THE IMPACT OF CASE MIX ON TIMELY ACCESS TO APPOINTMENTS FOR A PRIMARY CARE PHYSICIAN

January 30, 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 whose long term, holistic care the physician is responsible for. 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 the probability that demand exceeds the capacity, as a measure of access. We formulate the 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, in the long term, 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.

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 whose long term care she is responsible for. 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 (1999) point to several studies which show that patients who regularly see their own providers are 1) more satisfied with their care; 2) more likely to take medications correctly; 3) more likely to have problems correctly identified by their physician; and 4) less likely to be hospitalized. Continuity and coordination are especially important for vulnerable patients with a complex medical history and mix of medications [Nutting et al. (2003)].

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. We do this 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.

Our consideration of panel size and case-mix 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.

Our analysis 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. For example, new patients may join the practice, existing patients may move from the area, and patient preferences about who their PCP should be 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 propose an integer non-linear programming formulation for redesigning panels in a group practice. Our 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. We then use the analytical results to motivate heuristics, which will allow practice managers to 1) test various redesign options and, 2)infer which options are the least disruptive. A key advantage of our approach is that it 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, we review the relevant literature and in Section 3, we explain how we model case-mix. We motivate the panel redesign problem using an example involving 4 physicians in Section 4. Section 5 contains all the mathematical details and analytical results related to the panel redesign formulation. In Section 6, we describe our heuristics. In Section 7, 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 our results. Section 8 summarizes our 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 Mark Murray (Murray and Tantau 2000; Murray and Berwick 2003; Murray et al. 2007), attempted to promote same-day access for patients. Advanced access was based on a simple principle: it encouraged each physician "do today's work today". This meant that all appointments, regardless of their nature and urgency of request, were to be seen the same day by the patient's PCP. In traditional appointment systems, appointments are allowed to be booked into the future, whereas in advanced access this is discouraged. 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.

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 (Robinson and Chen, 2010); the impact of no-shows (LaGanga and Lawrence 2007, Muthuraman and Lawley 2008, and Chakraborty et al. 2010, Liu et al, 2010); the importance of considering patient preferences (Gupta and Wang, 2008; Wang and Gupta, 2011); and capacity allocation methods that allow practices to offer a blend of prescheduled (urgent) and same-day (urgent) appointments (Balasubramanian et al. 2011 and Qu et al, 2007).

We focus this review on the papers most relevant to our work on panel size and case-mix. Murray (2003) proposed six steps for clinics to implement advanced access. An important message of this work is that the primary lever for demand was the number of patients in a physician's panel. Murray et al. (2007) provide a simple algorithm to calculate the "right" panel size for physicians. Murray et al. (2007) 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.

Liu and D'Aunno (2012) develop queuing models to observe the productivity and cost-effectiveness

of increasing nurse practitioners (NPs) role in primary care. They compare 3 models: a regular solo-physician practice model, a supervision model and a shared-panel model. Supervision leads to a decrease in both productivity and cost-efficiency. However, the shared-panel model improves both of the outcomes significantly. By enabling capacity pooling, the team-based model is able to handle the variability in the demand and decrease the waiting times without any additional staffing cost. They suggest extending their model to investigate the effect of patient mix resulting in different visit frequencies as well as different consultation times.

Green and Savin (2008) 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, Savin and Murray (2007), where a newsvendor like model is proposed to determine the relationship between the size of a physician 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 the 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 in this paper is closest to the modeling framework of Green, Savin and Murray (2007). We extend their newsvendor like approach to include case-mix and also 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, we use overflow frequency as a measure of access, and then develop theoretical results 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. We demonstrate our results 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. For example, a panel where the majority of patients are young and healthy will have a different appointment profile compared to a panel consisting mostly of elderly patients with chronic conditions. 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) [Starfield et al, 1991] argued that understanding the role of patient clinical complexity in care utilization forms the cornerstone for effective resource planning and determining payment methods in healthcare.

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 [Murray et al., , Balasubramanian et al. 2010]. In this paper, we use the number of simultaneous chronic conditions a patient has 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 it. 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 [2011] show that the number of simultaneous chronic conditions is a strong predictor of the number of office visits. Comorbidity counts have also been used in the a new payment scheme for primary care proposed by the Minnesota Department of Health. Finally, statistical analysis of our 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 20,000 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.

Figure 1 shows mean and standard deviation of visit rates as a function of the number of patients under various counts of comorbidities. The data was simulated based on historical visits of 20,000 patients empanelled PCIM. Clearly, not only does the mean number of visits increase with the number of comorbidities, the variance does as well. For instance if a physician has 50 6-comorbidity patients then he will have 450 appointment requests on average each year. If he has same number of 0-comorbidity patients he will have only 75 yearly visits on average. The same trend is true for the standard deviation as well.

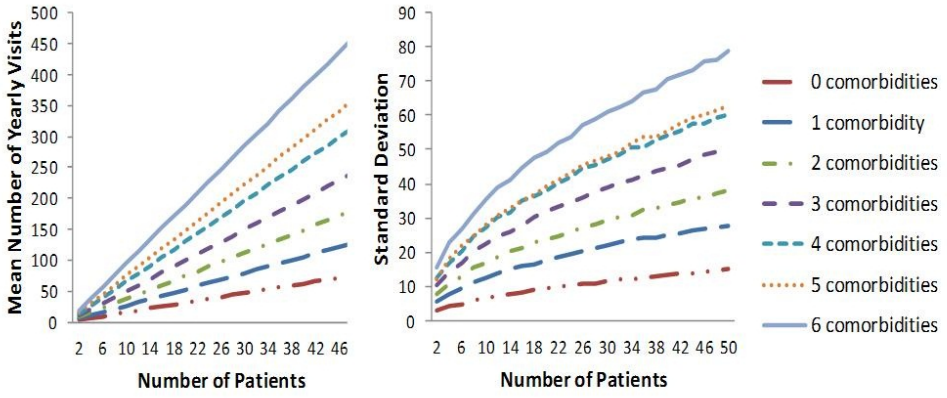


Figure 1: Mean and standard deviation of yearly visits for groups with different counts of comorbidities

4 Example of 4 physicians

The PCIM practice at the Mayo Clinic employs 39 physicians in total, many of whom work part time. We now consider an example of four physicians from PCIM with approximately the same panel size (1050 patients), but different case-mixes, based on comorbidity counts. These panel compositions are shown in Table 1.

Table 1: Four physicians at PCIM, Mayo Clinic and their patient case-mix based on comorbidity count.

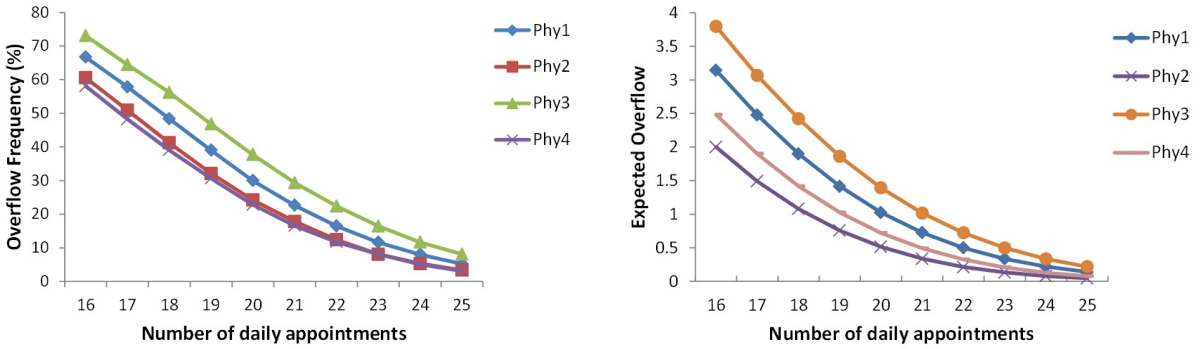
Physicians	0	1	2	3	4	5	6	7	Panel size
Physician 1	260	249	226	161	108	42	14	3	1063
Physician 2	299	293	212	147	77	26	6	1	1062
Physician 3	214	253	223	177	115	44	21	5	1053
Physician 4	290	296	218	145	84	27	12	5	1077

We next use a simple simulation to characterize the impact of case-mix on the overflow frequencies of the four physicians. Suppose there are $j = 1, ..., J$ physicians in the practice. Overflow frequency is the fraction of total samples in which the patients' visit requests exceed the available capacity of the physician. 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 7 for the exact values and how theses 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, and 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). To create a visit profile for each physician, we made use of p_i and n_{ij} values and simulated the demand distribution with 10000 realizations sampled from the binomial distributions.

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

The overflow frequency for the four physicians as a function of the total daily slots (capacity) is shown in Fig. 2 (a). For the same capacity, Physician 3 and Physician 1 have relatively high overflow frequencies. This is because there are more patients with two or more comorbidities in

their panels (see Table 1), and these patient groups generate a higher number of visits. Patients that are not seen either visit an unfamiliar physician or an ER, or may choose to wait to see the physician on another day. Thus, if overflow is high, both timely access and continuity are adversely affected. This graph shows that it is inappropriate for clinics to make capacity decisions based solely on panel size. Case-mix is also an important consideration. It is also clear that to keep overflow levels down to manageable levels, 20 or more appointment slots may be needed for each of the 4 physicians.



(a) Overflow Frequency

(b) Expected Overflow

Figure 2: Results for 4 physicians as a function of the daily capacity (appointment slots)

We also calculate the *expected overflow* which gives the average number of patients who did not get their requests fulfilled. In Fig. 2 (b) the expected overflow as a function of the physicians' capacity is shown. Not unsurprisingly, the overflow frequency and expected overflow are positively correlated. With 19 daily appointments, Physician 2 faces an overflow frequency of 32 percent, whereas the expected overflow is .75 patients a day. Such analysis allows practices to identify which physicians are overburdened. In the above case, it's clear that physicians 3 and 1 need to have their capacity enhanced – either by working extra hours in a day or by additional nurse practitioner or physician assistant support. The long-term option for practices is to redesign panels, by moving high demand, high variability patients from an overburdened physician to a physician with available capacity. We turn to this in the next section.

5 The Panel Redesign Formulation (PRF) and Analytical Results

In the last section, we saw two ways of measuring access: overflow frequency and expected overflow. The former gives probability that the demand exceeds available capacity while the latter counts the average number of patients whose requests went unfulfilled. The expected overflow gives more information compared to overflow frequency, but we saw from Fig. 2 that the two are correlated. An increase in overflow frequency implies that the expected overflow will also increase.

In this section, we provide a mathematical formulation 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. Later in the results section, we shall see again the positive correlation we have already observed between overflow frequency and the expected overflow.

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 lead eventually 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 previous section, n_{ij} denoted the number of patients from patient class i in physician j's 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 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 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 variance for physician j's panel are $\mu_j = \sum_{i=1}^{M} p_i x_{ij}$ and standard deviation $\sigma_j = \sqrt{\sum_{i=1}^{M} p_i(1-p_i)x_{ij}}$, respectively. Note that both the mean and standard deviation depend on the case-mix distribution given by the x_{ij} values for the physician. If we assume that the sum of M binomial random variables gives us a normal random variable, then O_j, the overflow for physician j is related to the percentile of the standard normal distribution, given by Φ, in the following way: $O_j = 1 - \Phi(\frac{C_j - \mu_j}{\sigma_j})$. Here C_j is the capacity of the physician, the total daily slots that she has available in a day. If panel sizes are sufficiently large (say 800-1000 patients), then normal approximation is reasonable since the total demand is the sum of 800-1000 Bernoulli random variables.

The goal is 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).

$$(PRF) \qquad \min\{\max\{O_1, O_2, ..., O_J\}\} \tag{1}$$

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

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

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

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

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

Note that PRF is an integer non-linear program. Further, the allocation (the set of x_{ij} values) influences both the mean and variance for each panel, which in turn affects $Z_j = \frac{C_j - \mu_j}{\sigma_j}$. Z_j is the transformation of normally distributed demand for physician j into the standard normal random variable format, with mean 0 and standard deviation 1. Intuitively, it gives the number of standard deviations that the capacity is distant from the mean of the panel demand. The greater the positive distance between C_j and μ_j and the smaller the σ_j, the lower the overflow frequency. When $C_j = \mu_j$, a system where the utilization (μ_j/C_j) is 100%, the overflow frequency is 0.5.

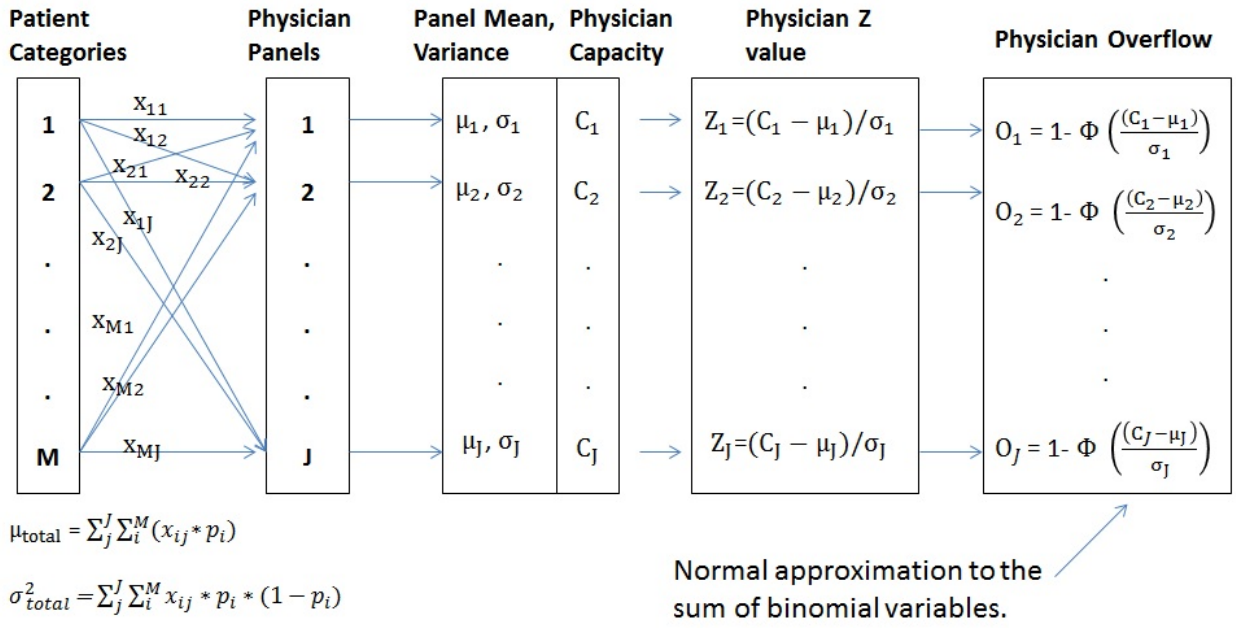


Figure 3: A visual summary of the panel redesign problem to minimize the maximum overflow

The formulation is described visually in Fig. 3. To develop intuition, we now make two important observations:

- **Observation 1:** The total mean and variance of the entire patient population given by $\mu_{total} = \sum_{j=1}^{J} \sum_{i=1}^{M} p_i x_{ij}$ and $\sigma^2_{total} = \sum_{j=1}^{J} \sum_{i=1}^{M} p_i(1-p_i)x_{ij}$. 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.

- **Observation 2:** The means and variances are not allocated independently of each other but are tied to the x_{ij} allocations. Recall that the number of patient requests from each patient category follows a binomial distribution. The variance of a binomial random variable is a constant probability times the mean (mean: np and variance: $np(1-p)$ for n independent trials, each trial with a probability p for success). Therefore, if more patients are assigned to a physician, the mean *and* the variance will increase for that physician's panel. In summary, O_j, μ_j and σ_j will increase (decrease) when x_{ij} increases (decreases) for any $i = 1...M$.

We are looking for allocations such that $\frac{C_1-\mu_1}{\sigma_1} = \frac{C_2-\mu_2}{\sigma_2} = ... = \frac{C_J-\mu_J}{\sigma_J}$, which implies that all

physicians will have identical O_j values. This may not happen exactly, since our decision variables, the x_{ij} values are integer. But we can be sure that overflow values will be almost equal, as stated formally below.

Claim 1. In an optimal allocation that minimizes the maximum overflow, $O_1 \approx O_2 \approx ... \approx O_J$.

To illustrate let us consider an optimal allocation in which physician H determines the maximum overflow O_H. This means that $Z_H = \frac{C_H - \mu_H}{\sigma_H}$ is highest among all the J physicians. Similarly, let physician L be the one with the lowest $Z_L = \frac{C_L - \mu_L}{\sigma_L}$ value and hence the lowest overflow value, O_L. Now, since the allocation is optimal it must be true that not a single patient from physician H's current panel can be transferred to physician L's panel. If this were to be possible, then the O_H value would decrease and this would contradict the fact that the current allocation is optimal.

This means that in an optimal allocation the difference between overflow values between physicians H and L, $O_H - O_L$, has to be small enough to ensure that even the transfer of a single patient cannot lower the optimal value, O_H.

Let Δ represent the decrease in physician H's overflow and the increase in physician L's overflow as a result of transferring a 0-comorbidity patient. We choose a 0 comorbidity patient because such a patient has the lowest probability of requesting an appointment, with $p_0 = 0.0062$. Shifting one 0-comorbidity patient would decrease physician j's mean by 0.0062 and variance by $0.0062 * (1 - 0.0062)$ and increase physician L's mean and variance by 0.0062 and $0.0062 * (1 - 0.0062)$ respectively. The effect on the overflow values of physicians H and L due to the transfer of this patient is very small – in the fifth of sixth decimal places, small enough to be negligible.

Now, if $O_H - O_L \leq \Delta$ – which it has to be since the solution is optimal – then, O_H and O_L must have been almost equal to begin with. And since O_H is the highest overflow value and O_L is the lowest overflow value, it must be true that $O_1 \approx O_2 \approx ... \approx O_J$. Therefore, for all practical purposes we will assume that we are looking for overflow values that are equal.

We consider two cases: the equal capacity case where $C_1 = C_2 = C_3.... = C_J$ and the unequal capacity case where $C_1 \neq C_2 \neq C_3.... \neq C_J$. Both cases exist in practice. 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.

5.1 The Equal Capacity Case

For the equal capacity case, suppose there exists an allocation that that divides mean and variance equally among the J physicians. In other words, suppose there exist feasible x_{ij} values such that for each physician, the mean $\mu_j = \mu_{total}/J$ and the variance $\sigma_j = \sigma^2_{total}/J$. In such an allocation, each physician has the same mean and each physician has the same variance, all physicians also have identical overflow frequency values since $\frac{C_1-\mu_1}{\sigma_1} = \frac{C_2-\mu_2}{\sigma_2} = ... = \frac{C_J-\mu_J}{\sigma_J}$. From Claim 1, such an allocation is optimal. This result is summarized in the theorem below:

Theorem 1. If $C_1 = C_2 = ... = C_J$ then any allocation in which $\mu_1 = \mu_2 = ... = \mu_J = \mu_{total}/J$ and $\sigma^2_1 = \sigma^2_2 = ... = \sigma^2_J = \sigma^2_{total}/J$ will minimize the maximum overflow frequency.

Theorem 1 can also be proved using a contradiction argument. Notice that if μ_{total} is not equally divided among physicians, then one physician will have a higher mean in comparison to the others. We have already noted in Observation 2 that a higher mean also implies a higher variance, so the physician's panel will also have a larger variance and standard deviation. Thus the overflow of the physician will also be higher than the others. Such an allocation cannot be optimal in relation to an allocation that divides the mean and variance equally among the J physicians.

Next, consider the allocation $x_{ij} = N_i/J$ – that is each physician j gets the same number of patients from each category i. To keep x_{ij} integer, we need to assume here that the number of patients in each class is a multiple of the total number of physicians in the practice. In practices with 5 or less physicians – a typical size in the US – this is a very mild assumption. Thus with $x_{ij} = N_i/J$ each physician gets the same number of patients from each category. Each panel thus has the same mean and variance, μ_{total}/J or σ^2_{total}/J. Naturally, the overflow for each physician is identical. It follows from Theorem 1 that $x_{ij} = N_i/J$ for all i and j minimizes the maximum overflow frequency.

Corollary 1. If $C_1 = C_2 = ... = C_J$ and the total number of patients in each category is a multiple of J then allocation $x_{ij} = N_i/J, \forall i, j$ is optimal.

Note that the "symmetric" allocation of Corollary 1 is *only one of many possible allocations* that follow the properties of Theorem 1. Nevertheless, 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 follow Theorem 1. We revisit this key point 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.

5.2 The Unequal Capacity Case

When the physicians have different number of slots available every day, it would seem appropriate to allocate patients keeping in mind the capacity a physician has. Greater capacity would imply a greater share of μ_{total} and σ^2_{total}. This is definitely true. 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 (we shall see specific examples in Section 7). 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 we have described in the equal capacity case, does not seem to be possible. Instead, our focus in this section is how the optimal objective can be approximated effectively. The approximation provides us with a reference or target overflow frequency, O_{ref}, which can be used the

heuristics proposed in the next section. We show that for all practical purposes O_{ref} is a good surrogate for the optimal overflow frequency O_{opt}.

5.2.1 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:

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

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. From Claim 1, 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 to write $Z_{opt} = Z_1 = Z_2 = Z_3 = ... = Z_J$. More precisely:

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

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

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

$$\begin{aligned} \sum_{i=1}^{J} \sigma_j * Z_{opt} &= \sum_{j=1}^{J} C_j - \sum_{j=1}^{J} \mu_j \\ Z_{opt} &= \frac{\sum_{i=1}^{J} C_j - \sum_{j=1}^{J} \mu_j}{\sum_{j=1}^{J} \sigma_j} \\ Z_{opt} &= \frac{C - \mu_{total}}{\sum_{j=1}^{J} \sigma_j} \end{aligned} \tag{8}$$

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

$$Z_{opt} = \frac{Z_{CP}}{R}, \text{where} \quad R = \frac{\sum_{j=1}^{J} \sigma_j}{\sqrt{\sigma_{total}^2}} \tag{9}$$

Note that since $\sigma_{total}^2 = \sum_j^J \sigma_j^2$, we can rewrite R as:

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

Notice that $R \geq 1$. This is because the sum of the square roots of J numbers (the numerator of R) is always greater than the square root of the sum (denominator of R). This means that $Z_{CP} \geq Z_{opt}$. The equality is tight when $R = 1$. 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 $Z_{ref} = \frac{Z_{CP}}{\sqrt{J}}$. If the capacities of the physicians are equal, then $Z_{opt} = Z_{ref}$ and if the capacities of the physicians are unequal, we have $\frac{Z_{CP}}{R} \geq \frac{Z_{CP}}{\sqrt{J}}$, which implies $Z_{opt} \geq Z_{ref}$.

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 the standard deviations of the J physicians stand in relation to each other – that relationship is captured by R.

From the above analysis, we have the following key result:

Lemma 2. $Z_{CP} \geq Z_{opt} \geq \frac{Z_{CP}}{\sqrt{J}}$

The corresponding overflows are given by $O_{CP} = 1 - \Phi(Z_{CP})$, $O_{opt} = 1 - \Phi(Z_{opt})$ and $O_{ref} = 1 - \Phi(Z_{ref})$ respectively. It follows from Lemma 2 that their relationship can be described as follows.

Theorem 2. $O_{CP} \leq O_{opt} \leq O_{ref}$

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 we do not have an exact method of computing O_{opt}, we will use $O_{ref} = 1 - \Phi(Z_{ref})$ as a surrogate for the optimal overflow. Note that O_{ref} is not a lower bound on the optimal overflow value. Rather we will demonstrate that $O_{ref} - O_{opt}$ is fairly small for most cases found in practice. Indeed, for the equal capacity case $R = \sqrt{J}, Z_{ref} = Z_{opt}$ and therefore $O_{ref} = O_{opt}$: the reference value is exactly equal to the optimal value.

5.2.2 $O_{ref} - O_{opt}$ for common cases in practice

To characterize $O_{ref} - O_{opt}$ 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 hours of work per day) is generally not common except in residency practices.

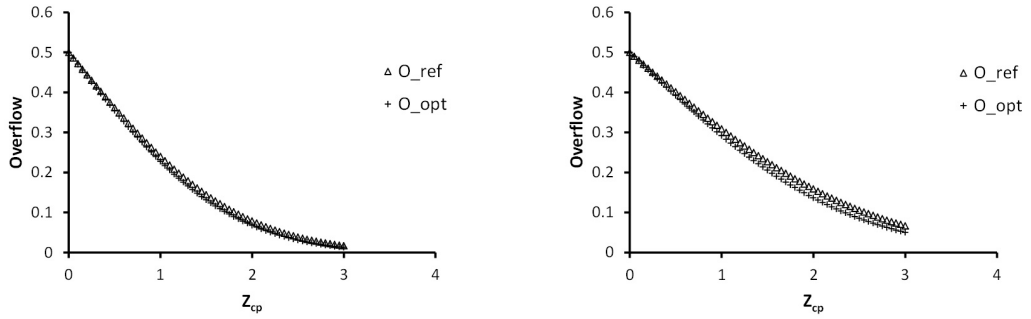
Although we do not know the optimal allocation of patients for above case, we can however state that the mean and variance allocated to the full time physician should be *roughly* four times that allocated to the quarter-time physician. We can say this because from Observations 1 and 2, we know that 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 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 represents more common 2-physician cases.

As an example, suppose we find that $Z_{CP} = 1.0$ 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, $Z_{ref} = \frac{Z_{CP}}{\sqrt{J}} = \frac{1.0}{1.414} = 0.707$. The optimal value, using $R = 1.34$ is $Z_{opt} = \frac{Z_{CP}}{R} = \frac{1.0}{1.34} = 0.746$. It follows that the reference overflow and optimal overflow are $O_{ref} = 1-\Phi(Z_{ref}) = 1-0.772 = 0.239$ and $O_{opt} = 1-\Phi(Z_{opt}) = 1-0.745 = 0.2227$ respectively. The difference is within 1%.

Figure 4 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 $Z_{CP} = 0$, 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 %.



(a) 2-physician case

(b) 4-physician case

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

To further reinforce the point we consider a 4 physician example. 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 $O_{ref} = 1 - \Phi(Z_{ref}) = 1 - \Phi(\frac{Z_{CP}}{sqrt(4)})$ 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 4 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.

6 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, we describe heuristics 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 we demonstrate with our heuristics, these two goals can be conflicting.

Before we explain 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, we try a different approach which involves all patient categories in the patient transfers. 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 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 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.

7 Results

7.1 Data Description

We use data from the Primary Care Internal Medicine (PCIM) practice at the Mayo Clinic in Rochester, MN. This practice empanels around 20,000 patients and employs 39 physicians. The 20,000 patients were empanelled with 39 physicians, many of whom worked part time. Panel data enabled us to identify which patient 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, we use the number of comorbidities 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:

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

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 Green et al. (2007) The values are listed in the Table 2 below.

Table 2: 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:

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

We will use this value to set the capacity of physicians in the test practices we create 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:

$$C_j = \lceil (L_j * p + 0.1 * L_j * p) \rceil$$

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 equal 100% would therefore be unsustainable. The additional slots ensure that there are few extra slots to buffer variability in demand. The above expression rounds up to the closest values, since the number of appointment slots per physician per day is typically 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.

7.2 Panel Redesign for Test Practices

Tables 3, 4, 5 and 6 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 figures, 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 $C_1 = C_2 = ... = C_J$, the allocation reduces to $x_{ij} = (N_i/J)$, which gives the optimal solution (see section 5). In the unequal capacity cases, $x_{ij} = (C_j/C) * N_i$ is a heuristic that we expect to perform well, but will not necessarily be optimal. For these cases, we use the reference overflow value 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 in each comorbidity group and as well as the total number of patients switched.

Table 3: Results for Test Practice 1: 2 Physicians with Equal Capacity. 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.

		Comorbidity Count													
		0	1	2	3	4	5	6	7	Panel Size	μ_j	σ_j	C_j	O_j	Utilization
Current	Phy 4	332	360	324	270	144	40	20	5	1495	21.99	21.58	24	0.33	0.92
	Phy 28	418	385	299	211	111	32	10	2	1469	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	1481	20.80	20.43	24	0.24	0.87
	Phy 28	375	373	311	241	127	36	15	4	1482	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	1307	20.81	20.42	24	0.24	0.87
	Phy 28	606	385	299	211	111	32	10	2	1656	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	1442	20.80	20.43	24	0.24	0.87
	Phy 28	425	392	306	218	118	39	16	7	1521	20.83	20.46	24	0.24	0.87
	# Switched	7	7	7	7	7	7	6	5	53					

In **Test Practice 1** shown in Table 3 , we see that while the two physicians have almost the same panel size and therefore the same capacity, 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 her patients would end up seeing Physician 28

most of the time. When the panels are redesigned, we see that their overflow values can be made even. Physician 4's overflow and utilization increase as she receives some of Physician 28'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 4), 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 switches 229 patients, but all of them are 0 comorbidity patients. Heuristic 2 switches only 62 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 Fig. 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.

Figure 6 shows the comparison of overflow frequency and utilization of the 4 physicians, with the 4 different case mixes resulting from 4 different approaches. The positive effects of the balancing the case mixes between the 4 physicians can clearly be observed from the figure. The heuristics and the capacity ratio algorithm balance the utilization and overflow frequency.

In **Test Practice 3** (Table 5), Physician 20 has more patients in her panel and also has more capacity compared to Physician 24. 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

Table 4: Results for Test Practice 2: 4 Physicians with Equal Capacity. 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.

		Comorbidity count													
		0	1	2	3	4	5	6	7	Panel Size	μ_j	σ_j	C_j	O_j	Utilization
Current	Phy 39	290	296	218	145	84	27	12	5	1077	14.73	14.47	17	0.28	0.87
	Phy 8	260	249	226	161	108	42	14	3	1063	15.54	15.26	17	0.35	0.91
	Phy 19	299	293	212	147	77	26	6	1	1062	14.10	13.86	17	0.22	0.83
	Phy 34	214	253	223	177	115	44	21	5	1053	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	1062	15.11	14.83	17	0.31	0.89
	Phy 8	266	272	220	157	96	34	14	3	1062	15.11	14.83	17	0.31	0.89
	Phy 19	265	273	219	158	96	35	13	4	1063	15.15	14.87	17	0.32	0.89
	Phy 34	266	274	220	158	96	36	12	4	1066	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	1144	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	1223	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	1094	15.13	14.86	17	0.31	0.89
	Phy 8	258	247	224	159	106	39	12	1	1046	15.14	14.87	17	0.31	0.89
	Phy 19	305	299	218	153	83	31	11	6	1106	15.11	14.84	17	0.31	0.89
	Phy 34	208	247	217	171	109	39	16	0	1007	15.15	14.87	17	0.32	0.89
	# Switched	8	8	8	8	8	8	7	7	62					

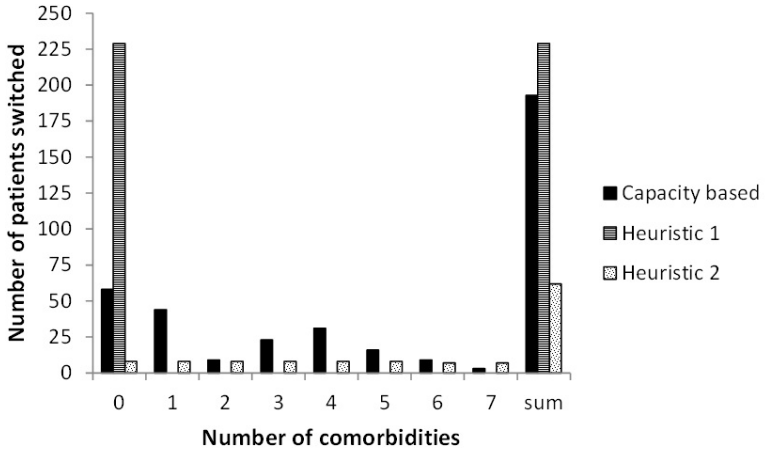
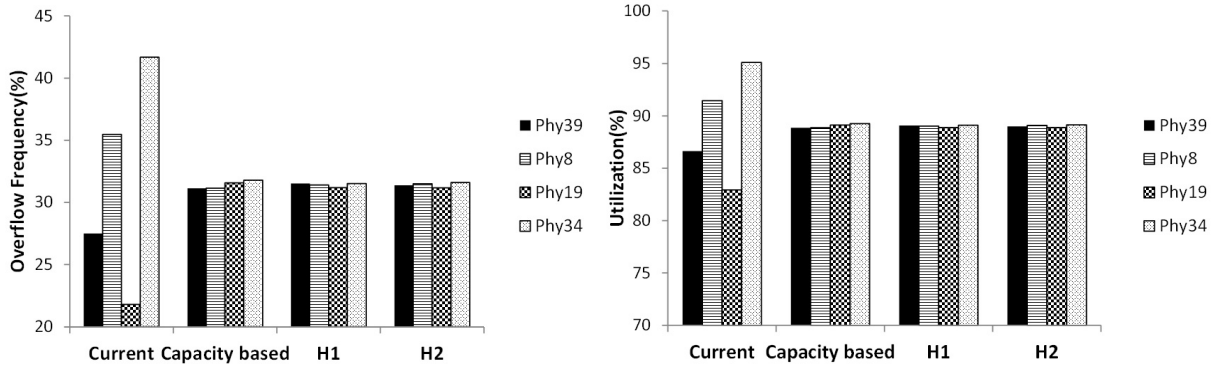


Figure 5: Comparison of the number of patients switched in Test Practice 2



(a) Overflow Frequency (b) Utilization

Figure 6: Results for Test Practice 2

Table 5: Results for Test Practice 3: 2 Physicians with unequal capacities. The Reference Overflow value, O_{ref} for this practice is 0.27. 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.

		Comorbidity Count													
		0	1	2	3	4	5	6	7	Panel Size	μ_j	σ_j	C_j	O_j	Utilization
Current	Phy 20	255	314	289	223	124	54	21	1	1281	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	1267	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	1109	18.26	17.90	21	0.26	0.87
	Phy 24	427	262	189	107	52	25	5	1	1068	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	1229	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					

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 and 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

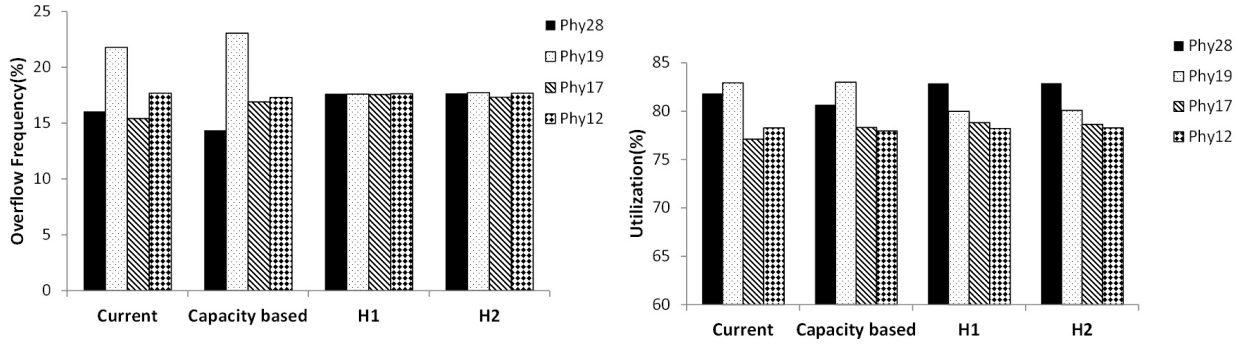
the overflow frequency does. Moreover Capacity Ratio switches 142 patients. Heuristic 1 and 2, on the other hand, produce overflows that are identical to the reference overflow (0.27). Heuristic 1 switches 172 healthy patients, while Heuristic 2 switches 52 patients in total from all the categories. Thus with regard to both overflow and patients switched, the H1 and H2 are better than Capacity Ratio.

Table 6: Results for Test Practice 4: 4 Physicians with unequal Capacities. 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.

		Comorbidity Count													
		0	1	2	3	4	5	6	7	Panel Size	μ_j	σ_j	C_j	O_j	Utilization
Current	Phy 28	418	385	299	211	111	32	10	2	1469	19.64	19.31	24	0.16	0.82
	Phy 19	299	293	212	147	77	26	6	1	1062	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	1465	19.36	19.03	24	0.14	0.81
	Phy 19	310	290	216	142	73	28	10	2	1071	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	1508	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	1479	19.89	19.56	24	0.18	0.83
	Phy 19	295	290	209	144	74	23	3	0	1038	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					

Finally as can be seen from Table 6, in **Test Practice 4**, we have four physicians with different panel sizes and capacity values. Notice, however, that the overflow and utilization values are not dramatically different to begin with (at least in relation to Test Practice 3), as can be seen from Fig. 7 as well. 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. (Should the practice be reluctant to change panels, our method also allows for practices to make the case that certain physicians should have extra capacity, or at least receive nurse practitioner or

physician assistant help in seeing patients.)



(a) Overflow Frequency

(b) Utilization

Figure 7: Results for Test Practice 4

As in Test Practice 4, 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. As before, Heuristic 1 affects only the healthy patients, while Heuristic 2 involves patients from all categories.

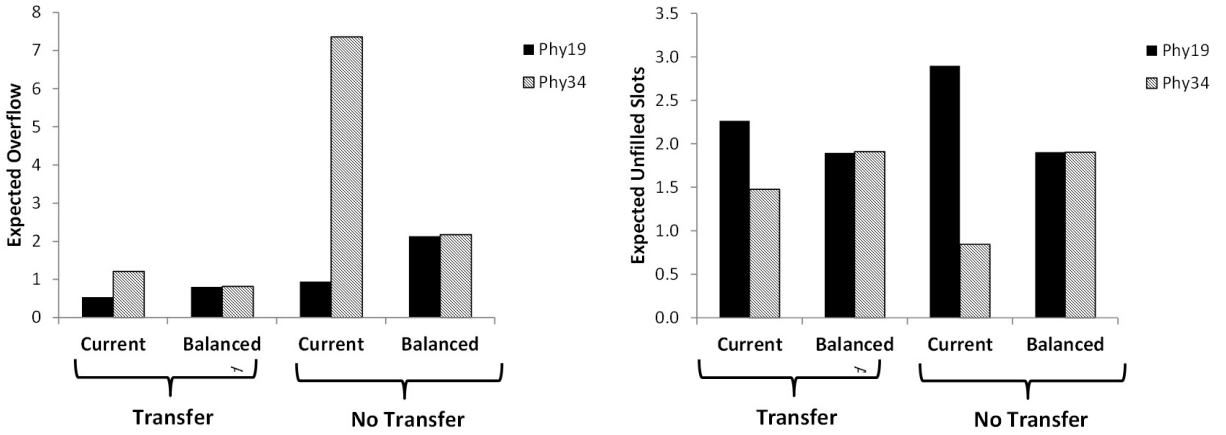
7.3 Impact on Other Measures

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 we do not have 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 8 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.



(a) Expected Overflow

(b) Expected Unfilled Slots

Figure 8: 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. 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 or the physicians need to work longer hours. We find similar results while testing other pairs of physicians, but in the interest of keeping the paper

concise we do not present these results.

8 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 how, in the long term, these imbalances in supply and demand can be minimized.

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. An important advantage of our approach is that all overflow frequency calculations can be easily carried out in an Excel spreadsheet. Panel redesign options can be easily tested, in a manner similar to Tables 4, 5, 6 and 7, and the least disruptive options of redesigning panels can be identified. Practices, however, should be cognizant of the fact physicians do differ from each other with regard to how frequently they prefer to see their patients. Any attempt at redesign should take into account such differences.

Should practices be reluctant to redesign panels, our framework will help to identify which physicians are capable of empanelling new patients or need additional capacity in the form of a nurse practitioner or a physician assistant. The results from section 7.3 suggest sharing patients between physicians reduces the overflow considerably, and is a viable alternative to panel redesign. Sharing reduces continuity, but it could be limited to same-day urgent requests for acute conditions or patients for whom quick access to a physician outweighs continuity. The overflow frequencies for the individual physicians can provide pointers as to which physicians should be paired together so that certain appointment requests can be flexibly shared. For example, the physician with the highest overflow frequency in the practice should ideally share his patients with the physician with

the lowest overflow frequency.

As mentioned earlier, our modeling approach is designed for aggregate level panel management decisions. We therefore do not consider a number of aspects relevant to appointment scheduling. For example, no-shows are not a part of our model. However, if a physician has low overflow frequency, backlogs will be small. Time to earliest available appointment will also be small and so will the no-show rates (see Green and Savin, 2008 for a more detailed discussion). Thus the improved design of panels can only reduce the impact of no-shows.

Also, in practice, non-urgent requests are booked in the future and urgent requests are addressed the same-day. In our model, all demand manifests itself for a given workday. While different appointment types are not explicitly considered, a high overflow frequency will be correlated with the inability to provide a non-urgent request with an appointment within a reasonable time frame.

Finally, capacity in our model refers to the number of appointment slots a physician has available. We have learned anecdotally from PCPs that they typically see 3 patients an hour (20 minutes per appointment). However patients with more comorbidities are likely to have longer appointments than healthy patients. We do not consider this reality. Our values for overflow frequency are therefore likely to be smaller than what those found in practice. However, in a relative sense, our approach will still correctly identify the imbalances in supply and demand across physicians.

Acknowledgements

This work was funded in part 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 AHRQ.

References

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

Chakraborty, S., Muthuraman, K., and Lawley, M., 2010, Sequential clinical scheduling with patient no-shows and general service time distributions, IIE Transactions, 42:5, 354-366.

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

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

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

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

LaGanga, L., and Lawrence, S., 2007, Clinic overbooking to improve patient access and increase provider productivity, Decision Sciences, 38:2, 2007.

Liu, N., D'Aunno, T., 2011. The Productivity and Cost-Efficiency of Models for Involving Nurse Practitioners in Primary Care: A Perspective from Queueing Analysis. Health Services Research, doi = 10.1111/j.1475-6773.2011.01343.x.

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

Murray M, Tantau C. Same-Day Appointments: Exploding the Access Paradigm. Fam Pract Manag. 2000;7(8): 45-50. Available at http://www.aafp.org/fpm/20000900/45same.html. Accessed April 28, 2008.

Murray, M., Davies, M., and Boushon, B., 2007, Panel size: How many patients can one doctor manage? Family Practice Management, available at: http://www.aafp.org/fpm/2007/0400/p44.html

Muthuraman, K., and Lawley, M., 2008, Stochastic overbooking model for outpatient clinical scheduling with no shows, IIE Transactions, 40 (9), 2008.

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

Nutting P, Meredith A. Goodwin, Flocke S, Zyzanski S, Kurt C. (2003) Continuity of Primary Care: To Whom Does It Matter and When? Ann Fam Med 1:149-155

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

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

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

Starfield B, Weiner J, Mumford L, Steinwachs D. Ambulatory care groups: a categorization of diagnoses for research and management. Health Serv Res. 1991;26:5374.

Wang, W., and Gupta, D., 2011, Adaptive appointment systems with patient preferences, Manufacturing and Service Operations Management, 13(3), 373-389.