Primary Care Practice Design under Case Mix:
Joint Consideration of Access to Care and Continuity of Care

Nan Liu, Ph.D.
Department of Health Policy and Management
Mailman School of Public Health
Columbia University
nl2320@columbia.edu

Asli Ozen
Department of Mechanical and Industrial Engineering
University of Massachusetts at Amherst
ozen@ecs.umass.edu

Hari J. Balasubramanian, Ph.D.
Department of Mechanical and Industrial Engineering
University of Massachusetts at Amherst
hbalasubraman@ecs.umass.edu
Abstract

Objective. To develop methods for evaluating access to care and continuity of care in commonly-used primary care delivery models adjusted for case-mixes; and to study how these two system performance measures change under panel (re)design and provider capacity pooling (i.e., patient sharing).

Data Sources and Study Design. The access to care is evaluated by the patient appointment delay, and the continuity of care is defined as the percentage of time a patient seen by her own primary care provider. Queuing analysis is performed for three commonly-used primary care practice designs, i.e., dedicated service design, partial pooling design and complete pooling design. Model parameters are estimated using data from Primary Care Internal Medicine Practice at the Mayo Clinic in Rochester, Minnesota.

Principal Findings. Case-mix not only affects, on average, how often patients need healthcare, but also how much time they need during each visit. Based on our data examples, panel redesign alone can improve the overall access of care by 75% without capacity pooling. Letting providers share patients with zero comorbidity, who only contribute to 11% of the total visits, can significantly improve access to care by 85% while continuity of care only decreases by 5%. More importantly, most improvement in access to care that can be achieved by patient sharing come from just sharing zero comorbidity patients.

Conclusions. Case-mix is a crucial factor to consider in primary care practice design. Both panel (re)design and capacity pooling can be effective strategies for primary care practice improvement. In particular, even a little capacity pooling can make a big difference.

Key Words. Practice design, primary care, access to care, continuity of care, queuing theory
**Introduction**

Primary care can prevent illness, improve health outcomes and reduce mortality (Starfield, Shi, and Macinko 2005). Providing communities with high-quality primary care is set as priority in many countries’ health care agenda. To build a successful primary care delivery system, access to care and continuity of care are two crucial cornerstones.

The concept *access* to care has a broad meaning (Aday, and Andersen 1974). Some researchers equate it to the availability of health system resources in an area, while others relate it to characteristics of the population, e.g., incomes, insurance coverage and attitudes toward medical care. Simply put, accessibility to primary care can be thought of as how easy for a patient to receive primary care when he/she needs it. Previous research has developed quantitative measures of accessibility, among which one of the most important measures is the appointment delay (Balasubramanian et al. 2010). The appointment delay refers to the time between a patient’s call for an appointment and her actual appointment date. The shorter the appointment delay, the earlier a patient can receive the medical service, and hence the more accessible the primary care service is.

The other crucial cornerstone for a successful primary care system is *continuity* of care. Saultz (2003) summarizes this concept in a hierarchical way: 1) informational continuity means patient information is transferred when she sees another provider; 2) longitudinal continuity of care refers to that patients receive most care from the same provider; 3) interpersonal continuity implies an ongoing relationship and trust exist between each patient and a personal physician. The most commonly-used concept for continuity of care is the longitudinal continuity of care, which is usually defined as the percentage of time that the patient is seen by her own primary care provider (Bice, and Boxerman 1977).

Ideally, a primary care practice would like to improve both access to care and continuity of care offered to its patients, but these two goals are often conflicting (Ozen, and Balasubramanian 2012). For example, many primary clinics aim to improve access to care and reduce appointment delays by implementing
In doing so, they try to provide a majority, if not all of, the patients with same-day appointments. To build up enough service capacity, they may choose to form practice teams with multiple providers, say two to three, sharing their patients. Though this pooling strategy does improve service capacity, it may lead to loss of continuity of care, because there is no guarantee that patients will always be seen by their own providers unless the patients’ situations dictate that or their own providers happen to have open slots during their visit. Indeed, among those Open Access trials that failed, many are because the loss of continuity as a price to pay for speedy access is just too high (Phan, and Brown 2009).

In the design of a primary care practice, case-mix is another crucial factor that needs to be accounted for. Case-mix refers to the type of patients served by a practice. Because different types of patients may have different visit frequencies as well as various demand for providers’ consultation time, case-mix directly influences both the “demand” and “supply” side of a primary care practice. For example, Potts et al. (2011) have calculated the “disease burden” of a physician’s panel by using the risk categories set for chronic diagnoses in order to decide on the support the physician needs from nurse practitioners (NPs). The goal of this paper is to develop methodologies to quantify and evaluate access to care and continuity of care in primary care practices, taking into account the impact of case-mixes.

Adding more patients in a physician’s panel increases the physician’s workload, and thus leads to longer appointment delays. The panel size here is the number of patients that a physician (group) is held accountable for (Murray, Davies, and Boushon 2007). Given the same panel size, a physician’s workload is larger if patient acuity level is higher because patients visit the clinic more often and each visit might also take longer time (Knox, and Britt 2004; Roos, Carriere, and Friesen 1998). To reduce appointment delays and improve practice, there are two major operational strategies. One is to take the advantage of economics of scales by forming a practice team and pooling service capacity together, but there would be a loss of continuity of care. Another strategy is via panel redesign, i.e., to reallocate patients to providers’ panels according to patient needs and individual provider’s service capacity so that the whole care team is
used more effectively (Balasubramanian et al. 2010); but reallocating patients may not be an easy task as it takes time and effort, and involves changing existing patient-PCP relationships. The qualitative effect of these two strategies is clear. The question, however, is how to quantify these effects \textit{ex-ante} and also adjusted for case-mixes.

In this article, we will use queueing theory to develop methods that enable us to conduct such quantitative analysis, which should provide useful information for practice change. Queueing theory concerns the study of wait lines (Gross, and Harris 1985). It can translate customer arrival characteristics and service patterns into measures of waiting experienced by the customers, e.g., average waiting time and the chance that customers will be delayed in the service process. In this paper, we measure access to care by appointment delays (i.e., wait time) and operationalize continuity of care by the percentage of patients who see their own primary care providers. Since we are interested in studying the relationship among panel size (which, to be discussed shortly, is directly related to patient appointment demand), provider service capacity and patient appointment delays, queueing theory is an ideal tool.

We consider three typical practice designs used in primary care. The first design is a \textit{dedicated service} model where patients only see their own providers. This design can also be viewed as a solo-practitioner service where the provider serves her own patients only. The second and third designs are both group practice models involving multiple providers working in the same team. The difference is that, in the second design, some patients have their dedicated providers while others see anyone available. In the third design, patients see any available provider. For each of these designs, we develop corresponding queuing models and derive performance measures for both access to care and continuity of care. We use data collected from the Mayo Clinic to populate our models and discuss how these measures change among designs. All these measures can be computed via closed-form formulas, and they can be easily evaluated using spreadsheet tools like Excel or even just calculators.
With the recent passage of the Patient Protection and Affordable Care Act, more than 30 million Americans are expected to gain health care coverage in the U.S. However, many areas in the nation are facing severe shortage in primary care workforce (U.S. Department of Health and Human Services 2009). Compounding the increase in patient volumes and the shortage of primary care workforce is the aging population and the epidemic of chronic diseases, which will likely give rise to more patients with multiple comorbidities requiring more physician time and resources. To reform primary care delivery in the U.S., many practices are engaged in transforming into Patient Centered Medical Homes (Patient-Centered Primary Care Collaborative 2012), one of the most important objectives being to form a coordinated and integrated care team that provides patient centered care. Yet, there is a lack of scientific and systematic methods that can inform the formation of such teams and the allocation of workload among different team members to achieve the best outcome. Our study provides a tool to assess the supply demand dynamics, conduct capacity planning and practice design for primary care teams.

**Methods**

**The Models**

We use queueing models to describe the operations in different primary care practice designs. As an example, consider a single physician’s practice. Patients in the physician’s panel call and request an appointment with the physician. To better understand our model, suppose for now that patients will take the earliest appointment slot available and the service time for each patient is deterministic with a common length (we will relax this assumption later). Thus the provider knows exactly when to schedule this patient upon her request. In particular, incoming appointment requests are registered to the provider’s work schedule in the order as they arrive. The provider’s schedule is the *queue* in our models. The queue here is not the waiting line of patients who physically wait in the clinic, but rather a *virtual* list for those who have not yet been seen by the provider.
During office hours, the provider sees patients and shortens the queue. When the practice is closed, no one joins or leaves the schedule, i.e., the queue remains intact. If we remove the non-office hours from the time horizon, we can view the provider’s work schedule as a continuous queueing process, where jumps and drops in this queue correspond to the arrival of an appointment request and the service completion of an actual patient, respectively.

In reality, patient preferences and punctuality, type of appointments (prescheduled versus same-day) and no-shows play an important role in practice operations. Some factors such as no-shows and types of appointments can be incorporated into queuing models, but require complex analysis; other factors such as patient preferences and punctuality are better addressed by more sophisticated frameworks, e.g., Gupta and Denton (2008) and Wang and Gupta (2011). These frameworks consider *intra-day* operations, while our goal is to evaluate the access to care and continuity of care in primary care practices *across days*. To that extent, our analysis is on a more strategic level, and thus we choose not to incorporate too many intra-day scheduling details in our models. Omitting these details leads to much more accessible mathematical frameworks that also provide quantifiable outcome measures for practical use. More importantly, several recent studies support the use of such simple queueing models in setups like ours (Green, and Savin 2008; Liu, and D'Anno 2012). In particular, by comparing with realistic simulation models that consider patient preference and other scheduling details, Green and Savin (2008) show that simple queueing models can yield relatively accurate estimate for panel sizes.

One interesting and innovative feature of our queuing models is that they can account for case-mixes. Case-mix refers to the type of patients in a physician’s panel, and it can be characterized by various attributes, such as age and gender and the chronic conditions afflicting the patient (Balasubramanian et al. 2010). The idea is to group patients into “categories,” and within each category patients have similar demand pattern and needs for providers’ time and resources. Using data from the Primary Care Internal Medicine Practice (PCIM) at the Mayo Clinic in Rochester, Minnesota, we will discuss how to categorize patients shortly.
We now describe the specifics of our models. The appointment rate of a patient is assumed to follow a Poisson Process with a rate $\lambda_i^0$ per day, for patient category $i$. If there are $N_i$ patients in category $i$, then the appointment rate from this category is $\lambda_i = N_i \times \lambda_i^0$. If there are $M$ patient categories, then the panel size is the number of patients in all categories, i.e., $N_1 + N_2 + \cdots + N_M$; and the joint arrival process is also a Poisson process whose arrival rate is the sum of those of its constituting arrival streams, i.e., $\lambda_1 + \lambda_2 + \cdots + \lambda_M$. The Poisson process is a widely used customer arrival model (Gross, and Harris 1985). It is especially reasonable in our outpatient primary care setting as patients requests usually arrive one at a time, and they can be treated independent of one another.

Since patients in different categories may require different amounts of service time, the service time of a physician also needs to be adjusted for the case-mix. For instance, a physician with more of higher acuity level patients in her panel should have a lower number of appointments per day to accommodate for longer service times. To adjust for case-mix, we calculate the average appointment duration for a physician by taking the weighted average of the service times from different categories of patients, where the weights correspond to the proportions of each patient category.

With the above model description in mind, we proceed to discussing the three practice designs we will investigate in this article. In Figure 1, the Greek letters $\lambda$ and $\mu$ represent the patient arrival rate and provider service rate, respectively.

[Insert Figure 1 here]

The first design is a dedicated service model where patients always see their own provider. This design can also be used in a multi-provider practice, where each provider practices as an independent single physician. The second design is a group service model with partial pooling of provider service capacity, where some patients have dedicated providers while others are flexible. In particular, dedicated patients to provider 1 have arrival rate $\lambda_1$ and they will wait as long as provider 1 is busy. Similarly, dedicated patients to provider 2 arrive at rate $\lambda_3$ and they will wait as long as provider 2 is busy. Another stream of
patients arriving at rate $\lambda_2$ are flexible patients; they will see any available provider and they wait only if both providers are busy. The third design is a group service model with complete provider capacity pooling, where patients will see any provider who is available.

Finally, we relax our service time assumption in our analysis. Recall that when describing our models, we suppose that the provider service time is deterministic. Under this service time assumption, we typically do not have closed-form expressions for the performance measures that we are interested in. For better tractability, we relax this assumption and assume that service times are random, which, in particular, follow exponential distributions. On the one hand, random service times bring some variability into the service process and seem to resemble practical settings better. On the other hand, queues with exponential service times are usually easier to analyze and often have closed-form expressions for their performance measures. Furthermore, previous studies show that variability in provider service time typically does not have a significant impact on the productivity of a practice, while the mean service time is a more important determinant (Liu, and D'Aunno 2012; Liu, Finkelstein, and Poghosyan 2012). For all reasons above, we will focus our analysis on queues with exponentially distributed service times.

**Data and Model Parameters**

We analyze the patient population (around 20,000 patients) empanelled at the Primary Care Internal Medicine Practice (PCIM) of the Mayo Clinic in Rochester, Minnesota. Our data constitute patient visits over three years 2003-06 to 39 physicians at PCIM. Detailed analysis on patient demand rates has been reported in an earlier study by Ozen and Balasubramanian (2012). We recapitulate the key results here for convenience. Their analysis reveals that comorbidity count (CC) is the strongest predictor for patient demand rate. Thus we divide patients based on the number of comorbidities they had. In all, there are 8 patient categories as patients with more than 7 comorbidities were extremely rare. Our categorization is consistent with earlier literature (Naessens et al. 2011) as well as practice guidelines set by governments.
(Minnesota Department of Health 2012). However, we should note that other categorization rules can also be used if deemed appropriate.

To estimate the daily demand rate of a patient from each category, we calculate the probability that a patient from a certain category will request an appointment on a given day, i.e., the total visits over a year for that category divided by the total patients in the category times the total workdays in a year. The appointment request for a day for one given category is simply the multiplication of this probability and the number of patients in that category. The total daily appointment request rate from a physician’s panel is the sum of daily appointment request rates from all categories.

To estimate the physician service rate, we use the idea of adjusted service times based on case-mix mentioned above. Since no time stamps data are available for us to estimate the length of provider consultation time, we set the following length based on the experiences of PCIM physicians. These lengths also seem to be consistent with those reported in the literature (Mechanic, McAlpine, and Rosenthal 2001). Patients with 0, 1 and 2 comorbidity count category require a 20 minute visit on average, whereas those that belong to higher comorbidity count categories require a 40 minute visit on average. Thus, we calculate the appointment duration for a physician by taking the weighted average of the service times. That is, we multiply the proportion of the patients that belong to the 0, 1 and 2 comorbidity count category with the required average appointment time (20 minutes) and add with the product of the proportion of those with higher comorbidity counts and the 40 minutes average. For example, if a physician has 50% of patients with lower comorbidity counts and 50% with higher, then the average appointment duration for this physician is $0.5 \times 20 + 0.5 \times 40 = 30$ minutes. Assuming eight hour work time every day, this physician can see on average 16 patients (= 8 hours ÷ 30 minutes) daily.

**Model Analysis**

Under our model assumptions, the first practice design becomes a simple M/M/1 queue and the third design is an M/M/2 queue; see Gross and Harris (1985) for the notation and analysis of such models. The
second design, however, is difficult to analyze. It is not amenable to the standard balance equation approach (Kulkarni 1995), due to the inclusion of flexible patients. Recently, Guo and Hassin (2012) study a two-server queuing system where some customers may place duplicate orders at both servers but will immediately withdraw one when they receive services from the other. They provide closed-form formulas to analyze such a system. A close examination of their work reveals that their model is equivalent to our second practice design where the flexible patients play the role of customers placing duplicate orders in the queueing system. Thus we can adopt their formulas to analyze our second design.

One of the primary benefits of using a queuing model is that it produces useful steady-state outputs. In our paper we only make use of only some of them: utilization of the physician, probability that a patient will be seen by her own provider (continuity of care measure) and average waiting time for the patients (access to care measure). All these measures can be calculated using closed-form formulas reported in the literature discussed above.

**Results**

**Impact of case-mix on provider utilization**

System utilization is an important measure for the workload placed on a service system; it is evaluated as the ratio of patient daily request rate and the provider daily service rate. A higher utilization level indicates a heavier workload, i.e., more percentage of time being spent by providers in seeing patients; however, it also comes with more congestion and longer patient wait. More importantly, as the system utilization increases, the customer wait does not increase linearly but rather exponentially (Green 2006). That is, when the system utilization is high, even a small disturbance, such as a slight increase in patient demand or drop in service rate, can significantly increase patient wait. Therefore, the system utilization is a crucial measure to monitor and control for in order to balance the utilization and congestion in a service system. In this section we discuss how case-mix can affect this important measure.
To illustrate, we use Mayo comorbidity count visit rates (see Table 1) to create seven hypothetical panels with the same system utilization 93.5% in Table 2. Recall that patients with different acuity may have different appointment demand rates, and they may also require different length of service times. Thus, it is not too surprising to observe that although these panels have the same system utilizations, their sizes are dramatically different due to different case-mixes. The largest panel is panel 5, in which a majority of the patients have no more than 4 comorbidities; in contrast, panel 4 is the smallest panel whose size is even smaller than a quarter of panel 5, and it is predominately occupied by patients with more than 4 comorbidities.

[Insert Table 1 here]

[Insert Table 2 here]

As mentioned above, one way to balance workload and improve practice is via panel redesign. That is, reassigning patients across panels in the long term to achieve identical workload proportions and thereby use existing capacity in the most efficient way possible (Balasubramanian et al. 2010). Here we use two real physician panels from Mayo Clinic Primary Care Internal Medicine (PCIM) to demonstrate the effect. The initial panel size and case-mixes are shown in Table 3a. Physicians 1 and 2 differ in their case-mixes, panel sizes, arrival and service rates and therefore utilizations. Physician 1 has a utilization of 94.8%, while physician 2 has a utilization of 99.6%. These differences can occur in practice due to reasons such as physician seniority, physician and patient preferences. As a result, patients of physician 2 will experience poorer access compared to those of physician 1. Now, what if the panels could be redesigned such that these two physicians had similar case-mixes? In this case, we balance panels simply by dividing the patients from each comorbidity count category equally among the two physicians. In doing so, the utilization of each physician equals at 97.2% (see Table 3b). In the next section, we will discuss how panel redesign affects the access to care and continuity of care measures.

[Insert Table 3 here]
Comparison of practice designs under different case-mixes

In this section, we compare the three practice designs introduced before (see Figure 1). Recall that in Design 1, the two physicians practice independently; while in Design 3, they form a provider team and share all their patients. In the former case we expect to see long waiting times (i.e., poor access) especially for a highly utilized physician, but the continuity of care is perfect for all patients. However, in the latter case, we expect the waiting times to decrease but continuity of care is no longer perfect. The patients may see one of two providers and hence continuity is 0.5 as opposed to 1 in the first case (0.5 means that patients will be seen by their own providers with 50% chance).

Between these two extremes of best continuity and best access is the partial pooling case, i.e., Design 2, where the providers form a team and share a subgroup of patients. Care for this group of patients could be provided by either provider; continuity for the shared patients is therefore 0.5. But each provider also retains a certain number of dedicated patients for whom continuity is 1. Thus, based on the number of patients shared and the number of patients dedicated, we can calculate an overall (weighted) continuity of care measure. If 50% of the total visits are shared by the two providers, and 25% are dedicated with each of the physicians, then the weighted continuity measure is $1 \times 0.25 + 1 \times 0.25 + 0.5 \times 0.5 = 0.75$.

In practice, it makes sense to provide greater levels of continuity to patients with multiple chronic conditions. Reid et al. (2010) and Coleman et al. (2010) discuss that in Group Health Practice during the reassignment of panels, when physicians were given the chance to choose patients to keep, they preferred the elderly and sicker patients. Compared to relatively healthy patients, these patients need a stronger bond with their PCP for better management of their health conditions.

To start with, we allow the providers to share only patients with zero comorbidity count, i.e., $CC = 0$, who are apparently healthier patients in the panel; all other patients still remain dedicated to their respective providers. We calculate access and continuity measures for this setting. Next we allow providers to share patients with $CC$ up to 1, thereby increasing the number of shared patients and again calculate access and
continuity measures. We proceed in the same way until all patients are shared by the two providers; this becomes Design 3. In our data, since CC range from 0 to 7, we have a total of 9 cases, including the two extreme cases (i.e., Designs 1 and 3).

Table 4 provides the waiting time and continuity measures for each of these 9 cases, for baseline panels and panels balanced via redesign introduced in the last section. Figure 2 summarizes the changes in access and continuity of care provided across all 9 cases for both the baseline and balanced panels. In Table 4, $W_1$ is the average appointment delay of patients dedicated to Provider 1; $W_2$ is the average appointment delay of patients shared by both providers; $W_3$ is the average appointment delay of patients dedicated to Provider 2. Clearly in the M/M/1 case, since no patients are shared, $W_2$ does not exist.

Similarly, since no patients are dedicated in the M/M/2 case, $W_1$ and $W_3$ do not exist. $W$ is a consolidated access measure for all patients, calculated as the weighted average of $W_1$, $W_2$ and $W_3$, where the weights are based on the proportion of the arrival rates for the dedicated and shared streams ($\lambda_1$, $\lambda_2$, $\lambda_3$). The unit of $W_1$, $W_2$ and $W_3$ is days.

[Insert Table 4 here]

In the baseline dedicated case (Design 1), Provider 2’s patients have average appointment delay of 13.8 days (see $W_3$), while Provider 1’s patients have an average delay of only 1.0 days (see $W_1$). This dramatic difference is due to the imbalance in the case-mixes of these two physicians, which results in 99.6% utilization for Provider 2 and 94.8% utilization for Provider 1, as discussed in the last section. This also signifies our earlier point that when utilization level is high, a slight increase will lead to dramatic increase in patient wait. Now, if we look at all patients, the average delay is 7.3 days in this case and the continuity of care is perfect. However, if we were able to redesign the panels of these two physicians and balance their workload, the utilization of both providers equals at 97.2% and the average appointment delay for all patients is reduced to 1.8 days (see Balanced Dedicated case). This is a 75% improvement in access to care.
Next, consider the Baseline panels when 0 Comorbidity Count (CC) patients – or apparently healthy patients – are shared by the 2-physician team. The access improves significantly for all patients (see $W_1$, $W_2$ and $W_3$), with the overall average delay reduced from 7.3 days to 1.1 days (85% reduction).

Interestingly, the overall continuity measure drops only marginally – from 1 to 0.95. Thus for a 5% drop in continuity of care in relatively healthy patients, we get an 85% improvement in access to care. For the balanced panels after panel redesign, we obtain similar findings. On top of the benefits generated from panel redesign, pooling 0 CC patients further reduces overall patient waiting time by additional 44% (from 1.8 days to 1.0 days) with only 5% drop in continuity measure.

A closer examination of Figure 2 reveals that when more patients are shared by the two physicians, access measures improve, but the improvement is not as significant as going from the dedicated to the 0 CC shared case. Furthermore, as more patients are shared, the Baseline and Balanced cases tend to get similar. When all patients are shared, they converge to Design 3 and have the same access and continuity measures.

Discussion

Our study is among the first to develop case-mix adjusted methods to evaluate continuity of care and access to care for primary care delivery models. We consider three commonly-used practice designs, namely dedicated service design, partial pooling design and complete pooling design. Our study highlights the importance of considering case-mix in primary care practice design. Case-mix not only affects, on average, how frequently patients need healthcare services, but also influences how much time/resources that a patient needs for each visit. Many primary care providers in the U.S. have panel sizes exceeding 2,000 patients regardless of the case-mix (Alexander, Kurlander, and Wynia 2005). Green et al. (2013) even look at alternative methods of delivering care on the supply side in order to increase the nationwide panel size of 2,500 patients to 5,000 patients due to the soon to increase demand in primary care. Our results suggest that such a seemingly one-size-fit-all approach does not work.
Providers can easily feel overwhelmed if their panels contain a relatively large number of patients with complicated conditions. It is crucial to take case-mix into account.

A practice typically has two strategies to improve access to care with available capacity. One is to create provider teams and pool service capacity on a certain group of patients. This strategy seems to be spreading fast in the U.S. as more and more primary care practices shift from solo-practice to group practice and physicians cover each other’s work in a care team. Indeed, the share of solo practices fell to 18 percent by 2008 from 44 percent in 1986, according to the AAFP’s 2008 member survey (Harris 2011). One important question that arises from such a practice shift is what kind of patients can be “shared.” Intuitively, patients who are relatively healthy can be shared because their cases are relatively simple and easy to handle. In our data examples, we use comorbidity counts (CC) as a measure for patient health status and the provider can choose who to share based on their CCs. We find that, letting providers to share patients with 0 CC, who only contribute to 11% of the total visits, can significantly improve access to care by 85% but continuity of care only decreases by 5%. More importantly, most of the benefits that can be generated by patient sharing come from just sharing zero comorbidities patients. In other words, a little flexibility can go a long way. Indeed, such ideas of using flexibility have been discussed in other non-healthcare contexts such as manufacturing (Jordan, and Graves 1995), and are shown to be effective in improving system efficiency.

The other strategy often used by practice is via panel redesign to balance workload among physicians. In our data examples, the two physicians have imbalanced panels at baseline. Panel redesign alone can improve the overall access of care by 75%. However, when a practice tries to redesign existing panels, it usually involves much effort related to redirecting and re-empanelling the patients; and such changes can take a long time and much effort (Balasubramanian et al. 2010). The reassignment experience at the Group Health practice in Seattle also illustrates these challenges (Coleman et al. 2010). Instead, if panels were to be designed proactively in the early phase of empanelling new patients rather than to be redesigned reactively after panels have been formed, the work might have been much easier and effective.
There are other strategies that a primary care practice can use to improve access to care. For instance, some practices choose to delegate certain tasks, e.g., preventive care and chronic care work, to non-physician members of the care team. A recent study examines how such task delegation affects the choice of panel size (Altschuler et al. 2012). In particular, it considers how much time a physician can save by task delegation, and then simply equates the physician available time with the time consumed by patients to derive the reasonable panel size, which usually gets expanded post task delegation. Altschuler et al. (2012) do not, however, consider the impact of such system changes on continuity or access to care. In contrast, our modeling framework can achieve both ends, i.e., considering task delegation and evaluating continuity and access to care. To do so, we just need to include only patient visits to the physician in our model analysis.

Our modeling framework is developed using queuing theory. It provides general and yet easy-to-use tools to model and analyze service systems when customer wait is an important focus of the problem. Despite its many merits, this method also has a few limitations in modeling a primary care practice. In particular, we assume that the service process is continuously running and “ignore” weekends when most practices are closed. We also assume that patients are always assigned to the earliest appointment slot available although it may not be the case in reality. Thus the appointment delay estimates generated by queueing models may underestimate the actual patient wait time. However, our analysis depicts how the appointment delay varies across different practice designs (see Figure 2), thereby enabling us to evaluate the relative changes in access to care. These relative changes perhaps provide more useful information compared to the absolute values of appointment delays when comparing practice designs.

Our study points to several future research directions. First, we use comorbidity count as a criterion for patient sharing. It is important for the clinical community to study how patient sharing affects health outcomes and develop guidelines for it, i.e., who to share or when to share. Second, it will be interesting to develop simulation models (Law, and Kelton 1991) rather than analytic models (like ours) to study different primary care practice designs. The advantage of a simulation model is that it can incorporate
more details and represent the reality better; however, it is usually developed based on a single facility, making its results difficult to generalize. Third, our models only consider primary care providers, e.g., physicians and nurse practitioners. There are many other important medical professional in a care team, e.g., medical assistants. It will be interesting to develop more comprehensive models to study the dynamics and patient flow through the whole care team. Last but not least, we only consider the effect of panel redesign. How to proactively develop panels in the early phase of building up a group practice remains an unexplored and yet very important research topic.

Reference


Figure 1. Practice Designs.
Table 1. Arrival rate per patient per day for each category.

<table>
<thead>
<tr>
<th>Comorbidity count</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.006</td>
<td>0.011</td>
<td>0.015</td>
<td>0.02</td>
<td>0.026</td>
<td>0.03</td>
<td>0.038</td>
<td>0.041</td>
</tr>
</tbody>
</table>
Table 2. Example of 7 hypothetical panels, with varying case-mixes, panel sizes, daily request rates and service rates. All the 7 panels have the same utilization of 93.5%.

<table>
<thead>
<tr>
<th>Panels</th>
<th>Comorbidity count</th>
<th>Panel Size</th>
<th>Daily Arrival Rate</th>
<th>Daily Service Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>160 150 226 200 142 42 14 3</td>
<td>937</td>
<td>15.53</td>
<td>16.60</td>
</tr>
<tr>
<td>2</td>
<td>100 100 226 161 108 40 50 20</td>
<td>805</td>
<td>14.98</td>
<td>16.05</td>
</tr>
<tr>
<td>3</td>
<td>50 50 140 140 5 30 85 90</td>
<td>590</td>
<td>13.67</td>
<td>14.57</td>
</tr>
<tr>
<td>4</td>
<td>5 5 5 5 25 50 100 125</td>
<td>320</td>
<td>11.34</td>
<td>12.17</td>
</tr>
<tr>
<td>5</td>
<td>425 350 275 200 110 13 2 1</td>
<td>1376</td>
<td>17.80</td>
<td>19.05</td>
</tr>
<tr>
<td>6</td>
<td>5 5 180 150 100 30 64 45</td>
<td>579</td>
<td>13.53</td>
<td>14.45</td>
</tr>
<tr>
<td>7</td>
<td>300 275 250 184 108 42 14 3</td>
<td>1176</td>
<td>16.89</td>
<td>18.05</td>
</tr>
</tbody>
</table>
Table 3. Case-mixes of physicians 1 and 2.

Part (a) Initial/Baseline panels

<table>
<thead>
<tr>
<th>Comorbidity count</th>
<th>Panel Size</th>
<th>Request Rate</th>
<th>Service Rate</th>
<th>Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>380</td>
<td>17.91</td>
<td>18.91</td>
<td>94.7%</td>
</tr>
<tr>
<td>1</td>
<td>372</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>269</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>187</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>610</td>
<td>17.38</td>
<td>17.45</td>
<td>99.6%</td>
</tr>
</tbody>
</table>

Part (b) Balanced Panels (after redesign)

<table>
<thead>
<tr>
<th>Comorbidity count</th>
<th>Panel Size</th>
<th>Request Rate</th>
<th>Service Rate</th>
<th>Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>305</td>
<td>17.68</td>
<td>18.18</td>
<td>97.2%</td>
</tr>
<tr>
<td>1</td>
<td>322</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>189</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>111</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balanced</td>
<td>1241</td>
<td>17.68</td>
<td>18.18</td>
<td>97.2%</td>
</tr>
</tbody>
</table>
### Table 4: Design comparison under the baseline and balanced panels.

<table>
<thead>
<tr>
<th>Comorbidity groups shared</th>
<th>% pooled</th>
<th>Weighted Continuity</th>
<th>Baseline Panels</th>
<th>Balanced Panels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>W₁  W₂  W₃  W</td>
<td>W₁  W₂  W₃  W</td>
</tr>
<tr>
<td>None (Dedicated)</td>
<td>0%</td>
<td>1.00</td>
<td>1.0 0.0 13.8</td>
<td>7.3 1.8 0.0 1.8</td>
</tr>
<tr>
<td>0</td>
<td>11%</td>
<td>0.95</td>
<td>1.0 0.9 1.2</td>
<td>1.1 0.9 1.1 1.0</td>
</tr>
<tr>
<td>0-1</td>
<td>30%</td>
<td>0.85</td>
<td>0.9 0.8 1.0</td>
<td>0.9 1.0 0.9 0.9</td>
</tr>
<tr>
<td>0-2</td>
<td>52%</td>
<td>0.74</td>
<td>0.9 0.8 0.9</td>
<td>0.9 0.9 0.9 0.9</td>
</tr>
<tr>
<td>0-3</td>
<td>73%</td>
<td>0.64</td>
<td>0.9 0.8 0.9</td>
<td>0.9 0.9 0.9 0.9</td>
</tr>
<tr>
<td>0-4</td>
<td>89%</td>
<td>0.55</td>
<td>0.9 0.9 0.9</td>
<td>0.9 0.9 0.9 0.9</td>
</tr>
<tr>
<td>0-5</td>
<td>96%</td>
<td>0.52</td>
<td>0.9 0.9 0.9</td>
<td>0.9 0.9 0.9 0.9</td>
</tr>
<tr>
<td>0-6</td>
<td>99%</td>
<td>0.50</td>
<td>0.9 0.9 0.9</td>
<td>0.9 0.9 0.9 0.9</td>
</tr>
<tr>
<td>0-7 (Pooled)</td>
<td>100%</td>
<td>0.50</td>
<td>0.0 0.9 0.0</td>
<td>0.9 0.0 0.9 0.9</td>
</tr>
</tbody>
</table>
Figure 2: The impact of partial pooling on access to care and continuity of care for both baseline and balanced panels. The x-axis ranges from the fully dedicated case (Design 1) to fully-pooled case (Design 3).