The impact of provider flexibility and capacity allocation on the performance of primary care practices

Hari Balasubramanian, Ana Muriel & Liang Wang

Flexible Services and Manufacturing Journal

ISSN 1936-6582
Volume 24
Number 4

Flex Serv Manuf J (2012) 24:422-447
DOI 10.1007/s10696-011-9112-5

Flexible Services and Manufacturing Journal

Editor-in-Chief
Hans-Otto Günther

Special Issue: Health Care Management
Editors: Sally Brailsford · Erhan Kozan · Marion S. Rauner

Available online
www.springerlink.com

Volume 24 • Number 4
December 2012

Springer

10696 • ISSN 1936-6582
24(4) 375–602 (2012)

Springer

Flex Serv Manuf J (2012) 24:422–447
DOI 10.1007/s10696-011-9112-5

The impact of provider flexibility and capacity allocation on the performance of primary care practices

Hari Balasubramanian · Ana Muriel · Liang Wang

Published online: 28 September 2011

Abstract The 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. In the past decade, a new paradigm called *advanced access* or *open access* has been adopted by practices nationwide to encourage physicians to "do today's work today." However, most clinics still reserve pre-scheduled slots for long lead-time appointments due to patient preference and clinical necessities. Therefore, an important problem for clinics is how to optimally manage and allocate limited physician capacities as much as possible to meet the two types of demand—pre-scheduled (non-urgent) and open access (urgent, as perceived by the patient)—while simultaneously maximizing timely access and patient–physician continuity. In this study we adapt ideas of manufacturing process flexibility to capacity management in a primary care practice. Flexibility refers to the ability of a primary care physician to see patients of other physicians. We develop generalizable analytical algorithms for capacity allocation for an individual physician and a two physician practice. For multi-physician practices, we use a two-stage stochastic integer programming approach to investigate the value of flexibility. We find that flexibility has the greatest benefit when system workload is balanced, when the physicians have unequal workloads, and when the number of physicians in the practice increases. We also find that partial flexibility, which

H. Balasubramanian (✉) · A. Muriel
Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst,
160 Governors Drive, Amherst, MA 01003, USA
e-mail: hbalasubraman@ecs.umass.edu
URL: http://people.umass.edu/hbalasub

L. Wang
Autonomous Earthmoving Equipment LLC, 21755 Interstate 45, BLDG 5, Spring, TX 77388, USA

restricts the number of physicians a patient sees and thereby promotes continuity, simultaneously succeeds in providing high levels of timely access.

Keywords Appointment scheduling · Open access · Flexibility · Capacity planning · Timely access · Continuity of care · Stochastic programming

1 Introduction

Primary care providers (PCPs) are typically the first point of contact between patients and health systems. They include general practitioners, family doctors, pediatricians, and geriatricians. 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.

Each primary care physician (or provider, in general) has a *panel* of patients, whose care she is responsible for. Long term, holistic care of patients is one of the cornerstones of primary care. In practice this translates to maximizing the number of visits with the patient's own PCP—maximizing continuity, in other words. Numerous studies have documented the importance of continuity to patients (O'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.

Patient–physician continuity is often in conflict with timely access. While a same day appointment may be available, it may be with a physician the patient is not familiar with. On the other hand, it may possible to see your own doctor, but the appointment may be weeks or months later. Rust et al. (2008) report that the inability to get a timely appointment to a primary care physician increases the likelihood of patients visiting emergency rooms. In the United States especially, the nationwide shortage of primary care physicians and the growing demand have made it difficult for practices to simultaneously provide patient–physician continuity and timely access.

To address this issue many primary care practices have tried implementing *advanced access* or *open access* (Murray and Berwick 2003; Murray et al. 2003). Advanced access promotes the concept that physicians should "do today's work today" rather than push appointments into the future. 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.

In this paper, we investigate the value of *physician flexibility* and its relationship to capacity allocation for a given workday under two streams of uncertain demand, prescheduled (non-urgent) and open access (urgent, as perceived by the patient). Flexibility refers to the ability of a physician to see patients of other physicians. Primary care physicians are inherently flexible; however, practices need to manage this flexibility effectively to strike a balance between timely access and continuity. Lower flexibility implies greater restriction on the number of physicians a patient can see and hence better continuity of care. But this may come at the cost of timely access. Greater flexibility, on the other hand, implies greater fragmentation of a patient's care (less continuity), but improved timely access for the patients.

The rest of the paper is organized as follows. In Sect. 2, we introduce concepts of flexibility, capacity allocation, and revenue maximization in the primary care setting in greater detail, and discuss the trade-off between timely access and continuity in fully flexible versus partially flexible (*2-chain*) practices. Section 3 reports on the literature related to operations research models applied to capacity setting and scheduling in open access practices, and flexibility in manufacturing and services. In Sect. 4, we present a stochastic integer programming model for capacity allocation and to test the value of flexibility. In Sect. 5, we present analytical results that give us insights into capacity allocation for single and two physician practices but that can be extended more generally. Section 6 presents the computational experiments and results. We present a summary of our main conclusions in Sect. 7.

2 Flexibility, capacity allocation and revenue maximization in primary care

2.1 Configurations of flexibility

Although the model we propose in the paper is capable of accommodating any flexibility configuration, we restrict our focus to the three configurations shown in Fig. 1. In Fig. 1 (a), patients may see any other physician (full flexibility). This configuration leads to the highest level of timely access as resources are pooled, but continuity suffers. In (b), patients can only see their own dedicated physician (no flexibility), which leads to the highest level of continuity, although timely access might not be guaranteed. Combining these two levels leads to configuration (c) partial flexibility, where patients and physicians are *chained* such that each patient in addition to having his/her own physician, also has one *auxiliary physician* (AP), but is not allowed to see any of the other physicians in the practice. We will refer to this configuration as *2-chain flexibility.*

While the 2-chain has been suggested in the manufacturing flexibility literature, its feasibility and implementation in primary care will involve redesigning a clinic's routine processes. The 2-chain concept is compatible with the concept of *team care* recommended by the Institute of Medicine (2001) in its report *Crossing the Quality Chasm: New Health System for the Twenty First Century*. Team care suggests that

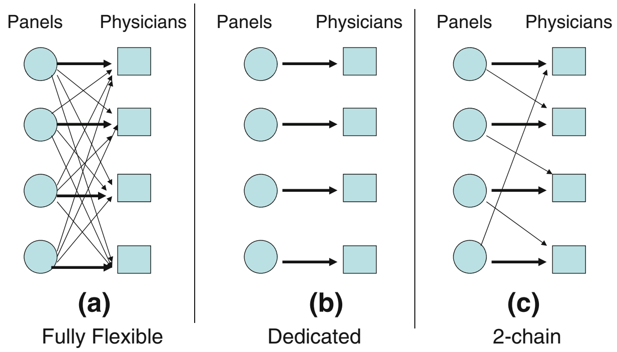

Fig. 1 Different flexibility configurations that tradeoff continuity with timeliness

patient's needs are coordinated by not just her physician but by a team—which may consist of other physicians, nurses and medical assistants. The challenge is in the design of these teams, since the 2-chain requires *overlapping* members who function as coordinators. This is because each physician in the 2-chain configuration will see patients of one other physician but his own patients may be seen by a third physician. While such a redesign may not be insurmountable, implementation may be dependent on a clinic's working structure and policies. Note, however, that clinics that operate under the fully flexible mode (which is typically the default in practice) require significantly greater amount of coordination.

Of the two demand streams, we assume that prescheduled (non-urgent) appointments are always seen by the patient's PCP. This is because such appointments typically require greater continuity (knowledge of the patient's history) and coordination. Flexibility thus does not apply to prescheduled appointments. However, the amount of capacity reserved in a day for prescheduled appointments by each physician has an effect on the number of open access appointments that can be seen. If the prescheduled demand is short, the leftover slots can be filled by open access appointments. If the open access demand exceeds the available capacity, the unsatisfied demand can be shared between physicians. The idea, however, is to limit this sharing while still providing acceptable levels of timely access.

2.2 Capacity allocation

Each physician has a fixed number of slots available in a day. If the physician works an 8 h day, this typically means 24 appointment slots, since each appointment is commonly allotted 20 min. Each physician also has prescheduled and open access demand distributions that have been identified using past data. The question we address is the following: given a certain flexibility configuration, how many slots should each physician make available for prescheduled appointments? In other words, we explore the clinic's decision of how many slots to set aside for open access appointments. We note here that the allocation decision is made at an aggregate level. That is, we assume that the demands are realized "instantly" and are fulfilled instantly as well if capacity is available (the exact allocation mechanism is a two-stage process, described in Sect. 4). In practice, demand gets realized over time and decisions have to be made without full knowledge of future demand. So our approach provides only an approximation of the actual capacity allocation process, in order to make high-level capacity reservation decisions.

2.3 Maximizing revenue

Our objective is to maximize the clinic's revenue. While this may seem a practice-centric measure, in effect the revenue consolidates both timely access and patient–physician continuity into a single function. The more prescheduled and open access patients seen, the higher the clinic's revenue and the better the timely access. Continuity is included in the function by adding a small deduction in revenue for every open access patient that is seen by an unfamiliar physician. (The magnitude of the deduction, however, is hard to determine, since the cost of seeing an unfamiliar physician is not easy to estimate.) Note as well that the level of continuity is mainly dictated by the flexibility configuration chosen.

Our revenue maximization also reflects the fact that, in a relative sense, open access appointments are more "valuable" than prescheduled appointments. This is because of three reasons. First, 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.

2.4 2-Chain versus full flexibility

Since full flexibility has more "outbound" links than 2-chain flexibility, it should have a better ability to absorb incoming demands and yield a higher timely access rate than 2-chain flexibility. This would be true for the dynamic setting of patient scheduling where allocation decisions are made as requests arrive, with limited future knowledge of the overall demand that will need to be serviced (Hippchen 2009). By contrast, in the aggregate demand setting captured by our two-stage stochastic integer programming approach, the patient allocation (second stage decisions) is only performed after the full system demand is realized (that is once the scenario is known). Thus the 2-chain, which indirectly links all the physicians, will in most cases manage to serve the same number of patients as a fully flexible practice where all physicians are linked to each other directly. To be sure, there are rare instances where full flexibility will clearly dominate. For instance, consider a practice with four physicians, where each has 10 slots left for open access, and the demands for open access are 20, 20, 0 and 0 respectively. In this extreme case, the 2-chain flexibility can only meet 30 open access demands, while the full flexibility can satisfy all of them. Since such an instance would occur with a low probability, from a statistical point of view, the 2-chain flexibility has almost the same effectiveness to absorb the demand as full flexibility.

This ability of the 2-chain to match full flexibility in timely access comes at the cost of increased diversion rates (in our model this is a measure of continuity). Since full flexibility has more "outbound" links than 2-chain flexibility, it should have a higher probability that the demand will be diverted to other physicians. In reality, however, a single patient redirection to an available physician, which can be made directly under full flexibility, may require redirecting several patients along the 2-chain if the initial patient's panel and available physician involved are not connected. For example, Fig. 2 shows a case of three physicians where each physician has 10 slots left for open access, and the demands are 16, 10 and 4 respectively. We can see that the total number of diversions under 2-chain flexibility is 12, but only 6 under full flexibility. Since 2-chain flexibility requires more "jumps" to shift the demands, the diversion rate of the 2-chain is higher than that of full flexibility in our model.

While the number of redirections is greater in the 2-chain, it is important to note that each patient will always see either one of two physicians. We believe this results in stronger continuity and efficiency from the perspective of both the patient (who could quickly get to be familiar and comfortable with both physicians) and the physician (who would be able to follow the other's panel relatively well and share cases with only one other physician). This becomes especially relevant as the practice size grows. For example, in a fully flexible 6-physician clinic, a patient may see any of six physicians, while in a 2-chain the number is never more than two. Thus, in our results, while the 2-chain does worse with respect to continuity as measured simply by the number of redirections, in practice it is likely to be better (or at least as good) in this regard when compared to full flexibility.

We also note that if the aggregate assumption that demands are realized and fulfilled instantly is relaxed (that is demand were to be realized at different points in time and not all at once) then the allocations of the 2-chain and the full flexibility would be different in the above example. In reality, clinics have to manage flexibility dynamically—decisions have to be made when demand has only been partially realized.

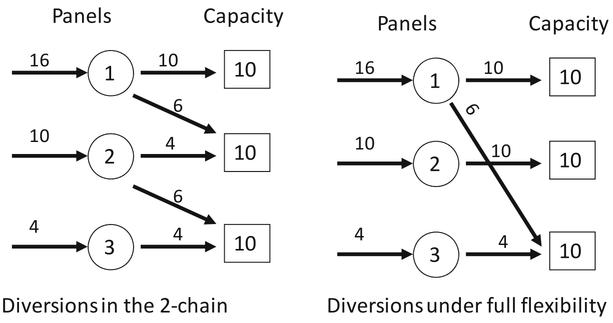

Fig. 2 An example of diversion process in 2-chain and full flexibility

3 Literature review

The application of operations research to healthcare is a growing area of research. We limit our review only to the most relevant papers in two topics. We first survey quantitative approaches that have thus far been used in the context of open access scheduling. Next, we review the principal results in the area of flexibility most related to our setting.

3.1 Operations research applied to open access

The literature on optimization approaches applied to open access is growing. The adoption of open access, which promises patients same-day appointments, has prompted a series of questions. 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 workloads 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 Sect. 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 ours is that patient preference drives the scheduling of prescheduled appointments, 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.

3.2 Literature related to flexibility

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. The latter point is mute in our aggregate analysis of the capacity allocation problem, but key to the dynamic clinic scheduling problem (see Hippchen 2009).

As in the case of cross-training in serial production lines (Hopp et al. 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 et al. (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 benefits of flexibility in increased sales and capacity utilization in multi-product, multi-plant production networks have been thoroughly studied by Jordan and Graves (1995) considering a single production period. They are the first to introduce the concept of chaining to achieve maximum benefits from limited flexibility configurations where each plant produces only a few of the products. 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 et al. 2006a, b). Chua et al. (2008) distinguish between range and response of flexible systems. Range refers to the set of demand scenarios that can be accommodated and response to the cost of doing so; that is, the cost of using secondary rather than primary resources for production/service. They show that upgrading system response outperforms improving system range. In the primary care setting, this means that systems where physicians can handle other physician's panels at lower additional cost should be preferred over those that can accommodate ever more extreme patient demand scenarios. This result suggests that the benefits of restricting the number of doctors that can see a particular patient (resulting in lower cost of using the secondary providers 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.

4 Model

We consider a general primary care practice with M physicians, each with N_i available appointment slots, $i = 1, 2,\ldots, M$. Let A be the set of all possible panel-physician links (i, j) such that 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; that is, the network of allowed open-access patient redirections within the practice. We assume that pre-scheduled patients are required to see their own physicians, and physician flexibility can only be used for the time-sensitive open access demand patients. This is the most relevant case in practice, since patient–physician continuity is highly beneficial to prescheduled appointments, in which major physicals or follow-ups of chronic conditions are performed.

Let R_i^p be the revenue associated with physician i, $i = 1, 2,\ldots, M$, 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 $D = (D_1^p, D_1^o, \ldots, D_M^p, D_M^o)$, 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, $s = 1, 2,\ldots, S$, where $S \equiv S_1^2 \times S_2^2 \times \cdots \times S_M^2$; that is, $P[D = (d_{1s}^p, d_{1s}^o, \ldots, d_{Ms}^p, d_{Ms}^o)] = q_s$.

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 pre-scheduled with physician i under demand realization s.

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

Finally, to indicate when some of the slots reserved for pre-scheduled appointments go unused and can be made available to open access demands, we introduce the following binary variables:

$$\phi_{iu_{is}} = 1 \quad \text{if } u_{is} \equiv \min\{d^p_{is}, N_i\} < N^p_i, \quad \text{otherwise, } \phi_{iu_{is}},$$
$$\text{for} \quad i = 1, 2, \ldots, M \quad \text{and} \quad s = 1, 2, \ldots, S.$$

Observe that $u_{is} = \min\{d^p_{is}, N_i\} \in \{0, 1, 2, \ldots, N_i\}$; therefore, the total number of binary variables $\phi_{iu_{is}}$ equals the total number of appointment slots in the practice, plus one more per physician, $N_1 + N_2 + \cdots + N_M + M$.

The objective is to maximize the expected revenue of satisfying prescheduled and open access appointments. We can formulate the problem as follows:

$$\text{Objective}: \qquad \text{Max} \sum_{s=1}^{S} \sum_{i=1}^{M} q_s \left[R^p_i x^p_{is} + \sum_{(i,j)\in A} R^o_{ij} x^o_{ijs} \right] \tag{1}$$

$$\text{Subject to}: \qquad N^p_i \leq N_i \quad \forall i = 1, 2, \ldots, M \tag{2}$$

$$N^p_i \leq d^p_{is} + N_i \phi_{iu_{is}} \quad \forall i = 1, 2, \ldots, M, \ s = 1, 2, \ldots, S \tag{3}$$

$$N^p_i \geq d^p_{is} \phi_{iu_{is}} \quad \forall i = 1, 2, \ldots, M, \ s = 1, 2, \ldots, S \tag{4}$$

$$x^p_{is} \leq N^p_i \quad \forall i = 1, 2, \ldots, M, \ s = 1, 2, \ldots, S \tag{5}$$

$$x^p_{is} \leq d^p_{is} \quad \forall i = 1, 2, \ldots, M, \ s = 1, 2, \ldots, S \tag{6}$$

$$\sum_{i:(i,j)\in A} x^o_{ijs} \leq N_j - d^p_{js} \phi_{ju_{js}} \quad \forall j = 1, 2, \ldots, M, \ s = 1, 2, \ldots, S \tag{7}$$

$$\sum_{i:(i,j)\in A} x^o_{ijs} \leq N_j - N^p_j + \phi_{ju_{js}} N_j \quad \forall j = 1, 2, \ldots, M, \quad s = 1, 2, \ldots, S \tag{8}$$

$$\sum_{j:(i,j)\in A} x^o_{ijs} \leq d^o_{is} \quad \forall i = 1, 2, \ldots, M, \ s = 1, 2, \ldots, S \tag{9}$$

$$\phi_{iu_{is}} \in \{0, 1\} \quad \forall i = 1, 2, \ldots, M, \ u_{is} = 0, 1, \ldots, N_i \tag{10}$$

$$N^p_i, x^p_{is}, x^o_{ijs} \geq 0 \quad \forall i, j = 1, 2, \ldots, M, \ (i, j) \in A, \ s = 1, 2, \ldots, S \ \text{and integer} \tag{11}$$

Equation 3 ensures that $\phi_{iu_{is}} = 1$ if $d^p_{is} < N^p_i$. Equation 4 ensures that $\phi_{iu_{is}} = 0$ if $d^p_{is} > N^p_i$. 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 $\phi_{ju_{js}} = 1$ and $\phi_{ju_{js}} = 0$ 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.

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, 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 Solak et al. (2010) for two-stage stochastic integer programming problems; see Sect. 6.

5 Analysis of dedicated and fully flexible group practices

Our objective in this section is to find the optimal number of slots, N_i^{p*}, to reserve for pre-scheduled appointments of each physician i in order to maximize the total expected revenue for practices with these simple management structures: (1) *Dedicated Practices* where physicians work independently and do not share any patients, and (2) *Fully-Flexible Physician Group Practices,* where all physicians share the open access demand from their panels.

When a practice with any number of physicians allows full flexibility in sharing both their pre-scheduled and open access demand streams, all the capacity in the system is pulled together to satisfy patient demand. Therefore, in the absence of redirection costs, the system is equivalent to that of a dedicated physician with the aggregate panel demand and aggregate capacities of the original system. Thus, for the flexible practice, we focus on the more interesting and most common case where only open access demand is shared.

5.1 Dedicated practice

In a dedicated practice, physicians can only serve the patients from their own panel. The system configuration is shown below in Fig. 3:

Given the number of slots, $N_i^p \in \{0, 1, 2, \ldots, N\}$, made available to pre-scheduled demand of physician i, the expected revenue from pre-scheduled appointments for physician i is:

$$ER_i^p(N_i^p) = \sum_{d_i^p=1}^{N_i^p} R_i^p \cdot d_i^p \cdot P(D_i^p = d_i^p) + R_i^p \cdot N_i^p \cdot P(D_i^p > N_i^p) \tag{12}$$

and the expected revenue from open access appointments for physician i is:

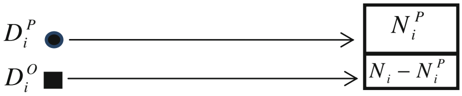

Fig. 3 System configuration for a dedicated practice

$$
\begin{aligned}
ER_i^o(N_i^p) = {} & \sum_{d_i^p=0}^{N_i^p} \left[\begin{array}{l} \displaystyle\sum_{d_i^o=1}^{N_i-d_i^p} R_i^o \cdot d_i^o \cdot P\left(D_i^p = d_i^p, D_i^o = d_i^o\right) + \\ \displaystyle\sum_{d_i^o=N_i-d_i^p+1}^{\infty} R_i^o \cdot (N_i - d_i^p) \cdot P\left(D_i^p = d_i^p, D_i^o = d_i^o\right) \end{array} \right] \\
& + \sum_{d_i^p=N_i^p+1}^{\infty} \left[\begin{array}{l} \displaystyle\sum_{d_i^o=1}^{N_i-N_i^p} R_i^o \cdot d_i^o \cdot P\left(D_i^p = d_i^p, D_i^o = d_i^o\right) + \\ \displaystyle\sum_{d_i^o=N_i-N_i^p+1}^{\infty} R_i^o \cdot (N_i - N_i^p) \cdot P\left(D_i^p = d_i^p, D_i^o = d_i^o\right) \end{array} \right]
\end{aligned} \tag{13}
$$

The total expected revenue from the panel of physician *i*, $ER_i(N_i^p)$, is equal to the sum of Eqs. 12 and 13. Our objective is to find the number of slots to make available to pre-scheduled appointments, N_i^{p*}, that maximizes the total expected revenue for each physician *i*, $i = 1, 2, \ldots, M$.

Dedicated Problem:

$$
\begin{aligned}
& \text{Max } ER_i^p(N_i^p) + ER_i^o(N_i^p) \\
& \text{Subject to:} \quad N_i^p \le N_i \\
& \qquad\qquad\qquad N_i^p \text{ is integer}
\end{aligned}
$$

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. In what follows, we show a stronger result, guaranteeing global optimality, by first showing that the objective function has diminishing returns under mild assumptions.

Proposition 1 *If $P[D_i^o \ge N_i - N_i^p | D_i^p \ge N_i^p]$ 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.*

Proof If an additional slot is made available to prescheduled appointments, then one more prescheduled appointment will actually be made only when the demand for prescheduled appointments is large. This, however, will come at the cost of foregoing an open access appointment if at the same time the demand for open access is sufficiently high. That is,

$$
\begin{aligned}
ER_i(N_i^p + 1) - ER_i(N_i^p) = {} & P[D_i^p \ge N_i^p + 1] \\
& \times \left(R_i^p - R_i^o P\left[D_i^o > N_i - (N_i^p + 1) | D_i^p \ge N_i^p + 1\right]\right)
\end{aligned}
$$

The first probability term is clearly non-increasing in N_i^p, and the second term is non-increasing since we require $P[D_i^o \ge N_i - N_i^p | D_i^p \ge N_i^p]$ to be non-decreasing in N_i^p. □

The above condition ($P[D_i^o \ge N_i - N_i^p | D_i^p \ge N_i^p]$ non-decreasing in N_i^p) simply requires that given that demand for prescheduled appointments is at least as large as the number of allocated slots N_i^p, the conditional probability of open access demand being greater than or equal to the remaining slots, $N_i - N_i^p$, does not decrease as N_i^p grows. In particular, the 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 has 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 $ER_i(N_i^p) - ER_i(N_i^p - 1) \geq 0$.

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 $P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p] \leq \frac{R_i^p}{R_i^o}$.*

Proof The proof of Proposition 1 shows that, as the number of slots made available to prescheduled appointments grows, the system revenue increases, with diminishing returns, if and only if $P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p] \leq \frac{R_i^p}{R_i^o}$. Since $P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p]$ is non-decreasing in N_i^p, once $P[D_i^o \geq N_i - (N_i^p + 1) | D_i^p \geq N_i^p + 1] > \frac{R_i^p}{R_i^o}$, the system revenue will decrease as N_i^p further increases. Therefore, the optimal number of slots is the largest integer for which $P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p] \leq \frac{R_i^p}{R_i^o}$. □

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, N_i^{p*}, does not depend on the distribution of pre-scheduled demand for physician i.

5.2 Fully flexible practice

In a fully flexible practice, open-access patients can be seen by any available physician (see Fig. 4).

In this case, the optimal number of slots to make available to prescheduled demand of the physicians, N_1^{p*} and N_2^{p*}, can still be found with a simple greedy algorithm as shown in Theorem 2. 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.

Proposition 2 *If $P\left[D_1^o + D_2^o > 2N - (N_1^p + \min(D_2^p, N_2^p) + 1) | D_1^p \geq N_1^p + 1\right]$ is non-decreasing in N_1^p and N_2^p, the difference in revenue associated with increasing*

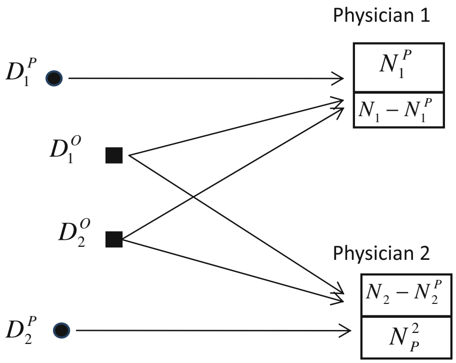

Fig. 4 System configuration for two physicians sharing open access demands

the number of prescheduled slots of physician 1 by one, $ER(N_1^p + 1, N_2^p) - ER(N_1^p, N_2^p)$*, is non-increasing in* N_1^p *and* N_2^p*.*

By symmetry, if $P\left[D_1^o + D_2^o > 2N - (\min(D_1^p, N_1^p) + N_2^p + 1) | D_2^p \geq N_2^p + 1\right]$ *is non-decreasing in* N_1^p *and* N_2^p*, the difference in revenue associated with increasing the number of prescheduled slots of physician 2 by one, that is,* $ER(N_1^p, N_2^p + 1) - ER(N_1^p, N_2^p)$*, is non-increasing in* N_1^p *and* N_2^p*.*

Proof To calculate de difference in expected revenue, we observe that offering one more appointment slot to physician 1 will only impact the revenue when it is actually used, that is, when $D_1^p \geq N_1^p + 1$. In that case, physician one will increase the revenue from prescheduled patients by R_1^p, but this may come at a loss of one open access patient if the number of open access requests exceeds the number of remaining slots. In mathematical form, we can write the difference as follows:

$$\begin{aligned} &ER(N_1^p + 1, N_2^p) - ER(N_1^p, N_2^p) = P\left[D_1^p \geq N_1^p + 1\right] \\ &\quad \left(R_1^p - R^o P\left[D_1^o + D_2^o > 2N - \left(N_1^p + \min\left(D_2^p, N_2^p\right) + 1\right) | D_1^p \geq N_1^p + 1\right]\right) \end{aligned} \tag{14}$$

This expression can be easily seen to be decreasing in both N_1^p and N_2^p under the given condition. □

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. The proposition shows that the revenue function exhibits decreasing returns in both N_1^p and N_2^p under those mild conditions; this can be interpreted as concavity of the discrete revenue function and it implies that a greedy algorithm, as stated in the following theorem, is optimal.

Theorem 2 *Under the conditions of Proposition 2, the optimal number of appointment slots for each physician i to make available to pre-scheduled patients in a two-physician flexible practice where the two physicians share only open access demands can be found using the following greedy algorithm:*

Step 1: Intialize $N_1^p = 0$ *and* $N_2^p = 0$.
Step 2: Calculate $F_1 = ER(N_1^p + 1, N_2^p) - ER(N_1^p, N_2^p)$ *and* $F_2 = ER(N_1^p, N_2^p + 1) - ER(N_1^p, N_2^p)$
Step 3: If $F_1 \geq F_2 > 0$*, then* $N_1^p := N_1^p + 1$ *and go to Step 2. If* $F_2 > F_1 > 0$*, then* $N_2^p := N_2^p + 1$ *and go to Step 2. If* $F_1 \leq 0$ *and* $F_2 \leq 0$*, then STOP. Else, if* $F_1 > 0$ *move on to Step 4, and if* $F_2 > 0$ *go to Step 5.*
Step 4: While $F_1 > 0$*, do* $N_1^p := N_1^p + 1$ *and update* $F_1 = ER(N_1^p + 1, N_2^p) - ER(N_1^p, N_2^p)$
Step 5: While $F_2 > 0$*, do* $N_2^p := N_2^p + 1$ *and update* $F_2 = ER(N_1^p, N_2^p + 1) - ER(N_1^p, N_2^p)$

The theorem is a direct consequence of the proposition. The difference in revenue can be easily calculated using expression (14).

Observe that Proposition 2 and Theorem 2 can be easily extended to the general case of M physicians that fully share open access demand. We can write the difference in revenue associated with increasing the number of prescheduled slots as follows.

$$\begin{aligned} & ER\left(N_1^p + 1, N_2^p, \ldots, N_M^p\right) - ER\left(N_1^p, N_2^p, \ldots, N_M^p\right) \\ & = P\left[D_1^p \geq N_1^p + 1\right]\Big(R_1^p - R^o P\Big[D_1^o + D_2^o + \cdots + D_M^o > MN \\ & \quad - \left(N_1^p + \min\left(D_2^p, N_2^p\right) + \cdots + \min(D_M^p, N_M^p) + 1\right) | D_1^p \geq N_1^p + 1\Big]\Big) \end{aligned}$$

This difference is again non-increasing in $(N_1^p, N_2^p, \ldots, N_M^p)$ under mild conditions that are satisfied in the case of independence and in most practical scenarios. As a consequence, a greedy algorithm such as the one in Theorem 2 that keeps increasing one pre-scheduled appointment slot at a time to the physician where it produces the highest system revenue will provide the optimal solution.

6 Computational experiments

We use the two stage stochastic integer program (SIP) to test the value of flexibility under the three different flexibility configurations—full flexibility, partial flexibility (2-chain) and no flexibility (dedicated)—under a number of settings. We will focus on three measures: system revenue, timely access rate and continuity rate. System revenue stands for the total expected revenue of meeting patient demands; timely access rate is the percentage of all patients, both prescheduled and open access, who get access to an appointment; and continuity rate presents the percentage of open access patients who see their own physician. Our model provides the optimal value of $N_1^{p*}, N_2^{p*}, \ldots, N_M^{p*}$ (the first stage decisions), the optimal allocation of patients to physicians and the optimal numbers of non-PCP diversions that should be made (second stage decisions).

We solve the SIP in the extensive form or deterministic equivalent form (Birge and Loveaux 1997) by creating random demand scenarios. The computational complexity of our model heavily depends on the number of scenarios, the most influential factor, and the number of physicians. Although our stochastic integer programming model can theoretically investigate the value of flexibility for any flexibility configuration with any number of physicians, the time-consuming nature of the optimization and evaluation makes it impractical. For tractability, we use concepts from the computationally effective sample average approximation method, proposed by Solak et al. (2010) for two-stage stochastic integer programming problems. The basic idea is to create a manageable number of samples/scenarios to produce an estimation of the optimal objective value and corresponding first stage solutions. To determine if the number of samples/scenarios is sufficient, we further run a large number of scenarios to have a precise estimation of the objective value based on the fixed first stage solution. This process is repeated over a number of replications to provide confidence intervals and statistical guarantees on the quality of the estimation. Using this method, we determined that 1,000 scenarios and 50 replications gives us narrow enough confidence intervals for both 3 physician and 6 physician cases. To allow for a fair comparison, the 2-chain, full flexibility and dedicated configurations use the same set of scenarios (analogous to the common random numbers approach in discrete event simulation).

Since the majority of practices in the US have six physicians or less, we conduct our experiments on 3 physician and 6 physician cases. Even larger practices, primary care clinics in academic medical centers for example, divide their physicians into smaller groups or teams.

In our experimental setting, each physician has twenty-four appointment slots in a day. This is because a typical appointment takes about 20 min and a physician's workday may be up to 8 h. In practice, this amount varies from physician to physician and from practice to practice. Our model can easily adjust for different capacities. We assume that the prescheduled and open access demands are independent of each other and are Poisson distributed.

Workload in our model is defined as the ratio of the expected total demand for the clinic and total available capacity. For instance, in a practice with three physicians, suppose each physician has a demand rate of 10 for prescheduled appointment and 14 for open access demand. The total expected demand is $10 \times 3 + 14 \times 3{=}72$, and the total capacity is $24 \times 3 = 72$, therefore, the clinic or system workload is 100%. A factor varying from 0.6 to 1.4 will be multiplied to the mean demand rate to generate different levels of system workload.

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, we assign the revenue of scheduling one pre-scheduled demand as 0.75, and 0.9 for seeing one open access patient. These values stand for the actual show rates. 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. Finally, the stochastic integer program uses a MIP tolerance gap of 0.01% Table 1 summarizes the parameters.

Table 1 Parameter settings for our computational experiments

Physician capacity	24 slots per day
Number of physicians in practice	3, 6
System workload	60%, 80%, 100%, 120%, 140%
Workload among physicians	Symmetric, asymmetric
Ratio of prescheduled to open access demand	10/14, 14/10, 6/18, 18/6
Scenarios for each replication	1,000
Number of replications	50
Revenue of seeing one pre-scheduled patient	0.75
Revenue of seeing one's own open access patient	0.90
Revenue of seeing one diverted open access patient	0.85
Relative MIP tolerance gap	0.01%

The results are divided into four main parts. We first discuss the value of flexibility in symmetric 3 and 6 physician clinics. "Symmetric" implies that physicians are identical with respect to both capacity and demand. Second, we discuss the value of flexibility in asymmetric 3 and 6 physician clinics. Physicians in this setting are identical with regard to their capacity, but vary with regard to their demands. This happens routinely in practice. Third, we compare the value of flexibility when the ratio of prescheduled (non-urgent) and open access (urgent) demands varies. This variation reflects differences in primary care clinic types—from urgent care clinics, where, as the name suggests, same-day urgent appointments are more prominent, to family medicine clinics, where prescheduled appointments are more common. Finally, we discuss the relationship between *Np* values, which determine how much capacity should be allocated for prescheduled appointments, and the level of system workload.

6.1 Value of flexibility in symmetric 3 and 6 physician cases

In the 3 and 6 physician cases, we set the ratio of mean prescheduled demand to mean open access demand for each physician to be 10/14 (in the tables we call this Symmetric 10/14). As discussed above, to achieve different levels of clinic workload, a factor varying from 0.6 to 1.4 is multiplied to the mean demand rate of each physician. We present results for 60, 80, 100, 120 and 140% cases. To illustrate, in the 120% workload case each physician has a mean prescheduled demand of $10 \times 1.2 = 12$ and an open access demand of $14 \times 1.2 = 16.8$

Tables 2, 3 and 4 give the measurement and comparison of 2-chain flexibility, full flexibility and dedicated case for the 3-physician case under different levels of system workload in the three dimensions of interest: system revenue, timely access rate and continuity rate.

Next, we present a graphical comparison (see Figs. 5, 6, 7 and 8) of the benefits of the 2-chain and full flexibility over the dedicated case in 3 and 6 physician cases for each of our measures (for the sake of brevity and ease of exposition we do not present tables for the 6-physician case).

Table 2 System revenue for different flexibility configurations in symmetric 10/14

System revenue Workload	60%	80%	100%	120%	140%
2-chain	35.1375	47.574	57.115	59.89385	61.1867
Full flex	35.1480	47.5819	57.1535	59.91734	61.1862
Dedicated	35.1185	46.8694	55.0977	58.63243	60.0828
2-chain versus dedicated	0.05%	1.50%	3.66%	2.15%	1.84%
Full versus dedicated	0.08%	1.52%	3.73%	2.19%	1.84%

Table 3 Timely access rate for different flexibility configurations in symmetric 10/14

Timely access rate Workload	60%	80%	100%	120%	140%
2-chain	99.95%	99.88%	95.29%	82.01%	70.10%
Full flex	99.98%	99.88%	95.29%	81.99%	70.08%
Dedicated	99.89%	98.40%	91.78%	80.72%	69.58%
2-chain versus dedicated	0.06%	1.50%	3.82%	1.59%	0.75%
Full versus dedicated	0.09%	1.50%	3.82%	1.58%	0.72%

Table 4 Continuity rate for different flexibility configurations in symmetric 10/14

Continuity Rate Workload	60%	80%	100%	120%	140%
2-chain	99.94%	98.24%	95.29%	97.03%	97.05%
Full Flex	99.94%	98.52%	96.41%	97.68%	97.71%
Dedicated	100.00%	100.00%	100.00%	100.00%	100.00%
2-chain versus dedicated	−0.06%	−1.76%	−4.71%	−2.97%	−2.95%
Full versus dedicated	−0.06%	−1.48%	−3.59%	−2.32%	−2.29%

We see that the highest benefit of both system revenue and timely access rate is achieved in the case where the system is balanced, i.e., when the expected demand equals the available capacity. The 2-chain and full flexibility cases are about 3–4% better than the dedicated case with regard to system revenue and timely access in the 3-physician case; and about 5–6% better in the 6-physician cases. Also, the graph of system performance improvement is not symmetric since the variability of demand at 140% workload is larger than 60% workload, presenting more opportunities for redirections.

The benefits of 2-chain flexibility are almost as high as those of full flexibility, with only a small detriment in terms of system revenue mainly due to increased patient redirections. The timely access rates of 2-chain flexibility and full flexibility are nearly the same no matter what the workload level of the system is. This is consistent with the results reported in the literature on flexibility in manufacturing

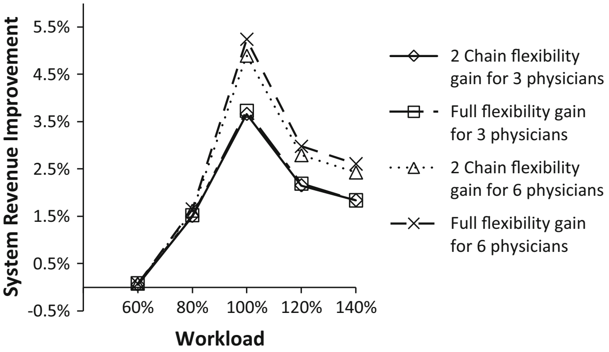

Fig. 5 Comparison of system revenue improvement between 3 and 6 physicians

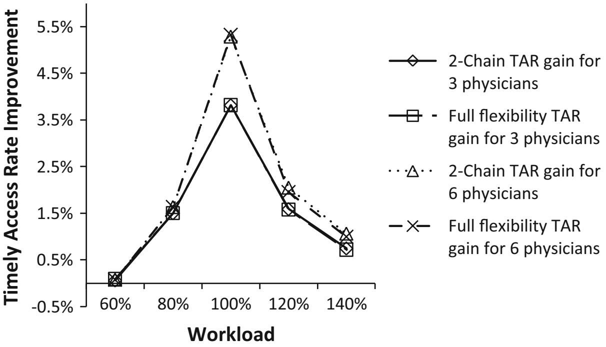

Fig. 6 Comparison of timely access rate (TAR) improvement between 3 and 6 physicians

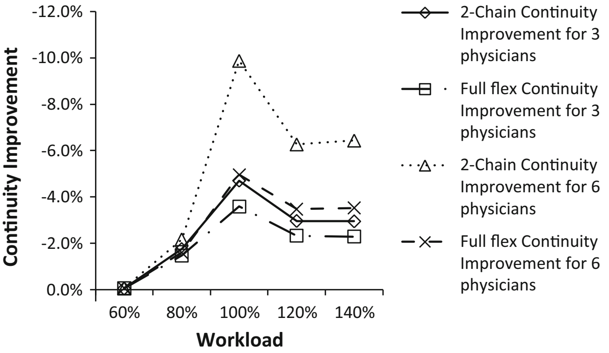

Fig. 7 Comparison of continuity improvement (over the dedicated case) between 3 and 6 physicians

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.

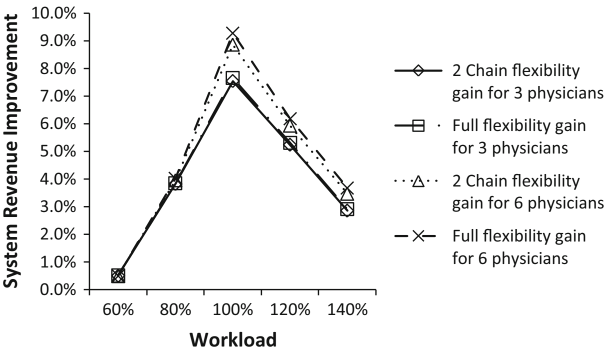

Fig. 8 Comparison of system revenue improvement between 3 and 6 physicians

The improvement in timely access comes at the cost of continuity (see Fig. 7). As discussed earlier, the 2-chain has a greater number of diversions (more "jumps" to ensure that the demand is directed along the chain) than full flexibility, and seemingly performs the worst with regard to continuity. However, the advantage of the 2-chain setting is that patients will always see one of two physicians; their care will be less fragmented. So although the loss in continuity for the 2-chain is 10% in the 6-physician case as compared to 5% under full flexibility, patients under full flexibility could end up seeing *any* of the 6 physicians in the practice. Care is thus significantly more fragmented and much harder to coordinate under full flexibility.

6.2 Value of flexibility in asymmetric 3 and 6 physician cases

The asymmetric case represents the situation where physicians have varying panel sizes and therefore varying appointment burdens. This is common in practice. Senior and well established physicians may have higher workloads since their panels are larger, while physicians who have been recently hired may have lower workloads. The benefits of flexibility are likely to be greater in the presence of such asymmetry. To test this, we create the following demand profiles for the 3 physician case. 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). Each physician still has 24 slots available in the day. Notice that although the individual physician workloads vary, the overall clinic or system workload is 100%. We use the same factors 0.6, 0.8, 1.2 and 1.4 to create varying levels of system workload, while still maintaining the imbalances between the physicians.

For the six physician case, we merely double the 3 physician case, thus retaining the imbalances. Thus there are two physicians with low workloads, two with balanced or full workloads and two with high workloads; the expected prescheduled and open access demands for each physician are identical to their 3 physician counterparts.

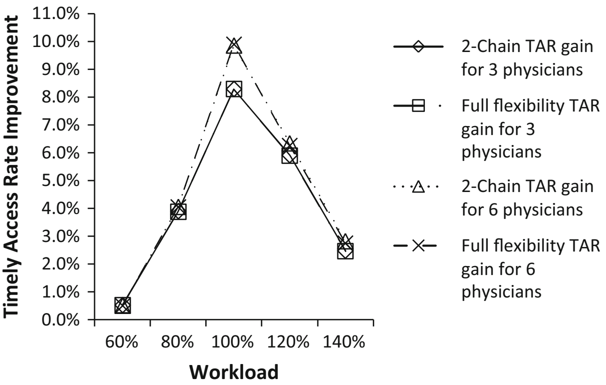

Fig. 9 Comparison of timely access rate (TAR) improvement between 3 and 6 physicians

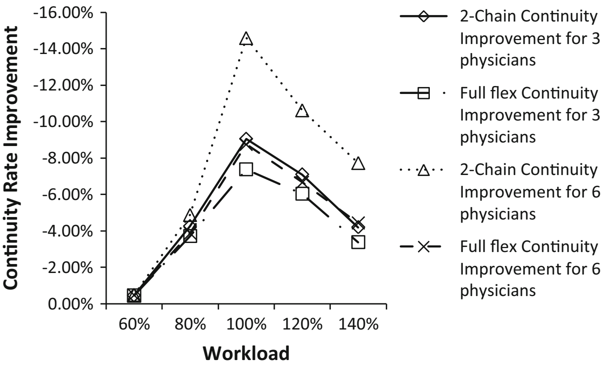

Fig. 10 Comparison of continuity rate (CR) improvement between 3 and 6 physicians

Figures 8, 9 and 10 present a graphical comparison of the benefits of the 2-chain and full flexibility over the dedicated case in 3 and 6 physician asymmetric cases for each of our measures. The results are similar to the symmetric cases except in one important respect. The magnitude of improvement in system revenue is nearly double that observed in the symmetric case: 8% in the 3 physician case and 10% in the 6 physician case, under 100% workload. The losses in continuity for the 2 chain are similarly greater when compared to the symmetric case. The differences between the 2-chain and full flexibility increase as the number of physicians increases. The difference is highest for the continuity rate.

6.3 Value of flexibility under different prescheduled and open access demand ratios

In the symmetric 3 and 6 physician cases, the ratio of prescheduled to open access demands was 10/14. In the asymmetric cases, it was 1/2 for each physician, even as their demand burdens varied. What effect, if any at all, does this ratio have on the

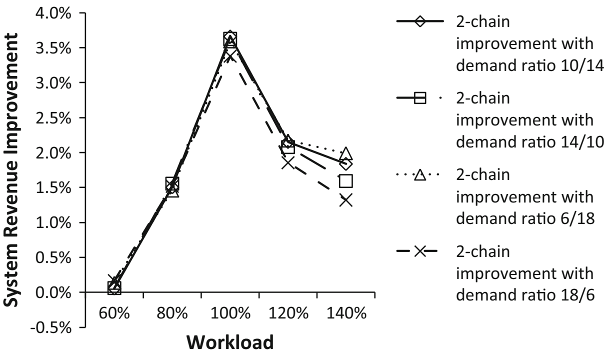

Fig. 11 2-chain flexibility improvement over the dedicated case under different ratios of prescheduled to open access demand for all symmetric cases

value of flexibility? The ratio is relevant for a practical reason. Primary care delivery varies substantially from clinic to clinic. A small town family medicine clinic that emphasizes continuity will likely have a high number of prescheduled appointments. At an urban urgent care clinic, walk-ins and same-day appointments may be more prominent than prescheduled appointments.

We test the value of flexibility under four different cases: 10/14 (already described), 14/10, 6/18 and 18/6. The first two represent only a slight skew in the ratio while the last two represent the extremes. 6/18 might represent an urgent care walk in clinic, while 18/6 might represent a well established family clinic. We test these ratios only on 3-physician symmetric cases. We expect to see similar trends in the 6-physician cases, with similar increases in revenue in going from 3 to 6 physicians as observed for the 10/14 ratio (see Figs. 5, 6 and 7).

Figure 11 shows the improvements obtained by the 2-chain configuration over the dedicated case under the four ratios of prescheduled to open access demand. The full flexibility case shows similar trends in improvement over the dedicated case. We observe that the system performs similarly under different demand ratios of prescheduled and open access appointments. The performance downgrades slightly when the demand ratio is 18/6—that is when the proportion of open access demand is reduced in relation to prescheduled demand. Since flexibility is only implemented in the open access phase, the benefit of using flexibility to balance the demands among physicians goes down due to lower in-bound open access demand.

Other system measures show the same properties. Although the absolute values of these metrics vary among different demand ratios due to the inequality of the revenues of the two types of demand, *the percentage improvements* of the flexible configurations over the dedicated case are not very sensitive to the change of the demand ratio between prescheduled and open access appointments. In particular, because of the higher value associated with meeting open access demand, the expected revenue is higher in cases where the open access demand is higher.

6.4 N_i^{p*} values and system workload

The results above discuss the value of flexibility. But what about the N_i^{p*} values, the first stage decisions in our model? What trends do they follow, if at all, and can the trends provide clues to capacity allocation decisions in practice? To understand this, we analyzed the total N_i^{p*} values of the clinic (that is the total capacity set aside by all physicians for prescheduled appointments), averaged over the 50 replications. Figure 12 shows the average N_i^{p*} 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' N_i^{p*} 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.

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

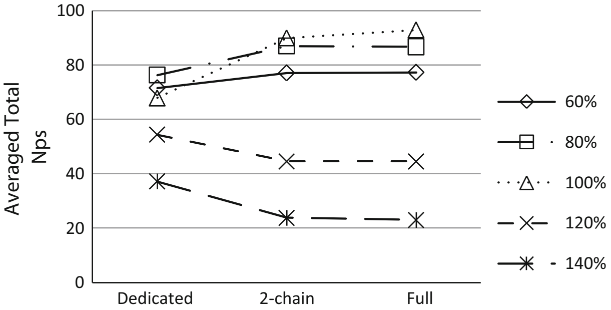

Fig. 12 Trends in averaged total *Np* values for 6 physicians with Prescheduled demands [6,8,10,6,8,10] and open access demands [12,16,20,12,16,20]

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.

7 Conclusions and future research

We have investigated the value of flexibility and its effect on capacity allocation for primary care practices. For dedicated and fully flexible cases, we develop analytical results, while for the general case with any flexibility configuration we develop a two-stage stochastic integer program. The results of our study confirm that introducing flexibility yields benefits even if there is a cost for using flexibility links. Similarly, we find that the benefits are the highest when the system is balanced, and decreasing for higher or lower levels of system workload. 2-chain flexibility yields almost all the benefits of full flexibility in terms of system revenue and timely access rate. While the number of patient redirections to alternative physicians is highest for the 2-chain, patients see only one of two physicians, and at the same time experience the same timely access benefits as full flexibility.

We also find that flexibility is more beneficial with increased number of physicians, and in the presence of asymmetry, that is when physicians have unequal workloads and flexibility can be used to balance supply and demand. The advantages of flexibility are not overly sensitive to the change of demand ratio between prescheduled and open access demands when physicians have equal workloads. Finally, our computational experiments show that the optimal capacity allocation decision under a flexibility configuration yields a specific structure. The optimal capacity to reserve for prescheduled appointments under flexible configurations tends to be higher for a system under a low workload and lower for a system under a high workload, as compared to the values obtained from the dedicated case.

Primary care practices are inherently flexible. Typically, the configuration most seen in practices is full flexibility. Our study provides capacity allocation guidelines for such practices. For smaller practices (consisting of two or three physicians), full flexibility may well be the best choice, since patients do not interact with too many physicians. However, for larger practices that are willing to redesign their team structure, the 2-chain is an attractive choice.

Several future research directions are possible. The models of flexibility in the paper emerged as a result of interactions that the authors had with clinics in the United States. These included primary care clinics in academic medical centers (in both rural and urban setting) as well as smaller family medicine clinics. However, a more formal empirical study of different primary care clinic types (urgent care, pediatric practice, adult primary care) that throws light on how flexibility is actually used in practice and how patients and physicians perceive continuity and timely access is necessary to validate our conclusions. For example, while full flexibility may be the default mode for many clinics, larger practices implement a different kind of partial flexibility, which we call *subgroups*. In the subgroup configuration, the physicians are divided into a smaller number of self-contained but fully flexible groups. A practice consisting of four physicians might be divided into two groups of two, functioning independently. Our preliminary comparisons of 2-chain and subgroups have revealed that they perform very close to each other, with the 2-chain having a consistent but small advantage. However, subgroups are attractive since they are easier to implement in practice.

Other research directions include the management of physician flexibility in a dynamic context, when allocation decisions are made as patients call in without full knowledge of future demand. Finally, a formal heuristic that suggests the optimal number of prescheduled appointments to reserve, based on the properties and analytical results discussed earlier, would complement this research.

Acknowledgments 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. We would also like to acknowledge two anonymous referees whose valuable comments helped improve the paper. Finally, our sincere thanks to Xiaoling Gao, doctoral student at the University of Massachusetts, who helped with editing of the results.

References

Atlas S, Grant R, Ferris T, Chang Y, Barry M (2009) Patient–Physician connectedness and quality of primary care. Ann Intern Med 150(5):325–326

Bennett K, Baxley E (2009) The effect of a carve out advanced access scheduling system on no show rates. Fam Med 41(1):51–56

Birge J, Loveaux F (1997) Introduction to stochastic programming. Springer, New York, NY

Chua GBA, Chou MC, Teo C-P (2008) On range and response: dimensions of process flexibility. Working paper, NSU

Edington M (eds) (2001) Crossing the quality chasm: a new health system for the 21st century. The Institute of Medicine Report. Technical report, National Academy Press, Washington DC

Gill JM, Mainous A (1999) The role of provider continuity in preventing hospitalizations. Arch Fam Med 7:352–357

Gill JM, Mainous A, Nsereko M (2000) The effect of continuity of care on emergency department use. Arch Fam Med 9:333–338

Graves SC, Tomlin BT (2003) Process flexibility in supply chains. Manage Sci 49(7):907–919

Green LV, Savin S (2008) Reducing delays for medical appointments: a queueing approach. Oper Res 56(6):1526–1538

Green LV, Savin S, Murray M (2007) Providing timely access to care: what is the right patient panel size? Jt Comm J Qual Patient Saf 33:211–218

Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. Oper Res 56(3):576–592

Gupta D, Potthoff S, Blowers D, Corlett J (2006) Performance metrics for advanced access. J Healthc Manag 51(4):246–259

Hippchen J (2009) Flexbility in primary care, Masters Thesis (Advisors: Hari Balasubramanian and Ana Muriel). Accessible at: http://people.umass.edu/hbalasub/FlexibilityThesis.pdf

Hopp W, Tekin E, Van Oyen MP (2004) Benefits of skill chaining in serial production lines with cross-trained workers. Manage Sci 50(1):83–98

Jordan WC, Graves SC (1995) Principles and benefits of manufacturing process flexibility. Manage Sci 41(4):577–594

Kopach R, DeLaurentis P, Lawley M, Muthuraman K, Ozsen L, Rardin R, Wan H, Intrevado P, Qu X, Willis D (2007) Effects of clinical characteristics on successful open access scheduling. Health Care Manage Sci 10:111–124

Liu N, Ziya S, Kulkarni V (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf Serv Oper Manage 12.2:347–365

Muriel A, Somasundaram A, Zhang Y (2006a) Impact of partial manufacturing flexibility on production variability. Manuf Serv Oper Manage 8(2):192–205

Muriel A, Zhang Y, Biller S (2006b) Impact of price postponement on capacity and flexibility investment decisions. Prod Oper Manage 15(2):198–214

Murray M, Berwick DM (2003) Advanced access: reducing waiting and delays in primary care. J Am Med Assoc 289(8):1035–1040

Murray M, Bodenheimer T, Rittenhouse D, Grumbach K (2003) Improving timely access to primary care: case studies of the advanced access model. J Am Med Assoc 289(3):1042–1046

O'Malley A, Cunningham P (2009) Patient experiences with coordination of care: the benefit of continuity and primary care physician as referral source. J Gen Intern Med 24(2):170–177.

Qu X, Rardin R, Williams JAS, Willis D (2007) Matching daily healthcare provider capacity to demand in advanced access scheduling systems. Eur J Oper Res 183(2):812–826

Robinson L, Chen R (2010) A comparison of traditional and open access policies for appointment scheduling. Manuf Serv Oper Manage 122:330–347

Rust G, Ye J, Baltrus P, Daniels E, Adesunloye B, Fryer GE (2008) Practical barriers to timely primary care access. Arch Intern Med 268(15):1705–1710

Solak S, Clarke J-P, Johnson E, Barnes E (2010) Optimization of R&D portfolios under endogenous uncertainty. Eur J Oper Res 207(1):420–433

Author Biographies

Dr. Hari Balasubramanian is an assistant professor of Industrial Engineering at the University of Massachusetts, Amherst. His research interests are in operations research applied to healthcare delivery. Specific application areas include capacity planning and scheduling in primary care, surgical suites and emergency departments. Dr. Balasubramanian has a PhD in Industrial Engineering from Arizona State University. He was a Research Associate at the Department of Health Sciences Research from August 2006-2008 at the Mayo Clinic in Rochester, Minnesota.

Prof. Ana Muriel is an associate professor of Industrial Engineering at the University of Massachusetts Amherst. Her research focuses on various aspects of logistics and supply chain management, and has recently branched out to applying some of the successful strategies there to the healthcare domain. Dr. Muriel has MS and PhD degrees from Northwestern University, and started her career at the Ross School of Business at the University of Michigan. She has also held visiting positions at the Olin School of Business and at the Economics Department in the Universidad de Salamanca. She is an associate editor of Naval Research Logistics and IIE Transactions on Scheduling and Logistics.

Liang Wang currently works as a General Manger at Autonomous Earthmoving Equipment LLC. He has a master's degree in Industrial Engineering from the University of Massachusetts, Amherst and a masters in Electrical Engineering in Huazhong University of Science and Technology, China. The work in this paper was part of Liang's master's thesis.