Optimizing Physician Processing Rates with Priority Queues

Abraham Seidmann
University of Rochester
avi.seidmann@simon.rochester.edu

Balaraman Rajan
University of Rochester
balaraman.rajan@simon.rochester.edu

Earl R Dorsey
John Hopkins University
ray.dorsey@jhmi.edu

Abstract
The typical family physician, or specialist, sees patients on a scheduled basis for physical exams, chronic care management, or a pre-determined follow-up. They also handle ‘emergency priority’ visits by their patients - handling unexpected clinical conditions such as flu, rapid blood pressure increase, or a sudden allergic reaction to a maintenance drug. This happens as the patients prefer care continuity and expect their regular physician to handle their urgent cases as well. We study the impact of having physicians handle these two kinds of patients’ visits at the same office and the impact of patients’ behavior on the physicians’ capacity and service level for patients. We show that mixing ‘emergency priority’ cases with regularly scheduled visits causes longer waiting times for the regular patients, and it reduces their overall satisfaction. We investigate how a physician can control this behavior by optimally using his levers of care – price and service rate.

1. Introduction

A frustrated patient once quoted “I know that some doctors in South Florida visit their patients (just) for about four minutes... Many times, the waiting room is so full that it is not unusual for a patient to wait three to four hours to be seen! Doctors are overwhelmed”[4]. Long waiting lines and waiting times have become a common problem across health systems. While part of the problem could be attributed to the capacity of the system, there are other inefficiency factors concerning the system.

But healthcare is not like industrial operations where productivity of physicians could be improved by just optimizing the workflow and speed. It is customer intensive and quality of care is somehow perceived as related to the amount of time a physician spends with a patient [6].

Patients visiting emergency rooms for non-emergency conditions is another often cited explanation which is a problem by itself. Cutting emergency room visits and reimbursements might be a great place to start health care cost control is a common belief [5]. While financial incentives can be put in place to motivate the hospitals to reduce ER visits, incorporating patient behavior could go a long way in reducing healthcare costs.

In this paper we focus on how a physician considers both quality of care and waiting time to optimally choose his service rate. While a high service rate will decrease the waiting time and hence the cost to a patient, spending too little a time would also decrease the quality of care and vice versa. On top of this model we incorporate priority queues to include the emergency visits and the negative externality caused by them.

2. Modeling Details

We consider a single physician based on the general notion that each patient prefers a specific specialist. We assume demand from patients follows a Poisson distribution and the specialist offers an exponential service rate resulting in an M/M/1 model, an assumption that is fairly common in healthcare literature [3]. We build on the model by Anand et al [1] to model the quality of care as an affine function of service rate. More specifically, we assume that the patient would prefer more time with the specialist and hence the benefit is decreasing in service rate. Like Anand et al we use \( \alpha \) to represent the proportionality constant.

We assume that the waiting costs are linear in waiting time. Since, the waiting times are inversely proportional to the service rate, we have a trade-off between consultation time and waiting time. On the one hand, more time spent with the patient increases their utility but on the other hand it also increases the waiting cost for patients which decreases their utility.
Thus, the specialist takes this trade-off into account and chooses an optimal service rate. We denote the coinsurance rate by $\beta$ which is the percentage of the actual expense that is required to be paid by the patient as part of their insurance plan. We consider the arrival of emergency patients to be a fraction of the arrival of regular patients ($\delta$). This is based on the assumption that a proportion of regular patients choose to visit the physician in an emergency room in addition to their regular visits.

Patients obtain a value from seeking treatment, and it is assumed to be a constant ‘$v$’. Thus, not everybody seeks treatment and only those for whom their total costs is less than the value that they obtain from treatment, seek treatment.

The utility $U(p, \mu)$ for regular patients seeking treatment is then given by

$$U(p, \mu) = m - \alpha*\mu - \beta*p - c\left(\frac{1}{\mu - (1+\delta)\lambda}\right) + \frac{1}{\mu}$$

where,

- $m$ is a constant (that includes value for seeking treatment $v$ minus any out of pocket costs)
- $\mu$ is the service rate of the physician assumed to be exponential
- $\alpha$ is the proportionality constant for quality of care
- $p$ is the price charged by the physician
- $\beta$ is the coinsurance rate (a fraction between 0 and 1)
- $\delta$ is the proportion of regular patients who seek emergency care
- $c$ is the waiting cost per unit of time and $(1+\delta)\lambda + \frac{1}{\mu}$ is the time spent by the patient in the system (assuming M/M/1 with priority for emergency patients [2])

We assume patients are uniformly distributed on a line, starting from the physician’s location, and form a direct correspondence between distance and $\lambda$. In other words if $U(p, \mu)$ is equated to zero and solved for $\lambda$, it would give the location of the farthest patient who would seek treatment from which we can find the arrival rates to the physician.

We assume the specialist to be a monopolist in his region and the objective function for the physician is simply

$$max_{[p, \mu]} \lambda(p, \mu)(1 + \delta) * p$$

We then find the optimal service rate and price charged by the physician.

3. Results and Insights from the base case

We first observe that if the physician ignores the emergency arrivals ($\delta$) and optimizes his service rate to provide both quality service and optimal waiting for the patients, then the queue explodes with the addition of emergency patients. Hence with limited capacity the physician is either forced to increase his service rate or decrease the number of patients he accepts (reduced panel size). If patients value their time more (higher waiting cost per unit time), then the physician, considering this congestion, sets his optimal service rate to be high even when discarding emergency cases and hence has enough additional capacity. The risk of patients choosing emergency service and clogging the system is thus mitigated with this buffer.

If the physician considers the optimization problem by also factoring in emergency patients ($\delta$), then he could still keep serving all his patients if his initial capacity was high enough. Thus if the waiting cost ($c$) for the patients is high and quality of care is also important ($\alpha$), there is excess capacity in the system and the physician will decide to increase his service rate (decreased quality of care) and also decrease his price to keep serving all his patients and keep them relatively happy. Because of the increased service rate, the effect of congestion on the number of patients seen by the physician is reduced. For a 20% rate of emergency visits, the physician might also need to increase the service rate by close to 20% to keep all his patients. However as more patients tend to choose emergency services, the physician is forced to serve less patients because of capacity and quality constraints.

But if patients do not value their time and hence $c$ is low, then the physician is already at high utilization and hence by incorporating emergency patients his optimal decision would be to serve less patients rather than adjusting his service rate and price. The physician thus suffers a loss in revenue because of loss of patients. Thus access to care rather than quality of care is affected based on the patients’ characteristics ($c$) and behavior ($\delta$). For instance for typical values of the parameters, a 20% rate of emergency visits could not only force the physician to see 17% less number of patients but also increase the waiting time of all the patients by about 6% even at revised optimal service rates and price.

4. Impact of distance on the optimal decisions

We now introduce transportation cost that might be incurred by the patient for each visit in the model. The transportation cost is perceived and is not limited to its traditional meaning. Thus it also includes the discomfort cost for critically ill patients, their care givers and so on. We analyze the impact of
transportation cost on the optimal decisions of the physician.

Ignoring emergency patients we see that traveling costs decrease the optimal service rate as the specialist tends to compensate for the traveling burden by spending more time with the patient. While the optimal service rate is affected by the transportation burden, optimal prices are not affected by it and the optimal price is the same with or without transportation cost. Similar to the scenario without transportation cost, when waiting cost (c) increases, the specialist tends to lower the price and increase the service rate.

5. Different service rates employed by the physician

Instead of assuming that the physician treats both emergency and regular patients the same way, we now allow the physician to choose different rates for these two visits and analyze the effect of $\delta$ (the proportion of regular patients who choose emergency care) on the service rates and waiting cost of regular patients. The waiting time in the system for regular patients with priority queuing is given by

$$\frac{\Delta}{\mu_r} + \frac{\lambda}{\mu_r^2} \left(1 - \frac{\mu_r}{\mu_r + \Delta} \right),$$

where $\mu_r$ and $\mu_u$ are the service rates for regular and emergency visits respectively.

In this case, since the physician can choose a different service rate for emergency visits, the physician optimally tends to choose a higher service rate for emergency visits. The intuitive reason is even though emergency patients are already seen on priority basis and hence their waiting costs are lower than that of regular patients, they indirectly affect the waiting time of regular patients. Hence an increase in service rate enables the physician to ease the system and see more patients. Although this compensates for some of the loss of patients because of congestion, the number of patients seen by the physician at optimality still reduces because of additional emergency patients, especially when the physician is working at capacity (low waiting costs for instance).

6. Summary and Conclusion

In this paper we investigate the effect of a physician taking care of both emergency and regular patients. While seeing their patients themselves, even in the case of emergency, provides care continuity to the patients, it imposes a negative externality on the regular patients whose waiting time increases.

We explore how a physician would optimally solve this problem under different scenarios. We see that if patients do not value their time, then the physician tends to run close to capacity maximizing his revenue. Hence, any additional emergency patients clog the system and the physician is forced to forego seeing some of his patients. Thus, it decreases the effective capacity of the physician as the number of patients seen by him reduces. Hence access to some patients might be reduced in order to compensate for the negative externality caused by the emergency patients.

If patients attach high value to their time, then physicians keep excess capacity to ensure patients wait less and with addition of emergency patients, they increase their service rate to still keep up with the demand from the patients. Hence, quality of care for everybody might reduce because of the new addition of emergency patients in order to reduce the negative externality caused by them. However, if the proportion of emergency patients increases, the physician is forced to reduce the number of patients coming to him.

We also observe that if the physician is allowed to choose different service rates for emergency and regular visits, then the optimal service rate for emergency visits is higher than the optimal service rate for regular visits. While emergency patients already have lower waiting costs as they are treated based on priority, the negative externality they cause on the regular patients may force the physician to spend less time with them.

7. References