 in equation (A.6). Then we

obtain

$$\begin{aligned}
\frac{\partial \Pi}{\partial \beta} &= -(r_2+c)\int_{\frac{\kappa}{\beta m}}^{1} m y_2 g(y_2) dy_2 \\
&\quad +(r_2+c)\int_{z^*}^{1} m y_2 g(y_2) \bar{F}\Big(\frac{\kappa-(1-\beta)mz^*}{n}\Big) \\
&\quad +(r_2+c)\int_{0}^{\frac{\kappa-(1-\beta)mz^*}{n}} \int_{\frac{\kappa-ny_1}{(1-\beta)m}}^{1} m y_2 g(y_2) dy_2 f(y_1) dy_1. \\
&\leq -(r_2+c)\int_{\frac{\kappa}{\beta m}}^{1} m y_2 g(y_2) dy_2 \\
&\quad +(r_2+c)\int_{z^*}^{1} m y_2 g(y_2) \bar{F}\Big(\frac{\kappa-(1-\beta)mz^*}{n}\Big) \\
&\quad +(r_2+c)\int_{0}^{\frac{\kappa-(1-\beta)mz^*}{\alpha n}} \int_{z^*}^{1} m y_2 g(y_2) dy_2 f(y_1) dy_1 \\
&= -(r_2+c)\int_{\frac{\kappa}{\beta m}}^{1} m y_2 g(y_2) dy_2 \\
&\quad +(r_2+c)\int_{z^*}^{1} m y_2 g(y_2) dy_2 \\
&\leq 0. \qquad \text{(A.6)}
\end{aligned}$$

Therefore, the revenue function decreases in β for Cases 2a and 2b. Note that $\beta \geq \max\{\kappa/(mz^*), 1-\kappa/(mz^*)\}$. Therefore, the optimal patient assignment under this scenario is $\alpha = 0$ and $\beta = \max\{\kappa/(mz^*), 1-\kappa/(mz^*)\}$.

A.3 Proof for Corollary 2.2.3

For $\beta \leq \kappa/(mz^*)$ and $(1-\beta) \leq \kappa/(mz^*)$, i.e. $\kappa_1^* = \kappa - \beta m z^*$ and $\kappa_2^* = \kappa - (1-\beta)mz^*$. Note that $(1-\kappa/(mz*)) \leq \beta \leq \kappa/mz^*$ and this scenario only exists when $0.5 \leq \kappa/(mz^*) \leq 1$.

$$
\begin{aligned}
\Pi = {} & r_1\alpha n E(Y_1) + r_2\beta m E(Y_2) - (r_1+\pi)\int_{\frac{\kappa-\beta m z^*}{\alpha n}}^{1}(\alpha n y_1 - \kappa + \beta m z^*) f(y_1)dy_1 \\
& - (r_2+c)\int_{z^*}^{1}(\beta m y_2 - \beta m z^*) g(y_2)dy_2 \bar{F}\Big(\frac{\kappa-\beta m z^*}{\alpha n}\Big) \\
& - (r_2+c)\int_{0}^{\frac{\kappa-\beta m z^*}{\alpha n}}\int_{\frac{\kappa-\alpha n y_1}{\beta m}}^{1}(\beta m y_2 - \kappa + \alpha n y_1) g(y_2) f(y_1)dy_2 dy_1 \\
& + r_1(1-\alpha)nE(Y_1) + r_2(1-\beta)mE(Y_2) \\
& - (r_1+\pi)\int_{\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}}^{1}((1-\alpha)ny_1 - \kappa + (1-\beta)mz^*) f(y_1)dy_1 \\
& - (r_2+c)\int_{z^*}^{1}((1-\beta)my_2 - (1-\beta)mz^*) g(y_2)dy_2 \bar{F}\Big(\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}\Big) \\
& - (r_2+c)\int_{0}^{\frac{\kappa-(1-\beta)mz^*}{(1-\alpha)n}}\int_{\frac{\kappa-(1-\alpha)ny_1}{(1-\beta)m}}^{1}((1-\beta)my_2 - \kappa + (1-\alpha)ny_1) dG(y_2)dF(y_1).
\end{aligned}
\tag{A.7}
$$

Therefore,

$$\begin{aligned}
\frac{\partial \Pi}{\partial \beta} &= r_2 mE(y_2) - (r_1+\pi)\int_{\frac{\kappa-\beta m z^*}{\alpha n}}^{1} m z^* f(y_1)dy_1 \\
&\quad -(r_2+c)\int_{z^*}^{1} m(y_2-z^*)g(y_2)dy_2 \bar{F}\Big(\frac{\kappa-\beta m z^*}{\alpha n}\Big) \\
&\quad -(r_2+c)\int_{0}^{\frac{\kappa-\beta m z^*}{\alpha n}} \int_{\frac{\kappa-\alpha n y_1}{\beta m}}^{1} m y_2 g(y_2) f(y_1) dy_2 dy_1 \\
&\quad -r_2 mE(y_2) + (r_1+\pi)\int_{\frac{\kappa-(1-\beta)m z^*}{(1-\alpha)n}}^{1} m z^* f(y_1)dy_1 \\
&\quad +(r_2+c)\int_{z^*}^{1} m(y_2-z^*)g(y_2)dy_2 \bar{F}\Big(\frac{\kappa-(1-\beta)m z^*}{(1-\alpha)n}\Big) \\
&\quad +(r_2+c)\int_{0}^{\frac{\kappa-(1-\beta)m z^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)n y_1}{(1-\beta)m}}^{1} m y_2 g(y_2) f(y_1) dy_2 dy_1 \\
&= (r_2+c)\int_{z^*}^{1} m y_2 g(y_2)dy_2 \Big[\bar{F}\Big(\frac{\kappa-(1-\beta)m z^*}{(1-\alpha)n}\Big) - \bar{F}\Big(\frac{\kappa-\beta m z^*}{\alpha n}\Big)\Big] \\
&\quad -(r_2+c)\int_{0}^{\frac{\kappa-\beta m z^*}{\alpha n}} \int_{\frac{\kappa-\alpha n y_1}{\beta m}}^{1} m y_2 g(y_2) f(y_1) dy_2 dy_1 \\
&\quad +(r_2+c)\int_{0}^{\frac{\kappa-(1-\beta)m z^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)n y_1}{(1-\beta)m}}^{1} m y_2 g(y_2) f(y_1) dy_2 dy_1. \qquad \text{(A.8)}
\end{aligned}$$

In addition,

$$\begin{aligned}
\frac{\partial^2 \Pi(\alpha,\beta,\kappa_1)}{\partial \beta^2} &= -(r_2+c)\Big[\int_{0}^{\frac{\kappa-\beta m z^*}{\alpha n}} \frac{(\kappa-\alpha n y_1)^2}{\beta^3 m} g\Big(\frac{\kappa-\alpha n y_1}{\beta m}\Big) f(y_1)dy_1 \\
&\quad + \int_{0}^{\frac{\kappa-(1-\beta)m z^*}{(1-\alpha)n}} \frac{(\kappa-(1-\alpha)n y_1)^2}{(1-\beta)^3 m} g\Big(\frac{\kappa-(1-\alpha)n y_1}{(1-\beta)m}\Big) dF(y_1)\Big] \\
&\quad \le 0
\end{aligned}$$

(A.9)

Therefore, the revenue function Π is concave in β. This results suggests that for any allocation of type-1 patients, there is an optimal allocation of type-2 patients when we set equation (A.8) equals zero. Consequently, for any given α, the optimal β is a function of α, i.e. $\beta^* = h(\alpha)$.Then we can also re-organize equation (A.8) and set it to

zero to obtain that

$$\begin{aligned}
&\int_{z^*}^{1} y_2 g(y_2) dy_2 \Big[\bar{F}\Big(\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}\Big) - \bar{F}\Big(\frac{\kappa-h(\alpha)mz^*}{\alpha n}\Big)\Big] \\
&- \int_0^{\frac{\kappa-h(\alpha)mz^*}{\alpha n}} \int_{\frac{\kappa-\alpha n y_1}{h(\alpha)m}}^{1} m y_2 g(y_2) f(y_1) dy_2 dy_1 \\
&+ \int_0^{\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)ny_1}{(1-h(\alpha))m}}^{1} m y_2 g(y_2) f(y_1) dy_2 dy_1 = 0. \qquad \text{(A.10)}
\end{aligned}$$

Take the derivative with respect to α for equation (A.10), and we can show that $h'(\alpha) < 0$. This suggests that the more type-1 patients are assigned to a panel, the fewer type-2 patients are assigned to the same panel.

Next we rewrite the revenue function by substituting β by $\beta^* = h(\alpha)$.

$$\begin{aligned}
\Pi \;=\;& r_1 n E(Y_1) + r_2 m E(Y_2) \\
&-(r_1+\pi)\int_{\frac{\kappa-h(\alpha)mz^*}{\alpha n}}^{1} (\alpha n y_1 - \kappa + h(\alpha)mz^*) f(y_1) dy_1 \\
&-(r_2+c)\int_{z^*}^{1} (h(\alpha)my_2 - h(\alpha)mz^*) g(y_2) dy_2 \bar{F}\Big(\frac{\kappa-h(\alpha)mz^*}{\alpha n}\Big) \\
&-(r_2+c)\int_0^{\frac{\kappa-h(\alpha)mz^*}{\alpha n}} \int_{\frac{\kappa-\alpha n y_1}{h(\alpha)m}}^{1} (h(\alpha)my_2 - \kappa + \alpha n y_1) g(y_2) f(y_1) dy_2 dy_1 \\
&-(r_1+\pi)\int_{\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}}^{1} ((1-\alpha)ny_1 - \kappa + (1-h(\alpha))mz^*) f(y_1) dy_1 \\
&-(r_2+c)\Big[\\
&\int_{z^*}^{1} ((1-h(\alpha))my_2 - (1-h(\alpha))mz^*) g(y_2) dy_2 \bar{F}\Big(\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}\Big) \\
&- \int_0^{\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)ny_1}{(1-h(\alpha))m}}^{1} ((1-h(\alpha))my_2 - \kappa) g(y_2) f(y_1) dy_2 dy_1. \\
&+ \int_0^{\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)ny_1}{(1-h(\alpha))m}}^{1} (1-\alpha)ny_1 g(y_2) f(y_1) dy_2 dy_1 \Big]. \qquad \text{(A.11)}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \Pi}{\partial \alpha} \;=\;& -(r_1+\pi)\int_{\frac{\kappa-h(\alpha)mz^*}{\alpha n}}^{1} ny_1 + h'(\alpha)mz^* f(y_1)dy_1 \\
& -(r_2+c)\int_{z^*}^{1} h'(\alpha)m(y_2-z^*)g(y_2)dy_2\bar{F}\Big(\frac{\kappa-h(\alpha)mz^*}{\alpha n}\Big) \\
& -(r_2+c)\int_{z^*}^{1} h(\alpha)m(y_2-z^*)g(y_2)dy_2 f\Big(\frac{\kappa-h(\alpha)mz^*}{\alpha n}\Big) \\
& \Big(\frac{h'(\alpha)mz^*\alpha n+(\kappa-h(\alpha)mz^*)n}{(\alpha n)^2}\Big) \\
& -(r_2+c)\int_{0}^{\frac{\kappa-h(\alpha)mz^*}{\alpha n}} \int_{\frac{\kappa-\alpha n y_1}{h(\alpha)m}}^{1} (h'(\alpha)my_2+ny_1)g(y_2)f(y_1)dy_2dy_1 \\
& +(r_1+\pi)\int_{\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}}^{1} ny_1 + h'(\alpha)mz^* f(y_1)dy_1 \\
& +(r_2+c)\int_{z^*}^{1} h'(\alpha)m(y_2-z^*)g(y_2)dy_2\bar{F}\Big(\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}\Big) \\
& +(r_2+c)\int_{z^*}^{1} (1-h(\alpha))m(y_2-z^*)g(y_2)dy_2 f\Big(\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}\Big) \\
& \Big(\frac{h'(\alpha)mz^*(1-\alpha)n+(\kappa-(1-h(\alpha))mz^*)n}{(1-\alpha)^2n^2}\Big) \\
& +(r_2+c)\int_{0}^{\frac{\kappa-(1-h(\alpha))mz^*}{(1-\alpha)n}} \int_{\frac{\kappa-(1-\alpha)ny_1}{(1-h(\alpha))m}}^{1} h'(\alpha)my_2+ny_1g(y_2)f(y_1)dy_2dy_1
\end{aligned}
\tag{A.12}
$$

As can be seen from the above expression, the second order derivative for the revenue function with respect to α becomes analytically intractable to determine if there is an unique optimum. We are able to show that for any given α the revenue function is concave in β, and for any given β the revenue function is concave in α. Upon examining the first order condition, a local optimal occurs at $\alpha = 0.5$ and $\beta = 0.5$ as this assignment sets equation (A.12) to zero. The strategy of assigning equal numbers for both types of patients to identical physicians is a practical solution. Therefore, we will compare the performance of this assignment to the optimal assignment under case 3 when $\kappa/(mz^*) \geq 0.5$.

A.4 Proofs for the Results in Table 2.1

We show the proof for each segment of $\kappa/(mz^*)$.

(a) When $0 < \kappa/(mz^*) \leq 0.5$

The optimal choice of β for case 1 is $\beta^* = 0.5$. Because the the revenue does not depend on α in case 1, we can compare the revenue function under the two cases using $(\alpha, \beta) = (0, 0.5)$.

The optimal expected revenue under case 1 is

$$\begin{aligned} \Pi^{(1)} &= r_1 nE(Y_1) + r_2 mE(Y_2) - (r_1 + \pi)nE(Y_1) \\ &\quad -(r_2 + c)\int_{\frac{2\kappa}{m}}^{1} (my_2 - 2\kappa)g(y_2)dy_2 \end{aligned} \tag{A.13}$$

Under scenario (a), the optimal expected revenue under case 2 occurs at $\alpha^* = 0$, and $\beta^* = 1 - \kappa/(mz^*)$. Let $\hat{\beta}$ denote the optimal allocation of type-2 patients (β^*) under case 2.

$$\begin{aligned} \Pi^{(2)} &= r_1 nE(Y_1) + r_2 mE(Y_2) \\ &\quad -(r_2 + c)\int_{\frac{\kappa}{\hat{\beta}m}}^{1} (\hat{\beta}my_2 - \kappa)g(y_2)dy_2 \\ &\quad -(r_1 + \pi)\int_{0}^{1} ny_1 f(y_1)dy_1 \\ &\quad -(r_2 + c)\int_{z^*}^{1} (1 - \hat{\beta})m(y_2 - z^*)g(y_2)dy_2. \end{aligned} \tag{A.14}$$

Therefore,

$$\begin{aligned} \Pi^{(1)} - \Pi^{(2)} &= -(r_2 + c)\int_{\frac{2\kappa}{m}}^{1} (my_2 - 2\kappa)g(y_2)dy_2 \\ &\quad +(r_2 + c)\int_{\frac{\kappa}{\hat{\beta}m}}^{1} (\hat{\beta}my_2 - \kappa)g(y_2)dy_2 \\ &\quad +(r_2 + c)\int_{z^*}^{1} (1 - \hat{\beta})m(y_2 - z^*)g(y_2)dy_2. \end{aligned} \tag{A.15}$$

First we observe that when $\hat{\beta} = 1 - \frac{\kappa}{mz^*} = 0.5$, $\Pi^{(1)} - \Pi^{(2)}$ in equation (A.15) = 0 because $2\kappa/m = \kappa/(\hat{\beta}m) = z^*$ and $(1-\hat{\beta})mz^* = \kappa$. Next, we observe that

$$\begin{aligned}\frac{\partial \Pi^{(1)} - \Pi^{(2)}}{\hat{\beta}} &= (r_2 + c)\int_{\frac{\kappa}{\hat{\beta}m}}^{1} my_2 g(y_2)dy_2 \\ &\quad -(r_2 + c)\int_{z^*}^{1} my_2 g(y_2)dy_2, \end{aligned} \tag{A.16}$$

which is greater than or equal to 0 when $\hat{\beta} \geq 0.5$. In other words, $\Pi^{(1)} \geq \Pi^{(2)}$ when $\hat{\beta} \geq 0.5$, which is always true because $\hat{\beta} = (1 - \kappa/(mz^*)) \geq 0.5$ in this scenario. Therefore, when $\kappa/(mz^*) \leq 0.5$, case 1's optimal patient allocation strategy (e.g. $\alpha = 0.5, \beta = 0.5, \kappa_1 = \kappa_2 = 0$) is a dominant assignment strategy.

(b) When $\kappa/(mz^*) \geq 0.5$ The local optimal expected revenue under case 3 when $\alpha = 0.5$ and $\beta = 0.5$ is

$$\begin{aligned}\Pi^{(3)} &= r_1 nE(Y_1) + r_2 mE(Y_2) - (r_1 + \pi)\int_{\frac{\kappa - 0.5mz^*}{0.5n}}^{1} (ny_1 - 2\kappa + mz^*)f(y_1)dy_1 \\ &\quad -(r_2 + c)\int_{z^*}^{1} (my_2 - mz^*)g(y_2)dy_2 \bar{F}\Big(\frac{\kappa - 0.5mz^*}{0.5n}\Big) \\ &\quad -(r_2 + c)\int_{0}^{\frac{\kappa - 0.5mz^*}{0.5n}} \int_{\frac{\kappa - 0.5ny_1}{0.5m}}^{1} (my_2 - 2\kappa + ny_1)g(y_2)f(y_1)dy_2 dy_1. \end{aligned} \tag{A.17}$$

The optimal expected revenue under case 2 is

$$\begin{aligned}\Pi^{(2)} &= r_1 nE(Y_1) + r_2 mE(Y_2) - (r_2 + c)\int_{z^*}^{1} (\frac{\kappa}{z^*}y_2 - \kappa)g(y_2)dy_2 \\ &\quad -(r_1 + \pi)\int_{\frac{2\kappa - mz^*}{n}}^{1} (ny_1 - 2\kappa + mz^*)f(y_1)dy_1 \\ &\quad -(r_2 + c)\int_{z^*}^{1} (1 - \frac{\kappa}{mz^*})m(y_2 - z^*)g(y_2)dy_2 \bar{F}\Big(\frac{2\kappa - mz^*}{n}\Big) \\ &\quad -(r_2 + c)\int_{0}^{\frac{2\kappa - mz^*}{n}} \int_{\frac{\kappa - ny_1}{m - \frac{\kappa}{z^*}}}^{1} ((1 - \frac{\kappa}{mz^*})my_2 - \kappa + ny_1)g(y_2)f(y_1)dy_2 dy_1. \end{aligned} \tag{A.18}$$

Hence,

$$
\begin{aligned}
\Pi^{(2)} - \Pi^{(3)} &= -(r_2+c)\int_{z^*}^{1}(\frac{\kappa}{z^*}y_2-\kappa)g(y_2)dy_2 \\
&\quad -(r_2+c)\Big[\int_{z^*}^{1}(1-\frac{\kappa}{mz^*})m(y_2-z^*)g(y_2)dy_2\bar{F}\Big(\frac{2\kappa-mz^*}{n}\Big) \\
&\quad -\int_{0}^{\frac{2\kappa-mz^*}{n}}\int_{\frac{\kappa-ny_1}{m-\frac{\kappa}{z^*}}}^{1}((1-\frac{\kappa}{mz^*})my_2-\kappa+ny_1)g(y_2)f(y_1)dy_2dy_1\Big] \\
&\quad +(r_2+c)\int_{z^*}^{1}(my_2-mz^*)g(y_2)dy_2\bar{F}\Big(\frac{2\kappa-mz^*}{n}\Big) \\
&\quad +(r_2+c)\int_{0}^{\frac{2\kappa-mz^*}{n}}\int_{\frac{2\kappa-ny_1}{m}}^{1}(my_2-2\kappa+ny_1)g(y_2)f(y_1)dy_2dy_1 \\
&= -(r_2+c)\Big[\frac{\kappa}{z^*}\int_{z^*}^{1}(y_2-z^*)g(y_2)dy_2F\Big(\frac{2\kappa-mz^*}{n}\Big) \\
&\quad +\int_{0}^{\frac{2\kappa-mz^*}{n}}\int_{\frac{\kappa-ny_1}{m-\frac{\kappa}{z^*}}}^{1}(my_2-\frac{\kappa}{z^*}y_2-\kappa+ny_1)g(y_2)f(y_1)dy_2dy_1 \\
&\quad -\int_{0}^{\frac{2\kappa-mz^*}{n}}\int_{\frac{2\kappa-ny_1}{m}}^{1}(my_2-2\kappa+ny_1)g(y_2)f(y_1)dy_2dy_1\Big].
\end{aligned}
\tag{A.19}
$$

Equation (A.19) contains both positive and negative terms. When $\kappa/(mz^*) = 0.5$, case 3 is identical to case 1 and therefore the equal assignment strategy (case 3) is optimal (i.e. $\Pi^{(2)} - \Pi^{(3)} < 0$ at $\kappa/(mz^*) = 0.5$). In addition, when $\kappa/(mz^*) = 1$, the specialization strategy is a dominant strategy (i.e. $\Pi^{(2)} - \Pi^{(3)} > 0$). Therefore, we evaluate whether the optimal patient allocation strategy changes in the value of $\kappa/(mz^*)$. Because a physician's capacity κ is a constant and z^* is tied to the uncertainty of demand distribution and cost parameters, we take the derivative of Equation (A.19) with respect to m.

$$
\begin{aligned}
\frac{\partial(\Pi^{(2)}-\Pi^{(3)})}{\partial m} &= (r_2+c)\frac{\kappa}{n}\int_{z^*}^{1}(y_2-z^*)g(y_2)dy_2 f\Big(\frac{2\kappa-mz^*}{n}\Big) \\
&\quad +(r_2+c)\frac{z^*}{n}\int_{z^*}^{1}(my_2-\frac{\kappa}{z^*}y_2+\kappa-mz^*)g(y_2)f\Big(\frac{2\kappa-mz^*}{n}\Big) \\
&\quad -(r_2+c)\int_0^{\frac{2\kappa-mz^*}{n}}\int_{\frac{\kappa-ny_1}{m-\frac{\kappa}{z^*}}}^{1} y_2 g(y_2)f(y_1)dy_2dy_1 \\
&\quad -\frac{z^*}{n}\int_{z^*}^{1}(my_2-mz^*)g(y_2)f\Big(\frac{2\kappa-mz^*}{n}\Big)dy_2dy_1 \\
&\quad +(r_2+c)\int_0^{\frac{2\kappa-mz^*}{n}}\int_{\frac{2\kappa-ny_1}{m}}^{1} y_2 g(y_2)f(y_1)dy_2dy_1 \\
&= -(r_2+c)\int_0^{\frac{2\kappa-mz^*}{n}}\int_{\frac{\kappa-ny_1}{m-\frac{\kappa}{z^*}}}^{1} y_2 g(y_2)f(y_1)dy_2dy_1 \\
&\quad +(r_2+c)\int_0^{\frac{2\kappa-mz^*}{n}}\int_{\frac{2\kappa-ny_1}{m}}^{1} y_2 g(y_2)f(y_1)dy_2dy_1
\end{aligned}
\tag{A.20}
$$

The value of Equation (A.20) ≥ 0 because $(2\kappa-ny_1)/m < (\kappa-ny_1)/(m-\kappa/z^*)$. Therefore, $(\Pi^{(2)}-\Pi^{(3)})$ is monotone increasing in m for $0.5 \leq \kappa/(mz^*) \leq 1$. Because $(\Pi^{(2)}-\Pi^{(3)}) < 0$ at $\kappa/(mz^*)=0.5$ and $(\Pi^{(2)}-\Pi^{(3)}) > 0$ at $\kappa/(mz^*)=1$, there is a cut-off point $\tilde{m}$ such that when $\kappa/(mz^*) \leq \kappa/(\tilde{m}z^*)$, equal assignment is better. When $\kappa/(mz^*) > \kappa/(\tilde{m}z^*)$, specialization strategy is better.

Appendix B

Proofs for Chapter 3

B.1 Learning Acceptance Probabilities

Let vectors $\alpha^\ell(t) = (\alpha_1^\ell(t), \cdots, \alpha_m^\ell(t))$ and $\beta^\ell(t) = (\beta_1^\ell(t), \cdots, \beta_b^\ell(t))$ denote estimated panel-ℓ physician and time-block acceptance probabilities after the t-th update. We assume that each patient's PCP is always acceptable to him/her and set $\alpha_\ell^\ell(t) = 1$ for each ℓ and t. This is not a requirement of our model but rather a reasonable assumption in the application domain. It does simplify some computations because the clinic only needs to estimate each panel's acceptance probabilities for the $(m-1)$ non PCPs. However, our approach can be reworked if it were important to consider a situation in which a patient's PCP would not be acceptable to him/her, but a non-PCP would be for the same time block. In practice, patients are free to change their designated PCPs as often as they wish. It is therefore unlikely that a patient will prefer not to visit his/her PCP at an acceptable time.

The estimated acceptance probabilities after the t-th update, $\alpha_i^\ell(t)$ (resp. $\beta_j^\ell(t)$), can be obtained as the relative frequency that panel ℓ's patients include physician i (time-block j) in the acceptable set. That is, $\alpha_i^\ell(t) = [\alpha_i^\ell(0)\theta + \sum_{z=1}^t N_{i,z}^\ell]/(\theta + t)$, where $\theta \geq 0$ is the prior count or the weight given to a subjective estimate of acceptance probability prior to information updating, and $N_{i,z}^\ell = 1$ if the z-th panel-ℓ arrival includes physician i in the acceptable set and 0 otherwise. Similarly, $\beta_j^\ell(t) = [\beta_j^\ell(0)\theta + \sum_{z=1}^t H_{j,z}^\ell]/(\theta + t)$, where $H_{j,z}^\ell = 1$ if time block j is included in the z-th panel-ℓ arrival's acceptable set, and 0 otherwise. A higher value of θ represents a higher level of confidence in the initial

estimates.

There can be a variety of ways to obtain initial estimates. For example, $\alpha_i^\ell(0)$ can be obtained by calculating $\alpha_i^\ell(0) = \sum_{k\neq\ell} n_{k,0}^\ell / n_0^\ell$ for $i \neq \ell$, where $n_{k,0}^\ell$ and n_0^ℓ represent respectively the total number of appointments with physician k and with all physicians booked by panel-ℓ patients. That is, when a patient booked with any one of the non PCPs, we may assume that each non PCP was equally acceptable to the patient. This is because we only observe actual bookings and not acceptable sets in the historical data. Similarly, $\beta_j^\ell(0) = h_{j,0}^\ell / n_0^\ell$ for time block j, where $h_{j,0}^\ell$ represents the total number of block j appointments by all patients in panel ℓ and n_0^ℓ is the total number of bookings from panel-ℓ patients in the historical records.

The updating procedure does not depend on which acceptable combination is actually booked by the clinic. It utilizes knowledge of the composition of acceptable sets from the web-based appointment request system. To better understand the accuracy of the estimating procedure, consider a primary-care physician who on average receives 25 booking requests each day. In this instance, 500 updates are reached in about 20 days of operation. After these many updates, the standard error of the estimated acceptance probability is reduced to less than 0.02 for each (i, j) combination. Twenty days is a relatively short time in our application domain during which aggregate patient choices are unlikely to change much. Therefore, our approach converges rapidly to true probabilities and remains accurate so long as these probabilities do not change too quickly.

Once the clinic learns each panel's acceptance probabilities for each physician and time-block combination, this information can be used as input to booking decisions described in Section 3.3.2. Because, we assume $\alpha_\ell^\ell = 1$, by the definition of acceptable combinations, we have

$$\alpha_I^\ell = \prod_{i\in I} \alpha_i^\ell \prod_{i\notin I} (1 - \alpha_i^\ell) \tag{B.1}$$

and

$$\beta_J^\ell = \big[\prod_{j\in J} \beta_j^\ell \prod_{j\notin J} (1 - \beta_j^\ell)\big] / \big[1 - \prod_{j=1}^{b} (1 - \beta_j^\ell)\big]. \tag{B.2}$$

B.2 Utility-Based Patients' Preference Model

Suppose that a clinic has the following utility-based model of patients' preferences. Each panel comprises of homogenous patients who assign the same aggregate utility $\mu^{\ell}_{i,j}$ to combination (i,j). However, in the k-th booking attempt of a patient from this panel, the utility that a particular patient derives is $U^{\ell}_{i,j} = \mu^{\ell}_{i,j} + \xi^{\ell}_{i,j}$, where $\xi^{\ell}_{i,j}$ is a random component that is independent of the mean utility. Note that the null choice, i.e. when a patient does not book, is denoted by the combination (0,0). Based on the patient-clinic web-based interface, the clinic's model of patients' utility has the following properties.

1. If $\mu^{\ell}_{i_1,j} \geq \mu^{\ell}_{i_2,j}$ for some j, then this implies $\mu^{\ell}_{i_1,z} \geq \mu^{\ell}_{i_2,z}$ for every z.
2. If $\mu^{\ell}_{i,j_1} \geq \mu^{\ell}_{i,j_2}$ for some i, then this implies $\mu^{\ell}_{z,j_1} \geq \mu^{\ell}_{z,j_2}$ for every z.

Let $u^{\ell}_{P,i}$ and $u^{\ell}_{B,j}$ denote the marginal mean utilities associated with physician i and time block j. Then, these two attributes are said to be additively independent (referred to as the AI property) if $\mu^{\ell}_{i,j} = k_1 u^{\ell}_{P,i} + k_2 u^{\ell}_{B,j}$, where $k_1 + k_2 = 1$. Lemma B.2.1 below proves that the AI property satisfies properties 1 and 2 of the patient choice model underlying the web-based appointment request systems.

Lemma B.2.1 *If $\mu^{\ell}_{i,j} = k_1 u^{\ell}_{P,i} + k_2 u^{\ell}_{B,j}$, where $k_1 + k_2 = 1$ for each (i,j) and ℓ, then properties 1 and 2 hold.*

Proof: From the AI property, $\mu^{\ell}_{i_1,j} \geq \mu^{\ell}_{i_2,j}$ implies that

$$\begin{aligned}
&k_1 u^{\ell}_{P,i_1} + k_2 u^{\ell}_{B,j} \geq k_1 u^{\ell}_{P,i_2} + k_2 u^{\ell}_{B,j} \\
&\Leftrightarrow u^{\ell}_{P,i_1} \geq u^{\ell}_{P,i_2} \\
&\Leftrightarrow k_1 u^{\ell}_{P,i_1} + k_2 u^{\ell}_{B,z} \geq k_1 u^{\ell}_{P,i_2} + k_2 u^{\ell}_{B,z} \\
&\Leftrightarrow u^{\ell}_{i_1,z} \geq u^{\ell}_{i_2,z}.
\end{aligned}$$

A similar set of arguments lead to property 2. Hence proved. #.

Estimating the Mean Utility

Given the additive form of the mean utility $\mu^{\ell}_{i,j}$, we can re-write $\mu^{\ell}_{i,j}$ as follows.

$$\mu^{\ell}_{i,j} = \sum_{n=1}^{m} \tilde{\alpha}^{\ell}_{n} I_{\{n=i\}} + \sum_{k=1}^{b} \tilde{\beta}^{\ell}_{k} I_{\{k=j\}} \qquad \forall\ i = 0, \cdots, m;\ j = 0, \cdots, b, \tag{B.3}$$

where $\tilde{\alpha}_i^\ell = k_1 u_{P,i}^\ell$, $\tilde{\beta}_j^\ell = k_2 u_{B,j}^\ell$ and $I_{\{\cdot\}}$ denotes an indicator function that takes value 1 if the expression within the braces is true and zero otherwise. Equation (B.3) allows us to estimate $\mu_{i,j}^\ell$ by fitting a linear model that estimates $\tilde{\alpha}_i^\ell$ and $\tilde{\beta}_i^\ell$ without knowing $u_{P,i}^\ell$, $u_{B,j}^\ell$, and their corresponding weights k_1 and k_2. Knowing $\mu_{i,j}^\ell$s, we can then estimate $p_{i,j}^\ell$s, as illustrated by a specific example below.

Suppose $U_{i,j}^\ell$ has a logistic distribution with mean $\mu_{i,j}^\ell$ and variance $\pi^2 {\sigma_{i,j}^\ell}^2/3$. Let ϑ^ℓ denote the utility of not booking for panel-ℓ patients. We assume this no-booking utility is identical for all patients in the same panel. If the utility of booking an appointment of a certain physician and time-block combination exceeds the utility of no booking, then the patient will include that physician and time-block combination in his/her acceptable set. Therefore, the (ℓ, k) patient would include combination (i, j) in his/her set of acceptable combinations if $U_{i,j}^\ell \geq \vartheta^\ell$. Consequently, the probability of a (i, j) physician and time-block combination being valued more than ϑ^ℓ by an arriving patient of panel ℓ is panel-ℓ's acceptance probability for that combination. That is,

$$p_{i,j}^\ell = P(U_{i,j}^\ell > \vartheta^\ell) = \frac{e^{-(\vartheta^\ell - \mu_{i,j}^\ell)/\sigma_{i,j}^\ell}}{1 + e^{-(\vartheta^\ell - \mu_{i,j}^\ell)/\sigma_{i,j}^\ell}} \;\forall\; i = 1, \cdots, m;\; j = 1, \cdots, b;\; \ell = 1, \cdots, m, \tag{B.4}$$

and we can obtain acceptance probabilities after calculating $\mu_{i,j}^\ell$s from (B.3) with the help of a linear model. Also note that equation (B.4) implies that if $\mu_{i,j}^\ell \geq \mu_{n,k}^\ell$, then $p_{i,j}^\ell \geq p_{n,k}^\ell$.

For the logistic utility function, we can alternatively obtain $p_{i,j}^\ell$ by using a generalized linear model as explained below. Re-write $U_{i,j}^\ell$ as

$$U_{i,j}^\ell = \mu_{i,j}^\ell + \xi_{i,j}^\ell, \tag{B.5}$$

where $\xi_{i,j}^\ell$ has a logistic distribution with mean 0 and variance $\pi^2 {\sigma_{i,j}^\ell}^2/3$. The mean utility $\mu_{i,j}^\ell$ can be related to $p_{i,j}^\ell$ by the following logistic link function.

$$\log \frac{p_{i,j}^\ell}{1 - p_{i,j}^\ell} = \frac{\mu_{i,j}^\ell}{\sigma_{i,j}^\ell} - \frac{\vartheta^\ell}{\sigma_{i,j}^\ell}. \tag{B.6}$$

Historical appointment data can be used to derive the proportion of time that each (i, j) physician and time-block combination is included in the acceptable set by panel-ℓ's

patients. Then, the linear form in equations (B.3) and (B.6) allows us to estimate $p^{\ell}_{i,j}$ by fitting a generalized linear model of the following form.

$$\begin{aligned} \log \frac{p^{\ell}_{i,j}}{1-p^{\ell}_{i,j}} &= -\frac{\vartheta^{\ell}}{\sigma^{\ell}_{i,j}} + \sum_{n=1}^{m} \frac{\tilde{\alpha}^{\ell}_{n}}{\sigma^{\ell}_{i,j}} I_{\{n=i\}} + \sum_{k=1}^{b} \frac{\tilde{\beta}^{\ell}_{k}}{\sigma^{\ell}_{i,j}} I_{\{k=j\}} \\ &= \bar{\vartheta}^{\ell} + \sum_{n=1}^{m} \bar{\alpha}^{\ell}_{n} I_{\{n=i\}} + \sum_{k=1}^{b} \bar{\beta}^{\ell}_{k} I_{\{k=j\}}. \end{aligned} \tag{B.7}$$

For estimating panel acceptance probabilities via a generalized linear model in equation (B.7), the total number of parameters to estimate is $(1+m)+(b-1) = m+b$. This is because $I_{\{n=i\}} = 0$ for all n immediately implies that $I_{\{k=j\}} = 0$ for all k. Put differently, a patient must have at least one acceptable physician for a non-null request.

Comparison of Utility-Based and Proposed Model

The total number of parameters that need to be estimated is the same in both the utility model and our approach. The utility model is not superior to our approach in terms of modeling flexibility. Moreover, our approach involves less work to estimate and update $p^{\ell}_{i,j}$ via $\alpha^{\ell}_{i}(t)$ and $\beta^{\ell}_{j}(t)$ than the effort required by a utility-based model via $\bar{\vartheta}^{\ell}$, $\bar{\alpha}^{\ell}_{i}$ and $\bar{\beta}^{\ell}_{j}$ (as explained above). Therefore, the proposed model with $p^{\ell}_{i,j} = \alpha^{\ell}_{i}\beta^{\ell}_{j}$ is a more natural choice for estimating acceptance probabilities.

In closing this section, we point out that a generalized linear model can be used to estimate probabilities $p^{\ell}_{i,j}(t)$ at each update epoch, or at regular intervals, by hypothesizing a linear relationship between $p^{\ell}_{i,j}(t)$ and factors such as the indices i and j through a logit link function. In order to use such a model, a utility-based argument is not required. However, the computational effort will not be different from what we estimated above. Therefore, we did not pursue this alternate approach for computing estimates of acceptance probabilities. A comprehensive review of generalized linear models can be found in McCullagh and Nelder (1989).

B.3 Convex Cost Structure for Unmet Same-Day Demand

Let $c(\cdot)$ be a convex increasing function. Then the marginal benefit of reserving a physician i's slot with no additional advance-book requests is

$$\begin{aligned}\hat{\Delta}(\bar{s}_i, \bar{s}_{-i}) &\doteq u_0(s) - u_0(s + e_{i,j}) \\ &= r_2 - (r_2 - r_2')F_i(\bar{\kappa}_i - \bar{s}_i - 1) - r_2' F(\bar{\bar{\kappa}} - \bar{s}_{-i} - \bar{s}_i - 1) \\ &\quad + E\big[c(X - \bar{\bar{\kappa}} + \bar{s}_{-i} + \bar{s}_i + 1) - c(X - \bar{\bar{\kappa}} + \bar{s}_{-i} + \bar{s}_i) | X > \bar{\bar{\kappa}} - \bar{\bar{s}} - 1\big].\end{aligned} \tag{B.8}$$

Because same-day patients do not have time preferences, the marginal benefit of saving a slot does not depend on which time block it belongs to.

Note that the last term in Equation (B.8) increases in $\bar{s}_i$ and $\bar{s}_{-i}$ because $c(\cdot)$ is a convex increasing function. In addition, CDF is a non-decreasing function. Therefore, for any given $\bar{s}_{-i}$, $\hat{\Delta}(\bar{s}_i, \bar{s}_{-i})$ increases in $\bar{s}_i$. Similarly, for any given $\bar{s}_i$, $\hat{\Delta}(\bar{s}_i, \bar{s}_{-i})$ increases in $\bar{s}_{-i}$. In this cost structure, the optimal decision is not to book a panel-ℓ patient in physician i's slot if $\hat{\Delta}(\bar{s}_i, \bar{s}_{-i}) > r^i_{1,\ell} + \pi_t$. Therefore, the protection level of each physician i also exists under a convex cost function for unmet same-day demand.

B.4 Example of No-Book States

We present an example in Figure B.1 that identifies no-book states. Our example clinic has two physicians with a capacity of 15 slots each. We assume independent and Poisson same-day demand distributions for the two panels with mean 5 patients per appointment day. Other parameters are $(r^1_{1,1}, r^2_{1,1}, r^2_{1,2}, r^1_{1,2}, r_2, r_2') = (20, 17, 18, 15.3, 20, 17)$ and $(\pi_t, c) = (4.25, 5)$. Note that π_t is invariant in t and $r^i_{1,\ell} + \pi_t \leq r_2 + c$ holds in this example. In Figure B.1, the open circles in left (resp. right) console represent the no-book states for physician-1's (resp. physician-2's) slots when the booking request comes from a patient of either panel. The squares in left console (respectively triangles in right console) represent physician 1's (resp. physician 2's) no-book states for panel-2 (resp. panel-1) patients. The crosses represent situations where the clinic will proceed to the second-step in the booking process following a request from a patient of either panel.

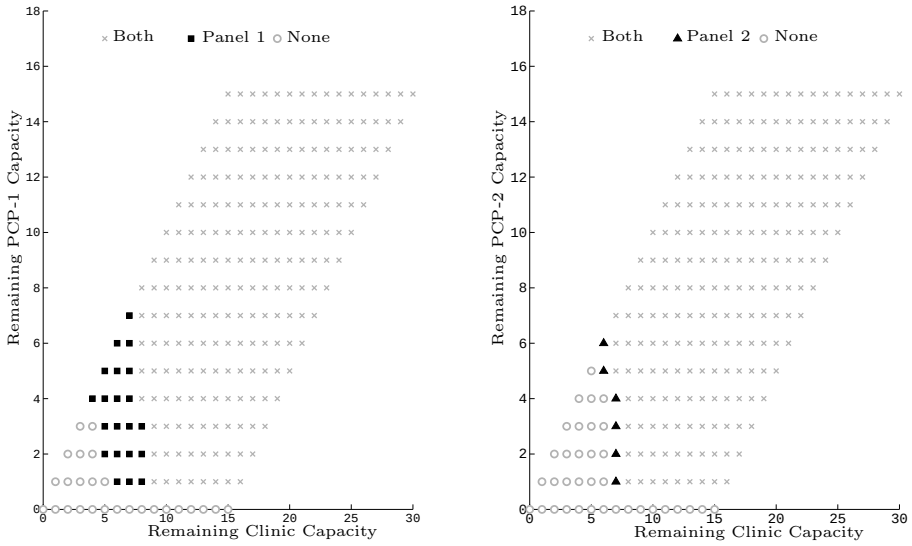


Figure B.1: Protection Levels by Physician in a Two-Physician Clinic Example

B.5 Proof of Proposition 3.4.2

The first statement in the proposition is proved by induction. Before we do so, recall that if $u_0(s) - u_0(s + e_{i,j}) > r^i_{1,\ell} + \pi_t$, then $s \in \hat{S}^{i,\ell}_t$, whereas if $u_{t-1}(s) - u_{t-1}(s + e_{i,j}) > r^i_{1,\ell} + \pi_t$, then $s \in S^{i,\ell}_t$. We will prove that $\hat{S}^{i,\ell}_t \subseteq S^{i,\ell}_t$ for $t \geq 1$ by showing that if $u_0(s) - u_0(s + e_{i,j}) > r^i_{1,\ell} + \pi_t$, then $u_{t-1}(s) - u_{t-1}(s + e_{i,j}) > r^i_{1,\ell} + \pi_t$ for each $t \geq 1$.

Because same-day patients are assumed to have no time preference, $u_0(s)-u_0(s+e_{i,j})$ is the same for all time-block j for a given physician i. Therefore, we write $u_0(s + e_{i,j})$ as $u_0(s + e_{i.})$ to highlight independence from j. We are now ready to prove the two statements in the proposition.

When t=1, $\hat{S}^{i,\ell}_1 = S^{i,\ell}_1$ by the definition of $\hat{S}^{i,\ell}_1$. Therefore, $\hat{S}^{i,\ell}_1 \subseteq S^{i,\ell}_1$ is trivially true. Next consider $t = 2$. Given that a panel-ℓ patient requests (i, j) physician and time-block combination in that period and $s \in \hat{S}^{i,\ell}_2$, we know that $u_0(s) - u_0(s + e_{i,j}) > r^i_{1,\ell} + \pi_2$. The decision rule in this period is to reject the patient's request if $u_1(s) - u_1(s + e_{i,j}) >$

$r^i_{1,\ell} + \pi_2$. From (3.3), we can calculate $u_1(s) - u_1(s + e_{i,j})$ as follows:

$$\begin{aligned} u_1(s) - u_1(s + e_{i,j}) &= \sum_{k=1}^{m} \lambda_1^k [u_1^k(s) - u_1^k(s + e_{i,j})] \\ &\quad + (1 - \sum_{k=1}^{m} \lambda_1^k)[u_0(s) - u_0(s + e_{i,j})]. \end{aligned} \tag{B.9}$$

The benefit of saving a slot of combination (i, j) in period 2 for a type k arrival in period 1 equals $u_1^k(s) - u_1^k(s + e_{i,j})$. This benefit is at least equal to $r^i_{1,\ell} + \pi_2$. This argument comes from the fact that if an (i, j) slot is saved for an advance-book arrival in period 1 and the system state is unchanged, then the clinic's benefit must be at least as much as what it can earn by saving the slot for a same-day patient's use, i.e. it must be greater than $r^i_{1,\ell} + \pi_2$. Therefore, $u_0(s) - u_0(s + e_{i,j}) > r^i_{1,\ell} + \pi_2$ implies $u_1^k(s) - u_1^k(s + e_{i,j}) > r^i_{1,\ell} + \pi_2$ for each $k = 1, \cdots, m$. By substituting $u_1^k(s) - u_1^k(s + e_{i,j})$ and $u_0(s) - u_0(s + e_{i,j})$ with the lower bound $r^i_{1,\ell} + \pi_2$ in equation (B.9), we obtain the inequality $u_1(s) - u_1(s + e_{i,j}) \geq u_0(s) - u_0(s + e_{i,j}) = u_0(s) - u_0(s + e_{i\cdot}) > r^i_{1,\ell} + \pi_2$. Therefore, $\hat{S}_2^{i,\ell} \subseteq S_2^{i,\ell}$.

Next assume that by induction hypothesis, we have for each $t = 1, \cdots, n + 1$, $u_{t-1}(s) - u_{t-1}(s + e_{i,j}) > r^i_{1,\ell} + \pi_{t+1}$. Then, one can argue that in period $n + 2$, the true no-book state depends on

$$\begin{aligned} u_{n+1}(s) - u_{n+1}(s + e_{i,j}) &= \sum_{k=1}^{m} \lambda_{n+1}^k [u_{n+1}^k(s) - u_{n+1}^k(s + e_{i,j})] \\ &\quad + (1 - \sum_{k=1}^{m} \lambda_{n+1}^k)[u_n(s) - u_n(s + e_{i,j})] \\ &> r^i_{1,\ell} + \pi_{n+2}. \end{aligned} \tag{B.10}$$

The underlying argument is the same as in the case when $t = 2$. Specifically, $u_n(s) - u_n(s + e_{i,j}) > r^i_{1,\ell} + \pi_{n+2}$ implies that $u_{n+1}^k(s) - u_{n+1}^k(s + e_{i,j}) > r^i_{1,\ell} + \pi_{n+2}$ for each $k = 1, \cdots, m$. Therefore, by induction, if $u_0(s) - u_0(s + e_{i\cdot}) > r^i_{1,\ell} + \pi_t$, then $u_{t-1}(s) - u_{t-1}(s + e_{i,j}) \geq \cdots \geq u_0(s) - u_0(s + e_{i,j}) = u_0(s) - u_0(s + e_{i\cdot}) > r^i_{1,\ell} + \pi_t$ for all $j = 1, \cdots, b$ and $t \geq 1$, and this implies that $\hat{S}_t^{i,\ell} \subseteq S_t^{i,\ell}$. The above completes a proof of the first statement of the proposition.

For the second statement, because $\pi_t \geq \pi_{t-1}$, we have that $u_0(s) - u_0(s + e_{i\cdot}) > r^i_{1,\ell} + \pi_t \geq r^i_{1,\ell} + \pi_{t-1}$, and consequently $\hat{S}_t^{i,\ell} \subseteq \hat{S}_{t-1}^{i,\ell}$. #

B.6 No-Book States for a Single-Physician Clinic

We show in this section that when there is a single physician labeled ℓ, $\hat{S}_t^\ell = S_t^\ell$. For simplicity, we omit the physician label and only use $s = (s_j)$, $j = 1, \cdots, b$ to represent the state of different time blocks, and use r_1 to represent the average revenue of an appointment ($r_1 \equiv r_{1,1}^1$). To prove that $\hat{S}_t^\ell = S_t^\ell$ is the same as proving that if $u_0(s) - u_0(s + e_j) > r_1 + \pi_t$, then $u_{t-1}(s) - u_{t-1}(s + e_j) > r_1 + \pi_t$, and vice versa. By Proposition 3.4.1, $u_{t-1}(s) - u_{t-1}(s + e_j) \geq u_0(s) - u_0(s + e_j)$. Therefore, $u_0(s) - u_0(s + e_j) > r_1 + \pi_t$ implies $u_{t-1}(s) - u_{t-1}(s + e_j) > r_1 + \pi_t$. Next we want to show that $u_{t-1}(s) - u_{t-1}(s + e_j) > r_1 + \pi_t$ implies $u_0(s) - u_0(s + e_j) > r_1 + \pi_t$.

Recall that we assume $\pi_t \geq \pi_n \; \forall \; n \leq t$. Therefore, the benefit of saving a slot for a future advance-book arrival is no more than the benefit of offering the slot to the current arrival (i.e. $r_1 + \pi_t \geq r_1 + \pi_n$). Consequently, the only case in which a request for a slot in block j will be denied under an optimal booking decision occurs when the benefit of saving the slot for a same-day request is higher than the benefit of offering it to the current patient. In other words, if the clinic's optimal decision is to deny the request for a slot in block j and save the slot for future use, then the marginal benefit of that slot is at most $u_0(s) - u_0(s + e_j)$. Hence, if $u_{t-1}(s) - u_{t-1}(s + e_j) > r_1 + \pi_t$, then $u_0(s) - u_0(s + e_j) \geq u_{t-1}(s) - u_{t-1}(s + e_j) \geq r_1 + \pi_t$. #

B.7 Proof of Proposition 3.4.3

With a single physician, we can simplify the revenue function as follows:

$$\begin{aligned} u_t(s) &= \lambda_t \sum_{J \in \mathcal{J}} p_J \max_{j \in J} \{r_1 + u_{t-1}(s + e_j), u_{t-1}(s) - \pi_t\} + (1 - \lambda_t) u_{t-1}(s) \\ &= u_{t-1}(s) - \lambda_t \pi_t + \lambda_t \sum_{J \in \mathcal{J}} p_J \max_{j \in J} \{(u_{t-1}(s + e_j) - u_{t-1}(s) + r_1 + \pi_t)^+\}, \end{aligned}$$

where p_J is the probability that an arriving patient's set of acceptable time blocks is J, $\mathcal{J}$ is the collection of all of acceptable sets of time blocks, and e_j is a vector with the j-th entry = 1 and the remaining entries = 0.

Given $\beta_j < \beta_k$ and $\phi(s + e_j) > \phi(s + e_k)$, we prove Proposition 3.4.3 by induction.

When $t = 1$, the revenue function can be written as

$$\begin{aligned} u_1(s) &= u_0(s) - \lambda_1\pi_1 + \lambda_1 \sum_{J\in\mathcal{J}} p_J \max_{j\in J}\{(u_0(s+e_j) - u_0(s) + r_1 + \pi_1)^+\} \\ &= u_0(s) - \lambda_1\pi_1 + \lambda_1\phi(s)(r_1 + \pi_1 - (r_2 + c)\bar{F}(\bar{\kappa} - \bar{s} - 1))^+, \end{aligned}$$

where $\bar{\kappa}$ and $\bar{s}$ are the clinic's total capacity and total number of slots booked, respectively.

Because $u_0(s+e_j) = u_0(s+e_k)$ and $\phi(s+e_j) > \phi(s+e_k)$, it follows that $u_1(s+e_j) \geq u_1(s+e_k)$. In Lemma B.7.1 (presented after the proof of Proposition 3.4.3 in this section), we show that if $\phi(s+e_j) > \phi(s+e_k)$, then this implies that $\phi(\tilde{s}+e_j) > \phi(\tilde{s}+e_k)$ for each $\tilde{s} \in \mathcal{S}$, where $\mathcal{S} = \{\tilde{s} : s_j \leq \tilde{s}_j,\ j = 1, \cdots, b\}$. Put differently, Lemma B.7.1 shows that if the clinic has a greater chance of meeting future demand when it books block j in state s, rather than block k, then this ordering is preserved for every state $\tilde{s}$ in which the remaining capacity in each block is no more than the corresponding remaining capacity in state s. Furthermore, because $u_0(\tilde{s}+e_j) = u_0(\tilde{s}+e_k)$, it follows immediately that $u_1(\tilde{s}+e_j) \geq u_1(\tilde{s}+e_k)$ for every $\tilde{s} \in \mathcal{S}$.

Next, by induction hypothesis, assume that $\beta_j < \beta_k$ and $\phi(s+e_j) > \phi(s+e_k)$ imply that for some t, $u_t(s+e_j) \geq u_t(s+e_k)$ and $u_t(\tilde{s}+e_j) \geq u_t(\tilde{s}+e_k)$ for all $\tilde{s} \in \mathcal{S}$. This also implies that $u_t(s+e_j+e_n) \geq u_t(s+e_k+e_n)$ for each n, and $\max_n u_t(s+e_j+e_n) \geq \max_{n'} u_t(s+e_k+e_{n'})$. In addition, $u_t(\tilde{s}+e_j+e_n) \geq u_t(\tilde{s}+e_k+e_n)$ for each n, and $\max_n u_t(\tilde{s}+e_j+e_n) \geq \max_{n'} u_t(\tilde{s}+e_k+e_{n'})$ for all $\tilde{s}$. Consider the ordering of blocks j and k in decision epoch $t+1$. This ordering can be obtained by suing the following arguments.

$$\begin{aligned} u_{t+1}(s+e_j) - u_{t+1}(s+e_k) &= \lambda_{t+1} \sum_{J\in\mathcal{J}} p_J [\max_{n\in J}\{u_t(s+e_j+e_n) + r_1, u_t(s+e_j) - \pi_t\} \\ &\quad - \max_{n'\in J}\{u_t(s+e_k+e_{n'}) + r_1, u_t(s+e_k) - \pi_t\}] \\ &\quad + (1-\lambda_{t+1})[u_t(s+e_j) - u_t(s+e_k)] \\ &\geq 0. \end{aligned}$$

Similarly,

$$\begin{aligned} u_{t+1}(\tilde{s}+e_j) - u_{t+1}(\tilde{s}+e_k) &= \lambda_{t+1} \sum_{J \in \mathcal{J}} p_J [\max_{n \in J} \{u_t(\tilde{s}+e_j+e_n) + r_1, u_t(\tilde{s}+e_j) - \pi_t\} \\ &\quad - \max_{n' \in J} \{u_t(\tilde{s}+e_k+e_{n'}) + r_1, u_t(\tilde{s}+e_k) - \pi_t\}] \\ &\quad + (1-\lambda_{t+1})[u_t(\tilde{s}+e_j) - u_t(\tilde{s}+e_k)] \\ &\geq 0. \end{aligned}$$

Both inequalities above follow immediately from the induction hypothesis. Hence in period $(t+1)$, $u_{t+1}(s+e_j) \geq u_{t+1}(s+e_k)$ and $u_{t+1}(\tilde{s}+e_j) \geq u_{t+1}(\tilde{s}+e_k)$ hold for all $\tilde{s}$. This completes the proof by induction.

Lemma B.7.1 *If $\beta_j < \beta_k$ and $\phi(s+e_j) > \phi(s+e_k)$, then $\phi(\tilde{s}+e_j) > \phi(\tilde{s}+e_k)$ $\forall$ $\tilde{s} \in \mathcal{S}$, where $\mathcal{S} = \{\tilde{s} : \tilde{s}_j \geq s_j,\ j = 1, \cdots, b\}$.*

Proof: From the definition of $\phi(s) = 1 - \prod_{j:s_j<\kappa_j}(1-\beta_j)$, it follows that the inequalities $\beta_j < \beta_k$ and $\phi(s+e_j) < \phi(s+e_k)$ hold if and only if the system state s has the following property: either $s_j + 1 < \kappa_j$ and $s_k + 1 = \kappa_k$, or $s_j + 1 = \kappa_j$ and $s_k + 1 = \kappa_k$. That is, when either state j has more remaining capacity than state k, or when the two states have exactly the same amount of remaining capacity.

Next we show that $\beta_j < \beta_k$ and $\phi(s+e_j) < \phi(s+e_k)$ together imply that $\phi(\tilde{s}+e_j) > \phi(\tilde{s}+e_k)$. If $\tilde{s}_k > s_k$, $\phi(\tilde{s}+e_k)$ does not exist because $\tilde{s}_k + 1 > \kappa_k$ from the fact that $s_k + 1 = \kappa_k$ is a necessary condition for $\phi(s+e_j) > \phi(s+e_k)$ to hold. The only case in which both $\phi(\tilde{s}+e_j)$ and $\phi(\tilde{s}+e_k)$ exist happens when $\tilde{s}_j > s_j$ and $\tilde{s}_k = s_k$. In this case $\tilde{s}_j + 1 \leq \kappa_j$, and $\tilde{s}_k + 1 = s_k$, which satisfies the required property above for $\phi(\tilde{s}+e_j) > \phi(\tilde{s}+e_k)$. Hence proved. #

References

McCullagh, P., J Nelder. 1989. *Generalized Linear Models.* Boca Raton, FL: Chapman and Hall.

Appendix C

Proofs for Chapter 4

C.1 Proof for Proposition 4.5.1

Note that $X_i \leq_{icx} X_i'$ for each $i = 1, \cdots, m$. Therefore, $E[c_0(X_i - Q_i(\boldsymbol{a}))^+] \leq E[c_0(X_i' - Q_i(\boldsymbol{a}))^+]$ for each i and $\pi(\boldsymbol{a}^*_{\boldsymbol{X}}) = \sum_{i=1}^u E[c_0(X_i - Q_i(\boldsymbol{a}^*_{\boldsymbol{X}})^+] \leq \sum_{i=1}^u E[c_0(X_i - Q_i(\boldsymbol{a}^*_{\boldsymbol{X'}})^+]$ $\leq \sum_{i=1}^u E[c_0(X_i' - Q_i(\boldsymbol{a}^*_{\boldsymbol{X'}})^+] = \pi(\boldsymbol{a}^*_{\boldsymbol{X'}})$.

C.2 Proof of Proposition 4.5.2

The result of Proposition 4.5.2 comes from a property of Schur-convex functions. Let $g(q_i) = E[c_o(X_i - q_i)^+]$, where q_i is the (deterministic) number of nurses who show up in unit i. It is straightforward to verify that $g(\cdot)$ is convex in q_i. Suppose $\boldsymbol{q} \leq_M \boldsymbol{q'}$, then $\sum_{i=1}^k g(q_i) \leq \sum_{i=1}^k g(q_i')\ \forall\ k = 1, \cdots, n$. Also, let $\bar{q} = (1/u)\sum_{i=1}^u q_i$. The inequality $\sum_{i=1}^u g(\bar{q}) \leq \sum_{i=1}^u g(q_i)$ holds for all convex functions (Marshall et al. 2011).

Note that each unit's demand are independent and identically distributed, therefore the cost function $\pi_i(\boldsymbol{a}) = E[c_0(X_i - q_i)^+]$ is the same convex cost function (convex in q_i) for all i. Based on the above Schur-convex function property, the best expected shortage cost ($\pi(\boldsymbol{a}^*) = \sum_{i=1}^u \pi_i(\boldsymbol{a}^*)$) is achieved when each unit's staffing level is equal (i.e. $q_i = Q_i(\boldsymbol{a}) = \sum_{t=1}^m a_t^{(i)} - \phi_i(\boldsymbol{a}) = q\ \forall\ i = 1, \cdots, m$.) See, Müller and Stoyan (2002).

C.3 Proof of Proposition 4.5.3

Let $\pi(\boldsymbol{a}) = \sum_{i=1}^{u} \pi_i(\boldsymbol{a}^{(i)})$. Because $\xi_t \leq_{icx} \xi'_t$, $\pi_i(\boldsymbol{a}^{(i)}|\boldsymbol{\xi}) = E[c_0(X_i - \sum_{t=1}^{m} a_t^{(i)} + \sum_{t=1}^{m} a_t^{(i)}\xi_t)^+] \leq E[c_0(X_i - \sum_{t=1}^{m} a_t^{(i)} + \sum_{t=1}^{m} a_t^{(i)}\xi'_t)^+] = \pi_i(\boldsymbol{a}^{(i)}|\boldsymbol{\xi}')$ for each i. Therefore $\pi(\boldsymbol{a}^*_{\xi}) = \pi(\boldsymbol{a}^*_{\xi}|\boldsymbol{\xi}) \leq \pi(\boldsymbol{a}^*_{\xi'}|\boldsymbol{\xi}) \leq \pi(\boldsymbol{a}^*_{\xi'}|\boldsymbol{\xi}') = \pi(\boldsymbol{a}^*_{\xi'})$.

C.4 Proof of Proposition 4.5.4

$\{B(p_t), p \in (0,1)\}$ is a family of independent random variables parameterized by p_t. Let $\bar{F}_t(k, p_t) = P(B(p_t) > k)$. Because $\bar{F}_t(k, p_t)$ is convex linear in p_t for each fixed k, $B(p_t)$ is SIL(st) – see Shaked and Shanthikumar (2007) for details. Next, we observe that the function $h(\boldsymbol{b}) = x + \sum_{t=1}^{m_i} b_t^{(i)} - m_i$, where x is a realization of random demand X_i, $b_t^{(i)}$ is a realization of $B(p_t^{(i)})$, $p_t^{(i)}$ is the t-th element of vector $\boldsymbol{p}^{(i)}$, and m_i is the cardinality of the vector $\boldsymbol{p}^{(i)}$, is an increasing *valuation* in $\boldsymbol{b}$. A function is said to be a valuation if it is both sub- and supermodular (Topkis 1998). Define $Z(\boldsymbol{p}) = h(B(p_1^{(i)}), \cdots, B(p_{m_i}^{(i)}))$. Then, it follows immediately that $\{Z(\boldsymbol{p}), \boldsymbol{p}^{(i)} \in (0,1)^{m_i}\}$ is SI-SchurCX(icx) – see Liyanage and Shanthikumar (1992, Theorem 2.12). This theorem also states that for an increasing convex function g, $\boldsymbol{p}^{(i)} \leq_M \boldsymbol{p'}^{(i)}$ implies that $E[g(h(B(p_1'^{(i)}), \cdots, B(p_{m_i}'^{(i)})))] \leq E[g(h(B(p_1^{(i)}), \cdots, B(p_{m_i}^{(i)})))]$. The statement of the proposition then follows from the fact that $c_0(\cdot)$ is an increasing convex function.

C.5 Proof of Corollary 4.5.5

Follow the logic in the proof of Proposition 4.5.4 in the previous section, and let $h(\boldsymbol{b}) = \sum_{t=1}^{m_i} b_t^{(i)}$. It is straightforward to argue that $h(\boldsymbol{b})$ is an increasing valuation in $\boldsymbol{b}$. Let $Z(\boldsymbol{p}) = h(B(p_1'^{(i)}), \cdots, B(p_{m_i}'^{(i)})) = \sum_{t=1}^{m_i} B(p_t^{(i)})$. $\{Z(\boldsymbol{p}), \boldsymbol{p}^{(i)} \in (0,1)^{m_i}\}$ is SI-SchurCX(icx). Therefore, by Liyanage and Shanthikumar (1992), for any increasing convex function g, $\boldsymbol{p}^{(i)} \leq_M \boldsymbol{p'}^{(i)}$ implies that $E[g(h(B(p_1'^{(i)}), \cdots, B(p_{m_i}'^{(i)})))] \leq E[g(h(B(p_1^{(i)}), \cdots, B(p_{m_i}^{(i)})))]$. In other words, if $\boldsymbol{p}^{(i)} \leq_M \boldsymbol{p'}^{(i)}$, then $\sum_{t=1}^{m_i} B(p'^{(i)}_t) \leq_{icx} \sum_{t=1}^{m_i} B(p_t^{(i)})$. Let $\bar{B}_i(\boldsymbol{p}) = \frac{1}{m_i}\sum_{t=1}^{m_i} B(p_t^{(i)})$ be the random proportion of nurses who are absent from work. Then $\boldsymbol{p}^{(i)} \leq_M \boldsymbol{p'}^{(i)}$ implies $\bar{B}_i(\boldsymbol{p'}) \leq_{icx} \bar{B}_i(\boldsymbol{p})$.