

# Capacity allocation and flexibility in primary care

Hari Balasubramanian, Ana Muriel, Asli Ozen, Liang Wang, Xiaoling Gao, Jan Hippchen

August 16, 2011

## Abstract

In this chapter, we discuss capacity allocation for primary care practices at three different levels of the capacity planning hierarchy. The goal in each case is to maximize two important but often conflicting metrics for any primary care practice are: (1) Timely Access and (2) Patient-physician Continuity. Timely access focuses on the ability of a patient to get access to a physician (or provider, in general) as soon as possible. Patient-physician continuity refers to building a strong or permanent relationship between a patient and a specific physician by maximizing patient visits to that physician.

At the highest level, the design of physician panels, we demonstrate the impact of case-mix, or the type of patients in a physician's panel, on the ability to provide timely access and continuity. Case mix can be considered using age and gender as predictors, or, when patient clinical data is available, using comorbidity counts. Using case-mix a practice can create overflow profiles for the physicians in the practice as a function of daily capacity and determine which physicians are overburdened. This in turn can point to opportunities for redesigning panels so that patients can see their own PCP as much as possible and redirections to unfamiliar physicians are minimized.

Panel redesign, however, involves changing existing patient-physician relationships. A viable alternative to redesign is managing the inherent flexibility of primary care physicians to see patients of other physicians. We study this flexibility at the aggregate (or tactical) as well as the dynamic (or operational) levels. The management of a flexible practice in the aggregate requires allocating capacity to two types of appointments: 1) prescheduled appointments which are booked in advance and require continuity with the patient's PCP; and 2) same-day or open access appointments which have to be fulfilled during the course of the day. We propose a framework, commonly observed in practice, in which the short notice open access appointments can be flexibly shared between physicians while mandating continuity for the prescheduled appointments. We show that greedy algorithms find the optimal capacity allocation under no flexibility (i.e. patients can only see their own physician) and under full flexibility (patients can see any physician in the practice). Using a two-stage stochastic integer programming model, we demonstrate the impact of flexibility on the ability to provide timely access to patients, measured by the number of patients seen a given workday. Specifically, we find that a partially flexible practice which restricts the number of physicians a patient sees to two but creates a closed chain between panels and physicians (a 2-chain) performs almost as well as the fully flexible practice with regard to timely access, without severely compromising continuity. The impact of flexibility increases as the number of physicians in the practice increases and as the demand loads between physicians are asymmetric or uneven. Our results also show that practices can heuristically determine their capacity allocation for prescheduled appointments depending on their flexibility configuration and overall system workload.

Finally, the implementation of flexibility at the level of a workday has to be made under partial demand information, since calls arrive dynamically over the course of a day. We outline a decision framework to evaluate the impact of flexibility in this dynamic case and discuss heuristics that practices can use to balance timely access and continuity.

# 1 Introduction

Primary care providers (PCPs) are typically the first point of contact between patients and health systems. Broadly they include family physicians, general internists, geriatricians and pediatricians. From a patient’s perspective, PCPs provide the majority of care they receive during their lifetime. They are responsible for a variety of health services including preventive medicine, patient education, routine physical exams, and the coordination of complex episodes in which patients are referred to medical specialties for secondary and tertiary care. The benefits of a strong primary care system are well documented in the clinical literature. Shi, Starfield and Macinko (2005) show that increased access to primary care 1) improves access to health services for relatively deprived population groups; 2) has a strong positive relationship with prevention and early management of health problems; and 3) leads to increased familiarity with patients and, consequently, to less wasteful expenditures due to inappropriate specialist care.

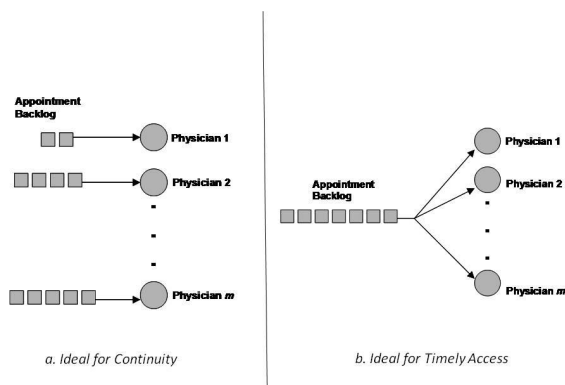


Figure 1: Dedicated versus pooled physician capacity illustrating the best cases for continuity and timely access respectively.

access, or the ability to secure an appointment quickly, is well known in the operations research literature. Rust et al. (2008) report that the inability to get a timely appointment to a primary care physician increases the likelihood of patients visiting the ER. This hinders the appropriate management of chronic diseases that could have been effectively treated in a primary care setting.

Patient-physician continuity is less familiar to an operations research audience. It is one of the hallmarks of primary care and promotes a long-term relationship between the patient and the physician. Numerous studies have documented the importance of continuity to patients [Malley and Cunningham (2009), Atlas et al. (2009)]. Gill and Mainous (1999) point to several studies which show that patients who regularly see their own providers are 1) more satisfied with their care; 2) more likely to take medications correctly; 3) more likely to have problems correctly identified by their physician; and 4) less likely to be hospitalized. Gill et al. (2000) show a link between lack of continuity and increased emergency department use. Continuity is especially important for vulnerable patients with a complex medical history and mix of medications [Nutting et al. (2003)] – patients with long standing chronic conditions (diabetes, hypertension and coronary artery disease for example). This forms a large percentage of the US population. The concept of a medical home, which has gained much traction in the last few years, emphasizes the need for continuity and long term coordination of care [see report, AAFP (2007), by the major family physician societies].

Despite its pivotal role, primary care currently pays much less than specialties, which has the effect of driving medical students away from pursuing careers in the former. This is one of main reasons for the current nationwide shortage in primary care. A recent study by the American College of Physicians (2006) reports that primary care in the United States “is at grave risk of collapse due to a dysfunctional financing and delivery system.” They also emphasize the growing demand for primary care; by 2015, “an estimated 150 million Americans will have at least one chronic condition”.

Timely access to care and patient-physician continuity, the two metrics important to primary care practices, have been adversely affected due to these broader trends. The focus on timely access, or the ability to secure an appointment quickly, is well known in the operations research literature.

Continuity is also beneficial for physicians, since their workloads are more focused [O’Hare and Corlett (2004)].

How are timely access and patient-physician continuity related to capacity planning and allocation in primary care? When it comes to access to appointments, the two measures are often in conflict. It may be possible for a patient to get a same-day appointment but not necessarily with her own physician. Alternatively, a patient may get to see her own physician but only weeks later. The two extremes are illustrated in Figure 1. Figure 1 (a) shows the situation where patients see only their own physician while in Figure 1(b) all physician resources are pooled. In the former continuity is perfect but timely access may be strained, while the latter results in high levels of timely access but patients may end up seeing unfamiliar physicians.

The focus of this chapter is on capacity planning and allocation for primary care practices at various levels of the planning hierarchy, to balance timely access and continuity. First, we consider the impact of size and composition of a physician’s *panel* on the ability to provide timely access and continuity. A panel refers to the set of patients whose care the PCP is responsible for. Next, we discuss how a multi-physician practice can manage the inherent flexibility of primary care physicians to see patients of other physicians to balance timely access and continuity. We study physician flexibility at the aggregate level where two uncertain demand streams: *prescheduled* (non-urgent) appointments and *open access* (same-day) appointments are sequentially realized and must share the same capacity. We also discuss the dynamic context, where decisions about capacity allocation under flexibility have to be made as patient requests for appointments arrive over the course of the day, that is, under incomplete demand information.

## 2 Literature Review

The application of operations research to appointment scheduling in healthcare is a growing area of research. We focus here only on the papers most relevant to our research in the primary care context.

Over the last decade the adoption of *advanced access*[Murray and Berwick, 2003], which urges practices to provide same-day appointments irrespective of the urgency of the request, has brought to the forefront questions regarding appointment system design. What should physician panel sizes be to allow open access? What if patients prefer to have appointments at some future time rather than see a doctor the same day? These questions have necessitated the use of queuing and stochastic optimization approaches that provide guidelines to practices. For instance, Green et al. (2007) investigate the link between panel sizes and the probability of ”overflow” or extra work for a physician under advanced access. They propose a simple probability model that estimates the number of extra appointments that a physician could be expected to see per day as function of her panel size. The principal message of their work is that for advanced access to work, supply needs to be sufficiently higher than demand to offset the effect of variability. Green and Savin (2008) use a queuing model to determine the effect of no-shows on a physician’s panel size. They develop analytical queuing expressions that allow the estimation of physician backlog as a function of panel size and no-show rates. In their model, no show rates increase as the backlog increases; this results in the paradoxical situation where physicians have low utilization even though backlogs are high – this is because patients that had to wait for long, do not show up.

Gupta et al. (2006) conduct an empirical study of clinics in the Minneapolis metropolitan area that adopted open access. They provide statistics on call volumes, backlogs, number of visits with own physician (which measures continuity) and discuss options for increasing capacity at the level

of the physician and clinic. Kopach et al. (2007) use discrete event simulation to study the effects of clinical characteristics in an open access scheduling environment on various performance measures such as continuity and overbooking. One of their primary conclusions is that continuity in care is affected adversely as the fraction of patients on open access increases. The authors mention provider groups (or physicians and support staff) working in teams as a solution to the problem. Robinson and Chen (2010) compare the performance of open access with that of a traditional appointment scheduling system. Their numerical analysis reveals that unless patient wait times to secure an appointment have marginal weights in the objective function and patient no-show rates are too small, open access is preferable to traditional scheduling systems. Liu et al. (2010) propose new heuristic policies for dynamic scheduling of patient appointments under no-shows and cancellations. They find that open access works best when patient load is relatively low.

The most closely related papers to our study are by Qu et al (2007) and Gupta and Wang (2008). Qu et al (2007) derive conditions under which a solution for the number of prescheduled appointments to reserve is locally optimal. In Section 5, we show a stronger result, guaranteeing global optimality, by first showing that our revenue maximization function has diminishing returns under mild assumptions. Gupta and Wang (2008) explicitly model many of the key elements of a primary care clinic. They consider scheduling the workday of a clinic in the presence of 1) Multiple physicians 2) Two types of appointments: same-day as well as non-urgent appointments 3) Patient preferences for a specific slot in a day and also a preference for physicians. The objective is to maximize the clinic’s revenue. They use a Markov Decision Process (MDP) model to obtain booking policies that provide limits on when to accept or deny requests for appointments from patients. In terms of flexibility, their clinic is fully flexible with regard to both non-urgent and urgent appointments. The principal difference between their model and the capacity allocation framework proposed in Section 5 is that patient preference drives the scheduling of prescheduled appointments in Gupta and Wang (2008), while we try to balance pre-scheduled demand and same-day demand through physician flexibility and an explicit consideration of its effect on timely access and continuity.

### 3 Panel Case-Mix

#### 3.1 Patient types

At the highest level, a physician builds a panel of patients. The physician’s appointment burden depends on the 1) size; and 2) case-mix or composition of the panel. A physician working full time may have 1500-2000 patients. Case-mix refers to the type of patients in the panel, and can be characterized by various patient attributes, such as age and gender and the chronic conditions afflicting the patient, which play an important role in determining the distribution of visits. For example, a panel where the majority of patients are young and healthy will have a different appointment profile compared to a panel consisting mostly of elderly patients with chronic conditions. Patient classification can be useful for clinics because they enhance a practice’s understanding of its population and disease trends, and allow it to design its care models effectively. Furthermore, Barbara Starfield’s seminal work about ACGs (Ambulatory Care Groups) [Starfield et al, 1991] argued that understanding the role of patient clinical complexity in care utilization forms the cornerstone for effective resource planning and determining payment methods in healthcare.

Age and gender is the simplest patient classification in absence of other data, yet is generally effective [Murray et al., , Balasubramanian et al. 2010]. Figure 2 illustrates the distribution of the fraction (or percentage) of total patients requesting appointments in a week for two categories - males (48-53 y.o.) and women (73-78 y.o.), based on on historical data from 2004-2006 (156 weeks), from the Primary Care Internal Medicine Practice (PCIM), Mayo Clinic. The two distributions

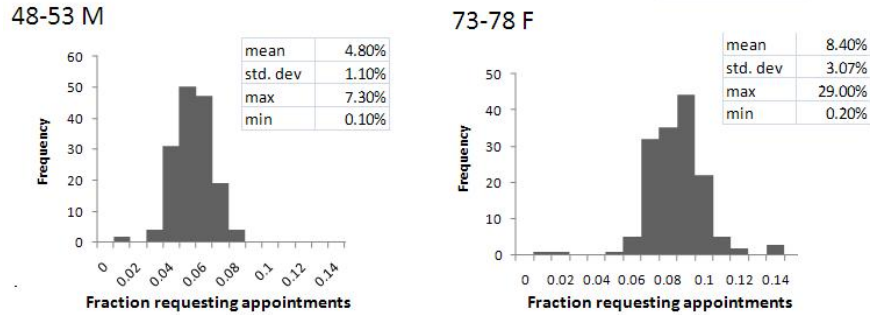


Figure 2: Histograms of the percentage (or fraction) of total patients requesting appointments in a week for two different patient age and gender categories.

show how appointment request rates can vary with gender and age. The distributions are different with regard to both mean and variance. There are 708 males 48-53 y.o (48-53 M) and 986 females 73-78 y.o (73-78 F) empanelled in the practice. 8.4% of all 73-78 F patients request for appointments on average in a week as opposed to 4.8% of all 48-53 M patients. The standard deviation 73-78 F (3.07%) is more than double that of 48-53 M (1.1%).

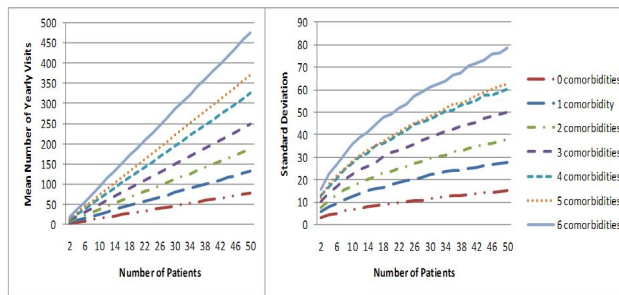


Figure 3: Mean and standard deviation of yearly visits for groups with different counts of comorbidities

Naessens et al [2011] show that more than an individual chronic condition such as diabetes or hypertension, it is the number of simultaneous chronic conditions (or *comorbidities*) that determines the consumption of healthcare costs. Furthermore, from a care point of view focusing on all comorbidities of a patient is more holistic than focusing in isolation on specific chronic conditions. Figure 3 shows mean and standard deviation of visit rates as a function of the number of patients under various counts of comorbidities. The data was simulated based on historical visits of 20,000 patients empanelled PCIM [Ozen and Balasubramanian, 2011]. Clearly, not only does the mean number of visits increase with the number of comorbidities, the variance does as well. For instance if a physician has 50 6-comorbidity patients then he will have 450 appointment requests on average each year. If he has same number of 0-comorbidity patients he will have only 75 yearly visits on average.

### 3.2 Example of 4 physicians

We now consider an example of four physicians with approximately the same panel size (1050 patients), but different case-mixes, based on comorbidity counts. These panel compositions are shown in Figure 4. The case-mix can be used to simulate the distribution of daily visits for

each physician, by sampling for each comorbidity count from historical data. Once the daily visit distribution is obtained, the *overflow* for a given daily appointment capacity can be calculated. Overflow is simply the fraction of total samples in which the patients’ visit requests exceed the available capacity of the physician. Patients that are not seen either visit an unfamiliar physician or an ER, or may choose to wait to see the physician on another day. Thus, if overflow is high, both timely access and continuity are adversely affected.

	Number of comorbidities								Panel Size
	0	1	2	3	4	5	6	7	
Physician 1	260	249	226	161	108	42	14	3	1063
Physician 2	299	293	212	147	77	26	6	1	1062
Physician 3	214	253	223	177	115	44	21	5	1053
Physician 4	290	296	218	145	84	27	12	5	1077

Figure 4: Four physicians at PCIM, Mayo Clinic and their patient case-mix based on comorbidity count.

– for example two years – and dividing it by the number of unique class  $i$  patients as well as the number of workdays in the two year period. The method is similar to the one proposed in Green et al. (2007). Next, suppose  $n_{ij}$  denotes the number of class  $i$  patients in physician  $j$ ’s panel. If we assume that each patient requests independently of others, then the total requests from each patient class for physician can be modeled as a binomial random variable, with mean  $n_{ij}p_i$  and variance  $n_{ij}p_i(1 - p_i)$ . Going further, the mean demand for the entire panel is given by  $\mu_j = \sum_{i=1}^M p_i n_{ij}$  and standard deviation  $\sigma_j = \sqrt{\sum_{i=1}^M p_i(1 - p_i)n_{ij}}$ . If we assume that the sum of  $m$  binomial random variables gives us a normal random variable, then  $O_j$ , the overflow for physician  $j$  is related to the percentile of the standard normal distribution, given by  $\phi$ , in the following way:  $O_j = 1 - \phi(\frac{C_j - \mu_j}{\sigma_j})$ . Here  $C_j$  is the capacity of the physician, that is the total daily slots that she has available in a day.

Note that this analysis is at the aggregate level – it does not consider the actual duration of appointments once patients are in the clinic, but tests whether the number of appointment slots (typically 20-minute slots) a physician plans to have available in a day is sufficient. It also assumes that all appointments are of the same type. In reality, some appointment requests (such as follow-up appointments) are for a future day, while some are same day requests. Nevertheless, if overflow is high for all appointments, then it is guaranteed that the timely access for both same-day as well as non-urgent future appointments with one’s own PCP will be adversely affected.

The overflow for the four physicians of Figure 5 as a function of the total daily slots (capacity) is shown in Figure 5. We calculated these overflow profiles using the binomial approximation described above, but it is also possible obtain the same curves by sampling from historical visit data. For the same capacity, Physician 3 and Physician 1 have relatively high levels of overflow. This is because there are more patients with two or more comorbidities in their panels (see Figure 4), and these patient groups generate a higher number of visits. This graph shows that it is inappropriate for clinics to make capacity decisions based solely on panel size. Case-mix is also an important consideration. It is also clear that to keep overflow levels down to manageable levels, 20 or more appointment slots may be needed for each of the 4 physicians.

Such analysis allows practices to identify which physicians are overburdened. In the above case, it’s clear that physicians 3 and 1 need to have their capacity enhanced – either by working extra hours in a day or by additional nurse practitioner support. The long-term option for practices is to redesign panels, by moving high demand, high variability patients from an overburdened physician

Overflow can also be modeled in the following way. First a practice determines  $p_i$ , the probability that a patient of class  $i$  will request an appointment on any given day. This can be obtained by calculating the total visits generated by all patients of the class  $i$  in the practice over a period of time

to a physician with available capacity. More details about the panel redesign approach are presented in Balasubramanian et al. [2010]. The paper shows that it is possible to improve the wait time and the number of redirections to unfamiliar physicians by more than 35%.

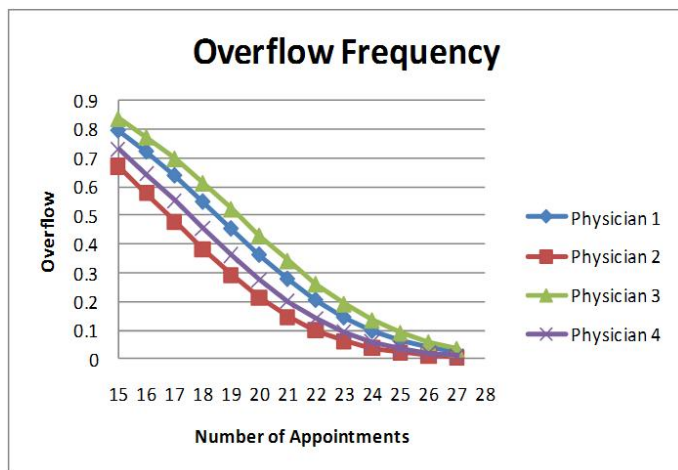


Figure 5: Overflow for the physicians as a function of the daily capacity (appointment slots)

new patients, and the turnover of existing patients. Patient surveys could be used to determine preferences and inclination towards change. In some cases, to minimize disruption, reassignment may simply be to another physician, whom the patient has seen almost as often as her own PCP, or to a physician within the same care team (if the care team consists of multiple physicians).

Another viable alternative to panel redesign is carefully managing physicians’ ability to see patients of other physicians, depending on whether the requests are urgent/same-day requests or non-urgent requests. This management of physician flexibility forms the content of the next two sections.

## 4 Flexibility in Primary Care

The inherent flexibility of primary care physicians to see patients from other panels gives practices another lever to provide timely access to care. Using this flexibility, of course, comes at a cost: the resulting loss of continuity when a patient sees unfamiliar physicians. How should practices be designed and managed to use this flexibility to better balance timely access and continuity?

A practice can achieve maximum continuity of care by mandating that patients should see only their own provider. This, however, hampers timely access to care. At the other extreme, a practice may allow patients to see any provider. This is ideal for timely access, but hampers continuity of care. The two extremes are shown in Figure 6 (a) and 2 (b). In the first case, the providers are dedicated while in the second the providers are fully flexible. Figures 6 (c), (d) and (e) show partially flexible configurations that offer a middle ground between (a) and (b). In each of them, a patient sees only one physician other than her own PCP. Figure 6 (c) is referred to as the 2-chain in the manufacturing flexibility literature [Jordan and Graves, 1995] and allows demand variation to be absorbed effectively by the entire practice. While the 2-chain is a concept new to healthcare,

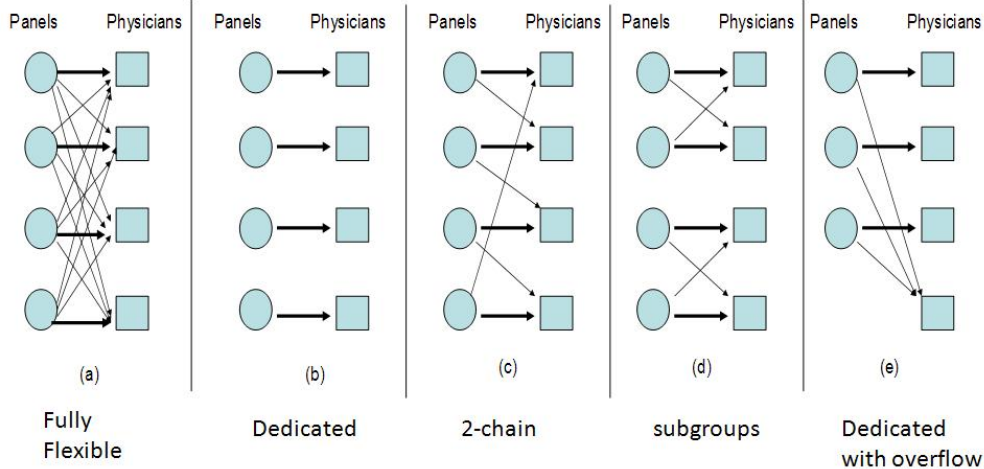


Figure 6: Figure illustrating different flexibility configurations that tradeoff continuity of care with timely access.

practices do use the subgroup configuration (Figure 6 (d)). Physicians here may be divided into independent, self-contained teams (such as in the Mayo Clinic and other academic primary care practices). The dedicated with overflow configuration of Figure 6 (e) is also common - here if the patient's PCP is unavailable, the patients tend to see an "overflow" physician or nurse practitioner (we have observed this setting at a small private practice as well as a community clinic in Western Massachusetts; academic medical centers also use this model).

In the ideal open access world, there are no appointment types, such as urgent and non-urgent. All appointments are treated identically and scheduled the same day with the patient's PCP. However, the reality is that clinics have a fraction of their schedule available for open access or urgent appointments. Such appointments, because of the perceived immediacy of need, are typically seen the same day, but not always by the patient's personal physician. The rest of the clinic's schedule consists of appointments booked a week or more in advance. We call these appointments prescheduled appointments. These non-urgent appointments are typically physicals or follow-up appointments for patients with chronic conditions. While the loss of continuity has to be minimized for all appointments, it can be appropriately sacrificed for urgent appointments needing immediate attention by introducing some form of flexibility. We will thus assume that flexibility applies only to urgent appointments. Non-urgent or prescheduled appointments are always seen by the patient's own provider.

We approach the design and management of the flexible practice at two different levels. At the tactical or aggregate level in primary care, the design of flexibility needs to consider capacity allocation for the two streams of uncertain demand, non-urgent (prescheduled) and urgent (open access), each with different requirements for timeliness and continuity; while at the operational or dynamic level, at which same-day, urgent appointments are booked as patients call over the day, allocation decisions have to be made in real time without full knowledge of demand.

In the remainder of this section, we first summarize the main lessons from the flexibility literature as they apply to the primary care context, and then consider the aggregate and dynamic cases in detail.

## 4.1 Flexibility Literature Perspective

Our study of flexibility in primary care practices builds upon the extensive literature on manufacturing flexibility and its more recent application to service systems and worker training and allocation. There are, however, key operational differences that make the application of flexibility to primary care worthy of further analysis: (1) two demand streams associated with each resource, where one (prescheduled demands) gets realized before the other (open access demands); (2) two conflicting objectives, timeliness and continuity of care; (3) no fixed cost associated with installing flexibility, but a loss in continuity for using it; (4) appointments are booked over time and thus future resource capacity is sequentially being allocated under partial demand information.

As in the case of cross-training in serial production lines (Hopp, Tekin, Van Oyen (2004)), flexibility improves efficiency in two main ways in the primary care environment. The first benefit is in what they refer to as capacity balancing: If physician panels are imbalanced with respect to the induced average number of visits to a physician per day, flexibility will allow the load to be shared between physicians, therefore improving overall timeliness of care and physician utilization. The second is in variability buffering: Even if the average workloads are balanced between physicians, variability in patient requests for a particular day/time will be better accommodated by a flexible environment. Hopp, Tekin, Van Oyen (2004) compare a strategy that balances capacity using the minimum amount of cross-training with the chaining of skills in the sequence of the serial line. They find that skill-chaining strategies are more robust, and more effective in variability buffering. The concept of chaining has received much attention since it was first introduced in the seminal work of Jordan and Graves (1995). In a single-period, multi-product, multi-plant production network, they show that the 2-chain (Figure 2 (c)) results in increased sales and capacity utilization, relative to the dedicated configuration (Figure 2(a)), comparable to those achieved by a fully flexible system (Figure 2(b)). That is, a few links, configured in the right way (2-chain) provide almost the same performance as the complete, fully flexible network. Furthermore, this strategic analysis has been extended recently to multi-stage supply chains (Graves and Tomlin (2003)), and to a make-to-order environment where flexibility is also used to hedge against operational variability (Muriel, Somasundaram and Zhang (2006)). Chua, Chu and Teo (2008) distinguish between range (the different demand scenarios that can be accommodated) and response (the cost of doing so; that is, the cost of using secondary rather than primary resources for production/service) of flexible systems. They show that upgrading system response (i.e., building systems where physicians can handle other physician's panels at lower additional cost) outperforms improving system range (creating systems that can accommodate ever more extreme patient demand scenarios). This result suggests that in the primary care setting, the benefits of restricting the number of doctors that can see a particular patient (resulting in lower cost of service because of familiarity and thus increased response) is likely to outweigh the higher range provided by a fully flexible team care practice where any doctor can see the patient.

A number of computational reports in the literature (e.g. Jordan and Graves (1995)) point out an increase in the marginal benefit associated with adding one more flexibility link (i.e. allowing one more panel to see a second physician) in forming the 2-chain, culminating with a markedly higher increase when the last link that closes the chain is put in place. Recently, Simchi-Levi and Wei (2011) prove that that is always the case and show that long chains are always superior to any other strategy where each product (panel) can be produced at two plants (can be assigned to two physicians.) This suggests that the larger practices will benefit most by managing their inherent flexibility in the form of a long chain.

## 5 Flexibility in the aggregate case

How much of the physician’s total daily workload should be dedicated to prescheduled versus urgent appointments? Well, this will depend on how much flexibility the practice allows when allocating urgent patient demand. We thus need to address this question under different flexibility configurations in Figure 2. This will also allow us to compare their resulting performance in terms of system revenue, continuity and timely access. For that purpose, we develop a 2-stage stochastic integer program that can accommodate any flexibility configuration, and greedy, but exact, algorithms to quickly calculate the optimal capacity allocations in dedicated and fully flexible systems. The analytical and experimental results and conclusions summarized here are from Balasubramanian et al. (2011).

**Two-stage capacity allocation model:** We solve this capacity allocation problem for a work-day using a 2-stage stochastic integer program, shown below. We consider a general primary care practice with  $M$  physicians, each with  $N_i$  available appointment slots,  $i = 1, 2, \dots, M$ . Let  $A$  be the set of all possible panel-physician links  $(i, j)$  such that the open access (same-day) requests of patients in panel  $i$  (i.e., physician  $i$ ’s panel) can be served by physician  $j$ . The set  $A$  represents the particular flexibility configuration under consideration. Let  $R_i^p$  be the revenue associated with physician  $i : i = 1, 2, \dots, M$  seeing one of his pre-scheduled patients, and  $R_{ij}^o$  be the revenue associated with physician  $j$  seeing an open-access patient of panel  $i$ , for any  $(i, j) \in A$ . The demand for prescheduled and open access appointments can be represented by a random vector  $D = (D_1^p, D_1^o, \dots, D_M^p, D_M^o)$  where the super-index  $p$  refers to prescheduled and  $o$  to open access, and the sub-index indicates the primary care physician.  $D$  follows a discrete distribution that assigns a probability  $q_s$  to each possible realization of demand, indexed by  $s$ ,  $s = 1, 2, \dots, S$ , where  $S \equiv S_1^2 \times S_2^2 \times \dots \times S_M^2$ . That is,  $P[D = (d_{1s}^p, d_{1s}^o), \dots, d_{Ms}^p, d_{Ms}^o] = q_s$ . We introduce the following capacity allocation variables:

$N_i^p$ : Number of slots allocated for pre-scheduled demand of physician  $i$ .

$x_{is}^p$ : Number of patients allocated to physician  $i$  under demand realization  $s$

$x_{ijs}^o$ : Number of open access patients of panel  $i$  assigned to physician  $j$  under demand realization  $s$ , for all  $i = 1, 2, \dots, M$  and  $(i, j) \in A$ .

The objective is to maximize the expected revenue of satisfying prescheduled and open access appointments. We use binary variables to capture whether the prescheduled demand for a physician is less or greater than the corresponding  $N_i^p$  value. Equation (3) ensures that  $\phi_{iu_{is}} = 1$  if  $d_{is}^p < N_i^p$ . Equation (4) ensures that  $\phi_{iu_{is}} = 0$  if  $d_{is}^p > N_i^p$ . Equations (5) and (6) limit the number of pre-scheduled appointments to the allocated capacity and the realized demand, respectively. Equations (7) and (8) ensure that the total open access appointments for any physician  $j$  do not exceed remaining capacity, when  $\phi_{iu_{is}} = 1$  and  $\phi_{iu_{is}} = 0$  respectively. Equation (9) limits the total number of open access appointments scheduled from a panel to the realized demand for such appointments from that panel. Equation (10) is the binary constraint.

$$(SIP) \quad \max \left\{ \sum_{s=1}^S \sum_{i=1}^m q_s \{ R_i^p x_{is}^p + \sum_{\{i,j\} \in A} R_{ij}^o x_{ijs}^o \} \right\} \quad (1)$$

$$s.t. N_i^p \leq N_i \quad \forall i \quad (2)$$

$$N_i^p \leq d_{is}^p + N_i \phi_{iu_{is}}, \quad \forall (i, s) \quad (3)$$

$$N_i^p \geq d_{is}^p \phi_{iu_{is}}, \quad \forall (i, s) \quad (4)$$

$$x_{is}^p \leq N_i^p \quad \forall (i, s) \quad (5)$$

$$x_{is}^p \leq d_{is}^p \quad \forall (i, s) \quad (6)$$

$$\sum_{i:(i,j) \in A} x_{ijs}^o \leq N_j - d_{js}^p \phi_{ju_{js}} \quad \forall (j, s) \quad (7)$$

$$\sum_{i:(i,j) \in A} x_{ijs}^o \leq N_j - N_j^p + \phi_{ju_{js}} N_j \quad \forall (j, s) \quad (8)$$

$$\sum_{i:(i,j) \in A} x_{ijs}^o \leq d_{is}^o, \quad \forall (i, s) \quad (9)$$

$$\phi_{iu_{is}} \in (0, 1), \quad \forall i, \forall u_{is} = 0, 1, \dots, N_i \quad (10)$$

$$N_i^p, x_{is}^p, x_{ijs}^o \geq 0, \quad \forall (i, j) : (i, j) \in A, \forall s \quad (11)$$

We note in the above revenue optimization that  $R_{ij}^o > R_{ij}^p$ . This is because in a relative sense, open access appointments are more “valuable” than prescheduled appointments, as explained in Balasubramanian et al. (2011). First, we note that open access appointments, because they have such short lead times, have much lower no-show rates. Second, if a prescheduled appointment results in a no-show, it can be substituted by an open access appointment, while the reverse is not possible at such a short notice. Third, because prescheduled appointments are made generally a week or more in advance, the patient is likely to be flexible about choice of the appointment day, and thus this may result in postponed but not lost demand if denied timely access. An open access patient, on the other hand, needs to see a physician immediately and hence is flexible in provider choice.

In the next two sections, we present analytical solutions to the capacity allocation problem for dedicated practices, where physicians can only see patients in their own panel, and fully flexible practices where open-access patients can be seen by any of the physicians in the practice. For large practices using partial flexibility such as the 2-chain configuration, unfortunately, the above stochastic program is too large to solve efficiently in practice. While the number of binary and integer variables is quite manageable, the sheer number of possible demand realizations makes the problem intractable. To overcome this issue, we will solve the problem using a computationally effective sample average approximation method proposed by S. Solak [34] for two-stage stochastic integer programming problems; see Section 5.1.

**The Dedicated Case:** In a dedicated practice, physicians can only serve the prescheduled and open access patients from their own panel. They need to decide, however, a maximum number of appointment slots to make available to prescheduled patients,  $N_i^p$ , so that enough capacity is reserved for the more lucrative open access ones. The system configuration is shown in Figure 7.

Let  $ER_i(N_i^p)$  be the total expected revenue from the panel of physician  $i$ , as a function of  $N_i^p \in 0, 1, 2, \dots, N$ . Our goal is to find the optimal value of  $N_i^p$ . The conditions for local optimality presented in Qu et al. (2007) for the problem of maximizing the expected number of patients consulted in a single-physician practice can be easily adapted to our revenue maximizing objective.

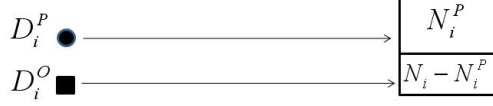


Figure 7: System configuration for a dedicated practice.

In Balasubramanian et al. (2011), we show a stronger result, guaranteeing global optimality, by first showing that the objective function has diminishing returns under mild assumptions.

**Proposition 1.** *If  $P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p]$  is non-decreasing in  $N_i^p$  the difference in revenue associated with increasing the number of prescheduled slots by one,  $ER_i(N_i^p + 1) - ER_i(N_i^p)$ , is non-increasing in  $N_i^p$*

The above condition holds when the demand for prescheduled and open-access appointments are independent of each other. Furthermore, it will be satisfied in most practical scenarios. Intuitively, for it to be violated, the probability of open access demand being large would need to significantly decrease as the demand for prescheduled appointments grows; that is, the demand for open access and prescheduled appointments would need to be heavily negatively correlated.

As a result of Proposition 1, we have that the expected revenue function exhibits diminishing returns, an analog of concavity for a discrete function, and thus its global maximum must occur at the largest integer  $N_i^p \leq N$  such that  $ER_i(N_i^p) - ER_i(N_i^p - 1) \geq 0$  leading to the following theorem.

**Theorem 1.** *If  $P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p]$  is non-decreasing in  $N_i^p$ , the optimal solution to the Dedicated Problem is the largest non-negative integer  $N_i^p \leq N$  such that  $P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p] \leq R_i^p / R_i^o$ .*

The optimal solution can thus be easily obtained by calculating that probability starting at  $N_i^p = 0$  and increasing one unit at a time until it exceeds the threshold  $R_i^p / R_i^o$ . A binary search could also be used. Observe that in the case of independent open-access and prescheduled demands, the optimal value of  $N_i^p$  does not depend on the distribution of pre-scheduled demand for physician  $i$ .

**Fully flexible practice:** In a fully flexible practice, open-access patients can be seen by any available physician. In this case, the optimal number of slots to make available to prescheduled demand of the physicians,  $N_1^{p*}$  and  $N_2^{p*}$  in the case of two physicians, can still be found with a simple greedy algorithm. This is because the revenue function again exhibits diminishing returns as the number of slots offered to prescheduled patients is increased. For ease of exposition, we assume that all physicians have the same capacity of  $N$  slots, and that the revenue of an open access appointment is identical for all physicians and panels and denoted by  $R^o$ . We first consider the case of two physicians,  $i$  and  $j$ . See Figure 8.

**Proposition 2.** *If  $P[D_i^o + D_j^o > 2N - (N_i^p + \min(D_j^p, N_j^p) + 1) | D_i^p \geq N_i^p + 1]$  is non-decreasing in  $N_i^p$  and  $N_j^p$ , the difference in revenue associated with increasing the number of prescheduled slots of physician  $i$  by one,  $ER_i(N_i^p + 1, N_j^p) - ER_i(N_i^p, N_j^p)$ , is non-increasing in  $N_i^p$  and  $N_j^p$ .*

Observe that, as in the dedicated case, the conditions will hold when open access and prescheduled demands are independent, and in any practical scenario except for contrived cases where the demands for prescheduled and open access appointments are severely negatively correlated. Since the revenue function exhibits decreasing returns in both  $N_i^p$  and  $N_j^p$  under those mild conditions, which can be interpreted as concavity of the discrete revenue function, a greedy algorithm that

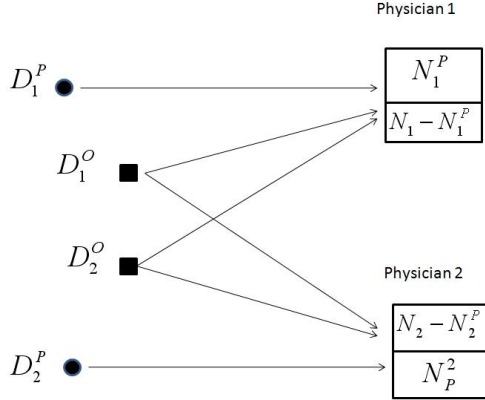


Figure 8: System configuration for a fully flexible practice.

keeps increasing one appointment slot at a time to the physician where it produces the highest system revenue will provide the optimal capacity allocation scheme.

Proposition 2, and therefore the optimality of a greedy algorithm, can be easily extended to the general case of  $M$  physicians that fully share open access demand. Full details of the theorems and the proofs are given in Balasubramanian et al. [2011].

## 5.1 Computational Experiments

The exact greedy algorithms allow us to find the optimal capacity allocation and system revenues for dedicated and fully flexible practices. To test the performance of partial flexibility configurations (see Figure 6), which promote continuity by restricting the number of doctors that a patient can be assigned to, we use the two stage stochastic integer program (SIP). In what follows we present a summary of the results emphasizing (1) the value of the 2-chain to improve open-access while keeping acceptable levels of continuity, and (2) how the optimal portion of clinic capacity reserved for open access changes as more flexibility is allowed when allocating open access demand; for full details, please see Balasubramanian et al. (2011).

### Value of Partial Flexibility

Following the findings of Bennett and Baxley (2009), we assume a typical no show rate for pre-scheduled demand of 25%, and a 10% no show rate for open access demand. Thus, an appointment slot given to an open access patient brings higher expected revenue, 0.9, as compare to revenue of 0.75 for scheduling one pre-scheduled patient. To encourage continuity in the system, we assume that there is a 0.05 cost of seeing patients from another physician's panel (the revenue of giving an appointment slot to one open access patient not from a physician's panel is therefore  $0.9 - 0.05 = 0.85$ ). While the no-show rates for the two types of appointments can be estimated from past data, the cost of diverting an open access patient to a non-PCP physician is very difficult to quantify. Furthermore, in a limited flexibility environment, where the patient only sees at most one physician beyond her PCP, the actual cost of redirection is minimal, very different from that occurring in a large, fully flexible practice where care is significantly more fragmented and much harder to coordinate. For that reason, rather than comparing the expected revenues obtained under the different configurations, we focus here on the resulting timely access rates (TAR). We define TAR as the percentage of all patients, both prescheduled and open access, who get access to an appointment.

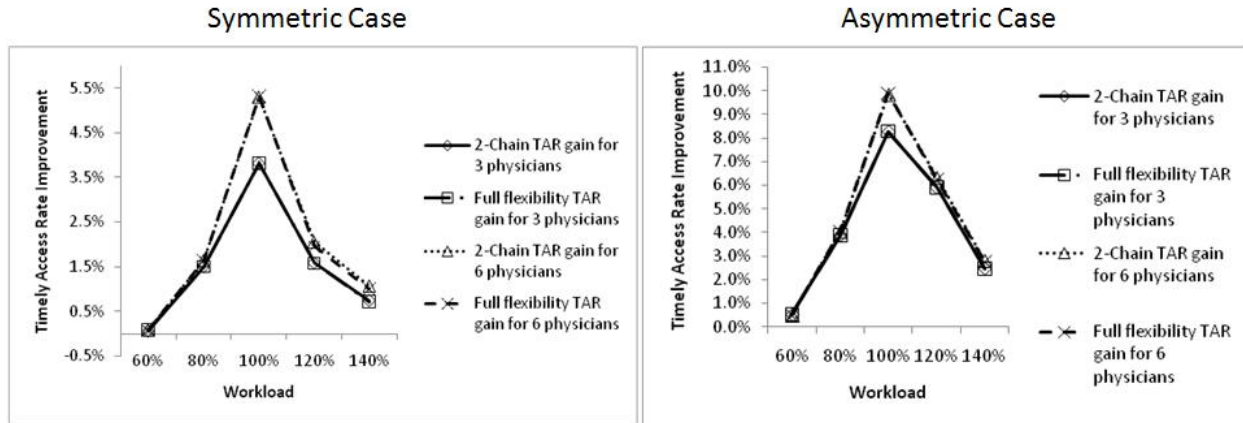


Figure 9: Comparison of timely access rate (TAR) improvement between 3 and 6 physicians in the symmetric and asymmetric cases

Figure 9 shows the gains, relative to a dedicated practice, of implementing partial flexibility (configured as a 2-chain) and full flexibility to share open access demand as system workload increases in practices with 3 and 6 physicians. Workload is defined as the ratio of the expected total demand for the clinic and total available capacity. Each physician has 24 appointment slots available in the day. The left graph, or symmetric case, involves a practice where all physicians face identical panel demand distributions (Poisson demands with a rate of 10 for prescheduled appointments and 14 for open access demand). The right graph, or asymmetric case, considers a practice where physicians have varying panel compositions and therefore varying appointment burdens. This is common in practice. Senior and well established physicians may have higher workloads since their panels are larger and include older, more complex patients, while physicians who have been recently hired may have lower workloads. In particular, we test a practice where: Physician 1 has an expected prescheduled demand of 6 and an expected open access demand of 12 (low workload); Physician 2 has an expected prescheduled demand of 8 and an expected open access demand of 16 (balanced or full workload); Physician 3 has an expected prescheduled demand of 10 and an expected open access demand of 20 (high workload). For the six physician case, we merely double the 3 physician case, thus retaining the imbalances.

The timely access rates of 2-chain flexibility and full flexibility are nearly the same no matter what the size and workload level of the system are. This is consistent with the results reported in the literature on flexibility in manufacturing settings. The difference is even lower in our healthcare setting, since we assume that prescheduled demand cannot be shared between physicians; flexibility can only be used for open access demand. We also observe, as in the manufacturing literature, that the gains accrued through flexibility increase significantly as: (1) the number of physicians increases from 3 to 6; and (2) the physicians have different workloads, i.e. in the asymmetric case, when flexibility helps not only to accommodate demand variability but also to balance physician workloads.

These results suggest that flexibility provides an important lever for practices to increase their ability to accommodate open access demand. Furthermore, the 2-chain configuration allows them to do so without severely compromising continuity and patient/physician bonds.

### Capacity Allocation

The results above discuss the value of flexibility. But how are capacity allocation decisions

affected by the flexibility configuration used? What trends do they follow, if at all, and can the trends provide clues to capacity allocation decisions in practice? In our model, the capacity allocation is decided with the optimal first stage variables,  $N_i^{p*}$ , which represent the capacity made available to prescheduled appointments. Figure 12 shows the average values for the entire clinic (that is for all the physicians) under different workloads and the three flexibility configurations for the 6 physician asymmetric case. We see the same trends by looking at the individual physicians' values (irrespective of the number of physicians, symmetry and prescheduled to open access demand ratios). Thus the figure summarizes our conclusions about  $N_i^{p*}$  values concisely.

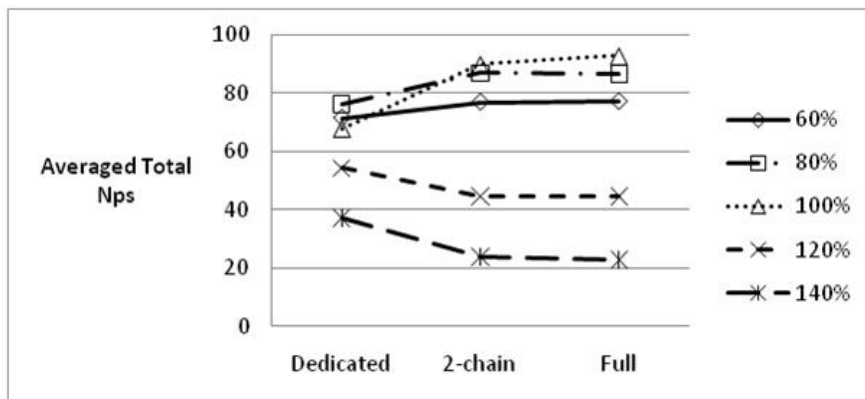


Figure 10: Trends in averaged total Np values for 6 physicians with Prescheduled demands [6,8,10,6,8,10] and open access demands [12,16,20,12,16,20]

In general, for the case of very low system workload, the total  $N_i^{p*}$  values for the dedicated and flexibility configurations, not surprisingly, are very close. Since the demands are so low, the values are likely to be fairly robust at this level. As the system or clinic workload increases to 80% and 100%, the clinic as a whole reserves more prescheduled appointments in the flexibility cases than the dedicated case. This is a direct consequence of flexibility: open access appointments can be absorbed effectively by pooling the (lower) remaining capacity of all physicians together. The effect is especially strong in the case of 100% workload: the dedicated case increases the capacity reserved for the more profitable and now more abundant open access patients ( $N - N_p$ ) relative to the lower workload cases, while the flexible configurations decrease it to allow for more of the now plentiful prescheduled patients and still meet open access demand through sharing any unused capacity.

In contrast, in the high system workload cases (120% and 140%), there is enough demand for the high revenue open access appointments to lower the total of the clinic. The flexibility cases have a lower total  $N_i^{p*}$  value than the dedicated case, reserving more capacity for open access. This is because there is a higher probability of using the additional capacity when physicians are able to see each others' open access appointments.

Thus, using the easily computable dedicated case  $N_i^{p*}$  as a reference, practices can heuristically determine their capacity allocation to be above or below the dedicated value, depending on their flexibility configuration and overall system workload.

## 6 Flexibility in the dynamic case

The above discussion of flexibility assumed that demands are instantly realized and fulfilled. In practice, however, allocation decisions for open access and same-day appointments have to be made without full realization of demand. At the beginning of the day all non-urgent appointments scheduled on physician calendars are known in advance, but calls for same-day appointments come throughout the day and have to be dynamically assigned to available physician slots. As before, the challenge is to balance timely access (minimize the number of denied same-day appointment requests) with continuity (ensuring patients see their own physician as much as possible).

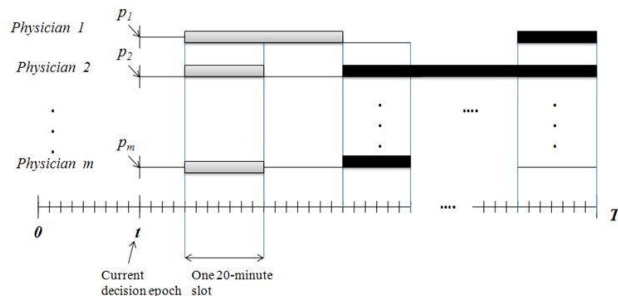


Figure 11: State of the system when certain slots may have been booked in a physician’s calendar in advance of the workday. These slots are shaded dark, while those booked the same day have a lighter shade. If a patient of panel 1 arrives, do we assign him/her to physician 1 with more available capacity or to physician 2 who has a greater likelihood of being idle on his/her only available slot?

Consider again a clinic with  $M$  physicians, each with a panel of patients for which he/she is the primary care physician (PCP). Depending on the flexibility configuration patients will be allowed to see one or more physicians. The time horizon begins when the clinic opens its open access appointments, and ends when the clinic stops taking further appointments. This will typically mean the entire duration of a day (7-8 hours). Calls for a given physician’s slots come with a certain probability ( $p_i$  for physician  $i$ ) at every time point during the horizon. Each physician’s calendar consists of successive 20-minute slots that are booked as calls come in during the day. In the ideal situation, all open access appointments are contiguous and occur during the same time of the workday. In practice, because of patient preferences for time slots, available same-day slots will be interspersed with prescheduled slots. The situation is shown in Figure 11.

The decision framework can be modeled in a finite horizon stochastic dynamic program. For the mathematical details see Hippchen (2009). At each time point  $t$ , if there is a request from physician panel  $i$ , the decision facing the clinic is whether this request should be 1) assigned to her PCP; 2) to some other physician (as allowed by the flexibility configuration) or 3) denied. If there are no requests at time  $t$  for any of the physicians then the action is to “do nothing”. The optimization problem is to choose the best action at each decision epoch to minimize total cost of denied requests and missed continuity (measured by the number of non-PCP diversions) for the day. The state of the system at any decision epoch  $t$  is represented by the number of open access patients booked in the future in each physician’s calendar. Denying a request incurs a cost – denied requests are a reflection of the lack of timely access to primary care, or the costs needed to provide care to these patients outside the regular hours of clinic operations.

What impact does flexibility have in the dynamic case? Recall that in the aggregate case, flexibility was beneficial in buffering against variability of demand. In the dynamic case, there is an

additional component of variability, since appointment requests arrive randomly over time. There is therefore greater opportunity for flexibility to meet demand imbalances at different points in time. On the other hand, patient calls require an immediate appointment allocation decision, under only partial demand information available at that point; this decreases the impact of flexibility, since allocation decisions that can be made optimally in the aggregate case may not be as effective in the dynamic case. These counteracting effects may be the reason why the benefits of flexibility are mostly identical in both the aggregate and dynamic cases. Our computational experience with the stochastic dynamic program (Hippchen, 2009) shows that the benefits of full and partial flexibility in the dynamic case produce the same percentage improvements in timely access rate shown in Figure X.

While the stochastic dynamic program can be used illustrate the impact of flexibility, primary care offices require easily implementable policies or heuristics that can be put into practice as calls for same-day appointments come in. Consider two contrasting policies in the fully flexible case: Primary First (PF) and Most Slots (MS). PF assigns incoming same-day calls to the patient’s PCP first, so long as slots are available. If PCP slots are not available for the day, it assigns the patient to the physician with the most available slots. MS, on the other hand, assigns an incoming patient call to the physician with the most slots available.

PF thus maximizes continuity, while MS utilizes physician slots more effectively and increases the number of patients seen per day. A hybrid approach that balances continuity with timely access would be assign to PCP so long as the difference the available slots between PCP and other physicians does not exceed a certain predetermined threshold. Our results show that the choice of heuristic also depends on the system workload or utilization. PF is the best choice in an underutilized as well as overutilized setting, while MS and the hybrid approach are better choices in systems where arrival rates and available capacity are relatively balanced.

## 7 Summary and Conclusions

In summary, we have discussed capacity allocation for primary care practices at three different levels of the capacity planning hierarchy. The goal in each case as been to maximize timely access and continuity. At the highest level, the design of physician panels, we demonstrate the impact of case-mix, or the type of patients in a physician’s panel, on the ability to provide timely access and continuity. Case mix can be considered using age and gender as predictors, or, when patient clinical data is available, using comorbidity counts. Using case-mix a practice can create overflow profiles for the physicians in the practice as function of daily capacity and determine which physicians are overburdened. This in turn can point to opportunities for redesigning panels so that patients can see their own PCP as much as possible and redirections to unfamiliar physicians are minimized.

Panel redesign involves changing existing patient-physician relationships. A viable alternative to redesign is managing the flexibility of physicians – the ability of physicians to see patients of other physicians. We discuss flexibility at aggregate as well as dynamic levels. The design of flexibility in the aggregate has to consider two types of appointments: 1) prescheduled appointments which are booked in advance and require continuity with the patient’s PCP; and 2) same-day or open access appointments which have to be fulfilled during the course of the day. We propose a framework – commonly observed in practice – in which the short notice open access appointments can be flexibly shared between physicians while mandating continuity for the prescheduled appointments. Using a two-stage stochastic integer programming model, we demonstrate the impact of flexibility on the ability to provide timely access to patients, measured by the number of patients seen a

given workday. Specifically, we find that the 2-chain partially flexible practice, which restricts the number of physicians a patient sees to two, performs almost as well as the fully flexible practice with regard to timely access. The impact of flexibility increases as the number of physicians in the practice increases and as the demand loads between physicians are asymmetric or uneven. Our results also show that practices can heuristically determine their capacity allocation for prescheduled appointments depending on their flexibility configuration and overall system workload.

Finally, the implementation of flexibility at the level of a workday has to be made under partial demand information, since calls arrive dynamically over the course of a day. We outline a decision framework to evaluate the impact of flexibility in this dynamic case and discuss heuristics that practices can use to balance timely access and continuity.

## Acknowledgements

This work was funded in part by the grant CMMI 1031550 from the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

American College of Physicians, 2006. The impending collapse of primary care and its implications for the state of the nation's healthcare. Technical report.

American Academy of Family Physicians (AAFP), American Academy of Pediatrics, American College of Physicians, American Osteopathic Association, Joint Principles of the Patient-Centered Medical Home, March 2007.

Atlas, S., Grant, R., Ferris, T., Chang, Y., and Barry, M., Patient-Physician connectedness and quality of primary care, *Annals of Internal Medicine*, 150 (5), 2009, 325-226.

Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., Wood, D., and Stahl, J., 2010, Improving clinical access and continuity using physician panel redesign, *Journal of General Internal Medicine*, 25 (10), 1109-15.

Balasubramanian, H., Denton, B., Lin, M., 2011 a, Managing physician panels in primary care, *Handbook of Healthcare Delivery Systems*, CRC Press (Taylor and Francis), Editor: Yuehwen Yih, 10-1.

Balasubramanian, H., Muriel, A., Wang, L., 2011 b, The impact of flexibility and capacity allocation on the performance of primary care practices, *accepted Flexible Services and Manufacturing Journal*.

Gill, J. M., Mainous, A., 1999. The role of provider continuity in preventing hospitalizations. *Archive of Family Medicine* 7, 352 - 357.

Gill, J. M., Mainous, A., Nsereko, M., 2000. The effect of continuity of care on emergency department use. *Archives of Family Medicine* 9, 333 - 338.

Graves, S. C., Tomlin, B. T., 2003. Process flexibility in supply chains. *Management Science* 49 (7), 907 - 919.

Green, L. V., Savin, S., 2008. Reducing Delays for Medical Appointments: A Queueing Approach. *Operations Research* 56(6), 1526 - 1538.

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

Gupta, D., Potthoff, S., Blowers, D., Corlett, J., 2006. Performance metrics for advanced

access. *Journal of Healthcare Management* 51(4), 246 - 259.

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

Hopp, W., Tekin, E., Van Oytten, M. P., 2004. Benefits of Skill Chaining in Serial Production Lines with Cross-Trained Workers. *Management Science* 50 (1), 83 - 98.

Hippchen, J., Flexibility in Primary Care, Masters Thesis, 2009 (Advisors: Hari Balasubramanian and Ana Muriel). Accessible at: <http://people.umass.edu/hbalasub/FlexibilityThesis.pdf>

Jordan, W. C., Graves, S. C., 1995. Principles and benefits of manufacturing process flexibility. *Management Science* 41 (4), 577 - 594.

Liu, N., Ziya, S., and Kulkarni, V., 2010, Dynamic scheduling of outpatient appointments under patient no-shows and cancellations, *Manufacturing and Services Operations Management*, 12.2, 347-365.

Muriel, A., Somasundaram, A., Zhang, Y., 2006. Impact of Partial Manufacturing Flexibility on Production Variability. *Manufacturing and Service Operations Management* 8(2), 192 - 205.

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

Naessens, J., Stroebel, R., Finnie, D., Shah, N., Wagie, A., Litchy, W., Killinger, P., O'Byrne, T., Wood, D., and Nesse, R., 2011, Effect of multiple chronic conditions among working-age adults, *American Journal of Managed Care*, 17(2), 118-122.

Qu, X., Rardin, R., Williams, J.A.S., Willis, D., Matching daily healthcare provider capacity to demand in advanced access scheduling systems, *European Journal of Operational Research*, 183(2), pp. 812-826.

Robinson, L., and Chen, R., 2010, A comparison of traditional and open access policies for appointment scheduling, *Manufacturing and Services Operations Management*, 12.2, 330-347.

Rust, G., Ye, J., Baltrus, P., Daniels, E., Adesunloye, B., Fryer, G. E., 2008. Practical Barriers to Timely Primary Care Access. *Archives of Internal Medicine* 268(15), 1705 - 1710.

Starfield, B., Macinko J., Shi, L., 2007. Quantifying the health benefits of primary care physician supply in the United States. *International Journal of Health Services* 37(1), 111 - 126.

Shi, L., Starfield, B., Macinko, J., 2005. Contribution of primary care to health systems and health. *The Milbank quarterly* 83(3), 457 - 502.