ABSTRACT

Concierge Medicine is a relatively new development in the U.S. Healthcare system and is designed and implemented, mostly by primary care physicians, to provide comprehensive care in a timely manner. Physicians often struggle with the decisions associated with adoption (implement or not), design (pricing and membership), and management (day-to-day execution) of these systems. The patients also struggle with decisions associated with signing up (or not) as it is predicated on the performance measures of complex service systems. We develop a simulation model that could be used by both the physicians and the patients to help them with these decisions. We demonstrate the effectiveness of this tool using data from a primary care physician in the Louisville, KY area.

1 INTRODUCTION

The U.S. Healthcare System, despite the billions of dollars spent on it every year, is surprisingly inefficient especially at the primary care level. When an unexpected need arises to see a family practitioner or a pediatrician, it is often the case that the next available appointment is more than a few days away. Not wanting to wait that long, the patients compromise and see a physician that is not familiar with their health record or a nurse practitioner and in either case, leave the system disappointed. In the extreme case, the patients end up going to the emergency room resulting in added burden (both cost and congestion) to a system that is already perceived to be inefficient. The primary care physicians on the other hand, feel overworked and underpaid, not happy with the amount of time they get to spend (about 10-15 minutes) with each patient, and certainly not happy with the care they provide to their patients. Out of this mutual discontent was born the system of concierge medicine, under which the patients pay an annual retainer to receive better (timely and comprehensive) service. The doctors, on the other hand, see fewer patients, are able to spend more time on each visit and also possess deeper knowledge on the health conditions of their patients.

Concierge medicine, also sometimes called “boutique” or “private physician” (O’Brien, 2013) is this alternative to traditional medical practices that is designed to better serve the needs of the patients and their doctors. Tom Blue, executive director of the American Academy of Private Physicians (a concierge
care trade group) is quoted in Gustke (2012) as stating that there were about 4,400 concierge doctors in
2012. Concierge medicine enables doctors to have some guaranteed revenue streams while simultaneously
increasing patient access and service times. Because the doctor has fewer patients, limited to 600 by
MDVIP (a group that supports concierge medicine), additional fees must be imposed. Patients each pay
an annual retainer to the doctor of at least $1,500. About $500 of this fee goes to MDVIP for electronic
health records and marketing support. The remainder supplements the doctor’s income to offset monies
lost from the reduction in patient visits.

A traditional practice may have 2,500 to 3,000 patients to a doctor, but since none of them pay an an-
nual fee, the physician is restricted to revenues received from patient visits. These revenues have been
steadily decreasing in the past years, mainly due to pressures from insurance companies and the govern-
ment (through Medicare and Medicaid), forcing the doctors to increase the daily number of patients they
see. MacStravic (2005) notes that many physicians are fed-up with this need to increase as it leads to inef-
fective care for the patients and lack of professional satisfaction for the physician. It is tempting for the
physician to consider implementing (Kihm, 2013) concierge medicine which provides some guaranteed
revenues while at the same time reducing the number of patients they see ensuring that they are able to
provide comprehensive care to them.

Signing up to be a concierge medicine patient is more expensive for the patient because of the annual
retainer (not covered by insurance), but the patient must balance this with the benefits of timely and com-
prehensive care received. A patient who values his/her time more or has a severely deteriorating health
condition (in both cases, the cost they attribute to waiting is high) might be willing to pay the annual re-
tainer in order to ensure a lower waiting time to see the doctor and to get more time with the doctor when
they do visit. However, it is an important financial and personal decision that they must make and should
take into consideration the impacts of that choice on health outcomes, waiting time, and utility.

In the context of concierge medicine, the physicians and the patients face very important and difficult
decisions and there is a need to develop decision support tools that can be of use to them. From a broader
research perspective, there are a number of research questions that need to be addressed: (i) what are the
conditions (customer heterogeneity, physician utilization, etc.) that should prompt a physician to consider
implementing concierge medicine? (ii) How much should the physician charge for a concierge medicine
patient and what service should they provide and to whom? (iii) Which customers (age, health condition,
waiting time cost rate) should sign up for concierge medicine? and (iv) what is the impact of concierge
medicine on social welfare? In this paper, we develop analytical and simulation models to answer these
research questions and provide decision support tools that can be used by both physicians and patients.

The rest of this paper is organized as follows. In the next section, we describe the main features of the
models that can answer the research questions. Section 3 describes existing research in this domain and in
section 4 we detail the results from analysis of the simulation models developed here. Section 5 describes
how these results can be useful for physicians and patients and we conclude in section 6 with a summary
of the results and directions for future research.

2 MODEL SETUP

2.1 Single-Physician vs. Multi-Physician

The primary healthcare offices in the U.S. either have a single physician taking care of all the patients or
many physicians that attend to the patients. In the multi-physician situations, when a patient needs to see a
doctor, they are assigned to the physician that is readily available. The single-physician setting leads to
continuity of care which is not guaranteed in the multi-physician setting. The implementation, if any, of
concierge medicine differs between these two contexts. In a single-physician setting, the doctor dictates
that either a patient signs up concierge medicine and continues to be in his/her practice or moves to a dif-
ferent primary care physician. In a multi-physician setting, a subset of the doctors offer concierge medi-
cine, and when a patient signs up for it, he/she is assigned to one of the concierge physicians and is seen
by that physician all the time. The patients that do not sign up for concierge medicine are seen by the non-concierge physicians in the traditional, ad-hoc manner. In this paper, we address the single-physician setting. A version of the multi-physician setting was addressed in Gavirneni and Kulkarni (2013).

2.2 Physician Perspective

The physician, first and foremost, must decide whether to transition into a concierge medicine setting. Since, we are focusing on a single-physician setting, the physician should then decide (i) how many patients she should have in her practice; (ii) how much she should charge to the patients that sign up; (iii) what services (e.g. time per visit) should the physician provide to the concierge patients; and (iv) what should be the mix (in terms of their demands for service) of patients. We will assume that the physician should be able to transition into the new setting without a change in the revenues and utilization. Further, the patients that do not sign up for concierge medicine are no longer part of our analysis. We do not specifically model where they go.

2.3 Patient Perspective

Patients should decide whether to sign up for concierge medicine or not. If they do sign up, they are part of the physician’s practice, will have to pay the additional fees and will receive the special service offered by the physician. Thus it is a trade-off between the additional fees and the reduction in the waiting cost experienced. We model the fact that the customers are heterogeneous in the cost they attribute to the waiting time. This could be a result of the patient’s age, his/her health condition, or just the patience they have in waiting. Gavirneni and Kulkarni (2013) model this heterogeneity in customer waiting cost rate using exponential and uniform distributions. However, in reality, this distribution can take any shape and using simulation gives us the ability to deal with a wide variety of waiting cost rate distributions. It is conceivable that the patients mostly focus on average waiting time (as modeled in Gavirneni and Kulkarni (2013)), but it is also conceivable that they are also concerned about the worst-case (say 90th percentile) waiting time they may expect to experience. Thus, in our simulation model, we will model the setting in which the customer waiting cost is a weighted combination of the average waiting time and the 90th percentile of the waiting time distribution. If the customer does not join the concierge medicine practice, then we assume that they go elsewhere for their services and that they are able to find a service provider that is able to mimic the service the patients previously received.

2.4 Society Perspective

While we will mostly focus on the physician and patient perspectives, the larger question of impact on society is never far from our mind. There is significant discussion in the media and the healthcare circles about the perceived social injustice associated with concierge medicine. The popular perception is that wealthy patients will receive better healthcare while the poor patients will receive worse care. Much of this discussion is based on qualitative perceptions, with little or no quantitative information. Gavirneni and Kulkarni (2013) model the setting in which the premium fees collected from the concierge patients are partially transferred, via a discount scheme, to the non-concierge patients, and show that this eventually leads to benefits for the entire society. Whether or not such a transfer is possible can lead to a significant discussion, but the issue of benefit to society can never be ignored when evaluating concierge medicine. If concierge medicine does in fact prove to be beneficial for the entire society, then policies can be designed and implemented to ensure that concierge medicine is more widely practiced.

3 PREVIOUS RESEARCH

Gavirneni and Kulkarni (2013) is one of the first studies to model concierge medicine from a queuing perspective, but position their research as applicable to not just healthcare, but the broader service economy. Their analysis is divided into two different settings, depending on how the customer reacts to the
implementation of a concierge option. The first one, which they call the *No Abandonment* case, is one in which the customer will not leave the service even if his/her costs are higher under the concierge setting. This would be representative of the typical multi-physician primary doctor's office. The second setting, which they call the *Abandonment* case, the customers are allowed to leave the system (they do not model where they go) if they are not happy with the concierge option. This would be akin to the single-physician service.

They first show that under *No Abandonment*, if the service provider wishes to ensure that all the customers are no worse-off, then the service provider cannot make any additional revenue from offering the concierge option. On the other hand, if the service provider wants to ensure that the customers are no worse off only in the average sense (i.e. some individual customer may be worse off, but taken as a whole group, they are no worse-off), then she can increase her profits by about 20-100%. Under the *Abandonment* case, it is imperative that all the remaining customers in the system are no worse off under the concierge option. Even so, the service provider can significantly increase her revenues. In all these cases, they show that there is a wide range of parameter settings in which, the customers and the service provider are no worse-off and that the total system cost is significantly reduced. By comparing this to the minimum possible system cost, they show that transition to the concierge option eliminates about 73% of the inefficiency in the system.

While their analytical results concretely establish the conditions under which the concierge option is attractive, they acknowledge that they were derived under many assumptions. In order to test whether these results are valid in real-world settings, they collected and analyzed the data associated with adoption of MDVIP service across the country. They show that concierge medicine is adopted in areas (categorized by zip codes) where the median income is larger, the population is older; and income has a larger variance. Their analysis showed that by providing the concierge option in these locations, the service providers are able to increase the revenues by about 14%. This analysis of the real-world data complements the analytical results and enables them to conclude that concierge option has the potential to significantly improve the performance of the service offering.

Despite the analytical and empirical contributions, their research suffers from many limitations. First, they model the queuing system as an M/G/1 process which unnecessarily restricts the arrival process to be Poisson. Second, they model the waiting cost heterogeneity in customers using exponential and uniform distributions which limit the degree of heterogeneity that can be analyzed. Third, they model the patient waiting cost as a function solely of the average waiting time, while in reality, it is probably a combination of the average waiting time and the worst-case (possibly the 90th percentile) waiting time. We develop models that overcome all these issues, but resort to simulation in order to achieve our research objectives.

### 4 PROBLEM SETTINGS

We model a single-physician primary care setting based on our interactions with a physician in the Louisville, KY area. She had previously operated in the traditional setting and recently transitioned into concierge medicine. However, she was not sure whether she made the right decision, whether she had the right panel size and composition of concierge patients, and whether the fee she was charging was appropriate. She encouraged us to develop a simulation/analytical model that could help answer her questions. In order to capture the real-world complications ignored by Gavirneni and Kulkarni (2013), we developed a simulation model in FlexSim HC 4.0.

#### 4.1 Traditional Setting

BusinessWire notes that many traditional practitioners had 2,500 or more patients in their panel. An interview with the concierge practitioner in Louisville, KY revealed that she had 3,500 patients in her prior traditional medicine practice. We will model both 2,500 and 3,500 patients in the analysis.

Journal Watchdog (2010) estimates that a traditional practice doctor spends 10 minutes with each patient. The remainder of their office visit is with nurses, other staff, or waiting. The career web site Indeed
Gavirneni, Kulkarni, Manikas, and Karageorge

(listed the average salary for a nurse at $73,000. We model the doctor’s service time on the patient as ~N(10,2). The PSR survey (2009) states a range of 4.0 – 5.0 support staff per full-time-equivalent (FTE) physician. Reeves (2002) reports that the Medical Group Management Associated gives a figure of 4.67 staff per FTE physician. Because this latter number is in the former range, we will use that to encompass the staff to support the doctor in traditional medicine. The concierge practice and traditional medicine practice will both have a receptionist as one of the FTE. However, we assign 3.67 more staff as nurses (at $73,000 annually per) to do some of the data collection and minor procedures in the traditional practice compared to the doctor performing all services herself in concierge medicine. 3.67 comes from the average FTE per physicians of 4.67 minus 1 for the receptionist. Further, given that we model traditional doctors as spending 10 minutes on average (standard deviation of 2 minutes) per patient, and the patients typically have a 30 minute exam, the remainder of time is spent with a nurse. To ensure that the highly skilled doctor is the constraint and not a more fungible nurse role, we need these extra FTEs in the traditional medical office. It may seem intuitive to have more staff for a larger panel size, however, patients will not see the doctor more quickly regardless of how many exam rooms and nurses are working. Our model looks at time from patient arrival until the examined by the doctor. If a patient felt they were not waiting by sitting in an exam room with or without a nurse, that would be necessitate investigation of differing staff sizes. We felt that the relevant metric is time until the primary physician examines the patient, therefore, having 3.67 nurses corresponds approximately to the number needed in parallel to ensure that the doctor always has a patient ready to see (if a nurse is spending 20 minutes with a patient, but the doctor takes 6 minutes to examine a patient, clearly, the doctor will have to wait for the next patient to be prepped if there are less than 3.33 in this example). Recall that 99.7% of patient exam times fall with 4 – 16 minutes. Indeed.com (retrieved May 10, 2013) lists the average national salary for a receptionist as $25,000. Both the traditional and concierge practices are assumed to have one receptionist at this salary. The nurse and receptionist salaries are likely gross rather than the fully loaded expense for the medical practice (i.e., including benefits). The PSR’s 2010 Practice Management STATS Quick Reference (Practice Support Resources, 2010) gives a benchmark of $454,878 to $597,028 for a family practice gross charges. This same survey lists the net earnings per family practice physician to be in the range $197,407 to $218,187. The average physician earns $207,797. Staff of 3.67 nurses for $267,910 (3.67 * $73,000) and $25,000 for the receptionist gives a total of $292,910 in labor costs. The average gross charges per practice from the PSR benchmark is $525,953.

For a panel of 2,500 patients, assuming a patient, on average, would see the doctor 3.19 times per year (Murray et al., 2007). This gives 7,975 patient visits during the 2,080 hour work year. This translates to approximately 15.65 minutes between patient arrivals (we use an exponential distribution for arrivals). We also model a panel of 3,500 patients, giving an inter-arrival time of 11.18 minutes. The PRS survey (Practice Support Resources, 2009) gives a family practice range of 2,338 to 3,118 patients, which are confirmation of the two panel sizes we are using.

Using the $525,953 average gross charges from the PRS survey (2009), along with the average panel size of 2,728 with 3.19 visits per for 8,702 visits per year, gives an average per visit charge of approximately $60.44. We will use this same per visit charge for both traditional and concierge medicine.

4.2 Concierge Medicine

MDVIP gives a patient quota of 600 (BusinessWire; MDVIP.com) to ensure more timely access to the doctor. Again assuming a patient, on average, will see the doctor 3.19 times per year gives 1,914 patient visits during the 2,080 hour work year. This translates to approximately 65 minutes between patient arrivals (we use an exponential distribution for arrivals). We also model a panel of 400 patients, the self-imposed quota of the doctor we interviewed. If the quota of patients is 400, the time between arrivals is set to 98 minutes.
The goal of better service from the doctor is that the doctor spends at least 30 minutes with a patient, with physicals taking 1.5 hours (Journal Watchdog). We model service time for the doctor as normal with mean 60 minutes, and a standard deviation of 15. In this model, the waiting area is set sufficiently high so there is no blocking. Patients who come will wait until they are served by the doctor. There is no balking or reneging.

The receptionist is modeled as taking ~U(5,15) minutes to register a patient. The receptionist escorts the patient to the examination room in ~U(1,2) minutes. These distributions seem reasonable, although we could not find prior studies or surveys that gave times for these two activities. After seeing the doctor, the patient leaves unescorted.

5 RESULTS AND ANALYSIS

We first use prior survey data to compare doctor income in traditional and concierge practices for varying panel sizes. Then, in 5.2, we show the results of our simulation runs to gain insight into the patient experience in terms of waiting times and waiting room occupancy.

5.1 Deterministic Analysis Based on Averages

From the parameters above, we can estimate doctor income based on the number of patient visits. We can also see how busy they are in a 40-hour work week. Figure 1 shows the panel sizes in the traditional and concierge settings that result in the same revenue for the physician. There is also a line depicting the panel sizes that result in the same utilization. Based on this graph, it is noteworthy that the current policy of dropping the panel size from 2500 to 400 and charging $1000 (patient pays 1500 of which 500 goes to MDVIP) as fees to the patients appears to be very reasonable since they result in similar revenues and utilizations. From here on, our analysis will focus on this transition.

5.2 Stochastic Analysis using Simulation

To understand the role of randomness in the performance of this system, especially the average and maximum waiting times experienced by the patients, we ran simulations of the system and compiled the relevant statistics. We used 2500 patients in the traditional setting and 400 patients in the concierge medicine system. The customer arrival rates were computed based on the 3.19 visits per year per patient and were modeled using a Poisson distribution. The service times were modeled using Normal distribution with coefficients of variation ranging from 0.05 to 0.30. We assumed that the utilization was similar in both setting and was about 90%.

We gathered statistics from one hundred simulation runs and computed the average waiting time and the maximum waiting time for both the settings. The average waiting times were 87.5 minutes and 11.4 minutes for the traditional and concierge settings. The maximum waiting times were 419.4 minutes and 36.2 minutes respectively for the traditional and concierge settings. Given that the concierge fees is $1500 per year and each patient visits the doctor 3.19 times per year, the fees per visit is about 470.2 (=1500/3.19) per visit. The difference in waiting times should more than compensate for these fees and thus we will be able to compute the costs associated with these performance measures for the patients that do sign up. Figure 2 depicts this data.

The physician we consulted with was located in the Louisville area in the zip code 40241. We gathered the census data on the income and age distribution of the people living in that region. Figure 3 depicts this data.
Figure 1: Panel sizes under the traditional and concierge settings that result in the same revenue.

Figure 2: The combinations of hourly costs attributed to average and maximum waiting times that encourage patients to sign up for concierge medicine.
As seen from Figure 3, the top 20% of the population had annual income of over 125,000 implying that the hourly wage for them is approximately $60. In that case, from Figure 2, we can see that the cost attributed to maximum waiting time is also around $60 per hour if the patient is to be induced to sign up for concierge medicine. This is perfectly understandable since, when a person is sick, he/she may be missing work and losing out on wages and other benefits. Based on the simulation results and the income distribution, it is very likely that the top 20% of the population will sign up for concierge medicine. Of course, the costs attributed to waiting (both average and maximum) are subjective in nature and vary quite a bit amongst people, the trade-off depicted in Figure 2 appears to quite effectively segment the customer base.

Based on this analysis, the physician feels very confident that she designed her concierge medicine practice very reasonably. The fact that she reduced her patient panel from 2500 to 400 and the $1500 fees that she charges the patient ensures that her revenues and utilization are similar across the two settings. In addition, the patients experience a significant reduction in average and maximum waiting times and given the income demographics of the region, there exists a significant proportion of the population that will indeed sign up for concierge medicine. A patient struggling with the issue of whether or not to sign up for concierge medicine can use Figure 2 as a tool to answer that question from an individual perspective. If the cost he/she attributes to waiting times falls above the line, then they should sign up for concierge medicine. If that is not the case, they should remain with the traditional system.

6 CONCLUSION

Concierge medicine is an attractive prospect for a traditional practice physician because having fewer patients allows them to spend more time with each patient. The annual retainer fee supplements the office visit charges so the doctor does not have adverse financial outcomes from moving to a smaller panel size. For patients, having the doctor spend an hour with them rather than 10 minutes, is very appealing. In ad-
dition, they also experience reduction in waiting times they experience. Patients that value their time highly will benefit from the concierge practice, and thus be willing to pay the retainer. In this study, we did not include different acuity rates. Patients with less severe issues may more likely not sign up for concierge medicine implying that the average demand per patient may go up under concierge medicine. Future research should account for the additional complexity. Concierge medicine is attractive to both patients and physicians and could lead to better health outcomes for the entire society. However, for that to happen, these practices should be implemented under the right circumstances, should be designed appropriately for the customer base, and managed effectively. This can only be done with a strong analytical foundation for answering these questions and we believe that this paper is a first step in that direction.

REFERENCES


AUTHOR BIOGRAPHIES

SRINAGESH GAVIRNENI is an associate professor of operations management in the Samuel Curtis Johnson Graduate School of Management at Cornell University. For the 2011-12 academic year, he was a visiting associate professor in the Nanyang Business School. His research interests are in the areas of supply chain management, inventory control, production scheduling, simulation, and optimization. His papers have appeared in Management Science, Manufacturing & Service Operations Management, Operations Research, Production and Operations Management, European Journal of Operational Research, Operations Research Letters, IIE Transactions, and Interfaces. Previously, he was an assistant professor in the Kelley School of Business at Indiana University, the chief algorithm design engineer of SmartOps, a software architect at Maxager Technology Inc., and a research scientist with Schlumberger. His under-
graduate degree from IIT-Madras is in mechanical engineering, and he has received an M.Sc. from Iowa State University and a Ph.D. from Carnegie Mellon University.

VIDYADHAR KULKARNI is a Professor in the Department of Statistics and Operations Research at the University of North Carolina – Chapel Hill. He received a B.Tech. from Indian Institute of Technology, Bombay, in Mechanical Engineering in 1976, and an MS and Ph.D. in Operations Research from Cornell University in 1980. He was a Visiting Assistant Professor in the College of Management, Georgia Institute of Technology in 1980-81. He joined the Department of Operations Research, University of North Carolina, in 1981, and has been there ever since. He chaired the department of Statistics and Operations Research from 2004 until 2009. His area of interest is stochastic modeling. He is the author of two textbooks on this topic.

ANDREW MANIKAS is an Assistant Professor in the College of Business at the University of Louisville. He earned his B.S. in Computer Science and M.B.A. in Materials and Logistics Management form Michigan State, and his Ph.D. from the Georgia Institute of Technology. Prior to that he was an instructor for supply chain optimization courses for i2 Technologies. Prior to that, he worked as a management consult for KPMG Peat Marwick, CSC, and Deloitte Consulting.

ALEXIS KARAGEORGE is a primary care physician. She graduated Summa Cum Laude with a Bachelor's Degree in Biology from the University of Louisville. In 1992, she received her Medical Doctorate Degree from the University of Louisville School of Medicine. She completed her Internship and Residency in Internal Medicine at the University of Cincinnati where she received both the Internal Medicine Residents' Teaching Award and First Place for the Department of Medicine Chairman's Presentation. She was selected to serve an extra year as Chief Medical Resident at which time she was awarded the coveted Visiting Fellowship at Addenbrooks Hospital in Cambridge, England. She is certified by the American Board of Internal Medicine and maintains certification for Advanced Cardiac Life Support. She is a member of the American Medical Association, Kentucky Medical Association and The Greater Louisville Medical Society.