2013

The Impact of Case-Mix on Timely Access to Appointments in a Primary Care Group Practice

Asli Ozen, *University of Massachusetts - Amherst*
Hari Balasubramanian, *University of Massachusetts - Amherst*

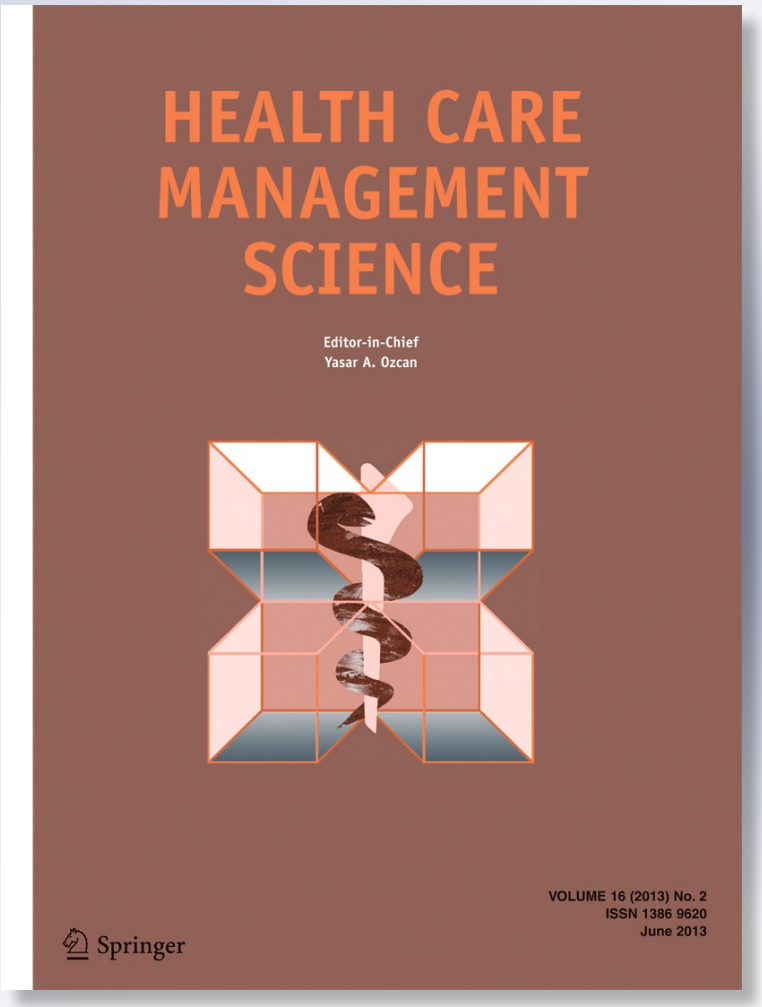
The impact of case mix on timely access to appointments in a primary care group practice

Asli Ozen & Hari Balasubramanian

Health Care Management Science

ISSN 1386-9620
Volume 16
Number 2

Health Care Manag Sci (2013)
16:101-118
DOI 10.1007/s10729-012-9214-y



Springer

Health Care Manag Sci (2013) 16:101–118
DOI 10.1007/s10729-012-9214-y

The impact of case mix on timely access to appointments in a primary care group practice

Asli Ozen · Hari Balasubramanian

Received: 20 May 2012 / Accepted: 13 September 2012 / Published online: 18 October 2012

Abstract At the heart of the practice of primary care is the concept of a physician panel. A panel refers to the set of patients for whose long term, holistic care the physician is responsible. A physician's appointment burden is determined by the size and composition of the panel. Size refers to the number of patients in the panel while composition refers to the case-mix, or the type of patients (older versus younger, healthy versus chronic patients), in the panel. In this paper, we quantify the impact of the size and case-mix on the ability of a multi-provider practice to provide adequate access to its empanelled patients. We use overflow frequency, or the probability that the demand exceeds the capacity, as a measure of access. We formulate problem of minimizing the maximum overflow for a multi-physician practice as a non-linear integer programming problem and establish structural insights that enable us to create simple yet near optimal heuristic strategies to change panels. This optimization framework helps a practice: (1) quantify the imbalances across physicians due to the variation in case mix and panel size, and the resulting effect on access; and (2) determine how panels can be altered in the least disruptive way to improve access. We illustrate our methodology using four test practices created using patient level data from the primary care practice at Mayo Clinic, Rochester, Minnesota. An important advantage of our approach is that it can be implemented in an Excel Spreadsheet and used for aggregate level planning and panel management decisions.

Keywords Panel size · Primary care · Continuity of care · Appointment scheduling

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

A. Ozen
e-mail: ozen@ecs.umass.edu

1 Introduction

Primary care providers (PCPs) are typically the first point of contact between patients and health systems. They include family physicians, general internists, and pediatricians. A primary care physician's *panel* refers to the patients for whose long term care she is responsible. Over time, the PCP becomes familiar with the patients in her panel and is therefore able to deliver more informed and holistic care, with a focus on prevention. This long-term patient-physician relationship, also termed as *continuity of care* is one of the hallmarks of primary care.

The benefits of continuity for both patients and physicians have been well documented in the clinical literature. Gill and Mainous [7] 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 their problems correctly identified by their physician; and (4) less likely to be hospitalized. Continuity and coordination are especially important for vulnerable patients with a complex medical history and mix of medications [19].

In practice continuity translates to maximizing patient-PCP matches when appointments are scheduled. But the ability of a PCP to provide continuity and timely access depends on (1) panel size, or the number of patients in her panel; and (2) case-mix, or the type of patients in the panel. For example, a panel consisting of mostly healthy patients will have a very different appointment burden compared to a panel consisting mostly of patients with chronic conditions.

In this paper, we characterize the interrelationship between panel size, case-mix and the individual capacities of physicians working in a group practice. This is done by measuring the *overflow frequency* of the physicians in relation to each other. The overflow frequency is the probability that the demand from a physician panel (i.e. patient requests for appointments in a day) will exceed the physician's capacity (i.e. the number of appointment slots a physician has available in a day). A high overflow frequency for a physician implies that patients in the panel will be unable to access their physician in a timely manner and are as a result more likely to visit an unfamiliar physician or emergency room. Thus a high overflow frequency implies that both timely access and continuity of care are compromised.

The consideration of panel size and case-mix in this paper, is particularly relevant given the acute shortage of PCPs in the United States. The demand for primary care continues to grow as the population ages and the prevalence of chronic conditions increases. Our approach allows practices to quantify their current supply and demand imbalances and use available capacity in the most efficient manner possible. Case-mix is an important consideration given that patient demographics and care needs vary from community to community and from one geographic region to another.

The analysis presented in this paper is at the aggregate planning level, where a practice has to decide how many and what type of patients are appropriate in each panel to ensure patients have adequate levels of access and continuity. In the long term, if imbalances in workload exist among the physicians, a practice may be interested in *redesigning* panels—that is in changing the size and case-mix of individual physician panels so that each physician's capacity is in balance with her demand. While this involves changing existing panel configurations, opportunities for redesign arise constantly in primary care (more details in Section 5). For example, new patients may join the practice, existing patients may move from the area, and patient preferences about their PCP may change over time. On the capacity side, a physician may leave the practice or retire, with the result that patients in that physician's panel now need to be reassigned. In residency practices found in academic medical centers, the turnover of residents every year provides constant opportunities for panel redesign. We discuss the feasibility of panel redesign in greater detail in Section 5.

We propose an integer non-linear programming formulation for redesigning panels in a group practice. The goal is to minimize the maximum overflow frequency over all physicians. Rather than prescribe exactly what practices should do, we derive analytical results to benchmark a practice's current performance. Then the analytical results are used to motivate heuristics, which will allow practice managers to: (1) test various redesign options and, (2) infer which options are the least disruptive. An important advantage of our approach is that all our analytical results can be implemented in Excel and used for aggregate level planning and panel management decisions.

The rest of the paper is organized as follows. In Section 2, the relevant literature is reviewed and in Section 3, we explain the modeling of case-mix. We motivate the panel redesign problem using an example involving 4 physicians in Section 4. The feasibility of panel redesign in practice is discussed in Section 5. Section 6 contains all the mathematical details and analytical results related to the panel redesign formulation. In Section 7, the heuristics are described. In Section 8, we explain how we used patient and panel data from the Primary Care Internal Medicine (PCIM) practice in Rochester, Minnesota to create four test practices to demonstrate the results. Section 9 summarizes the conclusions and explains the implications of our results for practices.

2 Literature review

Appointment scheduling in healthcare is an active and growing area of research. Over the last decade, the advanced access paradigm, made popular in clinical journals by Murray and Berwick [14], Murray and Tantau [15], and Murray et al. [16], attempted to promote same-day access for patients. In traditional appointment systems, appointments are allowed to be booked into the future, whereas in advanced access this is discouraged. All appointments, regardless of their nature and urgency of request, are to be seen the same day by the patient's PCP. In practice, most clinics follow a blend of traditional and advanced access scheduling. Clinical necessities (follow-ups for chronic conditions) and patient preferences require practices to allow the future booking of appointments, while at the same time enable same-day access for acute needs. Yet, whatever appointment system or blend a practice may follow,

effective access is possible only if the panel sizes of the physicians and their case-mixes are in balance with the available capacity, and the impact of variability is adequately addressed.

The operations research literature has in the last decade tackled a number of aspects related to appointment scheduling using stochastic optimization approaches. This includes an analytical comparison of traditional and advanced access appointment systems [23]; the impact of no-shows [5, 11, 12, 17]; the importance of considering patient preferences [10, 26]; and capacity allocation methods that allow practices to offer a blend of prescheduled (non-urgent) and same-day (urgent) appointments [2, 20].

We reiterate that the analysis presented in this paper is at the aggregate level. Thus we only focus this review on the papers most relevant to our work on panel size and case-mix. Murray and Berwick [14] proposed six steps for clinics to implement advanced access. An important message of this work is that the primary lever for demand is the number of patients in a physician's panel. Murray et al. [16] provide a simple algorithm to calculate the "right" panel size for physicians. They also mention other factors that might affect the workload of physicians like gender and age (panel case-mix) but do not provide any quantitative analysis. While the paper provides clinics with easily implementable policies to realize advanced access by resizing panels, there is no discussion on the impact of variability, an important factor in appointment scheduling.

Green and Savin [8] use queuing models and simulation to demonstrate the impact of panel size on the no-show rate, physician utilization, and the probability of getting a same-day appointment. They find that the backlog of appointments grows with panel size and as a result the no-show rate does as well, since patients booked well into the future will have a greater probability of no-show.

In Green et al. [9], a newsvendor like model is proposed to determine the relationship between the size of a physician's panel and the overflow frequency. Overflow frequency, as stated in Section 1, is the probability that the demand will exceed the available physician capacity. They assume that each patient in the panel has a probability p of requesting an appointment on any given day. This probability can be estimated from historical visit rates. Since each patient requests independently of each other, the demand for a panel of patients is a binomial random variable. Based on what the capacity of a physician is, the probability of overflow can then be easily calculated using the CDF of the binomial distribution.

The approach we take is closest to the modeling framework of Green et al. [9]. Their newsvendor like approach are extended to include case-mix and also to establish the interrelationship between multiple physicians working in a group practice. We first extend the binomial framework for modeling demand to consider different classes of patients. In our model, case-mix is represented by the number of simultaneous chronic conditions a patient has (more details in Section 3). Next, the overflow frequency is used as a measure of access, and then theoretical results are developed that will allow a group practice to benchmark their current performance. Finally we develop simple heuristics that will allow practices to test long-term panel redesign scenarios. The results are demonstrated using panel data from the primary care internal medicine (PCIM) practice at Mayo Clinic.

3 Patient classification

Patients can be characterized by various attributes, such as age and gender and the chronic conditions afflicting the patient. Our interest is in attributes that play an important role in determining the distribution of visits. In addition to operational and capacity planning reasons, patient classification can be useful for clinics because they enhance a practice's understanding of its population and disease trends, and allow it to design its care models effectively. Barbara Starfield's seminal work about ACGs (Ambulatory Care Groups) argued that understanding the role of patients' clinical complexity in care utilization forms the cornerstone for effective resource planning and determining payment methods in healthcare [25].

What classifications are the most effective in predicting appointment request rates? Age and gender is the simplest patient classification in absence of other data, yet is generally effective [1, 16]. In this paper, the number of simultaneous chronic conditions a patient has is used as a predictor of the number of visits. In clinical parlance, these conditions are *comorbidities*. Our choice is based on the following reasons. First, comorbidity counts have clinical relevance and are widely accepted by the primary care practices we have interacted with. Focusing on all comorbidities of a patient is more holistic than focusing in isolation on specific chronic conditions, and primary care was conceived to be a holistic approach rather than a disease specific approach. Secondly, our categorization has been used both in literature and practice. Naessens et al. [18] show that the number of simultaneous chronic conditions is a strong predictor of the number of office visits.

Table 1 Mean and standard deviation of visits in 2006, for patients with different counts of comorbidities

# of comorbidities	# of patients	Avg. visits/ pat/year	Std. dev.
0	6,524	1.72	2.88
1	6,980	2.74	4.56
2	5,819	3.82	6.25
3	4,179	5.16	8.56
4	2,370	6.82	9.95
5	989	7.67	10.72
6	346	9.62	13.14
7	84	11.17	13.39

Comorbidity counts have also been used in the new payment scheme for primary care proposed by the Minnesota Department of Health [13]. Finally, statistical analysis of the patient level data from Mayo Clinic (using classification and regression trees, CART) revealed the count of comorbidities as the strongest predictor of appointment request rates.

We note, however, that the models proposed in this paper can be applied to any patient classification. While patient classification is important, the central theme of this paper is not to find the "best" classification. Rather, it is to show the impact of patient classes on access measures. To illustrate the impact of comorbidity counts, we analyzed the patient population (around 27,000 patients) empanelled at the Primary Care Internal Medicine Practice (PCIM) at the Mayo Clinic in Rochester, Minnesota. Examples of commonly observed chronic conditions in patients included hypertension, depression, diabetes, osteoporosis, urinary tract infections, hyperlipidemia, coronary artery disease and otitis. We divided patients based on the number of comorbidities they had. In all there were 8 patient categories as patients with more than 7 comorbidities was extremely rare.

Table 1 above summarizes the number of patients, average number of visits and standard deviation for each comorbidity count category, based on historical visits in PCIM. Clearly, not only does the mean number of visits increase with the number of comorbidities, the standard deviation does as well. The standard deviations are higher than the means, suggesting significant variation in visit rates within each comorbidity count category.

4 Example of 4 physicians

In this section, we demonstrate the impact of case-mix using a simple simulation. In the general case, there are $j = 1, ..., J$ physicians in the practice. Suppose all patients empanelled in a practice have been categorized into $i = 1, ..., M$ patient classes. A patient of category i has a probability p_i of requesting an appointment on a given day. This probability will be higher for patients with multiple chronic conditions than for relatively healthy patients (see Section 8 for the exact values and how these probabilities are calculated). Next, suppose n_{ij} denotes the number of class i patients in physician j's panel. The total demand for the physician is the sum of the demand from each patient class. The demand from each patient class is a binomial random variable—with n_{ij} patients in patient class i and probability of class i patient requesting on a given day being p_i.

The p_i and n_{ij} values are used to generate binomial data realizations using random sampling and thereby simulate the total demand for each physician. If we know the total daily appointment slots a physician has available in a day, then the simulation can be used to calculate the utilization, overflow frequency, and the expected overflow for each physician. Utilization is simply the expected total demand divided by the total daily slots a physician has available in a day. Overflow frequency is the fraction of total realizations (each realization can be thought of as a day) in which the patients' visit requests for the day exceed the available capacity of the physician. Expected overflow is the average patient backlog (unfulfilled demand) at the end of each day.

As an example, consider the results of the simulation for four PCIM physicians at Mayo Clinic. The physicians have approximately the same panel size (around 1,060 patients), but different case-mixes: different patient numbers in the 8 comorbidity count categories. The panel compositions of each physician are shown in Table 2, as are the overflow frequency, expected overflow and utilization. All four physicians have a capacity of 17 slots. We use 10,000 realizations.

Notice that Physician 3 and Physician 1 have relatively high utilizations, overflow frequencies and expected overflows. This is because they have more patients with two or more comorbidities in their panels, and these patient groups generate a higher number of visits. High overflows result in (1) patients seeing an unfamiliar physician or visiting an emergency room (loss of continuity), or (2) longer wait times to secure an appointment (lack of timely access).

These results suggest that, in addition to using panel size, clinics may benefit by making capacity and allocation decisions based on case-mix. In the face of high overflows, physicians generally work longer hours. But this is not an appealing option, especially in primary care where reimbursements are low, and where more and more physicians are experiencing emotional exhaustion because of the number of patients they have to

Table 2 Case-mix, panel size and performance measures for 4 physicians, each with a capacity of 17 appointment slots per day, where PS: panel size; OF: overflow frequency in %; EO: expected overflow; Util: utilization in %

Phy	0	1	2	3	4	5	6	7	PS	OF	EO	Util
P1	260	249	226	161	108	42	14	3	1,063	30	3.64	92
P2	299	293	212	147	77	26	6	1	1,062	22	0.94	87
P3	214	253	223	177	115	44	21	5	1,053	35	7.36	95
P4	290	296	218	145	84	27	12	5	1,077	18	1.48	83

see [4]. The long-term option for practices is to *redesign panels*. This means changing case-mix proportions by reassigning patients across panels so that each physician's demand is in balance with her capacity.

5 Feasibility of panel redesign

Before describing the panel redesign formulation, it is important to discuss how feasible or useful such a framework is to practices, individual physicians and patients. Redesigning panels implies changing existing patient-physician relationships, and there appears to be a paradox. To improve timely access and continuity in the long run a practice has to invest in the short term disruption of existing-patient relationships. It is natural therefore to ask: how realistic is redesign in practice?

The feasibility of redesign would be a very valid concern if each patient in the panel was very loyal to the physician and had spent many years visiting the physician. Enforcing a break in that relationship would not be satisfactory to either the patient or the physician. But in practice, a panel is a lot more fluid. While there exist many patients who have spent years with the physician (we do not recommend that these relationships be disrupted), there also exist patients who are newly registered or are as yet uncommitted to their physician even though they have been assigned to a panel. It is these patients who would be amenable to redesign.

For example, in order to improve access to care, continuity and care coordination, Group Health practice of Seattle recently reduced panel sizes from 2,300 per physician to 1,800 per physician [21]. They hired new physicians and reassigned 500 patients per physician to either new physician or physicians who had available capacity. Patients were invited to an open house to meet their new physicians and surveys were used to identify patients who were willing to change their PCP.

In their papers, Reid et al. [22] and Coleman et al. [6] analyze the Group Health Clinic after the implementation. They used survey-based measures to quantify patient satisfaction and staff burnout. The results of the implementation were: (1) Staff burnout decreases since they find that emotional exhaustion becomes less frequent for physicians; (2) Patients' experience improves in terms of access to care and doctor-patient interactions (and this manifests itself in 29 % fewer emergency department (ED) visits and 11 % fewer hospitalizations); (3) During the reassignment, when physicians are given the chance to choose patients to keep in their panel, they prefer the elderly and sicker patients, who create a greater density of visits and need more continuity; and (4) Reassigned patients use primary care less, but there is no significant increase in their use of the ED.

While Group Health seems to have successfully achieved its redesign to improve patient centeredness, access and continuity, their reassignment of patients does not seem to have followed a quantitative basis. For example, how did the practice decide that 500 patients per physician (more than 20 % of the original panel size of 2,300) had to be reassigned? Could fewer patients have been reassigned or do panel sizes need to be even smaller? Quantitatively capturing the beneficial effects of redesign and the impact on the number of patients affected—which is the focus of this paper—will help individual physicians and the practice as a whole to make the choices that are most appropriate for them.

Indeed our experimental results (see Section 8.2) based on the primary care internal medicine practice (PCIM) at Mayo Clinic suggest that panel redesign will affect at most 5–8 % of the total patients (250 patients out of 4,300 total) in the practice. Furthermore, the number of patients affected can be as low as 2 % (less than 100 out of 4,300 total). So the very large majority of patient physician relationships will remain unaffected. Yet, the improvements in overflow frequency due to redesign are significant for the overburdened physicians in the PCIM practice. There is thus a strong incentive for overburdened physicians to consider redesign, since it improves access measures for their patients.

Furthermore, as Balasubramanian et al. [1] argue, redesign does not need to be carried out instantly as in the Group Health case, but can be achieved by most practices in the long term. Every practice has a natural attrition rate as well as a group of new patients wanting to join the practice. Patients' comorbidities can change

over time as well. Retiring physicians will need to transition their patients to newly hired physicians. These rates could be used, over time (a period of 1–2 years or perhaps more) to adjust case-mixes so that timely access and continuity are improved. Indeed we view the framework of this paper not as a strict prescription that dictates what practices should do. Rather it is an assessment tool, which practices can use to benchmark their current access and continuity levels on a quarterly or yearly basis and use whatever leverage they have to change panels.

6 The Panel Redesign Formulation (PRF) and analytical results

In this section, a mathematical formulation is provided to redesign physician panels in a multi-physician practice to minimize the maximum overflow frequency. We choose overflow frequency since it is a more tractable non-linear objective function than the expected overflow. It also allows us to derive properties that eventually allow near optimal solutions to be reached using simple heuristics.

We choose a minimax objective function over a summation function because even if the sum of overflow frequencies over all physicians in the practice is minimum, some physicians may still have higher overflow frequencies in relation to others. This will eventually lead to redirections to unfamiliar physicians and hence a loss of continuity. The minimax function, on the other hand, will ensure to the extent possible that each physician's panel demand is in balance with her capacity. We will also see in this section that identical overflow frequencies for all physicians does not mean that physician panels have to be identical in their case-mix proportions.

As discussed in the Section 4, n_{ij} denotes the number of patients from patient class i in physician j's current panel. The n_{ij} values over all J physicians and all M patient classes together describe the current panel design. However, the practice would like to *redesign* panels, that is determine new allocations from each patient class i to each physician panel j to minimize the maximum overflow frequency. Let x_{ij} be the number of patients to be assigned from patient class i to physician j. The constraints are that x_{ij} values should be integer and that all patients from each class have to be allocated, $\sum_{j=1}^{J} x_{ij} = N_i, \forall i = 1, ..., M$. Here N_i is the total number of class i patients (or category i patients) in the practice.

As before, the probability that a patient of class i requests for an appointment on any given day is p_i. If we assume that patients request independently of each other then the total demand for physician panel j from patient class i after reassignment is a binomial random variable with mean $x_{ij}p_i$ and variance $x_{ij}p_i(1-p_i)$. If we take the sum over all M patient classes, the mean and standard deviation of the total demand arising from physician j's panel are given by: $\mu_j = \sum_{i=1}^{M} p_i x_{ij}$ and $\sigma_j = \sqrt{\sum_{i=1}^{M} p_i(1-p_i)x_{ij}}$, respectively, $0 < p_i < 1$. Note that both the mean and standard deviation depend on the case-mix distribution given by the x_{ij} values for the physician. If panel sizes are sufficiently large (>800–1,000 patients), the total demand is the sum of as many Bernoulli random variables, and is likely to be well approximated by a normal distribution. We verified this statistically by applying the Kolmogorov–Smirnov (KS) goodness of fit test. The test was applied to total demand data generated using 10,000 random samples from the binomial demand distributions corresponding to the individual patient categories.

Let C_j denote the capacity of the physician, the total daily slots that she has available in a day. Then Z_j, the standard normal Z-score for physician j, is given by: $Z_j = \frac{C_j - \mu_j}{\sigma_j}$. Intuitively, the Z-score gives the number of standard deviations that the capacity is distant from the mean of the panel demand. If the percentile of the standard normal distribution is denoted by Φ, then the probability of overflow for physician j, O_j, is $O_j = 1 - \Phi(Z_j)$. The greater the positive distance between C_j and μ_j and the smaller the σ_j, the greater the Z_j value and the lower the overflow frequency O_j.

The goal is to optimize x_{ij} allocations to minimize $\max\{O_1, O_2, ..., O_J\}$—that is minimize the maximum overflow frequency over all physicians in the practice. The formulation is summarized below. We call it the panel redesign formulation (PRF).

$$\text{(PRF)} \quad \min_{x_{ij}}\{\max\{O_1, O_2, ..., O_J\}\} \tag{1}$$

$$\text{s.t.} \quad O_j = 1 - \Phi\{\frac{C_j - \mu_j}{\sigma_j}\}, \forall j = 1, ..., J \tag{2}$$

$$\mu_j = \sum_{i=1}^{M} p_i x_{ij} \quad \forall j = 1, ..., J \tag{3}$$

$$\sigma_j = \sqrt{\sum_{i=1}^{M} p_i(1-p_i)x_{ij}} \quad \forall j = 1, ..., J \tag{4}$$

$$\sum_{j=1}^{J} x_{ij} = N_i \quad \forall i = 1, ..., M \tag{5}$$

$$x_{ij} \geq 0 \text{ and integer } \forall (i, j) \tag{6}$$

Note that PRF is an integer non-linear program. The formulation is described visually in Fig. 1. The total mean and variance of the entire patient population given by $\mu_{\text{total}} = \sum_{i=1}^{M} p_i N_i$ and $\sigma^2_{\text{total}} = \sum_{i=1}^{M} p_i(1 - p_i)N_i$. The allocation problem is all about optimally partitioning the total population mean, μ_{total}, and variance, σ^2_{total} to individual physicians in the practice. The lever through which the partitioning is achieved are the x_{ij} values. The means and variances are not allocated independently of each other but are tied to the x_{ij} allocations. In other words, O_j, μ_j and σ_j will all increase (decrease) together when x_{ij} increases (decreases) for any $i = 1 \ldots M$.

Clearly, the maximum overflow will always be minimized if all the Z-scores and corresponding overflows can be made equal. Even if they cannot be made exactly equal, the differences in the overflows will be small enough to be negligible for large panel sizes. In other words, there is sufficient granularity in large panels (>800 patients) to smooth the overflows in the practice.

Consider, first, the equal capacity case, $C_1 = C_2 = C_3 \ldots = C_J$, which is relatively easy to understand. Since the physicians are all identical, then any allocation in which $\mu_1 = \mu_2 = \ldots = \mu_J = \mu_{\text{total}}/J$ and $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_J^2 = \sigma^2_{\text{total}}/J$ will minimize the maximum overflow frequency. Note the above statement refers to a *set of allocations*, not a particular one—the optimal overflow can be reached in multiple ways.

A special case is the allocation where each physician j gets the same number of patients from each category i. Mathematically, $x_{ij} = N_i/J, \forall i, j$. To maintain integrality of the decision variables in such a symmetric allocation, the number of patients in each category i should be a multiple of the number of physicians J in the practice. Even if this condition does not hold true, the general idea is that all physicians have nearly identical panel compositions. A symmetric allocation has practical benefits. Primary care physicians are the generalists of healthcare. Their training allows them to treat a wide variety of patients, ailments and chronic conditions. The $x_{ij} = N_i/J$ allocation maximizes diversity of patients in physician panels. This is especially important for panels of primary care residents in academic medical centers. Patient and diagnostic diversity is an essential education and training objective of a resident. Similarly private practices with a large number of relatively new physicians might benefit from introducing diversity in panels.

Practices, however, do not have to follow such symmetric allocations. Panels tend to grow more organically over time. In the interest of not disturbing existing patient-physician relationships, a practice may choose other allocations that are asymmetric yet in a manner that the overflows turn out to be identical. Thus, although the structure of allocations in the equal capacity case is obvious, the subtle point is that there are multiple optimal solutions. We revisit this theme again in the heuristics and results section. One of our objectives there is try to redesign panels with the minimum possible disruption to existing panels.

We next consider the more general unequal capacity case: $C_1 \neq C_2 \neq C_3 \ldots \neq C_J$. In academic medical centers, where physicians have research responsibilities, the unequal capacity case is more prevalent. But even in non academic small practices, with 3 or 4 physicians on staff (where majority of primary care in the US is delivered), physicians will often have different schedules or may work only part time. Physicians on the path to retirement also may gradually reduce their work hours.

6.1 The unequal capacity case

When the physicians have different number of slots available every day, it would seem appropriate to

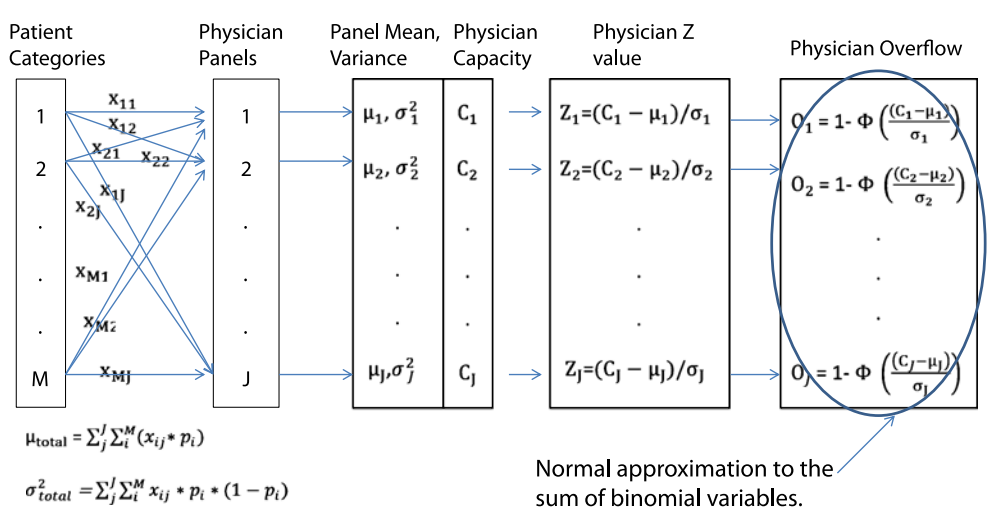


Fig. 1 A visual summary of the panel redesign problem to minimize the maximum overflow

allocate patients keeping in mind the capacity a physician has. Greater capacity would imply a greater share of μ_{total} and σ^2_{total}. However, the difficulty is in determining precisely *how much greater* that share should be for an optimal allocation. Let C be the total capacity of the clinic—total slots the clinic has available on a typical workday. Therefore $C = C_1 + C_2 + ... + C_J$. An allocation in proportion to the capacity is given by: $x_{ij} = (C_j/C) * N_i$ for all i and j. In other words, the number of patients from each category is proportioned in the ratio of an individual physician's capacity to the total clinic capacity. This seems an intuitive way of allocating patients and is an extension of the equal capacity case where each physician was assigned the same number of patients.

However, the allocation $x_{ij} = (C_j/C) * N_i$, while likely to be a good heuristic, is not guaranteed to give the optimal solution (specific examples in Section 8). This is because while the allocation of patients from each patient class increases linearly as the capacity increases, the objective function changes non-linearly. Indeed, a simple closed form expression for the optimal allocation, as described in the equal capacity case, may not be possible. It may be possible to solve PRF (at least numerically) by relaxing the integrality constraints on x_{ij}. However, rather than choosing this course, we approximate the optimal objective. This will give practices a reference or a target overflow frequency, O_{ref} to aim for when they redesign panels. We show that for all practical purposes O_{ref} is a good surrogate for the optimal overflow frequency O_{opt}. A practice can use O_{ref} to test various redesign options (multiple ways of reaching the optimal value), and choose whatever works best for them. This approach is less prescriptive than solving the non-linear program exactly to determine x_{ij} values. Furthermore, the calculation of O_{ref} can be achieved using an Excel spreadsheet, and therefore will be easy to implement in practice.

6.2 Deriving the reference overflow O_{ref}

Our method relies on relating overflow of individual physicians in the optimal allocation to the overflow of a hypothetical "combined physician". This combined physician (CP) is simply the aggregated system. In other words, the combined physician has a capacity of $C = C_1 + C_2 + ... + C_J$, a mean demand equal to μ_{total} and variance equal to σ^2_{total}. In such a practice, a physician can see the patients of any other physician—there is thus no concept of continuity. The standard normal value corresponding to the combined physician, Z_{CP} is given by:

$$Z_{\text{CP}} = \frac{C - \mu_{\text{total}}}{\sqrt{\sigma^2_{\text{total}}}} \tag{7}$$

Notice that the above expression can be easily obtained independently, without any knowledge of the x_{ij} values in the optimal allocation. We shall next try to relate the Z_{CP} value to the standard normal value Z_j for each physician j in an optimal allocation. Suppose μ_j, σ_j and Z_j represent the mean, standard deviation and Z value for physician j in an optimal allocation. For sufficiently large panel sizes, we know that the overflows of the physicians in an optimal allocation are approximately equal, which implies that the Z_j values will be approximately equal as well. So it is reasonable to write $Z_{\text{opt}} = Z_1 = Z_2 = Z_3 = ... = Z_J$. More precisely:

$$Z_{\text{opt}} = Z_j = \frac{C_j - \mu_j}{\sigma_j}, \forall j \tag{8}$$

$$\sigma_j * Z_{\text{opt}} = C_j - \mu_j, \forall j \tag{9}$$

If we add all the J equations, one for each physician, based on the equality above, we get:

$$\sum_{j=1}^{J} \sigma_j * Z_{\text{opt}} = \sum_{j=1}^{J} C_j - \sum_{j=1}^{J} \mu_j$$

$$Z_{\text{opt}} = \frac{\sum_{j=1}^{J} C_j - \sum_{j=1}^{J} \mu_j}{\sum_{j=1}^{J} \sigma_j} = \frac{C - \mu_{\text{total}}}{\sum_{j=1}^{J} \sigma_j} \tag{10}$$

From the expression for Z_{opt} and Z_{CP} (see Eqs. 7 and 8) we have the following result.

$$Z_{\text{opt}} = \frac{Z_{\text{CP}}}{R}, \text{where } R = \frac{\sum_{j=1}^{J} \sigma_j}{\sqrt{\sigma^2_{\text{total}}}} \tag{11}$$

Note that since $\sigma^2_{\text{total}} = \sum_{j=1}^{J} \sigma_j^2$, we can rewrite R as:

$$R = \frac{\sum_{j=1}^{J} \sigma_j}{\sqrt{\sum_{j=1}^{J} \sigma_j^2}} \tag{12}$$

Notice that $R \geq 1$. This is because the sum of J positive numbers (the numerator of R) is always greater than the square root of the sum of squares of the J numbers (denominator of R). This means that $Z_{\text{CP}} \geq Z_{\text{opt}}$. The equality is tight when $R = 1$ (i.e., the extreme case where one physician has all capacity and all demand, while all the others have none). We can also derive an

upper bound on R. The upper-bound $R = \sqrt{J}$ is realized when all the J numbers involved in the expression are equal, that is $\sigma_1 = \sigma_2 = ... = \sigma_J$. We define $Z_{\text{ref}} = \frac{Z_{\text{CP}}}{\sqrt{J}}$. If the capacities of the physicians are equal, then $Z_{\text{opt}} = Z_{\text{ref}}$ and if the capacities of the physicians are unequal, we have $\frac{Z_{\text{CP}}}{R} \geq \frac{Z_{\text{CP}}}{\sqrt{J}}$, which implies $Z_{\text{opt}} \geq Z_{\text{ref}}$.

Intuitively, R captures the decline in variability when demands and capacities are aggregated (the well known aggregation effect). The decline is highest when each physician has the same variance (and standard deviation). As physician panels become more and more unequal with regard to the variances allocated to them, R starts to approach 1 and Z_{CP} starts to approach Z_{opt}. Indeed, to calculate the optimal Z_{opt}, we do not need to know the exact standard deviation values of the individual physicians. But we need to know how the standard deviations of the J physicians stand in relation to each other—that relationship is captured by R. From the above analysis, the following key result is derived:

$$Z_{\text{CP}} \geq Z_{\text{opt}} \geq \frac{Z_{\text{CP}}}{\sqrt{J}} \tag{13}$$

The overflows corresponding to the Z-scores above are given by $O_{\text{CP}} = 1 - \Phi(Z_{\text{CP}})$, $O_{\text{opt}} = 1 - \Phi(Z_{\text{opt}})$ and $O_{\text{ref}} = 1 - \Phi(Z_{\text{ref}})$ respectively. The relationship between the overflows can be described as follows:

$$O_{\text{CP}} \leq O_{\text{opt}} \leq O_{\text{ref}} \tag{14}$$

O_{CP} can be interpreted as the overflow of a practice that has no concept of panels. Any physician in the practice can see any of the total patients in the practice. There is no continuity. Such sharing however has the benefit of capacity pooling and hence O_{CP} is the best overflow a practice can achieve—it is the lower bound. O_{opt} on the other hand is the overflow of each physician assuming that the physicians do not share their patients at all. This provides perfect continuity but the benefit of capacity pooling is lost. Practices usually lie between these two extremes. Thus the difference between O_{opt} and O_{CP} measures *the price of continuity*.

While there is no exact method of computing O_{opt}, $O_{\text{ref}} = 1 - \Phi(Z_{\text{ref}})$ is used as a surrogate for the optimal overflow. We will demonstrate that $O_{\text{ref}} - O_{\text{opt}}$ is fairly small for most cases found in practice. Indeed, for the equal capacity case $R = \sqrt{J}$, $Z_{\text{ref}} = Z_{\text{opt}}$ and therefore $O_{\text{ref}} = O_{\text{opt}}$: the reference value is exactly equal to the optimal value.

6.3 $O_{\text{ref}} - O_{\text{opt}}$ for common cases in practice

To characterize $O_{\text{ref}} - O_{\text{opt}}$ we must consider what values of R are reasonable in practice. Consider a 2-physician practice. When the physicians have identical capacities, we expect to see $\sigma_1 = \sigma_2$ in the optimal allocation and therefore $R = \sqrt{2} = 1.414$. The more unequal the physicians are with regard to their capacities, the more R starts to approach 1.

When the capacities of the two physicians are not equal, the optimal allocation is unknown. But the asymmetry in physician capacities can give us a hint of what the R value might be. Suppose one physician works full time and has 24 slots in a day (assuming an 8 hour day with 3 patients per hour, a typical workload for PCPs), while the other physician works only 6 slots in a day. This asymmetry in capacities is perhaps the limit of what might be observed in a practice—seeing 6 patients a day (about 2–3 h of work per day) is generally not common except in residency practices.

Although the optimal allocation of patients for the above case is not known, we can still state that the mean and variance allocated to the full time physician should be *roughly* four times that allocated to the quarter-time physician. This can be stated because, it is known that the mean and variance are tightly coupled through the x_{ij} values—they both increase and decrease together. So we have: $\mu_1 = 4\mu_2$ and $\sigma_1^2 = 4\sigma_2^2$. This gives us an R value of 1.34. So $R = 1.34$ represents (approximately) a fourfold variation in capacities for a 2-physician practice. R values smaller than this imply that one physician works a negligible amount of time daily in relation to the other. Capacities of 12 and 24 or 10 and 20 are more reasonable since some physicians may work full time while others may work only for half a day. For such cases $R \geq 1.34$. In general, all practical 2 physician cases are well represented by $1.34 \leq R \leq 1.414$.

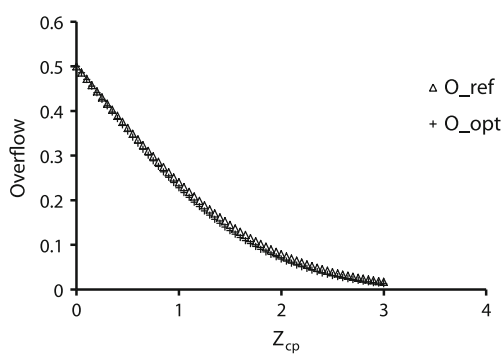
So a 2-physician practice which has $R = 1.34$ allows us to test the strength of our reference value O_{ref}. If O_{ref} approximates O_{opt} well for for this case, it will be even better for $R > 1.34$, which are more commonly observed.

As an example, suppose we find that $Z_{\text{CP}} = 1.0$ for a 2 physician practice with $R = 1.34$ (recall that Z_{CP} can be computed independently). If we don't know anything about the optimal allocation, our only option is to use the reference value, $Z_{\text{ref}} = \frac{Z_{\text{CP}}}{\sqrt{J}} = \frac{1.0}{1.414} = 0.707$. The optimal value, using $R = 1.34$ is $Z_{\text{opt}} = \frac{Z_{\text{CP}}}{R} = \frac{1.0}{1.34} = 0.746$. It follows that the reference overflow and optimal overflow are $O_{\text{ref}} = 1 - \Phi(Z_{\text{ref}}) = 1 - 0.772 =$

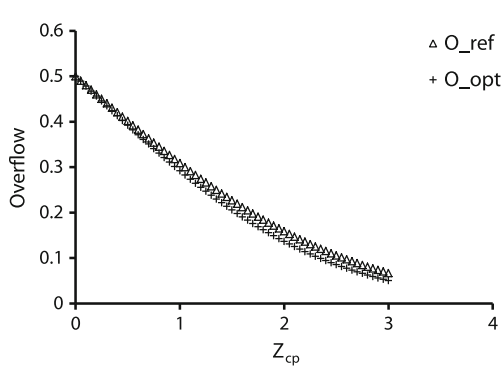
0.239 and $O_{\text{opt}} = 1 - \Phi(Z_{\text{opt}}) = 1 - 0.745 = 0.2227$ respectively. The difference is within 1 %.

Figure 2a below shows O_{ref} and O_{opt} as a function of Z_{CP}, which is varied from 0 to 3. The two lines are almost indistinguishable. At $Z_{\text{CP}} = 0$, when the aggregated demand equals the aggregated supply and the utilization is 100 %, both O_{ref} and O_{opt} are 0.5. The prediction is exact. As the overflow decreases, O_{ref} and O_{opt} differ from each other, with O_{ref} always being larger, but the difference never exceeds 1.3 %.

To further reinforce the point a 4 physician example is considered. Here we assume a sixteen-fold difference in capacities $C_1 = 4C_2 = 9C_3 = 16C_4$, which is an extreme limit on the capacity variation a practice is likely to have. Here the variance relationship will approximately be: $\sigma_1^2 = 4\sigma_2^2 = 9\sigma_3^2 = 16\sigma_4^2$. The R value for this setting is 1.825. We use $O_{\text{ref}} = 1 - \Phi(Z_{\text{ref}}) = 1 - \Phi(\frac{Z_{\text{CP}}}{sqrt(4)})$ as the reference value. If Z_{ref} works for well for this case, it will work even better for $1.825 < R \leq 2$. Figure 2b shows O_{ref} and O_{opt} as a function of Z_{CP} for the 4 physician example where $R = 1.825$. Here the difference between the two is slightly larger but O_{ref} is still within 2.5 % of O_{opt}. We have thus shown that O_{ref} is good surrogate for the optimal overflow O_{opt} for practical cases.



(a) 2-physician case



(b) 4-physician case

Fig. 2 Comparison of O_{ref} and O_{opt} as a function of Z_{CP} for the 2-physician and 4-physician example

6.4 Summary of contributions

In summary, the PRF formulation allows a practice to:

(1) Benchmark the access performance of each physician in the practice with other physicians as well as the reference overflow value.
(2) Capture the price of continuity (in terms of lost access). Specifically, the price of continuity for a practice is the difference between the reference or target overflow and the overflow of a practice in which all physicians together serve all the patients in the practice (no concept of a panel, but pooled capacity to meet the demand).
(3) Quantitatively evaluate and arrive at the least disruptive way of redesigning panels, since achieving the reference overflow is possible in many different ways (multiple optimal solutions). This allows a practice to quantify the minimum number of patients whose current PCP assignments will be affected if redesign were to be implemented.

Our heuristics and results, described in the next sections, quantitatively demonstrate each of these contributions and provides the foundation for a spreadsheet-based decision tool for aggregate level panel management decisions in a group practice.

7 Heuristics

In the last section, we have seen how a reference or target overflow can be determined for a group of physicians, and that this value is a good proxy for the optimal overflow for most practical scenarios. In this section, heuristics are described that practices can use to switch patients between panels so that this target overflow is achieved. Since switching patients disrupts existing patient-PCP relationships, a practice will be keen to (1) minimize the number of patients that are switched; (2) ensure that patients with the greatest continuity needs (for example a patient with multiple chronic conditions) are not switched. As it is demonstrated with our heuristics, these two goals can be conflicting.

Before explaining our heuristics, it is important to note that we assume that patient categories are ranked in non-decreasing order, based on their p_i values, which determines the visit rate of that patient category. In our classification method for instance, 0 comorbidity patients have the lowest visit rate, 1 comorbidity patients have the next lowest visit rate and so on.

To use the patient switching heuristics, practices start with an initial solution, for example the practice's current case mix or current panel design. Next, the

overflow value for each of the physicians is computed based on the initial solution. The physicians are ranked in decreasing order of their overflow values. A patient of the lowest visit category (the group with 0 comorbidities in our case) is then selected from the panel of the physician with the highest overflow and is now assigned to the panel of the physician with the lowest overflow. The overflow values for the two physicians are updated. If maximum overflow for the practice is greater than the reference overflow value (calculated as described in the previous section), another patient from the lowest visit category is transferred. If the physician with the highest overflow has no more patients in the lowest visit category, we move to the patient category with the next lowest visit rate and transfer a patient to the physician with the least overflow. This process of transferring patients is continued until the difference between maximum overflow of the practice and the reference overflow is small enough. We call this Heuristic 1, or H1.

Notice that in H1, we may have to shift a very large number of patients from low visit rate categories to achieve identical overflows in the practice. This may not be a bad strategy since relatively healthy patients have a lower chance of having formed a strong bond with the PCPs and are therefore more likely to change their PCPs.

In Heuristic 2, or H2, a different approach which involves all patient categories in the patient transfers is analyzed. As before we start with the current panel design and identify the physicians with the highest and lowest overflow values. We then transfer one patient from the patient category with the lowest visit rate to begin with, update the overflow values of the two physicians and again identify the physicians with the highest and lowest overflow values. If the current value of maximum overflow and the reference overflow is still large, we switch—in contrast to H1—a patient from the category with the next lowest visit rate. Thus we move from one category to the next, whereas in Heuristic 1, we tried to exhaust all possibilities in the lowest visit category. In Heuristic 2, patients are more evenly moved across the different categories, but more importantly *fewer* patients are moved in relation to Heuristic 1. The downside is that patients with chronic conditions who are more likely to have a strong relationship with their PCP will also be transferred in Heuristic 2.

While H1 and H2 lie at two ends of the spectrum, a practice manager can be more creative in his transfer choices. Patient and physician surveys as well as past visit patterns can be used to make more intelligent transfer choices that minimize disruption. In practice, patient reassignment is a dynamic process, which will be carried out over a period of time, as new patients are empanelled in the practice, when physicians leave or retire (thus leaving their panel to be reassigned among still working physicians). In addition, practices can use surveys to determine the willingness of patients to change their PCPs, thus creating a pool of patients who are amenable to changing their PCPs.

8 Case study

8.1 Data description

We use data from the Primary Care Internal Medicine (PCIM) practice at the Mayo Clinic in Rochester, MN. This practice empanels around 27,000 patients and employs 39 physicians. Many of these physicians worked part time. Panel data enabled us to identify which patients belonged to which physician. Patient level data included the number and type of chronic conditions afflicting each patient as well as the number of visits for each patient for 3 years (2004, 2005 and 2006). The list of chronic conditions included commonly occurring diseases such as hypertension, depression, diabetes, osteoporosis, urinary tract infections, hyperlipidemia, coronary artery disease and otitis. As discussed before, the number of comorbidities are used to come up with patient categories and this gives us 8 patient categories in all. To determine the p_i values for each comorbidity count, we first determine A_i, which is the total number of appointment visits for all patients with i comorbidities in the population for a long period of time, say a year. If N_i denotes all patients with i comorbidities, and if there are T workdays in a year, then:

$$p_i = \frac{A_i}{N_i * T}. \tag{15}$$

Assuming there are 250 workdays in a typical year, we are now able to calculate the per day request probability p_i for each patient category. The method is similar to the one proposed in [9]. The values are listed in the Table 3 below.

Table 3 Binomial p_i values for each patient category

p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7
0.006242	0.010571	0.014907	0.019914	0.025987	0.029823	0.03799	0.041167

It is also possible to calculate the p value for the entire population. If A is the total visits generated by the total population of N patients, then:

$$p = \frac{A}{N * T} = 0.0143 \tag{16}$$

This value will be used to set the capacity of physicians in the test practices created based on our data. The idea is to replicate the default process by which practices typically assign capacity—they recognize that capacity should increase with panel size, but generally do not consider case-mix in how they determine capacity. Thus, if a physician's panel size is L_j, then the physician's capacity, C_j, is assigned as follows:

$$C_j = \lceil (L_j * p + 0.1 * L_j * p) \rceil \tag{17}$$

The physician is given 10 % more slots than the mean demand $L_j * p$. Setting it equal to the mean—as many practices might, since they remain unaware of the impact of variance—would mean that each physician's utilization would be 100 %, leading to an unsustainable system. The 10 % additional slots ensure that there are a few extra slots to buffer variability in demand. Yet the utilization of the physician will still be sufficiently close to 100 %, as it is for most primary care physicians practicing in the US today. The above expression rounds up to the closest values, since the number of appointment slots per physician per day is typically an integer. We note that our approach can work with any other capacity inputs as well.

Our goal is not to obtain results specific to Mayo Clinic data. Rather it is to use the data to generate a series of "test" practices with 2 and 4 physicians, with different case-mixes to illustrate the impact of case-mix and our heuristics. The majority of practices in the US have 5 physicians or less, so our practice sizes are appropriate. Furthermore, larger practices tend to be divided into smaller self-contained subgroups to ensure continuity. We note, however, that our method is not computationally constrained in any way and can address larger practices as well.

8.2 Panel redesign for test practices

Tables 4, 5, 6 and 7 provide detailed results for our 4 test practices. The table format allows a reader to see the panels, case mixes and corresponding measures clearly. We consider the equal and unequal capacity case and under each we test a 2 physician case and a 4 physician case. In the first two test practices, the physicians have approximately the same panel sizes and hence the same capacity. In the next two, physicians have different panel sizes and hence have different capacities. The capacities are calculated as described above, based on panel size only. The physicians are numbered based on the original Mayo Clinic data (which had 39 physicians) to distinguish them from each other. We note that any combination of the 39 physicians from the data set can be considered in a similar way.

In the tables, we present panel case mixes before and after redesign, the corresponding means and variances for each panel, the overflow and the utilization for each physician. We also present panels designed based on the (1) Capacity Ratio (2) Heuristic 1 and (3) Heuristic 2. Note that the capacity ratio rule allocates patients from each category i to each physician j as follows: $x_{ij} = (C_j/C) * (N_i)$, where $C = \sum_{j=1}^{J} C_j$ is total capacity of the clinic. In the equal capacity case, when

Table 4 Results for Test Practice 1: 2 physicians with equal capacity

		Comorbidity count								Panel	μ_j	σ_j^2	C_j	O_j	Utilization
		0	1	2	3	4	5	6	7	size					
Current	Phy 4	332	360	324	270	144	40	20	5	1,495	21.99	21.58	24	0.33	0.92
	Phy 28	418	385	299	211	111	32	10	2	1,469	19.64	19.31	24	0.16	0.82
	# switched	0	0	0	0	0	0	0	0	0					
Capacity ratio	Phy 4	375	372	312	240	128	36	15	3	1,481	20.80	20.43	24	0.24	0.87
	Phy 28	375	373	311	241	127	36	15	4	1,482	20.83	20.46	24	0.24	0.87
	# switched	43	12	12	30	16	4	5	2	124					
Heuristic 1	Phy 4	144	360	324	270	144	40	20	5	1,307	20.81	20.42	24	0.24	0.87
	Phy 28	606	385	299	211	111	32	10	2	1,656	20.81	20.47	24	0.24	0.87
	# switched	188	0	0	0	0	0	0	0	188					
Heuristic 2	Phy 4	325	353	317	263	137	33	14	0	1,442	20.80	20.43	24	0.24	0.87
	Phy 28	425	392	306	218	118	39	16	7	1,521	20.83	20.46	24	0.24	0.87
	# switched	7	7	7	7	7	7	6	5	53					

O_{ref} for this practice is 0.24. The number of patients switched is provided as a separate row under each heuristic. The total patients switched by a particular heuristic appears under the panel size column

$C_1 = C_2 = ... = C_J$, the allocation reduces to $x_{ij} = (N_i/J)$, which gives the optimal solution (see Section 6). In the unequal capacity cases, $x_{ij} = (C_j/C) * N_i$ is a heuristic that is expected to perform well, but will not necessarily be optimal. For these cases, reference overflow values are used as the benchmark for comparisons.

In both Heuristic 1 and Heuristic 2, we start with the current panels or current case-mix and switch patients (as described in the previous section) until the required maximum overflow value is reached. In each heuristic (including the Capacity Ratio), we list the number of patients switched from each comorbidity group as well as the total number of patients switched.

In **Test Practice 1** shown in Table 4, while the two physicians have almost the same panel size and therefore the same capacity (24), differences in their case-mix result in significantly different overflow values. Physician 4 would therefore be unable to provide timely access and continuity to her patients. It is quite likely that the patients of Physician 4 that are unable to secure an appointment would end up seeing Physician 28. When the panels are redesigned, their overflow values can be made even. Physician 28's overflow and utilization increase as she receives some of Physician 4's patients.

The Capacity Ratio heuristic which is optimal for this practice evens the case-mix differences between the physicians and in the end results in similar panel sizes as before. However, in order for the two physicians to achieve the allocation suggested by Capacity Ratio, 124 patients need to be switched—this includes a number of high comorbidity patients. Heuristic 1 achieves identical overflows by starting with the original case mix and then transferring 0 comorbidity (healthy) patients from Physician 4 to Physician 28. As mentioned before, these patients are more likely to accept a PCP change. Notice that Heuristic 1 results in very different panel sizes as a result. Heuristic 2, on the other hand, switches patients evenly across categories but this does mean that higher comorbidity patients will be switched. The total patients switched however is only 53, about half of what Heuristic 1 requires. The panel sizes are different after Heuristic 2, but the difference is not as drastic as that produced by Heuristic 1.

For **Test Practice 2** (Table 5), all four physicians have a capacity of 17 and approximately the same panel size. These are the same four physicians whom we used to motivate the paper in Section 4. We see here too Physicians 34 and 8 have significantly higher overflow. The Capacity Ratio heuristic evens out the differences but this comes at a cost of shifting 193 patients. Heuristic 1

Table 5 Results for Test Practice 2: 4 physicians with equal capacity

		Comorbidity count								Panel	μ_j	σ_j^2	C_j	O_j	Utilization
		0	1	2	3	4	5	6	7	size					
Current	Phy 39	290	296	218	145	84	27	12	5	1,077	14.73	14.47	17	0.28	0.87
	Phy 8	260	249	226	161	108	42	14	3	1,063	15.54	15.26	17	0.35	0.91
	Phy 19	299	293	212	147	77	26	6	1	1,062	14.10	13.86	17	0.22	0.83
	Phy 34	214	253	223	177	115	44	21	5	1,053	16.16	15.85	17	0.42	0.95
	# switched	0	0	0	0	0	0	0	0	0					
Capacity based	Phy 39	266	272	220	157	96	34	14	3	1,062	15.11	14.83	17	0.31	0.89
	Phy 8	266	272	220	157	96	34	14	3	1,062	15.11	14.83	17	0.31	0.89
	Phy 19	265	273	219	158	96	35	13	4	1,063	15.15	14.87	17	0.32	0.89
	Phy 34	266	274	220	158	96	36	12	4	1,066	15.17	14.90	17	0.32	0.89
	# switched	58	44	9	23	31	16	9	3	193					
Heuristic 1	Phy 39	357	296	218	145	84	27	12	5	1,144	15.14	14.89	17	0.32	0.89
	Phy 8	194	249	226	161	108	42	14	3	997	15.13	14.85	17	0.31	0.89
	Phy 19	461	293	212	147	77	26	6	1	1,223	15.11	14.87	17	0.31	0.89
	Phy 34	51	253	223	177	115	44	21	5	889	15.15	14.84	17	0.32	0.89
	# switched	229	0	0	0	0	0	0	0	229					
Heuristic 2	Phy 39	292	298	220	147	86	30	14	7	1,094	15.13	14.86	17	0.31	0.89
	Phy 8	258	247	224	159	106	39	12	1	1,046	15.14	14.87	17	0.31	0.89
	Phy 19	305	299	218	153	83	31	11	6	1,106	15.11	14.84	17	0.31	0.89
	Phy 34	208	247	217	171	109	39	16	0	1,007	15.15	14.87	17	0.32	0.89
	# switched	8	8	8	8	8	8	7	7	62					

O_{ref} for this practice is 0.31. The number of patients switched is provided as a separate row under each heuristic. The total patients switched by a particular heuristic appears under the panel size column

switches 229 patients, which constitutes 5 % of the total patients, but all of them are 0 comorbidity patients. Heuristic 2 switches only 62 patients (only 1.5 % of the total patients) but this does include a few high comorbidity patients. The difference in the number of patients switched (from each patient category and in total) can clearly be observed from Table 5. Notice that both Heuristic 1 and Heuristic 2 are able to reach the overflow values that the Capacity Ratio allocation produces, which is optimal in this equal capacity case. Both the capacity ratio algorithm and the heuristics are able to balance the utilization and overflow frequency.

In **Test Practice 3** (Table 6), Physician 20 has more patients in her panel and also has more capacity (21) compared to Physician 24 (15). However, the former's overflow is more than double the latter's. There is a clear case for panel redesign here, since Physician 20's current capacity of 21 slots per day is already quite high and mostly likely cannot be increased anymore. This is especially true since primary care physicians are responsible for numerous other non-visit tasks during the day, such as attending phone calls, coordinating with specialists her patient might have recently visited and so on. The Capacity Ratio reduces the imbalance in panel workloads somewhat but clearly does not provide the optimal solution. Notice that the utilizations (which are calculated using the mean demands and the capacity of the physician) are perfectly balanced under Capacity Ratio, but the overflows are not. This is because the utilization (μ_j/C_j) does not consider variance but the overflow frequency does. Moreover Capacity Ratio switches 142 patients. Heuristic 1 and 2, on the other hand, produce overflows that are almost identical to the reference overflow (0.264). Heuristic 1 switches 172 healthy patients, while Heuristic 2 switches 52 patients in total from all the categories. Thus with regard to both overflow and patients switched, the H1 and H2 are better than Capacity Ratio.

Finally as can be seen from Table 7, in **Test Practice 4**, there are four physicians with different panel sizes and capacity values (24, 17, 15 and 14 respectively). Notice, however, that the overflow and utilization values are not dramatically different to begin with (at least in relation to Test Practice 3). In this case, the practice may decide that no redesign is required. We note here that our approach and presentation of performance measures will help practices come to such a conclusion.

As in Test Practice 3, we note that Capacity Ratio is a good heuristic and reduces the imbalance between physicians but does not give the optimal overflow. It also requires that 129 patients be moved, despite the fact that overflow differences between the physicians are not significant. Heuristic 1 and 2 are more effective in reducing the overflow, but also move fewer patients compared to Capacity Ratio. Heuristic 1 switches 2 %, whereas Heuristic 2 only changes 0.5 % of the total patients. As before this is because, Heuristic 1 affects only the healthy patients, while Heuristic 2 involves patients from all categories.

8.3 Quantifying the price of continuity

We can also measure the price of continuity by quantifying the gap between O_{ref} and O_{CP} in terms of the number of new patients who can can be empanelled. Recall that O_{ref} is a surrogate for the best possible access that the physicians in the practice can provide with the available capacity once panels are redesigned

Table 6 Results for Test Practice 3: 2 physicians with unequal capacities

		Comorbidity count								Panel	μ_j	σ_j^2	C_j	O_j	Utilization
		0	1	2	3	4	5	6	7	size					
Current	Phy 20	255	314	289	223	124	54	21	1	1,281	19.33	18.97	21	0.35	0.92
	Phy 24	255	262	189	107	52	25	5	1	896	11.64	11.45	15	0.16	0.78
	# switched	0	0	0	0	0	0	0	0	0					
Capacity ratio	Phy 20	297	336	278	192	102	46	15	1	1,267	18.01	17.69	21	0.24	0.86
	Phy 24	213	240	200	138	74	33	11	1	910	12.96	12.73	15	0.28	0.86
	# switched	42	22	11	31	22	8	6	0	142					
Heuristic 1	Phy 20	83	314	289	223	124	54	21	1	1,109	18.26	17.90	21	0.26	0.87
	Phy 24	427	262	189	107	52	25	5	1	1,068	12.71	12.51	15	0.26	0.85
	# switched	172	0	0	0	0	0	0	0	172					
Heuristic 2	Phy 20	247	306	282	216	117	47	14	0	1,229	18.26	17.92	21	0.26	0.87
	Phy 24	263	270	196	114	59	32	12	2	948	12.71	12.50	15	0.26	0.85
	# switched	8	8	7	7	7	7	7	1	52					

The reference overflow value, O_{ref} for this practice is 0.264. The number of patients switched is provided as a separate row under each heuristic. The total patients switched by a particular heuristic appears under the panel size column

Table 7 Results for Test Practice 4: 4 physicians with unequal capacities

		Comorbidity count								Panel	μ_j	σ_j^2	C_j	O_j	Utilization
		0	1	2	3	4	5	6	7	size					
Current	Phy 28	418	385	299	211	111	32	10	2	1,469	19.64	19.31	24	0.16	0.82
	Phy 19	299	293	212	147	77	26	6	1	1,062	14.10	13.86	17	0.22	0.83
	Phy 17	274	245	189	98	52	23	11	1	894	11.57	11.37	15	0.15	0.77
	Phy 12	244	233	162	107	46	27	9	2	830	10.96	10.77	14	0.18	0.78
	# switched	0	0	0	0	0	0	0	0	0					
Capacity ratio	Phy 28	426	399	297	194	98	37	12	2	1,465	19.36	19.03	24	0.14	0.81
	Phy 19	310	290	216	142	73	28	10	2	1,071	14.24	14.00	17	0.23	0.83
	Phy 17	259	242	181	118	60	22	7	1	890	11.75	11.55	15	0.17	0.78
	Phy 12	240	225	168	109	55	21	7	1	826	10.91	10.73	14	0.17	0.78
	# switched	28	33	11	27	15	8	6	1	129					
Heuristic 1	Phy 28	458	385	299	211	111	32	10	2	1,508	19.89	19.56	24	0.18	0.83
	Phy 19	219	293	212	147	77	26	6	1	981	13.60	13.37	17	0.18	0.80
	Phy 17	315	245	189	98	52	23	11	1	934	11.82	11.63	15	0.18	0.79
	Phy 12	243	233	162	107	46	27	9	2	829	10.95	10.77	14	0.18	0.78
	# switched	81	0	0	0	0	0	0	0	81					
Heuristic 2	Phy 28	418	387	301	213	112	34	11	3	1,479	19.89	19.56	24	0.18	0.83
	Phy 19	295	290	209	144	74	23	3	0	1,038	13.61	13.39	17	0.18	0.80
	Phy 17	278	246	190	99	54	24	13	1	905	11.79	11.60	15	0.17	0.79
	Phy 12	244	233	162	107	46	27	9	2	830	10.96	10.77	14	0.18	0.78
	# switched	4	3	3	3	3	3	3	1	23					

The reference overflow value, O_{ref} for this practice is 0.177. The number of patients switched is provided as a separate row under each heuristic. The total patients switched by a particular heuristic appears under the panel size column

and assuming that physicians do not see each others patients. In contrast, O_{CP} is the overflow of a practice in which all panel demand is aggregated and all physician capacity is pooled. The latter provides improved access to care (lower wait times) but at the expense of continuity. (Since $O_{\text{CP}} \leq O_{\text{ref}}$ from Section 6.)

If a practice cares more about access to care than continuity, then how many patients could it have added if O_{CP} is allowed to increase and until it equals O_{ref}? In other words, if the access performance as measured by overflow frequency is held constant, how many more patients can a pooled practice with no concept of continuity empanel compared to a dedicated practice where patients only see their own PCP? We quantify the number of patients that could be empaneled for each of the test practices in Section 8.2.

In Test Practice 1, which consists of 2 equal capacity physicians, 127 new patients could have been empaneled if the current O_{CP} value of 0.16 is allowed to increase to the O_{ref} value of 0.24. For this calculation, we assume that the new patients added have the same comorbidity mix that the practice currently has. For example, 750 of the 2963 total patients (around 25 %) in Test Practice 1 were 0 comorbidity patients. Since this may be a fair reflection of the demographics of the neighborhood in which the practice is located, we assume that 25 % of the 127 new patients that the practice can empanel will also be 0-comorbidity patients. Similar calculations apply for other comorbidity counts (Table 8).

The addition of new patients implies a loss of continuity since any physician in the practice can see any patient. There is no single PCP who coordinates the patients' care. In a fee-for-service system, where physicians are reimbursed based on the number of visits, the revenues for the practice will increase as will the

Table 8 Price of continuity in terms of number of patients

Capacity	Test practice	No. of physicians	Total capacity	Ocp as %	Oref as %	Patients added
Equal	1	2	48	16	24	127
	2	4	68	17	31	255
Unequal	3	2	36	18	26	100
	4	4	70	3	18	500

overall ability to access physicians, but patient centeredness and possibly physician satisfaction will likely to decrease.

Among the two 4-physician practices, Test Practice 2 will be able to add 255 patients at the expense of continuity while Test Practice 4 will be able to add 500 patients at the expense of continuity. This difference is because of two reasons. First, comorbidity counts are higher in Test Practice 2 compared to Test Practice 4. Second, utilization and overflow values are lower in Test Practice 4 to begin with, allowing for a greater number of patients to be added.

Thus our framework allows a practice to look at extremes of best possible continuity and best possible access and make their empanelment decisions accordingly.

8.4 Impact on other measures

We have so far investigated overflow frequency and utilization. We now look at Expected overflow (EO) and Expected unfilled slots (EU). Expected overflow, which was explained in Section 4, represents the average number of patients who were not able to get appointments. Expected unfilled slots tells us how under-utilized each physician is. To test the impact on these two measures, we choose Physicians 19 and 34, from Test Practice 2. Both these physicians have equal capacity (17) and before their panels are redesigned, their overflow frequencies were 0.22 and 0.42 respectively. We calculate EO and EU for both physicians before redesign (Current) and after redesign (Balanced). The heuristic used for redesign is Capacity-ratio, which gives an optimal allocation since the two physicians have the same capacity.

Since there are no closed form expressions for EO and EU, we simulate 10,000 realizations of demand, sampled from the binomial distributions appropriate for each patient category. Each realization represents a day in the model. If the physicians have any backlog it is transferred to the next day. We also investigate the impact of *sharing or transferring* patients. That is if a physician has capacity available after seeing her own patients, then she is allowed to see the other physician's patients (if the other physician has a backlog), at the expense of continuity. We compare this case against the dedicated case, where the physicians do not share or transfer their patients; that is, they maintain continuity at the expense of timely access.

Figure 3 clearly shows the benefits of redesign (Balanced versus Current). The benefits are especially significant when the two physicians do not share their patients (the No Transfer case). If the physicians are not allowed to transfer patients and case mixes remain the same then the resulting expected overflow is almost unsustainable (for Physician 34 especially), resulting in poor access. Panel redesign produces more even EO profiles when sharing is allowed (Transfer case), but the difference is not as significant as in the no-transfer case. We notice here that sharing of patients mitigates the poor timely access problem. The unevenness in expected unfilled slots between physicians is leveled with the balanced case mixes.

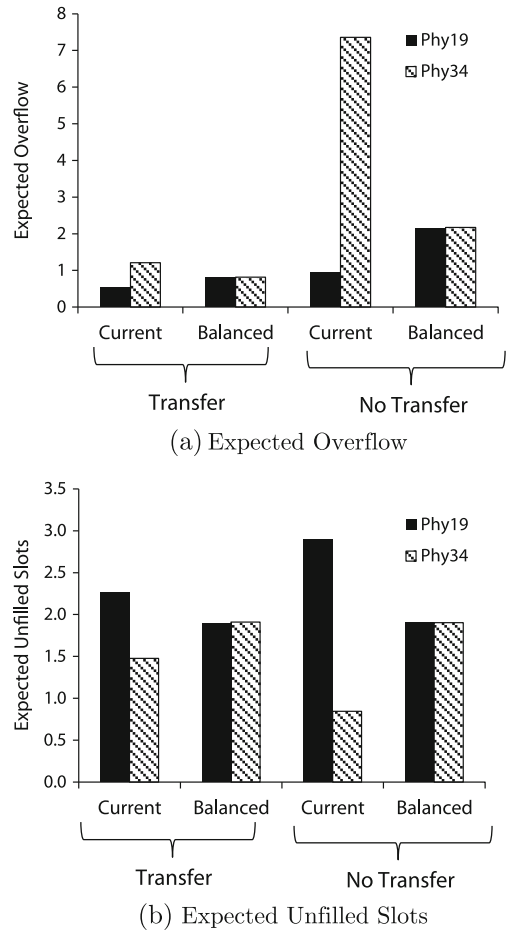


Fig. 3 Results for 2 physicians with equal capacity

These results suggest that even if the practices are unwilling to redesign panels, sharing of patients between physicians is a viable alternative, especially if a practice consists of 2–3 physicians. Moreover the sharing can be restricted to same-day requests for which continuity is not always necessary or desired by the patients. While this is not the ideal scenario, access is improved at the cost of continuity of care. If the physicians are keen on providing continuity then it is clear that the panels have to be redesigned. We find similar results while testing other pairs of physicians,

but in the interest of keeping the paper concise these results are not presented.

9 Conclusions and implications for practice

In summary, we have shown that case-mix is an important consideration in primary care. Physicians with the same panel size but different case-mixes can have very different overflow frequencies. We have characterized how overflow frequencies can vary from physician to physician and demonstrated, using actual data from a primary care practice, how these imbalances in supply and demand can be minimized in the long term.

To implement our results, a practice will have to collect appointment request rates of its patient population from historical data. Two to three years worth of visit data should be sufficient to classify patients according to their visit patterns. With the increasing use of electronic records, such data should be easily available. Practices can use the opportunity to update information about currently active patients and obtain more precise information about panel sizes.

Once this assessment is complete, practices can then begin to benchmark their current performance by comparing the overflow frequencies of the physicians in relation to one another and in relation to the reference overflow derived in this paper. Panel redesign options can be easily tested, in a manner similar to Tables 4, 5, 6, 7 and the least disruptive options of redesigning panels can be identified. In general clinics should be aware that O_{ref} values of 0.3 or above, which result in high utilization, should be avoided.

All overflow frequency calculations derived in this paper can be easily carried out in an Excel spreadsheet. The American Academy of Family Physicians (AAFP) has an Excel spreadsheet tool for panel size calculations [16]. However, it uses only the mean and does not consider case-mix or the impact of variance. The Excel tool provided by Green et al. [9] allows practices to decide on panel size for a single physician based on overflow frequency. The impact of variance is also considered in their calculations. The results in this paper extends the Green et al. [9] framework, to allow for an Excel tool that (1) quantifies the impact of case-mix; (2) calculates a benchmark overflow value for a group practice; and (3) allows for testing of various panel redesign options in the long term. A preliminary version of our Excel spreadsheet is available for free at people.umass.edu/hbalasub/PanelDesignSpreadsheet.xlsx.

Our model does have limitations, which provide opportunities for future investigation and model refinement. We do not consider seasonality and day of week effects on overflow frequencies. Savin [24] analyzes the effect of seasonality and day-to-day variability in a primary care practice and observes that variations can be quite high. To model this effect, he adjusts the probability that a patient requests an appointment for a specific day or month. Our category specific p_i values can also be adjusted depending on the time of the year or day of the week. Savin suggests that to cope with such variations practices will either have to adjust panel sizes, or flexibly adjust the capacities of the physicians. In addition, practices can leverage the benefits of working in groups—an aspect we consider in this paper. In peak seasons or busy days, urgent same-day requests could be flexibly shared by a small group of 2–3 physicians. As Section 8.4 shows, such flexibility can improve access; the compromises in continuity will be small so long as provider team is small. For more details on how same-day flexibility can be designed to balance access and continuity under different utilization levels, we point to Balasubramanian et al. [2, 3].

Another extension worth considering is whether physician practice style has an impact on visit rates and consequently overflow frequencies. This can happen if some physicians schedule more follow-up visits than others on average. Recall that in the current model, demand is controlled by the p_i values for the comorbidity categories, which in turn is decided by the total number of visits from each category over a long period (2–3 years). Now, as an example, if we were able to determine—through new empirical data and appropriate statistical tests—that physician j scheduled twice as many visits for high comorbidity count patients compared to physician k, then the p_i values for that category would accordingly have to be physician specific. So not only do higher comorbidity patients have higher visit rates (which is indeed the case and is the premise of our paper), but some physicians schedule more visits for these patients than others, with implications for the overflow frequency. This is an interesting direction for future work, and would require careful collection of new physician-specific appointment data.

As mentioned earlier, our modeling approach is designed for aggregate level panel management decisions. While we do not explicitly consider different appointment types, such as prescheduled and same-day, a high overflow frequency will be correlated with the inability to provide access for both types of appointments. In the same way, although no-shows are not a part of our model, well designed panels can only reduce the impact of no-shows, by improving time to earliest available appointments. See [8] for a discussion. Finally, patients with more comorbidities are more likely to have longer appointments than healthy patients. Our values for

overflow frequency are therefore likely to be slightly smaller than those found in practice. However, in a relative sense, our approach will still correctly identify the imbalances in supply and demand across physicians. If anything, redesign will have an even greater effect.

Acknowledgements This work was funded by the grant R03 HS 018795 from the Agency of Healthcare Research and Quality (AHRQ). 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 AHRQ.

References

1. Balasubramanian H, Banerjee R, Denton B, Naessens J, Wood D, Stahl J (2010) Improving clinical access and continuity using physician panel redesign. J Gen Intern Med 25(10):1109–1115
2. Balasubramanian H, Muriel A, Wang L (2011) The impact of flexibility and capacity allocation on the performance of primary care practices. Flex Serv Manuf J. doi:10.1007/s10696-011-9112-5
3. Balasubramanian H, Muriel A, Ozen A, Wang L, Hippchen J, Gao X (2013) Capacity allocation and flexibility in primary care. In: Denton B (ed) Handbook of healthcare operations management. Springer US
4. Bodenheimer T, Pham HH (2010) Primary care: current problems and proposed solutions. Health Aff 29(5):799–805
5. Chakraborty S, Muthuraman K, Lawley M (2010) Sequential clinical scheduling with patient no-shows and general service time distributions. IIE Trans 42(5):354–366
6. Coleman K, Reid RJ, Johnson E, Hsu C, Ross TR, Fishman P, Larson E (2010) Implications of reassigning patients for the medical home: a case study. Annals of Family Medicine 8(6):493–498
7. Gill JM, Mainous A (2010) The role of provider continuity in preventing hospitalizations. Arch Fam Med 7:352–357
8. Green LV, Savin S (2008) Reducing delays for medical appointments: a queueing approach. operations research. Oper Res 56(6):1526–1538
9. Green LV, Savin S, Murray M (2007) Providing timely access to care: what is the right patient panel size? Joint Comm J Quality Saf 33:211–218
10. Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. Oper Res 56(3):576–592
11. LaGanga L, Lawrence S (2007) Clinic overbooking to improve patient access and increase provider productivity. Decis Sci 38(2):251–276
12. Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf Serv Op 12(2):347–364
13. MDoH (2012) Health care homes payment methodology development. http://www.health.state.mn.us/healthreform/homes/payment/index.html. Accessed 20 Sep 2012
14. Murray M, Berwick DM (2003) Advanced access: reducing waiting and delays in primary care. J Am Med Assoc 289(8):1035–1040
15. Murray M, Tantau C (2000) Same-day appointments: exploding the access paradigm. Fam Pract Manag 7(8):45–50
16. Murray M, Davies M, Boushon B (2007) Panel size: how many patients can one doctor manage? Fam Pract Manag 14(4):44–51
17. Muthuraman K, Lawley M (2008) Stochastic overbooking model for outpatient clinical scheduling with no shows. IIE Trans 40(9):820–837
18. Naessens J, Stroebel R, Finnie D, Shah N, Wagie A, Litchy W, Killinger P, O'Byrne T, Wood D, Nesse R (2011) Effect of multiple chronic conditions among working-age adults. American Journal of Managed Care 17(2):118–122
19. Nutting P, Goodwin M, Flocke S, Zyzanski S, Stange K (2003) Continuity of primary care: to whom does it matter and when? Ann Fam Med 1(3):149–155
20. Qu X, Rardin R, Williams J, Willis D (2006) Matching daily healthcare provider capacity to demand in advanced access scheduling systems. Eur J Oper Res 183(2):812–826
21. Reid RJ, Fishman P, Yu O, Ross TR, Tufano JT, Soman M, Larson E (2009) Patient-centered medical home demonstration: a prospective, quasi-experimental, before and after evaluation. The American Journal of Managed Care 15(9): 71–87
22. Reid RJ, Coleman K, Johnson EA, Fishman P, Hsu C, Soman M, Trescott CE, Erikson M, Larson E (2010) The group health medical home at year two: cost savings, higher patient satisfaction, and less burnout for providers. Health Aff 29(5):835–843
23. Robinson L, Chen R (2010) A comparison of traditional and open access policies for appointment scheduling. Manuf Serv Oper Manag 12(2):330–347
24. Savin S (2006) Managing patient appointments in primary care. In: Hall RW (ed) Patient flow: reducing delay in healthcare delivery. Springer US, pp 123–150
25. Starfield B, Weiner J, Mumford L, Steinwachs D (1991) Ambulatory care groups: a categorization of diagnoses for research and management. Health Serv Res 26(1):53–74
26. Wang W, Gupta D (2011) Adaptive appointment systems with patient preferences. Manuf Serv Op 37(1):111–126