

1

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

5 **Hari Balasubramanian · Ana Muriel · Liang Wang**

6
7 © Springer Science+Business Media, LLC 2011

8 **Abstract** The two important but often conflicting metrics for any primary care
9 practice are: (1) *Timely Access* and (2) *Patient-physician Continuity*. Timely access
10 focuses on the ability of a patient to get access to a physician (or provider, in
11 general) as soon as possible. Patient-physician continuity refers to building a strong
12 or permanent relationship between a patient and a specific physician by maximizing
13 patient visits to that physician. In the past decade, a new paradigm called *advanced*
14 *access* or *open access* has been adopted by practices nationwide to encourage
15 physicians to “do today’s work today.” However, most clinics still reserve pre-
16 scheduled slots for long lead-time appointments due to patient preference and
17 clinical necessities. Therefore, an important problem for clinics is how to optimally
18 manage and allocate limited physician capacities as much as possible to meet the
19 two types of demand—pre-scheduled (non-urgent) and open access (urgent, as
20 perceived by the patient)—while simultaneously maximizing timely access and
21 patient-physician continuity. In this study we adapt ideas of manufacturing process
22 flexibility to capacity management in a primary care practice. Flexibility refers to
23 the ability of a primary care physician to see patients of other physicians. We
24 develop generalizable analytical algorithms for capacity allocation for an individual
25 physician and a two physician practice. For multi-physician practices, we use a two-
26 stage stochastic integer programming approach to investigate the value of flexi-
27 bility. We find that flexibility has the greatest benefit when system workload is
28 balanced, when the physicians have unequal workloads, and when the number of
29 physicians in the practice increases. We also find that partial flexibility, which

A1 H. Balasubramanian (✉) · A. Muriel
A2 Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst,
A3 160 Governors Drive, Amherst, MA 01003, USA
A4 e-mail: hbalasubraman@ecs.umass.edu
A5 URL: <http://people.umass.edu/hbalasub>

A6 L. Wang
A7 Autonomous Earthmoving Equipment LLC, 21755 Interstate 45, BLDG 5, Spring, TX 77388, USA

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

32
 33 **Keywords** Appointment scheduling · Open access · Flexibility · Capacity
 34 planning · Timely access · Continuity of care · Stochastic programming
 35

36 1 Introduction

37 Primary care providers (PCPs) are typically the first point of contact between
 38 patients and health systems. They include general practitioners, family doctors,
 39 pediatricians, and geriatricians. From a patient's perspective, PCPs provide the
 40 majority of care they receive during their lifetime. They are responsible for a variety
 41 of health services including preventive medicine, patient education, routine physical
 42 exams, and the coordination of complex episodes in which patients are referred to
 43 medical specialties for secondary and tertiary care.

44 Each primary care physician (or provider, in general) has a *panel* of patients,
 45 whose care she is responsible for. Long term, holistic care of patients is one of the
 46 cornerstones of primary care. In practice this translates to maximizing the number of
 47 visits with the patient's own PCP—maximizing continuity, in other words.
 48 Numerous studies have documented the importance of continuity to patients
 49 (O'Malley and Cunningham 2008; Atlas et al. 2009). Gill and Mainous (1999) point
 50 to several studies which show that patients who regularly see their own providers
 51 are (1) more satisfied with their care; (2) more likely to take medications correctly;
 52 (3) more likely to have problems correctly identified by their physician; and (4) less
 53 likely to be hospitalized. Gill et al. (2000) show a link between lack of continuity
 54 and increased emergency department use.

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

64 To address this issue many primary care practices have tried implementing
 65 *advanced access* or *open access* (Murray and Berwick 2003; Murray et al. 2003).
 66 Advanced access promotes the concept that physicians should “do today's work
 67 today” rather than push appointments into the future. In the ideal open access world,
 68 there are no appointment types, such as urgent and non-urgent. All appointments are
 69 treated identically and scheduled the same day with the patient's PCP.

70 However, the reality is that clinics have a fraction of their schedule available for
 71 open access or urgent appointments. Such appointments, because of the perceived
 72 immediacy of need, are typically seen the same day, but not always by the patient's
 73 personal physician. The rest of the clinic's schedule consists of appointments

74 booked a week or more in advance. We call these appointments *prescheduled*
 75 *appointments*. These non-urgent appointments are typically physicals or follow-up
 76 appointments for patients with chronic conditions.

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

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

98 2 Flexibility, capacity allocation and revenue maximization in primary care

99 2.1 Configurations of flexibility

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

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



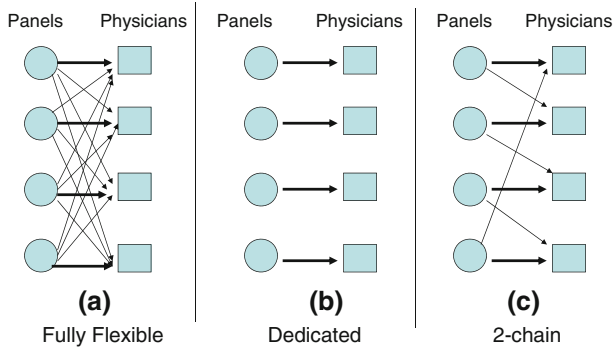


Fig. 1 Different flexibility configurations that tradeoff continuity with timeliness

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

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

135 2.2 Capacity allocation

136 Each physician has a fixed number of slots available in a day. If the physician works
 137 an 8 h day, this typically means 24 appointment slots, since each appointment is
 138 commonly allotted 20 min. Each physician also has prescheduled and open access
 139 demand distributions that have been identified using past data. The question we
 140 address is the following: given a certain flexibility configuration, how many slots
 141 should each physician make available for prescheduled appointments? In other
 142 words, we explore the clinic's decision of how many slots to set aside for open
 143 access appointments. We note here that the allocation decision is made at an
 144 aggregate level. That is, we assume that the demands are realized "instantly" and
 145 are fulfilled instantly as well if capacity is available (the exact allocation mechanism

146 is a two-stage process, described in Sect. 4). In practice, demand gets realized over
147 time and decisions have to be made without full knowledge of future demand. So
148 our approach provides only an approximation of the actual capacity allocation
149 process, in order to make high-level capacity reservation decisions.

150 2.3 Maximizing revenue

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

160 Our revenue maximization also reflects the fact that, in a relative sense, open
161 access appointments are more "valuable" than prescheduled appointments. This is
162 because of three reasons. First, open access appointments, because they have such
163 short lead times, have much lower no-show rates. Second, if a prescheduled
164 appointment results in a no-show, it can be substituted by an open access
165 appointment, while the reverse is not possible at such a short notice. Third, because
166 prescheduled appointments are made generally a week or more in advance, the
167 patient is likely to be flexible about choice of the appointment day, and thus this
168 may result in postponed but not lost demand if denied timely access. An open access
169 patient, on the other hand, needs to see a physician immediately and hence is
170 flexible in provider choice.

171 2.4 2-Chain versus full flexibility

172 Since full flexibility has more "outbound" links than 2-chain flexibility, it should
173 have a better ability to absorb incoming demands and yield a higher timely access
174 rate than 2-chain flexibility. This would be true for the dynamic setting of patient
175 scheduling where allocation decisions are made as requests arrive, with limited
176 future knowledge of the overall demand that will need to be serviced (Hippchen
177 2009). By contrast, in the aggregate demand setting captured by our two-stage
178 stochastic integer programming approach, the patient allocation (second stage
179 decisions) is only performed after the full system demand is realized (that is once
180 the scenario is known). Thus the 2-chain, which indirectly links all the physicians,
181 will in most cases manage to serve the same number of patients as a fully flexible
182 practice where all physicians are linked to each other directly. To be sure, there are
183 rare instances where full flexibility will clearly dominate. For instance, consider a
184 practice with four physicians, where each has 10 slots left for open access, and the
185 demands for open access are 20, 20, 0 and 0 respectively. In this extreme case, the
186 2-chain flexibility can only meet 30 open access demands, while the full flexibility
187 can satisfy all of them. Since such an instance would occur with a low probability,

188 from a statistical point of view, the 2-chain flexibility has almost the same
189 effectiveness to absorb the demand as full flexibility.

190 This ability of the 2-chain to match full flexibility in timely access comes at the
191 cost of increased diversion rates (in our model this is a measure of continuity). Since
192 full flexibility has more “outbound” links than 2-chain flexibility, it should have a
193 higher probability that the demand will be diverted to other physicians. In reality,
194 however, a single patient redirection to an available physician, which can be made
195 directly under full flexibility, may require redirecting several patients along the
196 2-chain if the initial patient’s panel and available physician involved are not
197 connected. For example, Fig. 2 shows a case of three physicians where each
198 physician has 10 slots left for open access, and the demands are 16, 10 and 4
199 respectively. We can see that the total number of diversions under 2-chain flexibility
200 is 12, but only 6 under full flexibility. Since 2-chain flexibility requires more
201 “jumps” to shift the demands, the diversion rate of the 2-chain is higher than that of
202 full flexibility in our model.

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

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

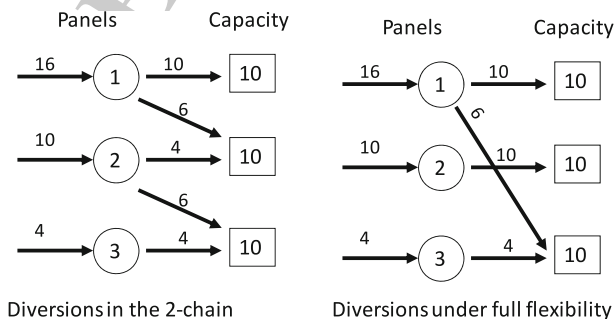


Fig. 2 An example of diversion process in 2-chain and full flexibility

220 3 Literature review

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

226 3.1 Operations research applied to open access

227 The literature on optimization approaches applied to open access is growing. The
228 adoption of open access, which promises patients same-day appointments, has
229 prompted a series of questions. What should physician panel sizes be to allow open
230 access? What if patients prefer to have appointments at some future time rather
231 than see a doctor the same day? These questions have necessitated the use of
232 queuing and stochastic optimization approaches that provide guidelines to
233 practices. For instance, Green et al. (2007) investigate the link between panel
234 sizes and the probability of “overflow” or extra work for a physician under
235 advanced access. They propose a simple probability model that estimates the
236 number of extra appointments that a physician could be expected to see per day as
237 function of her panel size. The principal message of their work is that for advanced
238 access to work, supply needs to be sufficiently higher than demand to offset the
239 effect of variability. Green and Savin (2008) use a queuing model to determine the
240 effect of no-shows on a physician’s panel size. They develop analytical queuing
241 expressions that allow the estimation of physician backlog as a function of panel
242 size and no-show rates. In their model, no show rates increase as the backlog
243 increases; this results in the paradoxical situation where physicians have low
244 workloads even though backlogs are high—this is because patients that had to wait
245 for long, do not show up.

246 Gupta et al. (2006) conduct an empirical study of clinics in the Minneapolis
247 metropolitan area that adopted open access. They provide statistics on call volumes,
248 backlogs, number of visits with own physician (which measures continuity) and
249 discuss options for increasing capacity at the level of the physician and clinic.
250 Kopach et al. (2007) use discrete event simulation to study the effects of clinical
251 characteristics in an open access scheduling environment on various performance
252 measures such as continuity and overbooking. One of their primary conclusions is
253 that continuity in care is affected adversely as the fraction of patients on open access
254 increases. The authors mention provider groups (or physicians and support staff)
255 working in teams as a solution to the problem. Robinson and Chen (2010) compare
256 the performance of open access with that of a traditional appointment scheduling
257 system. Their numerical analysis reveals that unless patient wait times to secure an
258 appointment have marginal weights in the objective function and patient no-show
259 rates are too small, open access is preferable to traditional scheduling systems. Liu
260 et al. (2010) propose new heuristic policies for dynamic scheduling of patient
261 appointments under no-shows and cancellations. They find that open access works
262 best when patient load is relatively low.



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

280 3.2 Literature related to flexibility

281 Our study of flexibility in primary care practices builds upon the extensive literature
 282 on manufacturing flexibility and its more recent application to service systems and
 283 worker training and allocation. There are, however, key operational differences that
 284 make the application of flexibility to primary care worthy of further analysis: (1)
 285 two demand streams associated with each resource, where one (prescheduled
 286 demands) gets realized before the other (open access demands); (2) two conflicting
 287 objectives, timeliness and continuity of care; (3) no fixed cost associated with
 288 installing flexibility, but a loss in continuity for using it; (4) appointments are
 289 booked over time and thus future resource capacity is sequentially being allocated
 290 under partial demand information. The latter point is mute in our aggregate analysis
 291 of the capacity allocation problem, but key to the dynamic clinic scheduling
 292 problem (see Hippchen 2009).

293 As in the case of cross-training in serial production lines (Hopp et al. 2004)
 294 flexibility improves efficiency in two main ways in the primary care environment.
 295 The first benefit is in what they refer to as *capacity balancing*: If physician panels
 296 are imbalanced with respect to the induced average number of visits to a physician
 297 per day, flexibility will allow the load to be shared between physicians, therefore
 298 improving overall timeliness of care and physician utilization. The second is in
 299 *variability buffering*: Even if the average workloads are balanced between
 300 physicians, variability in patient requests for a particular day/time will be better
 301 accommodated by a flexible environment. Hopp et al. (2004) compare a strategy
 302 that balances capacity using the minimum amount of cross-training with the
 303 chaining of skills in the sequence of the serial line. They find that skill-chaining
 304 strategies are more robust, and more effective in variability buffering. The benefits
 305 of flexibility in increased sales and capacity utilization in multi-product, multi-plant
 306 production networks have been thoroughly studied by Jordan and Graves (1995)

307 considering a single production period. They are the first to introduce the concept of
 308 chaining to achieve maximum benefits from limited flexibility configurations where
 309 each plant produces only a few of the products. Furthermore, this strategic analysis
 310 has been extended recently to multi-stage supply chains (Graves and Tomlin 2003),
 311 and to a make-to-order environment where flexibility is also used to hedge against
 312 operational variability (Muriel et al. 2006a, b). Chua et al. (2008) distinguish
 313 between range and response of flexible systems. Range refers to the set of demand
 314 scenarios that can be accommodated and response to the cost of doing so; that is, the
 315 cost of using secondary rather than primary resources for production/service. They
 316 show that upgrading system response outperforms improving system range. In the
 317 primary care setting, this means that systems where physicians can handle other
 318 physician's panels at lower additional cost should be preferred over those that can
 319 accommodate ever more extreme patient demand scenarios. This result suggests that
 320 the benefits of restricting the number of doctors that can see a particular patient
 321 (resulting in lower cost of using the secondary providers because of familiarity, and
 322 thus increased response) is likely to outweigh the higher range provided by a fully
 323 flexible team care practice where any doctor can see the patient.

324 **4 Model**

325 We consider a general primary care practice with M physicians, each with N_i
 326 available appointment slots, $i = 1, 2, \dots, M$. Let A be the set of all possible panel-
 327 physician links (i, j) such that patients in panel i (i.e., physician i 's panel) can be
 328 served by physician j . The set A represents the particular flexibility configuration
 329 under consideration; that is, the network of allowed open-access patient redirections
 330 within the practice. We assume that pre-scheduled patients are required to see their
 331 own physicians, and physician flexibility can only be used for the time-sensitive
 332 open access demand patients. This is the most relevant case in practice, since
 333 patient-physician continuity is highly beneficial to prescheduled appointments, in
 334 which major physicals or follow-ups of chronic conditions are performed.

335 Let R_i^p be the revenue associated with physician i , $i = 1, 2, \dots, M$, seeing one of
 336 his pre-scheduled patients, and R_{ij}^o be the revenue associated with physician j seeing
 337 an open-access patient of panel i , for any $(i, j) \in A$. The demand for prescheduled
 338 and open access appointments can be represented by a random vector
 339 $D = (D_1^p, D_1^o, \dots, D_M^p, D_M^o)$, where the super-index p refers to prescheduled and
 340 o to open access, and the sub-index indicates the primary care physician. D follows
 341 a discrete distribution that assigns a probability q_s to each possible realization of
 342 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,
 343 $P[D = (d_{1s}^p, d_{1s}^o, \dots, d_{Ms}^p, d_{Ms}^o)] = q_s$.

344 We introduce the following capacity allocation variables.

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

346 x_{is}^p : Number of patients pre-scheduled with physician i under demand
 347 realization s .

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

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

$$\phi_{iuis} = 1 \text{ if } u_{is} \equiv \min\{d_{is}^p, N_i\} < N_i^p, \text{ otherwise, } \phi_{iuis},$$

$$\text{for } i = 1, 2, \dots, M \text{ and } s = 1, 2, \dots, S.$$

354 Observe that $u_{is} = \min\{d_{is}^p, N_i\} \in \{0, 1, 2, \dots, N_i\}$; therefore, the total number of
 355 binary variables ϕ_{iuis} equals the total number of appointment slots in the practice,
 356 plus one more per physician, $N_1 + N_2 + \dots + N_M + M$.

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

$$\text{Objective : Max } \sum_{s=1}^S \sum_{i=1}^M q_s \left[R_i^p x_{is}^p + \sum_{(i,j) \in A} R_{ij}^o x_{ijs}^o \right] \tag{1}$$

360 Subject to : $N_i^p \leq N_i \quad \forall i = 1, 2, \dots, M$ (2)

362 $N_i^p \leq d_{is}^p + N_i \phi_{iuis} \quad \forall i = 1, 2, \dots, M, s = 1, 2, \dots, S$ (3)

364 $N_i^p \geq d_{is}^p \phi_{iuis} \quad \forall i = 1, 2, \dots, M, s = 1, 2, \dots, S$ (4)

366 $x_{is}^p \leq N_i^p \quad \forall i = 1, 2, \dots, M, s = 1, 2, \dots, S$ (5)

368 $x_{is}^p \leq d_{is}^p \quad \forall i = 1, 2, \dots, M, s = 1, 2, \dots, S$ (6)

370 $\sum_{i:(i,j) \in A} x_{ijs}^o \leq N_j - d_{js}^p \phi_{juis} \quad \forall j = 1, 2, \dots, M, s = 1, 2, \dots, S$ (7)

372 $\sum_{i:(i,j) \in A} x_{ijs}^o \leq N_j - N_j^p + \phi_{juis} N_j \quad \forall j = 1, 2, \dots, M, s = 1, 2, \dots, S$ (8)

374 $\sum_{j:(i,j) \in A} x_{ijs}^o \leq d_{is}^o \quad \forall i = 1, 2, \dots, M, s = 1, 2, \dots, S$ (9)

376 $\phi_{iuis} \in \{0, 1\} \quad \forall i = 1, 2, \dots, M, u_{is} = 0, 1, \dots, N_i$ (10)

378 $N_i^p, x_{is}^p, x_{ijs}^o \geq 0 \quad \forall i, j = 1, 2, \dots, M, (i, j) \in A, s = 1, 2, \dots, S$ (11)
 and integer

380 Equation 3 ensures that $\phi_{iuis} = 1$ if $d_{is}^p < N_i^p$. Equation 4 ensures that $\phi_{iuis} = 0$ if
 381 $d_{is}^p > N_i^p$. Equations 5 and 6 limit the number of pre-scheduled appointments to
 382 the allocated capacity and the realized demand, respectively. Equations 7 and 8
 383 ensure that the total open access appointments for any physician j do not exceed
 384 remaining capacity, when $\phi_{juis} = 1$ and $\phi_{juis} = 0$ respectively. Equation 9 limits
 385 the total number of open access appointments scheduled from a panel to the
 386 realized demand for such appointments from that panel. Equation 10 is the binary
 387 constraint.

388 In the next two sections, we present analytical solutions to the capacity allocation
 389 problem for dedicated practices, where physicians can only see patients in their
 390 own panel, and fully flexible practices where open-access patients can be seen by
 391 any of the physicians in the practice. For large practices, unfortunately, the above
 392 stochastic program is too large to solve efficiently in practice. While the number of
 393 binary and integer variables is quite manageable, the sheer number of possible
 394 demand realizations makes the problem intractable. To overcome this issue, we will
 395 solve the problem using a computationally effective sample average approximation
 396 method proposed by Solak et al. (2010) for two-stage stochastic integer
 397 programming problems; see Sect. 6.

398 **5 Analysis of dedicated and fully flexible group practices**

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

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

412 **5.1 Dedicated practice**

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

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

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

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

Fig. 3 System configuration for a dedicated practice



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

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

425 Dedicated Problem:

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

426 The conditions for local optimality presented in Qu et al. (2007) for the problem
 428 of maximizing the expected number of patients consulted in a single-physician
 429 practice can be easily adapted to our revenue maximizing objective. In what
 430 follows, we show a stronger result, guaranteeing global optimality, by first showing
 431 that the objective function has diminishing returns under mild assumptions.

432 **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*
 433 *in revenue associated with increasing the number of prescheduled slots by one,*
 434 *$ER_i(N_i^p + 1) - ER_i(N_i^p)$, is non-increasing in N_i^p .*

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

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

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

444 The above condition ($P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p]$ non-decreasing in N_i^p) simply
 445 requires that given that demand for prescheduled appointments is at least as large as

446 the number of allocated slots N_i^p , the conditional probability of open access demand
 447 being greater than or equal to the remaining slots, $N_i - N_i^p$, does not decrease as N_i^p
 448 grows. In particular, the condition holds when the demand for prescheduled and
 449 open-access appointments are independent of each other. Furthermore, it will be
 450 satisfied in most practical scenarios. Intuitively, for it to be violated, the probability
 451 of open access demand being large would need to significantly decrease as the
 452 demand for prescheduled appointments grows; that is, the demand for open access
 453 and prescheduled appointments would need to be heavily negatively correlated.

454 As a result of Proposition 1, we have that the expected revenue function has
 455 diminishing returns, an analog of concavity for a discrete function, and thus its
 456 global maximum must occur at the largest integer $N_i^p \leq N$ such that $ER_i(N_i^p) -$
 457 $ER_i(N_i^p - 1) \geq 0$.

458 **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*
 459 *solution to the Dedicated Problem is the largest non-negative integer $N_i^p \leq N$ such*
 460 *that $P[D_i^o \geq N_i - N_i^p | D_i^p \geq N_i^p] \leq \frac{R_i^o}{R_i^p}$.*

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

467 The optimal solution can thus be easily obtained by calculating that probability
 468 starting at $N_i^p = 0$ and increasing one unit at a time until it exceeds the threshold
 469 R_i^o/R_i^p . A binary search could also be used.

470 Observe that in the case of independent open-access and prescheduled demands,
 471 the optimal value, N_i^{p*} , does not depend on the distribution of pre-scheduled demand
 472 for physician i .

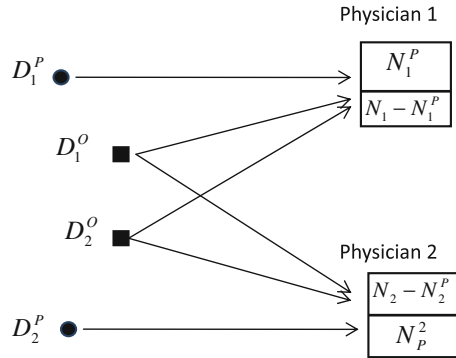
473 5.2 Fully flexible practice

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

476 In this case, the optimal number of slots to make available to prescheduled
 477 demand of the physicians, N_1^{p*} and N_2^{p*} , can still be found with a simple greedy
 478 algorithm as shown in Theorem 2. For ease of exposition, we assume that all
 479 physicians have the same capacity of N slots, and that the revenue of an open access
 480 appointment is identical for all physicians and panels and denoted by R^o . We first
 481 consider the case of two physicians.

482 **Proposition 2** *If $P[D_1^o + D_2^o > 2N - (N_1^p + \min(D_2^p, N_2^p) + 1) | D_1^p \geq N_1^p + 1]$ is*
 483 *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



484 the number of prescheduled slots of physician 1 by one, $ER(N_1^p + 1, N_2^p) -$
 485 $ER(N_1^p, N_2^p)$, is non-increasing in N_1^p and N_2^p .

486 By symmetry, if $P[D_1^o + D_2^o > 2N - (\min(D_1^p, N_1^p) + N_2^p + 1) | D_2^p \geq N_2^p + 1]$ is
 487 non-decreasing in N_1^p and N_2^p , the difference in revenue associated with increasing
 488 the number of prescheduled slots of physician 2 by one, that is, $ER(N_1^p, N_2^p + 1) -$
 489 $ER(N_1^p, N_2^p)$, is non-increasing in N_1^p and N_2^p .

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

$$ER(N_1^p + 1, N_2^p) - ER(N_1^p, N_2^p) = P[D_1^p \geq N_1^p + 1] (R_1^p - R^o P[D_1^o + D_2^o > 2N - (N_1^p + \min(D_2^p, N_2^p) + 1) | D_1^p \geq N_1^p + 1]) \quad (14)$$

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

499 Observe that, as in the dedicated case, the conditions will hold when open access
 500 and prescheduled demands are independent, and in any practical scenario except for
 501 contrived cases where the demands for prescheduled and open access appointments
 502 are severely negatively correlated. The proposition shows that the revenue function
 503 exhibits decreasing returns in both N_1^p and N_2^p under those mild conditions; this can
 504 be interpreted as concavity of the discrete revenue function and it implies that a
 505 greedy algorithm, as stated in the following theorem, is optimal.

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

510 *Step 1: Initialize $N_1^p = 0$ and $N_2^p = 0$.*
 511 *Step 2: Calculate $F_1 = ER(N_1^p + 1, N_2^p) - ER(N_1^p, N_2^p)$ and $F_2 = ER(N_1^p, N_2^p + 1) -$*
 512 *$ER(N_1^p, N_2^p)$*
 513 *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*
 514 *$N_2^p := N_2^p + 1$ and go to Step 2. If $F_1 \leq 0$ and $F_2 \leq 0$, then STOP. Else, if $F_1 > 0$*
 515 *move on to Step 4, and if $F_2 > 0$ go to Step 5.*
 516 *Step 4: While $F_1 > 0$, do $N_1^p := N_1^p + 1$ and update $F_1 = ER(N_1^p + 1, N_2^p) -$*
 517 *$ER(N_1^p, N_2^p)$*
 518 *Step 5: While $F_2 > 0$, do $N_2^p := N_2^p + 1$ and update $F_2 = ER(N_1^p, N_2^p + 1) -$*
 519 *$ER(N_1^p, N_2^p)$*

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

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

$$ER(N_1^p + 1, N_2^p, \dots, N_M^p) - ER(N_1^p, N_2^p, \dots, N_M^p) \\ = P[D_1^p \geq N_1^p + 1] \left(R_1^p - R^o P \left[D_1^o + D_2^o + \dots + D_M^o > MN \right. \right. \\ \left. \left. - (N_1^p + \min(D_2^p, N_2^p) + \dots + \min(D_M^p, N_M^p) + 1) | D_1^p \geq N_1^p + 1 \right] \right)$$

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

532 6 Computational experiments

533 We use the two stage stochastic integer program (SIP) to test the value of flexibility
 534 under the three different flexibility configurations—full flexibility, partial flexibility
 535 (2-chain) and no flexibility (dedicated)—under a number of settings. We will focus
 536 on three measures: system revenue, timely access rate and continuity rate. System
 537 revenue stands for the total expected revenue of meeting patient demands; timely
 538 access rate is the percentage of all patients, both prescheduled and open access, who
 539 get access to an appointment; and continuity rate presents the percentage of open
 540 access patients who see their own physician. Our model provides the optimal value
 541 of $N_1^{p*}, N_2^{p*}, \dots, N_M^{p*}$ (the first stage decisions), the optimal allocation of patients to
 542 physicians and the optimal numbers of non-PCP diversions that should be made
 543 (second stage decisions).

544 We solve the SIP in the extensive form or deterministic equivalent form (Birge
 545 and Loveaux 1997) by creating random demand scenarios. The computational
 546 complexity of our model heavily depends on the number of scenarios, the most

547 influential factor, and the number of physicians. Although our stochastic integer
 548 programming model can theoretically investigate the value of flexibility for any
 549 flexibility configuration with any number of physicians, the time-consuming nature
 550 of the optimization and evaluation makes it impractical. For tractability, we use
 551 concepts from the computationally effective sample average approximation method,
 552 proposed by Solak et al. (2010) for two-stage stochastic integer programming
 553 problems. The basic idea is to create a manageable number of samples/scenarios to
 554 produce an estimation of the optimal objective value and corresponding first stage
 555 solutions. To determine if the number of samples/scenarios is sufficient, we further
 556 run a large number of scenarios to have a precise estimation of the objective value
 557 based on the fixed first stage solution. This process is repeated over a number of
 558 replications to provide confidence intervals and statistical guarantees on the quality
 559 of the estimation. Using this method, we determined that 1,000 scenarios and 50
 560 replications gives us narrow enough confidence intervals for both 3 physician and
 561 6 physician cases. To allow for a fair comparison, the 2-chain, full flexibility and
 562 dedicated configurations use the same set of scenarios (analogous to the common
 563 random numbers approach in discrete event simulation).

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

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

574 Workload in our model is defined as the ratio of the expected total demand for the
 575 clinic and total available capacity. For instance, in a practice with three physicians,
 576 suppose each physician has a demand rate of 10 for prescheduled appointment and
 577 14 for open access demand. The total expected demand is $10 \times 3 + 14 \times 3 = 72$,
 578 and the total capacity is $24 \times 3 = 72$, therefore, the clinic or system workload is
 579 100%. A factor varying from 0.6 to 1.4 will be multiplied to the mean demand rate
 580 to generate different levels of system workload.

581 Following the findings of Bennett and Baxley (2009), we assume a typical no
 582 show rate for pre-scheduled demand of 25%, and a 10% no show rate for open
 583 access demand. Thus, we assign the revenue of scheduling one pre-scheduled
 584 demand as 0.75, and 0.9 for seeing one open access patient. These values stand for
 585 the actual show rates. To encourage continuity in the system, we assume that there
 586 is a 0.05 cost of seeing patients from another physician's panel (the revenue of
 587 giving an appointment slot to one open access patient not from a physician's panel
 588 is therefore $0.9 - 0.05 = 0.85$). While the no-show rates for the two types of
 589 appointments can be estimated from past data, the cost of diverting an open access
 590 patient to a non-PCP physician is very difficult to quantify. Finally, the stochastic
 591 integer program uses a MIP tolerance gap of 0.01% Table 1 summarizes the
 592 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%

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

606 6.1 Value of flexibility in symmetric 3 and 6 physician cases

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

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

618 Next, we present a graphical comparison (see Figs. 5, 6, 7 and 8) of the benefits
 619 of the 2-chain and full flexibility over the dedicated case in 3 and 6 physician cases
 620 for each of our measures (for the sake of brevity and ease of exposition we do not
 621 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%

622 We see that the highest benefit of both system revenue and timely access rate is
 623 achieved in the case where the system is balanced, i.e., when the expected demand
 624 equals the available capacity. The 2-chain and full flexibility cases are about 3–4%
 625 better than the dedicated case with regard to system revenue and timely access in the
 626 3-physician case; and about 5–6% better in the 6-physician cases. Also, the graph of
 627 system performance improvement is not symmetric since the variability of demand
 628 at 140% workload is larger than 60% workload, presenting more opportunities for
 629 redirections.

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

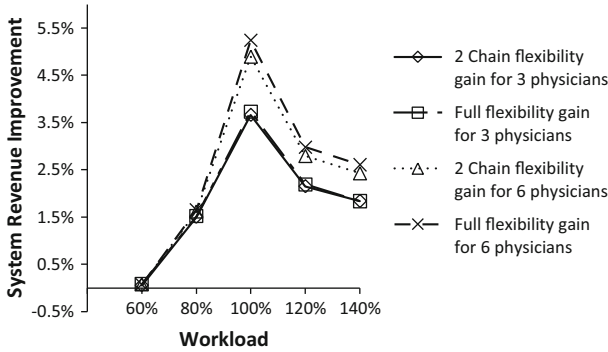


Fig. 5 Comparison of system revenue improvement between 3 and 6 physicians

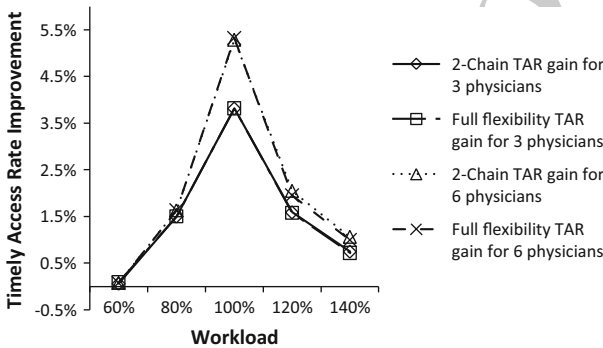


Fig. 6 Comparison of timely access rate (TAR) improvement between 3 and 6 physicians

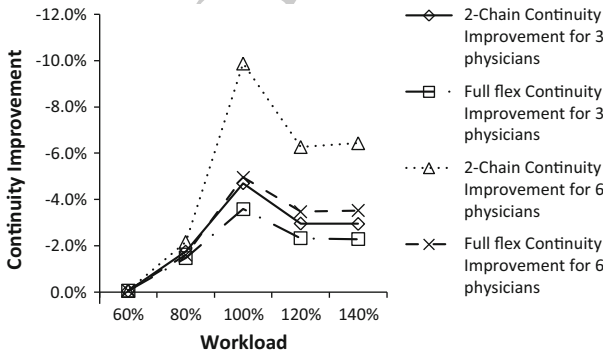


Fig. 7 Comparison of continuity improvement (over the dedicated case) between 3 and 6 physicians

635 settings. The difference is even lower in our healthcare setting, since we assume that
 636 prescheduled demand cannot be shared between physicians; flexibility can only be
 637 used for open access demand.

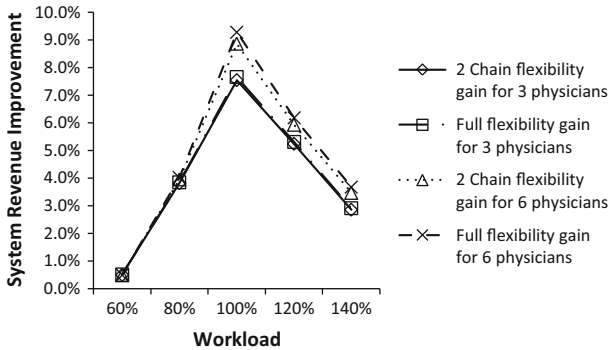


Fig. 8 Comparison of system revenue improvement between 3 and 6 physicians

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

647 6.2 Value of flexibility in asymmetric 3 and 6 physician cases

648 The asymmetric case represents the situation where physicians have varying panel
 649 sizes and therefore varying appointment burdens. This is common in practice.
 650 Senior and well established physicians may have higher workloads since their
 651 panels are larger, while physicians who have been recently hired may have lower
 652 workloads. The benefits of flexibility are likely to be greater in the presence of such
 653 asymmetry. To test this, we create the following demand profiles for the 3 physician
 654 case. Physician 1 has an expected prescheduled demand of 6 and an expected open
 655 access demand of 12 (low workload); Physician 2 has an expected prescheduled
 656 demand of 8 and an expected open access demand of 16 (balanced or full workload);
 657 Physician 3 has an expected prescheduled demand of 10 and an expected open
 658 access demand of 20 (high workload). Each physician still has 24 slots available in
 659 the day. Notice that although the individual physician workloads vary, the overall
 660 clinic or system workload is 100%. We use the same factors 0.6, 0.8, 1.2 and 1.4 to
 661 create varying levels of system workload, while still maintaining the imbalances
 662 between the physicians.

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

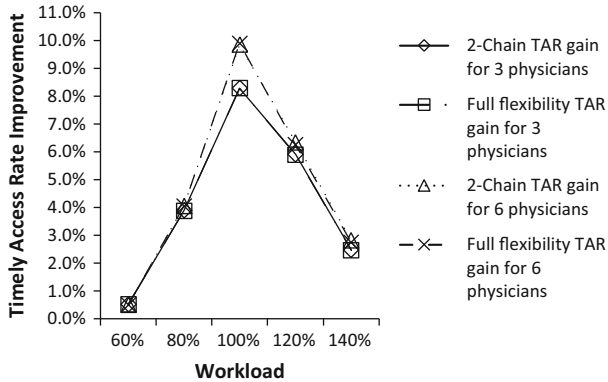


Fig. 9 Comparison of timely access rate (TAR) improvement between 3 and 6 physicians

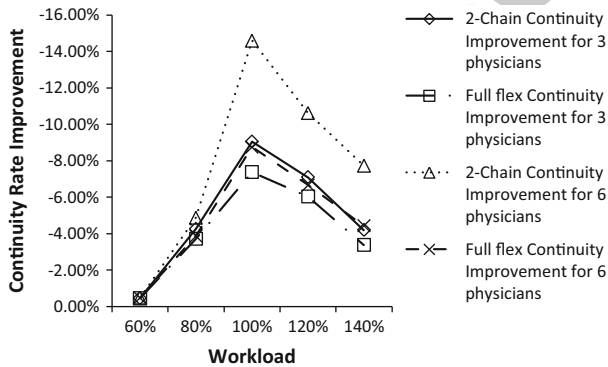


Fig. 10 Comparison of continuity rate (CR) improvement between 3 and 6 physicians

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

677 6.3 Value of flexibility under different prescheduled and open access demand
 678 ratios

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

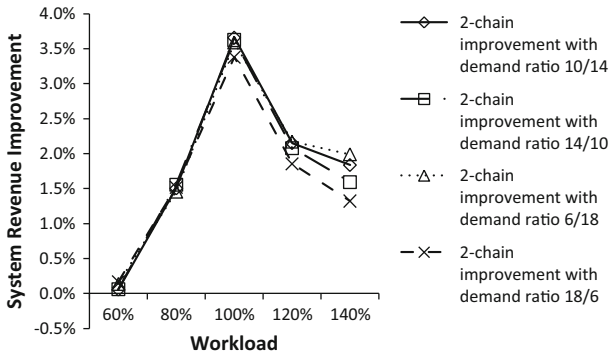


Fig. 11 2-chain flexibility improvement over the dedicated case under different ratios of prescheduled to open access demand for all symmetric cases

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

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

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

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

712 6.4 N_i^{p*} values and system workload

713 The results above discuss the value of flexibility. But what about the N_i^{p*} values, the
 714 first stage decisions in our model? What trends do they follow, if at all, and can the
 715 trends provide clues to capacity allocation decisions in practice? To understand this,
 716 we analyzed the total N_i^{p*} values of the clinic (that is the total capacity set aside by
 717 all physicians for prescheduled appointments), averaged over the 50 replications.
 718 Figure 12 shows the average N_i^{p*} values for the entire clinic (that is for all the
 719 physicians) under different workloads and the three flexibility configurations for the
 720 6 physician asymmetric case. We see the same trends by looking at the individual
 721 physicians' N_i^{p*} values (irrespective of the number of physicians, symmetry and
 722 prescheduled to open access demand ratios). Thus the figure summarizes our
 723 conclusions about N_i^{p*} values concisely.

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

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

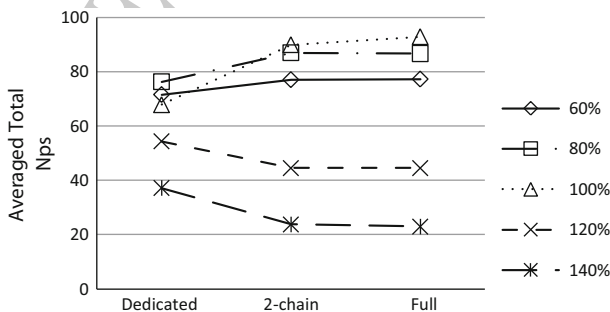


Fig. 12 Trends in averaged total Np values for 6 physicians with Prescheduled demands [6,8,10,6,8,10] and open access demands [12,16,20,12,16,20]

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

746 7 Conclusions and future research

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

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

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

775 Several future research directions are possible. The models of flexibility in the
 776 paper emerged as a result of interactions that the authors had with clinics in the
 777 United States. These included primary care clinics in academic medical centers (in
 778 both rural and urban setting) as well as smaller family medicine clinics. However, a
 779 more formal empirical study of different primary care clinic types (urgent care,
 780 pediatric practice, adult primary care) that throws light on how flexibility is actually
 781 used in practice and how patients and physicians perceive continuity and timely
 782 access is necessary to validate our conclusions. For example, while full flexibility
 783 may be the default mode for many clinics, larger practices implement a different
 784 kind of partial flexibility, which we call *subgroups*. In the subgroup configuration,

785 the physicians are divided into a smaller number of self-contained but fully flexible
 786 groups. A practice consisting of four physicians might be divided into two groups of
 787 two, functioning independently. Our preliminary comparisons of 2-chain and
 788 subgroups have revealed that they perform very close to each other, with the 2-chain
 789 having a consistent but small advantage. However, subgroups are attractive since
 790 they are easier to implement in practice.

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

796 **Acknowledgments** This work was funded in part by the grant CMMI 1031550 from the National
 797 Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this
 798 material are those of the authors and do not necessarily reflect the views of the National Science
 799 Foundation. We would also like to acknowledge two anonymous referees whose valuable comments
 800 helped improve the paper. Finally, our sincere thanks to Xiaoling Gao, doctoral student at the University
 801 of Massachusetts, who helped with editing of the results.

802 **References**

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

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

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

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

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

812 Gill JM, Mainous A (1999) The role of provider continuity in preventing hospitalizations. *Arch Fam Med*
 813 7:352–357

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

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

817 Green LV, Savin S (2008) Reducing delays for medical appointments: a queueing approach. *Oper Res*
 818 56(6):1526–1538

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

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

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

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

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

829 Jordan WC, Graves SC (1995) Principles and benefits of manufacturing process flexibility. *Manage Sci*
 830 41(4):577–594

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

- 834 Liu N, Ziya S, Kulkarni V (2010) Dynamic scheduling of outpatient appointments under patient no-shows
835 and cancellations. *Manuf Serv Oper Manage* 12.2:347–365
- 836 Muriel A, Somasundaram A, Zhang Y (2006a) Impact of partial manufacturing flexibility on production
837 variability. *Manuf Serv Oper Manage* 8(2):192–205
- 838 Muriel A, Zhang Y, Biller S (2006b) Impact of price postponement on capacity and flexibility investment
839 decisions. *Prod Oper Manage* 15(2):198–214
- 840 Murray M, Berwick DM (2003) Advanced access: reducing waiting and delays in primary care. *J Am*
841 *Med Assoc* 289(8):1035–1040
- 842 Murray M, Bodenheimer T, Rittenhouse D, Grumbach K (2003) Improving timely access to primary care:
843 case studies of the advanced access model. *J Am Med Assoc* 289(3):1042–1046
- 844 O'Malley AS, Cunningham PJ (2008) Patient experiences with coordination of care: the benefit of
845 continuity and primary care physician as referral resource. *J Gen Intern Med* 24(2):170–177.
- 846 Qu X, Rardin R, Williams JAS, Willis D (2007) Matching daily healthcare provider capacity to demand
847 in advanced access scheduling systems. *Eur J Oper Res* 183(2):812–826
- 848 Robinson L, Chen R (2010) A comparison of traditional and open access policies for appointment
849 scheduling. *Manuf Serv Oper Manage* 12.2:330–347
- 850 Rust G, Ye J, Baltrus P, Daniels E, Adesunloye B, Fryer GE (2008) Practical barriers to timely primary
851 care access. *Arch Intern Med* 268(15):1705–1710
- 852 Solak S, Clarke J-P, Johnson E, Barnes E (2010) Optimization of R&D portfolios under endogenous
853 uncertainty. *Eur J Oper Res* 207(1):420–433
- 854

855 Author Biographies

856
857 **Dr. Hari Balasubramanian** is an assistant professor of Industrial Engineering at the University of
858 Massachusetts, Amherst. His research interests are in operations research applied to healthcare delivery.
859 Specific application areas include capacity planning and scheduling in primary care, surgical suites and
860 emergency departments. Dr. Balasubramanian has a PhD in Industrial Engineering from Arizona State
861 University. He was a Research Associate at the Department of Health Sciences Research from August
862 2006–2008 at the Mayo Clinic in Rochester, Minnesota.

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

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

874