

# OPTIMAL STAFFING POLICY AND TELEMEDICINE

## ABSTRACT

*In this paper, we study the optimal strategy of a specialty hospital in providing traditional face-to-face consultations via experts and remote medical services via tele-specialists. We model the whole system as a queuing problem and provide the optimal staffing policy for this hospital by taking into account the various cost components, such as those for staffing, mistreatment, and waiting. We also find the optimal investment in telemedicine technology that offers the best trade-off between the quality and accuracy of telemedicine services and the cost of technology. Under certain conditions, the hospital does not offer any telemedicine services. When it does, it may or may not invest in the most advanced technology available. Finally, we provide the optimal tele-specialist policy of which patients to treat remotely via telemedicine and which patients to refer to the experts for a face-to-face consultation. We show that a policy of treating all patients via tele-medicine is never optimal.*

## Keywords

Telemedicine, technology investment, queuing system

## Introduction

Enormous potential exists to improve health services throughout the world by using information and communication technologies to expand access to primary, secondary, and tertiary care, raise quality, increase efficiency, and decrease costs (Bashshur 2002, Brown 2003, Pew 2005). By providing greater access to medical expertise, telemedicine can reduce the geographical variability of diagnosis and clinical management (Heinzelmann et al. 2005). In particular, tele-consultations have been shown to change diagnoses and management recommendations, and also to reduce the long waiting times associated with access to high-demand specialty care (Kedar et al. 2003). A recent report by the Institute of Medicine, *Crossing the Quality Chasm*, states that “patients should receive care whenever they need it and in many forms, not just face-to-face visits...access to care should be provided over the Internet, by telephone, and by other means” (Borst et al., p. 61).

Some of the top experts in the United States provide online second opinions for important diagnoses and treatments. The Cleveland Clinic, a well-known multi-specialty academic medical center in Cleveland, has established its online interface, the eClevelandClinic, to serve patients who need advice and possibly a major intervention, but who cannot easily access the doctors in person. The electronic service is limited to life-threatening and life-altering conditions that can be safely assessed online, such as a new cancer diagnosis, cardiac procedures, joint replacements, and neurological problems. Patients provide a personal medical history and the original diagnosis as well as other relevant materials such as test results, MRI, films, x-rays, and a consent form. Three Harvard University teaching hospitals have initiated a similar service at [econsults.partners.org](http://econsults.partners.org). In addition, several leading oncologists render medical opinions at [www.mdexpert.com](http://www.mdexpert.com) and charge fees as high as \$3,200.

Despite the recent popularity of telemedicine, there is little research that analyzes its economic viability with respect to alternative approaches using sound methodologies. Instead, research on telemedicine is mainly conducted to answer the question “Can we do this?”, while leaving the important question “Should we do this?” unanswered (Kedar et al. 2003). Cost information about telemedicine applications are very preliminary and are often concerned with making a case to proceed further (Hailey 2005). Several economic studies indicate avoidance of travel or patient transfer as one of the most salient benefits of telemedicine, but most of these studies have methodological limitations, and the generalizability of findings is rather limited (Hailey et al. 2002, Hailey et al. 2004). Consequently, evaluating the effectiveness of telemedicine services is difficult at best.

Even when cost-effectiveness of telemedicine is evident, relevant frameworks that aid in defining the appropriate scope and application of telemedicine in different settings do not exist. In other words, the important question of “How should we do this?” is also unanswered. This is an important limitation because the adoption of telemedicine involves making decisions on complicated matters. For example, when a specialty hospital considers offering health care to remote locations via telemedicine, it should first decide on how much to invest in technology and training while taking into account concerns

about both quality and cost. Furthermore, it should employ the appropriate number of qualified personnel who will serve via its face-to-face and telemedicine channels. Referrals among experts that serve face-to-face and via telemedicine should also be coordinated. Unless these important decisions are made with adequate understanding of the related economic issues, adoption of telemedicine can easily fail.

The main objective of this paper is to provide a decision framework and insights on when and how a specialty hospital should adopt telemedicine. Our main contributions to the literature are the following. First, while taking into account various cost factors, we find the optimal staffing policy for the hospital (i.e., the number of face-to-face traditional experts and tele-specialists that need to be employed). Second, we determine to what extent the hospital should invest in telemedicine technology to improve the skills of its tele-specialists. Third, we find the optimal referral policy for the tele-specialists (i.e., when they should attempt to treat a patient via telemedicine after the initial diagnosis and when they should invite the patient to the hospital). Given the cost of mistreatment, a model of the tele-specialist's effectiveness, and the relative costs of service via the two channels, it is possible to identify a referral policy that is optimal for the system as a whole.

Before proceeding further, it is important to provide a brief discussion on the way health care services are structured. Typically, one observes a two-tiered structure where primary care physicians serve as gatekeepers to medical specialists. In many countries (such as Australia, Norway, and the United Kingdom) general practitioners act as gatekeepers for specialty services. Although private medical insurance has not traditionally used gatekeepers in the United States, many managed care plans use gatekeeper arrangements, either by requiring a referral from a specified primary care physician before consulting a specialist (Glied 2001) or by fully insuring provision only if it is supplied, or authorized, by the responsible health maintenance organization (Malcomson 2007).

The tiered provision of health care may take many forms. In some countries with a shortage of expert doctors (such as Australia), governments are in the process of introducing a novel system where patients would be first directed to call the tele-specialist (Nader 2007). It is possible that the tele-specialist could be a highly trained nurse, who can diagnose and treat relatively uncomplicated cases with the use of advanced technology. An example of an advanced technology is two-way interactive television that is typically used for interactive consultations with specialists at urban medical centers. There are many peripheral devices which can aid in an interactive examination. For instance, an otoscope allows a physician to "see" inside a patient's ear; a stethoscope allows the consulting physician to hear the patient's heartbeat. With adequate training and technology, a tele-specialist can determine quickly whether the patient should be directed to the expert doctor or whether she should treat the patient herself. This is akin to the triage system now prevalent in the Accident & Emergency (AE) departments of hospitals worldwide, with the differences being (i) there is now a highly skilled tele-specialist able to also treat the patients, and (ii) the patient does not personally see the tele-specialist but is attended to over the phone. The tiered arrangement is implemented to better utilize the expert doctors' time as well as other resources in the hospital system, as well as preventing non-emergency patients from clogging the system. We also incorporate this tiered provision of health care services into our analysis.

The rest of the paper is organized as follows. We review the literature in the next section. We then set up the model and then provide a sensitivity analysis that demonstrates the effects of various parameters on the optimal policy of staffing and technology investment. A discussion of the results, limitations, and future work is provided in the last section.

## Literature Review

From a methodological perspective, this paper is related to the literature on capacity planning in queuing networks. Halfin and Whitt (1981) establish heavy-traffic stochastic limits for multi-server queues in which the number of servers is allowed to increase along with the traffic intensity. Whitt (1992) shows that by using a square-root staffing principle, one can generate the same limiting regime as in (Haflin et al. 1981). Borst et al. (2004) use a similar framework for asymptotic optimization of a queuing system. Given a cost function involving both waiting and staffing costs, they demonstrate the asymptotic optimality of the square root staffing principle. We apply their approximation to a two-tier queuing network. We use the square-root staffing rule to find the optimal number of tele-specialists and experts. Taking the technological skill level and referral policy of tele-specialists as the decision variables, we then determine the total cost of operating such a system and minimize that cost.

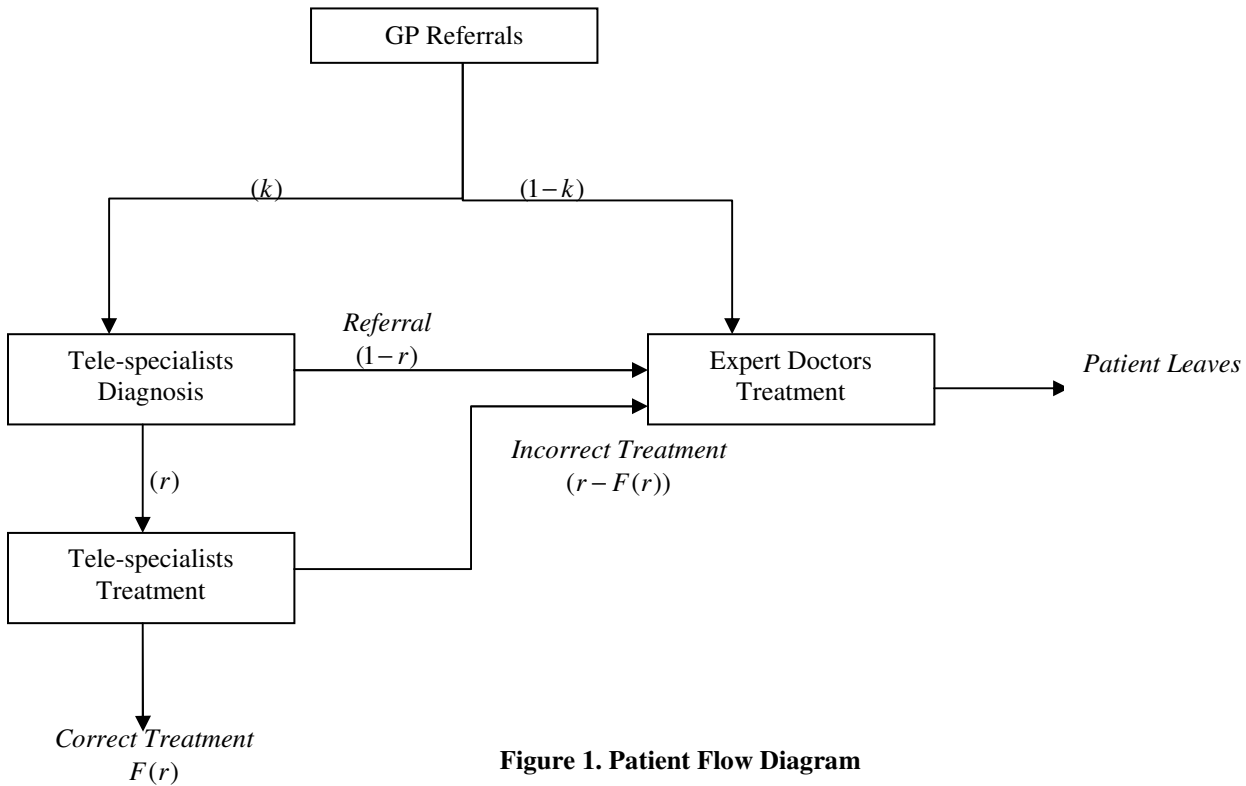
Shumsky and Pinker (2003) determine the optimal referral policy for a deterministic two-tiered system and, using a principal-agent model, study the incentive issues that affect gatekeepers' performance. Although we do not consider the incentive problem here, we adopt their treatment function in our analysis (as discussed in the next section). We also incorporate queuing effects by relaxing their assumption that the firm always achieves the exogenous waiting time goals.

Our paper closely follows (Haflin et al., 1981) where the authors consider a similar two-tier system to optimize the total system costs. They analyze how the cost-minimizing referral rate varies with changes in arrival and service rates, and show that as the arrival rate increases, the optimal referral rate converges to the optimal referral rate for the deterministic case. One

of their findings is that if the gatekeepers' (tele-specialists in our paper) skill level is not high, a direct access system is always the preferred option. This is a very important observation because, as pointed out in the previous section, the gatekeeper of the telemedicine system envisaged by some governments will be a highly trained and skilled nurse and not a doctor. So, the natural questions that arise are (i) what skill level should the gatekeepers attain, and (ii) is it worth investing in? Further, there is a high likelihood for incorrect diagnosis and mistreatment of a patient by such a tele-specialist, potentially leading to litigation costs if the patient goes to a court of law. Hence, it is very important that these personnel are highly skilled, and there needs to be a program to train these gatekeepers and also to invest in the latest technology to help bring them up to a satisfactory level to operate the new technology, diagnose the patients' condition, and treat them, if necessary. Obviously this involves capital investment. In this paper we introduce training costs and technology investment costs to the model of (Haflin et al., 1981), and analyze whether a telemedicine system consisting of the traditional experts and tele-specialists is preferred over the usual direct access system given the cost of technology.

## The Model

As noted in the previous section, we consider a specialty hospital to which referrals are made by the General Practitioners (GPs). The hospital may employ a pool of experts who provide traditional face-to-face consultations and a pool of tele-specialists who provide medical services remotely. In the current model, a tele-specialist need not be either an expert doctor or a generalist. Figure 1 illustrates the queuing network model considered here.



**Figure 1. Patient Flow Diagram**

Patients are always seen first by the General Practitioners (GP). Upon an unsuccessful outcome with the GP, the patients either call the tele-specialist or request a personal appointment with one of the expert doctors in the specialty hospital. We assume that the proportion of patients seeking tele-specialists (denoted by  $k$ ) depends on the skill level of the tele-specialists (denoted by  $b$ ), where the skill level  $b$  is a function of the investment in technology. In other words, the hospital can improve its tele-specialists' ability to successfully diagnose and treat remote patients through costly investment in technology. Any patient remotely contacting the specialty hospital waits for an available tele-specialist. The tele-specialist first tries to diagnose the patient's ailment. If she decides that she can also treat the patient then she does so; otherwise, the patient immediately gets referred to the expert pool to be physically seen by an expert at the hospital. Expert doctors are assumed to provide the correct treatment to the patients' full satisfaction. On the other hand, remote interactions with the tele-specialists may end unsuccessfully, in which case the system incurs a cost for mistreatment. The mistreatment by the tele-specialists is modeled as in (Hasija et al.2005). We assume that patients contacting the tele-specialists have ailments at complexity/difficulty level  $X$ , which is a random variable following a uniform distribution on the unit interval  $[0, 1]$ . Higher complexity levels imply more difficult cases which require higher skill levels. Given the complexity level  $x$ , we adopt the treatment function  $f(x)$ , as defined by (Shumsky et al. 2003), to be the probability that a tele-specialist successfully treats the

patient. We also assume infinite waiting space at both the tele-specialists and the expert doctors. Specifically, we use the following notation:

|                                |   |
|--------------------------------|---|
| $s$ :                          | Index to denote tele-specialist.  |
| $e$ :                          | Index to denote experts.  |
| $n_i$ :                        | Number of tele-specialists ( $i = s$ ) or experts ( $i = e$ ). <b>Decision variables.</b>   |
| $\lambda$ :                    | Arrival rate of patients to the GP.   |
| $b \in [0,1]$ :                | The skill level of tele-specialists. <b>A decision variable.</b>  |
| $k(b) = 1 - e^{-\kappa_1 b}$ : | Proportion of patients, referred by GP, that choose tele-specialists over experts, a function of skill level, $b$ .   |
| $r \in [0,1]$ :                | The fraction of patients referred by the tele-specialist to the experts. <b>A decision variable</b>   |
| $\mu_d(b)$ :                   | A tele-specialist's diagnosis rate, a function of $b$ .   |
| $\mu_t(b)$ :                   | A tele-specialist's diagnosis and treatment rate, a function of $b$ .   |
| $\mu_e$ :                      | An expert doctor's treatment rate.  |
| $n_s$ :                        | The number of tele-specialists in the system.   |
| $n_e$ :                        | The number of expert doctors in the system.   |
| $w_i$ :                        | The respective costs per unit time to the system for making a patient wait at $i = s, e$ .  |
| $c_i$ :                        | The respective unit staffing costs at $i = s, e$ .  |
| $m$ :                          | Inconvenience cost to a customer due to the failed treatment at the tele-specialist.  |
| $W(n, \lambda, \mu)$ :         | The expected waiting time in a $M/M/n$ system with $\lambda$ and $\mu$ as the respective arrival and service rates.   |
| $TC_i$ :                       | The total cost per unit time for the subsystem, ( $i = s, e$ ) which is a function of $n, b$ and $r$ .  |
| $STC$ :                        | The system total cost, a function of $n, b$ and $r$ .   |
| $f(r) = b(1-r)$ :              | The treatment function as defined in [10, 17]   |
| $F(r) = \int_0^r f(x)dx$ :     | Expected fraction of patients treated successfully by the tele-specialists. It is easy to see that $r - F(r)$ is the fraction mistreated and $1 - F(r)$ is the fraction referred to expert doctors. |
| $C(b) = C_1[e^{c_2 b} - 1]$ :  | The training and technology investment cost to improve the skill level of tele-specialists.   |

Note that the training and technology investment cost function  $C(b)$  is assumed to be increasing in the skill level of the tele-specialists ( $b$ ) at a decreasing returns to scale. We further assume that there are theoretical maximum rates of diagnosis and treatment for the tele-specialists which is attained when  $b = 1$ . Accordingly, we have  $\mu_d(b) = \mu_1 b$  and  $\mu_t(b) = \mu_2 b$ , where  $\mu_1$  and  $\mu_2$  are the respective theoretical maximum rates. Please also note that, to maintain consistency, we assume all the cost parameters are given in cost per unit time.

It is clear that the telemedicine system as defined above is a queuing network. The flow into the two subsystems of tele-specialists and expert doctors are Poisson due to the assumptions made above. However, while the service time distribution is assumed to be exponentially distributed for the expert doctors, it is not so for the tele-specialists. A tele-specialist's service time is exponentially distributed when she diagnoses a patient and immediately refers the patient to the experts, and is the sum of two exponential distributions when she decides to treat the patient following the diagnosis. So, effectively the service time distribution of a tele-specialist is considered as a double exponential distribution (i.e., a mixture of exponential distributions). Similar to (Hasija et al. 2005), we use two approximations to analyze the queuing network described above.

(i) We assume that the effective service rate of a tele-specialist is approximately exponentially distributed with the effective rate given by

$$\bar{\mu}(b) = \left( \frac{1-r}{\mu_d(b)} + \frac{r}{\mu_t(b)} \right)^{-1}. \quad (1)$$

(ii) The approximation (i) above makes the subsystems of the tele-specialist and expert doctor pools to be  $M/M/n$  queuing systems where the arrival rates into these subsystems are as shown in Figure 1. One of the decisions to be made is the staffing levels of tele-specialists and expert doctors. For finding these optimal staffing levels, we use the heuristic suggested by (Borst et al. 2004). For a brief description of the heuristic we refer the reader to [10] which shows the closeness of the heuristic solution to the exact optimal solution. We present below the relevant formulas:

For a  $M/M/n$  queuing system with offered load  $\rho = \lambda/\mu$ , (Borst et al. 2004) shows the asymptotically optimal staffing level to be

$$\hat{n} = \rho + y^*(c, w)\sqrt{\rho}, \quad (2)$$

where  $c$  and  $w$  are the unit cost rate of staffing and the unit cost rate of waiting. The above expression has the interpretation that  $\rho$  is the minimum staffing level required and the second term is the safety staffing level due to the uncertainty in the system. Here,  $y^*(c, w)$  is the minimizer of the function

$$\alpha(y, c, w) = cy + \frac{w\pi(y)}{y}, \quad (3)$$

where

$$\pi(y) = \left[ 1 + \frac{y\Phi(y)}{\varphi(y)} \right]^{-1} \quad (4)$$

and  $\varphi(y)$  and  $\Phi(y)$  are the probability distribution function and the cumulative distribution function of the standard normal distribution, respectively. Thus,

$$y^*(c, w) = \arg \min_{y>0} \alpha(y, c, w). \quad (5)$$

It is known (Hasija et al. 2005) that  $\alpha(c, y, w)$  is unimodal, and hence  $y^*(c, w)$  can be determined using a numerical procedure. Further, the approximate total cost of staffing and waiting is known to be

$$TC^{app}(\rho, c, w) = c\rho + \alpha(y^*(c, w), c, w)\sqrt{\rho}. \quad (6)$$

In the sequel, we will use expression (6) in our total cost calculations.

We are now ready to derive the total cost function. The system now consists of two  $M/M/n$  queuing systems. From queuing theory, we now have

$$TC_s(n_s, b, r) = \lambda k(b)w_s W(n_s, \lambda k(b), \bar{\mu}(b)) + c_s n_s + m\lambda[r - F(r)] + C(b), \quad (7)$$

$$TC_e(n_e, b, r) = \lambda(1 - k(b)F(r))w_e W(n_e, \lambda(1 - k(b)F(r)), \mu_e) + c_e n_e \quad (8)$$

and the combined total cost function is

$$TC = TC_s(n_s, b, r) + TC_e(n_e, b, r). \quad (9)$$

If we use the approximations (2) and (6) in (9), the total cost function just becomes a function of  $b$  and  $r$  only. Thus, the total cost function (9) approximates to

$$STC^{app}(b, r) = \begin{cases} TC^{app}(\rho_e, c_e, w_e), & \text{if } b = 0 \\ TC^{app}(\rho_s, c_s, w_s) + TC^{app}(\rho_e, c_e, w_e), & \text{otherwise} \end{cases} \quad (10)$$

where

$$\rho_s = \frac{\lambda k}{\bar{\mu}(b)} \quad (11)$$

and

$$\rho_e = \frac{\lambda[1 - k(b)F(r)]}{\mu_e}. \quad (12)$$

## Analysis

In this section, we first characterize the optimal decisions and then analyze the effect the model parameters on these decisions. We note that no investment in telemedicine ( $b = 0$ ) implies that the system will have no tele-specialists and will have only expert doctors to attend to the patients ( $r = 0$ ). We find the following results.

**Proposition 1.** The optimal levels of investment in telemedicine and referrals from tele-specialists to the experts, given by  $b^*$  and  $r^*$ , respectively, is one of the following three combinations:

- (i)  $b^* = 0$  and  $r^* = 0$ , implying that no tele-specialists should be employed.
- (ii)  $r^*$  is the solution(s) satisfying the first order condition  $\frac{\partial}{\partial r} STC^{app}(b, r) = 0$ , while  $b^*$  is given by the solution(s) of  $\frac{\partial}{\partial b} STC^{app}(b, r) = 0$ .

or

- (iii)  $r^*$  is the solution(s) satisfying the first order condition  $\frac{\partial}{\partial r} STC^{app}(b, r) = 0$  and  $b^* = 0$ .

**Proof.** The proof is based on usual calculus-based arguments studying the behavior of first and second order conditions. We find that the second derivative of the cost function with respect to  $r$  is not always positive or negative. Further, we find that the first derivative with respect to  $r$  is always positive evaluated at extreme point  $r = 1$  but it could be positive or negative evaluated at extreme point  $r = 0$ . Combining these facts, we know that the optimal  $r$  value must be either zero or the internal  $r$

value(s) that satisfies the first order conditions. A similar analysis with respect to  $b$  shows that the first derivative evaluated at extreme points 0 and 1 can be positive or negative; therefore, the optimal  $b$  value could be 0, 1 or an internal solution satisfying first order conditions. For a complete proof, please contact the authors of this paper.

Proposition 1 characterizes the optimal solution, which can be found by simply comparing the values of the total cost function at the extreme values for the decision variables along with any value that satisfies the first-order conditions. This is illustrated in a numerical example in the next section to obtain the optimal solution and also to perform a sensitivity analysis. Further, we observe  $r^* < 1$ , which implies that the tele-specialists never completely replace expert doctors. This should be expected since tele-specialists are prone to mistreatments while the experts are not.

We use the implicit function theorem to derive the theoretical effect of the model parameters on the optimal outcomes. This method gives us comparative statistics, which enable us to determine the change in the optimal  $b$  and  $r$  values when one of the model parameters is altered. This analysis is very valuable in terms of determining whether more or less technology investment ( $b$ ) is needed and/or how the proportion of patients treated by tele-specialists ( $r$ ) should change as a result of . Proposition 2 outlines the results of this analysis (obtained by applying the implicit function theorem to  $STC^{app}$ ). Please note that we limited the comparative statistics analysis to internal solutions only; that is, for  $0 < r < 1$  and  $0 < b < 1$ .

**Proposition 2.** The (internal) optimal referral ( $r$ ) value is *increasing* in diagnosis plus treatment rate of tele-specialists ( $\partial r/\partial \mu_t > 0$ ), the waiting cost at experts ( $\partial r/\partial w_e > 0$ ) and the cost of staffing of experts ( $\partial r/\partial c_e > 0$ ), but is *decreasing* in the cost of staffing of tele-specialists ( $\partial r/\partial c_s < 0$ ), the mistreatment cost ( $\partial r/\partial m < 0$ ), the diagnosis rate of tele-specialists ( $\partial r/\partial \mu_d < 0$ ), the treatment rate of experts ( $\partial r/\partial \mu_e < 0$ ) and the waiting cost at tele-specialists ( $\partial r/\partial w_s < 0$ ).

Proposition 2 states that as  $c_s$ ,  $m$ ,  $\mu_d$ ,  $\mu_e$  and  $w_s$  increase, the optimal  $r$  value decreases. This implies that for large values of these parameters, it is optimal for tele-specialists to treat less difficult cases – and thus a lower proportion of patients. On the other hand, as  $c_e$ ,  $\mu_t$  and  $w_e$  increase, the optimal  $r$  value increases as well. Hence, for large values of these parameters, tele-specialists should treat more difficult cases which will result in a higher proportion of patients.

Please note that in Proposition 2, the implicit assumption is that the optimal  $b$  value remains the same. Of course, this will not be true when the model parameters change unless  $b$  is an external solution. Therefore, we need to keep this fact in mind when using Proposition 2. Nevertheless, combining the results of Propositions 2 and 3 (presented below) will give us valuable insights into the change in optimal strategies when model parameters change.

**Proposition 3.** The (internal) optimal technology ( $b$ ) value is *increasing* in the cost of staffing of tele-specialists ( $\partial b/\partial c_s > 0$ ), the waiting cost at tele-specialists ( $\partial b/\partial w_s > 0$ ) and the arrival rate of patients ( $\partial b/\partial \lambda > 0$ ) but is *decreasing* in the cost of staffing of experts ( $\partial b/\partial c_e < 0$ ), the treatment rate of experts ( $\partial b/\partial \mu_e < 0$ ), the waiting cost at experts ( $\partial b/\partial w_e < 0$ ) and the cost of technology parameters ( $\partial b/\partial C_1 < 0$ ,  $\partial b/\partial C_2 < 0$ ). Also, the optimal technology level is decreasing in the mistreatment cost if the technology level is already low but increasing if the level is already high ( $\partial b/\partial m < 0$  for small  $b$  and  $> 0$  for large  $b$ ). The optimal technology level is increasing in diagnosis and diagnosis plus treatment rates of tele-specialists at low technology levels but decreasing at high levels ( $\partial b/\partial \mu_d$ ,  $\partial b/\partial \mu_t$  are both  $> 0$  for small  $b$  and  $< 0$  for large  $b$ ).

Proposition 3 states that as  $c_s$ ,  $\lambda$  and  $w_s$  increase, the optimal  $b$  value increases, implying that for large values of these parameters, the hospital should invest more in tele-medicine technology. Please note that increase in these parameters would increase total cost. Hence, the hospital would (partially) negate this increase in total cost by investing in better technology. On the other hand, as  $c_e$ ,  $\mu_e$ ,  $C_1$ ,  $C_2$  and  $w_e$  increase, the optimal  $b$  value decreases. Thus, for large values of these parameters, investment in better tele-medicine technology is not economically justified.

## Conclusion and Future Research Directions

Telemedicine is of great importance at a time when healthcare costs are increasing at a faster rate than inflation, while some remote/rural areas in even developed countries lack good quality specialty healthcare services. Therefore, in this paper, we have studied a specialty hospital that provides traditional face-to-face medical consultation (via experts) as well as remote telemedicine (via tele-specialists) services. Specifically, using queuing theory, we have provided optimal staffing policies that determine the number of experts and number of tele-specialists the hospital should employ. Our analysis reveals that, although it does not make economic sense to always have tele-specialists, a specialty hospital should always employ experts.

Another important decision we have analysed is the optimal technology level in telemedicine. Presumably, the hospital can invest in technology that will raise the skill level and diagnosis/treatment accuracy of tele-specialists. For instance, the hospital can invest in state-of-the-art videoconferencing technology and/or powerful medical equipment to provide better

telemedicine service to patients. Of course, the more advanced the technology becomes, the more expensive it is going to be. Therefore, we find the optimal investment in technology level that provides the optimal balance between cost and service.

The last major decision we provide is regarding the behavior of tele-specialists. When patients call/contact tele-specialists, these tele-specialists have two choices after analyzing patient information: either to attempt treating the patient if it seems feasible or refer him to one of the experts for an in-person consultation. Therefore, the important question is “What percentage of patients should be treated by the tele-specialists and what percentage should be referred to the experts?” In this paper, we provided this optimal referral rate to the experts by the tele-specialists, which is dependent on the complexity of the ailment.

In the analysis, we have studied the effects of various parameter changes on the optimal decisions. For example, we show that as the cost of staffing tele-specialists goes up, the hospital should hire fewer tele-specialists. However, the investment in technology should be raised, in effect replacing people with technology. In another analysis, we found that increases in the arrival rate of patients necessitate a larger investment in technology and hiring more tele-specialists and experts. On the other hand, the optimal ratio of patients to receive tele-specialist treatment first increases and then decreases with the arrival rate.

There are a number of avenues for future research. We first would like to conduct a numerical analysis to better understand the intuition behind the results. One immediate extension to the current model would be to include the General Practitioner (GP) as a decision maker, in effect creating a 3-tier process: GP, tele-specialist and expert. Another idea would be to include public benefits, costs and policies in our model. For example, the hospital could be required to provide a minimum medical service level to remote areas, in which case tele-specialists would be hired even if it doing so were not economically feasible.

## References

1. Bashshur, R.L., “Telemedicine and Health Care,” *Telemedicine Journal and E-Health* (8:1), 2002, pp. 5-12.
2. Borst, S., Mandelbaum, A. and Reiman, M. I., “Dimensioning Large Call Centers,” *Operations Research* (52:1), 2004, pp. 17-34.
3. Brown SJ., “Next Generation Telecare and its Role in Primary and Community Care,” *Health and Social Care in the Community* (11:3), 2003, pp. 459-462.
4. Committee on the Quality of Health Care in America, Institute of Medicine., *Crossing the quality chasm: A new health system for the 21<sup>st</sup> century*. Washington, DC: National Academy Press, 2001.
5. Glied, S., “Managed Care,” In A.J. Culyer and J.P. Newhouse, eds., *Handbook of Health Economics*, Vol. 1A. Amsterdam: Elsevier Science, 2000.
6. Hailey, D., “The Need for Cost-Effectiveness Studies in Telemedicine,” *Journal of Telemedicine and Telecare*, (11:8), 2005, pp. 379-383.
7. Hailey, D., Roine, R., and Ohinmaa, A., “Systematic Review of Evidence for the Benefits of Telemedicine,” *Journal of Telemedicine and Telecare* (8), 2002, pp. 1-7.
8. Hailey, D., Roine, R., and Ohinmaa, A., “Study Quality and Evidence of Benefit in Recent Assessments of Telemedicine,” *Journal of Telemedicine and Telecare* (10:6), 2004, pp. 318-324.
9. Halfin, S., Whitt, W., “Heavy-Traffic Limits for Queues with Many Exponential Servers,” *Operations Research* (29:3), 1981, pp. 567-587.
10. Hasija, S., Pinker, E. J. and Shumsky, R. A., “Staffing and Routing in a Two-Tier Call Center,” *International Journal of Operational Research* (1:1/2), 2005, 8-29.
11. Heinzlmann, P.J., Lugn, N.E., Kvedar, J.C., “Telemedicine in the Future,” *Journal of Telemedicine and Telecare* (11:8), 2005, pp. 384-390.
12. Kedar, I., Ternullo, J.L., Weinrib, C.E., Kelleher, K.M., Brandling-Bennett, H., Kvedar, J.C., “Internet Based Consultations to Transfer Knowledge for Patients Requiring Specialised Care: Retrospective Case Review,” *British Medical Journal* (326), 2003, pp. 696-699.
13. Mair, F.S., Haycox, A., May, C., and Williams, T., “A Review of Telemedicine Cost-Effectiveness Studies,” *Journal of Telemedicine and Telecare* (6), 2000, pp. 38-40.
14. Malcomson, J.M., “Health Service Gatekeepers,” *RAND Journal of Economics* (35:2), 2004, pp. 401-421.
15. Nader, C., “Triage by telephone,” *The Age* (27), January 2007.
16. Pew Internet and American Life Project., “The Future of the Internet: In a Survey, Technology Experts and Scholars Evaluate Where the Network is Headed in the Next Ten Years,” Washington, DC: Pew Internet, 2005. See [http://www.pewinternet.org/pdfs/PIP\\_Future\\_of\\_Internet.pdf](http://www.pewinternet.org/pdfs/PIP_Future_of_Internet.pdf) (last checked 4 January 2007).
17. Shumsky, R. A. and Pinker, E. J., “Gatekeepers and referrals in service,” *Management Science* (38:5), 2003, pp. 708-723.
18. Whitt, W., “Understanding the efficiency of multi-server service systems,” *Management Science* (38:5), 1992, pp. 708-723.