The Productivity and Cost-Efficiency of Models for Involving Nurse Practitioners in Primary Care: A Perspective from Queueing Analysis

Nan Liu and Thomas D'Aunno

Objective. To develop simple stylized models for evaluating the productivity and cost-efficiencies of different practice models to involve nurse practitioners (NPs) in primary care, and in particular to generate insights on what affects the performance of these models and how.

Data Sources and Study Design. The productivity of a practice model is defined as the maximum number of patients that can be accounted for by the model under a given timeliness-to-care requirement; cost-efficiency is measured by the corresponding annual cost per patient in that model. Appropriate queueing analysis is conducted to generate formulas and values for these two performance measures. Model parameters for the analysis are extracted from the previous literature and survey reports. Sensitivity analysis is conducted to investigate the model performance under different scenarios and to verify the robustness of findings.

Principal Findings. Employing an NP, whose salary is usually lower than a primary care physician, may not be cost-efficient, in particular when the NP's capacity is under-utilized. Besides provider service rates, workload allocation among providers is one of the most important determinants for the cost-efficiency of a practice model involving NPs. Capacity pooling among providers could be a helpful strategy to improve efficiency in care delivery.

Conclusions. The productivity and cost-efficiency of a practice model depend heavily on how providers organize their work and a variety of other factors related to the practice environment. Queueing theory provides useful tools to take into account these factors in making strategic decisions on staffing and panel size selection for a practice model.

Key Words. Practice models, nurse practitioner, primary care, queueing theory

The Institute of Medicine has identified timely care as one of the six key aims for improving health in the United States (Institute of Medicine 2001). However, shortages of primary care physicians (PCPs) make it difficult for

many patients to access care in a timely fashion (Blumenthal 2004; Arvantes 2007; Halsey 2009). With the recent passage of the health care reform bill, more than 30 million Americans may be included in health insurance plans. This could, however, exacerbate the existing imbalance between primary care needs and supply; demand for care could surge due to an increasing number of patients covered by health insurance, and waiting times could rise. In short, the goal of timely care may not be met despite the expansion of health insurance coverage.

This article focuses on how to improve timely access to care in a cost-efficient way. We first operationalize “timeliness” by defining patient *waiting time* as the time elapsed from the point at which a patient makes an appointment to the point at which his or her medical service starts. We call this *appointment waiting time* (AWT). Long AWTs can potentially increase disease severity, resulting in more intensive treatment and higher costs. They can also constrain access to care. In states like Massachusetts, many clinics have stopped accepting new patients due to long AWTs for existing patients. To a large extent, cutting AWTs is equivalent to improving access to care.

To achieve the goal of timely care, there are generally two ways to reduce patient AWTs. One is to reduce the average consultation time that a PCP spends with patients so that more patients can be seen in one day. The other way is simply to hire more PCPs. However, adjusting the average consultation time of PCPs by shortening visits might lead to reduced quality of service and patient dissatisfaction (Dugdale, Epstein, and Pantilat 1999), while hiring more PCPs is complicated by current shortages. Hence, alternative solutions must be sought.

Nurse practitioners (NPs) are advanced practice nurses whose scope of practice includes health promotion, disease prevention, health education, and counseling, as well as the diagnosis and management of acute and chronic diseases (American Academy of Nurse Practitioners 2007). NPs can independently manage 80 percent of patients’ primary care needs (Kreitzer, Kligler, and Meeker 2009), and extensive studies show that NPs provide high-quality care (Horrocks, Anderson, and Salisbury 2002; Laurant et al. 2009). Clinical outcomes do not differ for patients who receive services from NPs compared

Address correspondence to Nan Liu, Ph.D., Assistant Professor, Department of Health Policy and Management, Mailman School of Public Health, Columbia University, 600 W. 168th St., 6th Floor, New York, NY 10032; e-mail: nl2320@columbia.edu. Thomas D’Aunno, Ph.D., is with the Department of Health Policy and Management, Mailman School of Public Health, Columbia University, New York, NY.

to PCPs. Some studies conclude that care provided by NPs leads to better outcomes; patients appear to be as satisfied—and in some cases, more satisfied—with NP care as compared to PCPs.

This literature supports engaging more NPs in primary care with even more autonomy. In fact, this way of expanding the primary care workforce has been recognized as a feasible and effective solution to solve the PCP shortage problem in Canada (Canadian Nurse Association 2006). Nonetheless, we know relatively little about how to organize the work of NPs and PCPs in primary care settings to maximize overall *productivity* and *cost-efficiency*. Laurant et al. (2005) indicate that nurse–doctor substitution has the potential to reduce doctor’s workload and overall health care costs, but this benefit may not be realized for several reasons. First, nurses are not granted sufficient autonomy and doctors continue to provide the types of care that have been transferred to nurses. Second, nurses may generate new demands where previously there were none, for example, referrals. Third, cost savings due to salary differentials between nurses and doctors may be offset by nurses’ longer consultation times. However, until now, there has not been a general analytic framework for considering these factors.

Using a systems modeling approach, we develop simple stylized models to investigate the productivity and cost-efficiencies of different health care delivery models (practice models) where NPs and PCPs work together in a salaried environment. These stylized models enable us to *quantitatively* examine the impact of those factors considered in Laurant et al. (2005). We study three models: (I) a solo-physician practice model, which serves as our benchmark; (II) a supervision model where NPs manage a certain amount of office visits under the supervision of PCPs and make referrals to PCPs if necessary; and (III) a shared-panel model where NPs work collaboratively with PCPs and provide care to patients with full autonomy. Model II resembles a practice where the NP and the PCP each maintains a panel of patients. Model III represents a group practice where providers work as a team. These models are useful in generating insights to the following questions related to staffing, productivity, and cost-efficiency.

- Is it always cost-efficient to employ NPs because they are paid less than PCPs?
- What factors affect the productivity and cost-efficiency of each practice model? And how?
- How can one improve the productivity and cost-efficiency of a practice model?

This article addresses these questions by comparing the productivity and cost-efficiencies of different models across various parameter settings. These settings capture how providers organize their work, and the settings that result in good performance can be used to inform how a practice model can be designed or adjusted to reduce waiting and improve access while keeping costs contained. However, it only makes sense to conduct such comparisons if the care provided in these models is “equivalent” in some sense. We thus consider that two patients have received “equivalent” care if their AWTs are the same. Indeed, setting a requirement on AWTs can be regarded as an assurance for quality of care, because reducing AWTs improves timeliness-to-care and patient attendance rates and hence helps maintain continuity of care (Penchansky and Thomas 1981; Bean and Talaga 1995).

Following Green and Savin (2008), we define the productivity of a practice model to be the *maximum* number of patients that can be accounted for by the model under a given AWT requirement; cost-efficiency is then measured by the corresponding annual cost per patient in that model under the same AWT requirement. Therefore, productivity is a measure of the maximum access to care that can be provided by a particular practice model, while cost-efficiency can be regarded as the *minimum* average annual cost (or more broadly speaking, medical resources) of serving a patient. The lower this cost, the higher the cost-efficiency.

To develop an evaluation framework for the productivity and cost-efficiencies of these models while controlling for AWTs, we use a systems modeling method called *queueing theory* (Gross and Harris 1985). Queueing theory is an advanced mathematical modeling technique (see Appendix) that can estimate waiting times given customer arrival information and provider service patterns, and hence it is an ideal tool for our analysis. The next section describes how our models are developed using queueing theory.

METHODS

Our goal is to compare different practice models at a strategic level and generate insights for primary health care delivery rather than to model actual operations in a real practice; hence, we use simple stylized models for analysis. We assume that there is one PCP and, at most, one NP in each model considered. Extension of models with multiple providers can also be analyzed similarly.

Figure 1: Practice Models

The Models and Analysis

System Dynamics in General. Figure 1 shows the three models. The general system dynamics work as follows. Patients call the health care facility for appointments in a random fashion. Their names are registered on the appointment schedule (i.e., sitting in the queue) in the order that they call. There are no capacity limits on the appointment schedule; that is, the clinic will not reject patient appointment requests. We assume the PCP and NP consultation times are two sequences of independent and identically distributed nonnegative random variables. Patients will arrive on time. After the provider sees one patient, he or she will take the next one on the schedule (according to each model's setup). Patients are discharged after consultation.

Notice that the queue modeled here is for *appointments*, while not for actual patients waiting in the clinic to see the provider. The queue can be thought of as a list of patient names that appear on the appointment schedule, and patients are physically elsewhere waiting for their actual appointment times.

We acknowledge that certain features in practice, for example, patient punctuality and preferences on appointment dates, are not fully captured in our models. Queueing studies often make these simplifications to keep models technically tractable; more important, if the purpose is to conduct a strategic level analysis and not to model daily operations in clinics, these are reasonable simplifications (Green and Savin 2008; Liu and Ziya unpublished data). Some other factors such as patient mix and nonattendance behavior (no-shows), though not modeled here, can be accommodated using our framework.

Evaluation of the Productivity and Cost-Efficiency. We need to set a common standard on the average AWT to compare these different models. Without the loss of generality, let this waiting time requirement be “not to exceed T days” (T can be any positive number depending on needs). Denote the average AWT under model i by W_i . Thus, the timeliness-to-care requirement is simply that $W_i \leq T$, for $i = \text{I, II, and III}$.

We evaluate different practice models in a *salaried* environment, and therefore staffing costs are the major cost components in our analysis, that is, annual salaries for PCPs and NPs, denoted by S_P and S_N , respectively. Indeed, one major economic incentive to involve NPs in primary care delivery is that they are less expensive to employ compared to PCPs (Laurant et al. 2005). Other costs related to resource utilization, for example, tests and investigations ordered, and prescriptions, are assumed the same per patient per year in different models, as many studies show no differences between nurse-led care and physician-led care (Laurant et al. 2009).

Suppose that each patient will call the clinic, that is, generate an appointment request (*demand*), according to a Poisson process (PP) with rate λ_0 calls (*arrivals*) per day. The PP is widely used in modeling random arrivals of patients when they arrive one at a time, and it is empirically verified to work well (Green 2006). Then if the clinic serves N patients in total (i.e., its panel size is N), the demand process for this clinic is also a PP with daily rate $\lambda = N\lambda_0$, assuming each patient behaves independently. We choose this arrival model because it can be estimated with few data requirements; it has also been used in other queueing studies that carry out similar strategic-level analyses (Green and Savin 2008; Liu and Ziya unpublished data).

In modeling provider consultation (*service*) processes, we use two service time distributions that have different variability in service times. This is motivated by the fact that variability in consultation length can vary across practices. A commonly used measure for such variability is the coefficient of variation (CV) of consultation times, that is, the ratio of standard deviation and mean ($=\sigma/\mu$). Empirical studies show that such CV values range from approximately 0.35 to 0.85 (Cayirli, Veral, and Rosen 2006). Therefore, we consider models with both high and low CV values, respectively. First, we consider *Markovian queueing models*, where both PCP and NP consultation times follow exponential distributions. This setting is plausible when the CV values of these service times are close to one, a fairly high value indicating large variability. However, if service times have smaller variability, Markovian models usually underestimate the productivity and cost-efficiency. This motivates our analysis of *non-Markovian models* which

assume deterministic service time distributions with zero variability in service times ($CV = 0$).

By considering both Markovian and non-Markovian models, we are able to generate bounds for system performance: non-Markovian models provide upper bounds for productivity and cost-efficiency, whereas Markovian models yield lower bounds. More important, this enables us to investigate the robustness of model performance by examining whether the choice of the most productive and/or cost-efficient model is sensitive to service time variability.

Given the waiting time requirement that the average AWT should not exceed T days, each model has a *unique* maximum number of patients that can be accounted for, as AWT strictly increases when the panel size increases. It follows from the previous definition that this maximum panel size is the productivity. Dividing the total annual staffing costs of model i by its productivity, we obtain its annual cost per patient, which measures its cost-efficiency.

Model I—Solo Physician Model. Model I considers a single-physician's office, where a PCP serves patients alone. The average daily arrival rate is λ patients/day, and the service rate of the PCP is μ_2 patients/day. Patients are served in a *first-come-first-serve* (FCFS) order. In this case, the number of patients waiting in the appointment schedule (including the one in service) can be modeled as an M/M/1 (M/D/1) queue, if service times follow an exponential (deterministic) distribution. These two models are the most basic ones in queueing theory and can be evaluated easily.

Model II—Supervision Model. In Model II, the NP provides care to certain patients under the supervision of the PCP and makes referrals to the PCP when needed. In this model, $100p\%$ ($0 \leq p \leq 1$) of the office visits generated from the population are first handled by the NP, and the rest are handled by the PCP. We call p the *substitution ratio*. The *referral rate*, that is, the proportion of patients who initially come to see the NP and then are referred to the PCP is denoted by r ($0 \leq r \leq 1$).

In this model, the PCP and the NP maintain their own individual appointment schedules, respectively. The patients coming to the NP are registered on the NP's schedule; others will be registered on the PCP's schedule. The service rates of the NP and the PCP are μ_1 and μ_2 , respectively, and the service order is FCFS.

Model III—Shared-Panel Model. Model III is a practice model where the PCP and the NP share the same panel of patients and jointly manage their appointment schedules. In this model, the NP is assumed to practice with full autonomy and requires no PCP supervision. Patients will be seen by any available provider. This mechanism of integrating capacity among providers is known in queueing theory as “capacity pooling,” which can utilize resource more efficiently and reduce patient waiting time. The feasibility and benefit of capacity pooling are discussed below. Studying Model III is also of theoretical interest: it yields a performance upper bound for Model II and hence provides a useful benchmark for evaluation.

Most of these models either have closed-form solutions or can be numerically evaluated. The only exception is Model III with deterministic service time distribution, for which we carried out a discrete-event simulation study (Law and Kelton 1991). See detailed analyses of these models in the Appendix.

Model Parameters

We need the following parameters as model inputs: the demand rate of an individual patient λ_0 , the NP service rate μ_1 and the PCP service rate μ_2 , the substitution ratio p , the referral rate r , and the annual salary of the PCP S_P and that of the NP S_N . We extracted these parameters from earlier literature or survey results.

First, we considered visits to general and family practice, internal medicine, and pediatrics as primary care visits. According to the 2006 National Ambulatory Medical Care Survey (NAMCS) data (Cherry et al. 2008), the annual number of such visits made per person in the United States is 2.9, which is translated to $\lambda_0 = 0.008$ visit per person per day.

Based on the 1989–1998 NAMCS data, Mechanic, McAlpine, and Rosenthal (2001) found that the mean length of office visits to PCPs was around 16–18 minutes in 1998 in the United States. Recently, Konrad et al. (2010) interviewed 128 PCPs in the United States and confirmed that, on average, these PCPs allocated 18 minutes for a routine visit and 32 minutes for a new patient appointment. Therefore, we set $\mu_2 = 20$ patients per day, that is, 24 minutes allocated to a patient on average if the PCP spends 8 hours seeing patients each day.

Nurse practitioners typically have longer consultation lengths (and hence lower service rates) than PCPs (Laurant et al. 2009). However, there could be significant variation in how much their consultation times differ (Kin-

nersley et al. 2000; Shum et al. 2000; Venning et al. 2000). Instead of fixing NP service rate μ_1 , we considered three possible values for it: 12, 15, and 18 patients per day, corresponding to 0.6, 0.75, and 0.9 in terms of the PCP–NP consultation time ratio, respectively. This setup is also consistent with empirical findings in Kinnersley et al. (2000).

Regarding the substitution ratio p , no literature seems to provide an estimate. The value of p depends on how workload is allocated between the PCP and NP in a clinic, and it can also be indirectly influenced by NP scope of practice since its value is likely to increase (or can be more easily adjusted) if NP scope of practice is expanded (Eibner et al. 2009). To estimate the potential cost savings by encouraging greater use of NPs, Eibner et al. (2009) varied the percentage of office visits that are dealt with by NPs and estimated bounds for such savings. Following a similar strategy, we conducted a sensitivity analysis using different values of p (from 20 to 60 percent). Similarly, we tested different values for the referral rate r (from 0 to 25 percent) to study the impact of referrals. To model PCP supervision, we considered three levels of intensity: (1) no supervision: the NP has full autonomy and the PCP is not required to supervise; (2) low level: the PCP spends half of a patient consultation slot every day to supervise; and (3) high level: the PCP spends one full consultation slot every day to supervise.

We obtained salary data from the website of the 2008 Occupational Employment Statistics (U.S. Bureau of Labor Statistics 2008). According to the Bureau's Standard Occupational Classification (SOC), the most relevant job titles to the PCP and the NP are Family and General Practitioner (SOC code 291062) and Registered Nurses (SOC code 291111). The annual mean wages for these two job titles on the national level (May 2008) were retrieved from the website. Accordingly, we set $S_P = 161,490$ dollars per year and $S_N = 65,130$ dollars per year.

Without loss of generality, we set $T = 1$ day for all models in our numerical study; that is, the average AWT is required not to exceed 1 day. We note that other values for the aforementioned model parameters can also be used depending on the context of interest, but the insights generated should not change much.

RESULTS

Model I is our benchmark model for which we considered its performance under both Markovian and non-Markovian settings. For Model II, we con-

ducted an extensive sensitivity analysis by considering 648 different scenarios [2 service time distributions $\times$ 3 NP service rates (μ_1) $\times$ 6 substitution ratios (p) $\times$ 6 referral rates (r) $\times$ 3 supervision levels]. For Model III, we varied the NP service rate and tested its performance under both service time distribution assumptions. Detailed numerical results are presented in the Appendix.

The productivity for Model I is estimated to be 2,380–2,440 patients, and this is consistent with those reported elsewhere (Green and Savin 2008; Balasubramanian et al. 2010; Liu and Ziya unpublished data), validating our use of queueing models to estimate productivity. The estimated productivity for models with two providers (Model II and Model III) ranges from 2,400 to 4,600 patients, less than the sum of two solo PCPs' productivity. These are reasonable estimates because NPs have longer consultation times than PCPs.

Comparison between Markovian and Non-Markovian Models

A careful comparison of Markovian and non-Markovian models reveals that both performances within a practice model and comparative results across these models seem not sensitive to the variability in provider consultation times, implying that our results can be generalizable to different consultation time distributions. We present details on such comparisons in the Appendix. For brevity, the numerical results discussed below are all based on Markovian models unless otherwise specified.

The Relationships among Productivity, Cost-Efficiency, and Total Costs

We note that a more cost-efficient model does not necessarily provide greater productivity: the total annual costs consumed may differ among models. Table 1 compares Model I (solo physician model) and Model II (supervision model) assuming $\mu_1 = 15$, $p = 20$ percent, $r = 0$ percent, and low level of supervision. Model I is more cost-efficient than Model II (\$67.83 versus \$77.45 per patient per year), but Model II's productivity is 19 percent greater than Model I (2,926 versus 2,381 patients) and its total annual cost is also 40 percent higher (\$226,620 versus \$161,490 per year). If we double the staffing level in Model I and consider two solo PCPs practicing separately (just like two separate queues), then both the productivity (4,762 patients, 62 percent higher than that of Model II) and the total annual costs (\$322,980 per year, 43 percent higher than that of Model II) double while the cost-efficiency remains the same. In this case, which model to choose and how to staff depends on the tradeoff between productivity (i.e., access to care that can be

Table 1: Performance Comparison between Models I and II (Assuming Exponential Service Time Distributions, $\mu_1 = 15$, $p = 20\%$, $r = 0\%$ and Low Level of Supervision)

	<i>Productivity</i>	<i>Total Annual Staffing Costs</i>	<i>Cost-Efficiency</i>
Model I (one PCP practice)	2,381	\$161,490	\$67.83/patient/year
Model I (two separate PCP practices)	4,762	\$322,980	\$67.83/patient/year
Model II (one PCP + one NP)	2,926	\$226,620	\$77.45/patient/year

provided) and total annual costs. However, if the total annual costs are the same across models, then productivity and cost-efficiency are “equivalent” performance measures: larger productivity implies higher cost-efficiency, and vice versa.

Performance of Models II and III

Supervision leads to a decrease in productivity and cost-efficiency of Model II (supervision model), as PCPs need to spend extra time in supervising NPs. However, this impact seems negligible (see the Appendix). Referrals also result in a decrease in cost-efficiency, as they increase the workload of PCPs. Such impact is more significant when the PCP handles a larger proportion of the demand, that is, when the substitution ratio p is smaller.

Figure 2 shows the relative cost-efficiency (i.e., ratio of cost-efficiencies) of Model II (assuming no supervision and $\mu_1 = 15$) versus Model I (solo physician model). For a fixed referral rate, the cost-efficiency of Model II first increases in substitution ratio p and then decreases. We observe a similar trend for productivity. The reason is that, when p is too low, too many patients go to see PCPs; when p is too high, a disproportionately large amount of visits are handled by NPs. Neither is efficient in capacity utilization, as either the PCP or the NP is “overused,” leaving the other provider “underused.” More interestingly, the cost-efficiency when $p = 40$ percent or 50 percent always dominates those under other values of p , regardless of the value of referral rate r . This suggests a dominating impact of p on the performance of Model II over other factors.

With the same two providers, Model III (shared-panel model) outperforms Model II (supervision model) with 3–73 percent improvement in productivity and 3–42 percent improvement in cost-efficiency. As noted above,

Figure 2: Relative Cost-Efficiency of Model II versus Model I under Different Values of Substitution Ratio p and Referral Rate r (Relative Cost-Efficiency Is Defined as the Ratio of the Average Cost of Model II to That of Model I; Substitution Ratio p is the Percentage of the Office Visits First Handled by the NP; Referral Rate r is the Proportion of Patients Who Initially Come to See the NP and Then are Referred to the PCP.)

Model III provides an upper performance bound for Model II, and this superiority is largely due to the effect of capacity pooling, which we discuss below.

DISCUSSION

Involving NPs in a practice team and exerting their full capabilities is a promising way to expand primary care workforce. However, the productivity and cost-efficiency of practice models involving NPs depend on a variety of factors.

Determinants for the Productivity and Cost-Efficiency of Model II

The productivity and cost-efficiency of Model II (supervision model) depend on a full collection of model parameters, including substitution ratio p , referral rate r , level of PCP supervision and NP service rate μ_1 . First, the overall productivity improves when NP service rate μ_1 increases. In particular, the optimal productivity (assuming $r = 0$ percent and no PCP supervision) when $\mu_1 = 12, 15$, and 18 patients per day is $3,747, 4,136$, and $4,510$ patients, respectively. Substitution ratio p is another most important determinant for the per-

formance of Model II. It captures the proportion of initial office visits handled by the NP, and it can be regarded as a measure of how workload is allocated among providers.

Model II (supervision model) is always more productive than Model I (solo physician model) but not necessarily more cost-efficient. We compare them to identify under what values of p , employing NPs is more cost-efficient than a solo-PCP practice. Table 2 shows such comparative results under a variety of scenarios (assuming $\mu_1 = 15$). The results under other NP service rates are similar.

Table 2 reveals that, to make Model II more cost-efficient, substitution ratio p should be at least 30 percent. This implies that hiring an NP is cost-efficient only if NPs handle at least 30 percent of the initial workload. As discussed in Laurant et al. (2004), “gains for efficiency of service can be achieved only if GP give up providing the types of care that they have delegated to nurses and instead invest their time in activities that only doctors can perform.” Our modeling results lend support to this statement and *quantitatively* characterize how much work should be delegated to NPs to achieve cost-efficiency.

Table 2: The More Cost-Efficient Choice between Model I and Model II (Assuming $\mu_1 = 15$ Patients/day)

p	$r = 0\%$	5%	10%	15%	20%	25%
a. Comparison of Model I and Model II with no supervision						
20%	I	I	I	I	I	I
30%	II	II	I	I	I	I
40%	II	II	II	II	II	II
50%	II	II	II	II	II	II
60%	I	I	I	I	I	I
b. Comparison of Model I and Model II with low level of supervision						
20%	I	I	I	I	I	I
30%	II	I	I	I	I	I
40%	II	II	II	II	II	II
50%	II	II	II	II	II	II
60%	I	I	I	I	I	I
c. Comparison of Model I and Model II with high level of supervision						
20%	I	I	I	I	I	I
30%	I	I	I	I	I	I
40%	II	II	II	II	II	I
50%	II	II	II	II	II	II
60%	I	I	I	I	I	I

The optimal choice of substitution ratio p , which leads to the highest productivity and cost-efficiency of Model II, heavily depends on other model parameters. Notably, one most important factor is NP service rate μ_1 . As expected, when μ_1 increases from 12 to 15 to 18 patients per day, the optimal values for p (when $r = 0$ percent) also increases from approximately 38 to 43 to 47 percent, respectively. It is interesting to note why the productivity of Model II peaks at these values. Suppose that $\mu_1 = 15$, $\mu_2 = 20$, and $r = 0$ percent. From an efficiency point of view, a “fair” distribution of patients between the NP and the PCP would be based on their service rates, that is, the NP should serve $15/(15 + 20) = 43$ percent of the patients. The optimal value for p also depends on the referral rate r . Assuming other parameters kept unchanged, the PCP’s workload increases in r . Hence, to keep an optimal workload allocation, one needs to adjust (in fact, increase) the value for p . For example, assuming $\mu_1 = 15$, $\mu_2 = 20$, and $r = 25$ percent, the optimal p is about 48 percent, compared with 43 percent when $r = 0$ percent.

Other factors that could affect the performance of Model II include the referral rate r and the level of PCP supervision. When the workload distribution among providers is disproportional to their capacity, referrals can exacerbate this imbalance. In particular, referrals lead to a higher reduction in both productivity and cost-efficiency of Model II when PCP has already handled a larger percentage of visits. Interestingly, supervision does not seem to have a significant impact on the performance of Model II. Compared with the situation without supervision, the model with a high level of supervision has no more than a 5 percent drop in either productivity or cost-efficiency in all scenarios we tested. Considering that, in practice, PCPs usually spend even less time in supervision (e.g., by simply doing a monthly or quarterly review), the impact of supervision on productivity and cost-efficiency seems negligible.

In summary, the most influential determinants for the performance of Model II are provider service rates and substitution ratio which determine workload allocation among providers. One could try to achieve an optimal workload allocation by panel design (Balasubramanian et al. 2010). However, it may not be easy to implement such a design that exactly splits patient demand optimally, and even if one could, randomness in demand makes it impossible to keep the delivery system running efficiently and optimally every day. To resolve this issue, Model III suggests a possible solution: to develop a shared-panel, encourage a team-based approach, and integrate the capacity of providers by allowing capacity pooling.

Value of Capacity Pooling

To quantify the value of capacity pooling, we compare Model II (supervision model) to Model III (shared-panel model). One implicit assumption in Model III is that all patients will receive adequate care from the provider they see, and no referrals are made. To make a fair comparison, we set the referral rate $r = 0$ and assume no supervision in Model II.

Figure 3 shows the comparison assuming that $\mu_1 = 15$. Compared to Model II with substitution ratio $p = 50$ percent, Model III generates a much broader access to care (635 more patients covered or equivalently an 18 percent increase in productivity) under the same timeliness-to-care requirement and annual staffing costs. Such gain in the productivity of Model III over that of Model II can be regarded as the “value” of capacity pooling. When the substitution ratio p deviates from its optimal value 43 percent, capacity pooling improves productivity significantly (42, 23, 18, and 42 percent improvement when $p = 0.2, 0.3, 0.5,$ and 0.6 , respectively). However, when p is close to 43 percent, Model II performs almost as cost-efficient as Model III. Considering that Model III sets a performance upper bound for a two-provider model, these findings accentuate the importance of an appropriate workload allocation when using Model II.

Capacity pooling could improve access to care and reduce waiting time without incurring additional staffing costs. However, it may not be implementable or desirable in practice due to issues related to NP scope of practice, patient preference, or clinical concerns such as disruption of continuity of

Figure 3: Productivity of Model II (Supervision Model) and Model III (Shared-Panel Model)

care. Though a full scale of capacity pooling may not be feasible, a partial implementation of it on certain patients can still be possible, in particular those who are strongly against or cannot afford waiting, for example, walk-ins, those requesting a same-day appointment, and urgent care patients. These patients may not have strong preference for providers and would opt to see anyone available (Venning et al. 2000; Balasubramanian et al. 2010). Pooling providers' capacity to serve these patients could be a useful strategy to improve efficiency in care delivery.

CONCLUSION AND EXTENSIONS FOR FUTURE WORK

We develop simple stylized queueing models for evaluating the productivity and cost-efficiencies of different practice models to involve NPs in primary care and investigate what affects the performance of these models and how. These queueing models can also be used by primary care clinics to make strategic decisions on staffing and panel size selection. In summary, the productivity and cost-efficiency of different practice models depend heavily on model parameters, which reflect how providers organize their work and a variety of other practice environment-related factors such as mean consultation lengths of providers, referral rate and physician supervision intensity. Using data reported in the literature, we find that employing an NP, whose salary is usually lower than a PCP, may *not* be cost-efficient in certain situations, in particular when NPs' capacity is underutilized. An appropriate workload allocation (e.g., by panel design) is important in maintaining and improving the cost-efficiency of a practice model with multiple providers at work. Cost-efficiency and productivity of a practice model seem insensitive to the variability in provider consultation lengths controlling for their means. Capacity pooling among providers could be a helpful strategy to improve efficiency in care delivery.

There are several ways to extend our work. First, our models can be extended to consider patient no-shows, a frequently encountered problem by many clinics. There are many possible factors associated with no-shows, but one particularly important predictor appears to be patient AWTs (Bean and Talaga 1995; Liu, Ziya, and Kulkarni 2010). As queueing models deal with waiting time, they can be extended to consider AWT-dependent no-shows. Second, our models can be used to investigate the impact of patient mix. In particular, patient visit frequency to a clinic and provider consultation times typically depend on the patient mix (Knox and Britt 2004; Cayirli, Veral, and

Rosen 2006). Built upon such dependency, our proposed models can be used to examine how patient mix affects the productivity and cost-efficiency of the clinic. Third, capacity pooling may only be applied to a limited set of patients as some others are “loyal” to their own providers. Our models can be extended to analyze such situations by making proper modifications on the arrival process. Last, though applications of queueing models have been shown helpful in many industrial settings (Gross and Harris 1985), further validation of their use in health care by directly comparing queueing model estimates with collected descriptive data in the field is important.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: We are grateful to Professors Robyn Gershon, Linda Green, and Patricia Stone for review and comments on previous versions of this paper. We also are grateful to the audience at the Brown-bag Seminar held in the Department of Health Policy and Management at Columbia University on October 27, 2010.

Disclosures: None.

Disclaimers: None.

REFERENCES

- American Academy of Nurse Practitioners. 2007. *FAQs about Nurse Practitioners*. Austin, TX: American Academy of Nurse Practitioners.
- Arvantes, J. 2007. “Survey Confirms Growing Demand for Primary Care Physicians.” *American Academy of Family Physicians News Now*, October 16.
- Balasubramanian, H., R. Banerjee, B. Denton, J. Naessens, D. Wood, and J. Stahl. 2010. “Improving Clinical Access and Continuity Using Physician Panel Redesign.” *Journal of General Internal Medicine* 25 (10): 1109–15.
- Banks, J. 1999. “Introduction to Simulation.” In *The Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, pp. 7–13. New York: ACM.
- Bean, A., and J. Talaga. 1995. “Predicting Appointment Breaking.” *Journal of Health Care Marketing* 15 (1): 29–34.
- Blumenthal, D. 2004. “New Steam from an Old Cauldron—The Physician-Supply Debate.” *New England Journal of Medicine* 350 (17): 1780–7.
- Canadian Nurse Association. 2006. *Nurse Practitioners: The Time Is Now: A Solution to Improving Access and Reducing Wait Times in Canada*. Ottawa, ON, Canada: Canadian Nurse Association.

- Cayirli, T., E. Veral, and H. Rosen. 2006. "Designing Appointment Scheduling Systems for Ambulatory Care Services." *Health Care Management Science* 9 (1): 47–58.
- Cherry, D., E. Hing, D. Woodwell, and E. Rechtsteiner. 2008. *National Ambulatory Medical Care Survey: 2006 Summary*. National Health Statistics Reports, no. 3. Hyattsville, MD: Division of Health Care Statistics, U.S. Department of Health & Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
- Dugdale, D. C., R. Epstein, and S. Z. Pantilat. 1999. "Time and the Patient-Physician Relationship." *Journal of General Internal Medicine* 14 (S1): 34–40.
- Eibner, C. E., P. S. Hussey, M. S. Ridgely, and E. A. McGlynn. 2009. *Controlling Health Care Spending in Massachusetts: An Analysis of Options*. Santa Monica, CA: RAND Corp.
- Fishman, G. S. 2001. *Discrete-Event Simulation: Modeling, Programming, and Analysis*. New York: Springer Verlag.
- Green, L. 2002. "How Many Hospital Beds?" *Inquiry* 39 (4): 400–12.
- . 2006. "Queueing Analysis in Healthcare." In *Patient Flow: Reducing Delay in Healthcare Delivery*, edited by R. W. Hall, pp. 281–308. New York: Springer-Verlag.
- Green, L. V., and V. Nguyen. 2001. "Strategies for Cutting Hospital Beds: The Impact on Patient Service." *Health Services Research* 36 (2): 421–41.
- Green, L. V., and S. Savin. 2008. "Reducing Delays for Medical Appointments: A Queueing Approach." *Operations Research* 56 (6): 1526–38.
- Gross, D., and C. Harris. 1985. *Fundamentals of Queueing Theory*. New York: John Wiley and Sons, Inc.
- Halsey, A. 2009. "Primary-Care Doctor Shortage May Undermine Reform Efforts." *The Washington Post*, June 20.
- Horrocks, S., E. Anderson, and C. Salisbury. 2002. "Systematic Review of Whether Nurse Practitioners Working in Primary Care Can Provide Equivalent Care to Doctors." *British Medical Journal* 324 (7341): 819–23.
- Institute of Medicine. 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, D.C.: National Academy Press.
- Kinnersley, P., E. Anderson, K. Parry, J. Clement, L. Archard, P. Turton, A. Stainthorpe, A. Fraser, C. C. Butler, and C. Rogers. 2000. "Randomised Controlled Trial of Nurse Practitioner Versus General Practitioner Care for Patients Requesting "Same Day" Consultations in Primary Care." *British Medical Journal* 320 (7241): 1043–8.
- Knox, S. A., and H. Britt. 2004. "The Contribution of Demographic and Morbidity Factors to Self-Reported Visit Frequency of Patients: A Cross-Sectional Study of General Practice Patients in Australia." *BMC Family Practice* 5 (1): 17–23.
- Konrad, T. R., C. L. Link, R. J. Shackelton, L. D. Marceau, O. von dem Knesebeck, J. Siegrist, S. Arber, A. Adams, and J. B. McKinlay. 2010. "It's about Time: Physicians' Perceptions of Time Constraints in Primary Care Medical Practice in Three National Healthcare Systems." *Medical Care* 48 (2): 95–100.

- Kreitzer, M. J., B. Kligler, and W. C. Meeker. 2009. "Health Professions Education and Integrative Health Care." *Journal of Science and Healing* 5 (4): 212–27.
- Laurant, M., R. Hermens, J. Braspenning, B. Sibbald, and R. Grol. 2004. "Impact of Nurse Practitioners on Workload of General Practitioners: Randomised Controlled Trial." *British Medical Journal* 328 (7445): 927–30.
- Laurant, M., D. Reeves, R. Hermens, J. Braspenning, R. Grol, and B. Sibbald. 2005. "Substitution of Doctors by Nurses in Primary Care." *The Cochrane Library* 5: CD001271.
- Laurant, M., M. Harmsen, H. Wollersheim, R. Grol, M. Faber, and B. Sibbald. 2009. "The Impact of Nonphysician Clinicians: Do They Improve the Quality and Cost-Effectiveness of Health Care Services?" *Medical Care Research and Review* 66 (6): 36S–89S.
- Law, A., and W. Kelton. 1991. *Simulation Modeling and Analysis*. New York: McGraw-Hill.
- Liu, N., S. Ziya, and V. Kulkarni. 2010. "Dynamic Scheduling of Outpatient Appointments under Patient No-Shows and Cancellations." *Manufacturing & Service Operations Management* 12 (2): 347–64.
- Mechanic, D., D. D. McAlpine, and M. Rosenthal. 2001. "Are Patients' Office Visits with Physicians Getting Shorter?" *New England Journal of Medicine* 344 (3): 198–204.
- Penchansky, R., and J. W. Thomas. 1981. "The Concept of Access: Definition and Relationship to Consumer Satisfaction." *Medical Care* 19 (2): 127–40.
- Shum, C., A. Humphreys, D. Wheeler, M. A. Cochrane, S. Skoda, and S. Clement. 2000. "Nurse Management of Patients with Minor Illnesses in General Practice: Multicentre, Randomised Controlled Trial." *British Medical Journal* 320 (7241): 1038–43.
- U.S. Bureau of Labor Statistics. 2008. Employment and Wages from Occupational Employment Statistics (OES) survey [accessed on January 12, 2010]. Available at <http://bls.gov/oes/>.
- Venning, P., A. Durie, M. Roland, C. Roberts, and B. Leese. 2000. "Randomised Controlled Trial Comparing Cost Effectiveness of General Practitioners and Nurse Practitioners in Primary Care." *British Medical Journal* 320 (7241): 1048–53.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

- Appendix SA1: Author Matrix.
- Appendix S1: Introduction to Queueing Theory.
- Appendix S2: Analysis of the Models.
- Appendix S3: Numerical Results of All Tested Cases.

Appendix S4: Comparison of Markovian and Non-Markovian Models.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.