Optimizing patient contact: the tradeoff between serving a physician's office practice and hospital referrals [1]

Abstract

We study a planning problem that applies to any private practice physician who accepts hospital consultations in the United States health care system. Medical specialists who maintain independent office practices also serve the public by accepting service referrals from community hospitals that would otherwise lack such specialties. Accepting a referral requires a trip to the hospital, takes time away from the office practice, generates revenue but arrives randomly, requires an uncertain number of follow-up hospital visits, and results in a variable daily time commitment. Using empirical data from a neurologist who accepts referrals from a New York City hospital, we developed and implemented solutions to three related problems: the ideal number of hours per day to plan for hospital rounds, the conditions for which patient referrals should be rejected, and the order in which new referrals and follow-up patients should be seen during hospital rounds. This paper describes the solution methodologies, develops effective heuristics, addresses the sensitivity to parameters that change, and evaluates the impact of the models as well as the implementation issues that we encountered in practice. We conclude that with proper planning, private-practice medical specialists can serve community hospitals in a sustainable way. We also propose one way that American hospitals can viably counterbalance the punitive impact of longer lengths of stay in the United States' DRG (Diagnostic-Related Group) system for the best interests of patient care.

1 Introduction

Our subject healthcare facility is a 235-bed community hospital in a borough of New York City. The hospital provides standard medical care to admitted inpatients with consultations by various medical specialists when appropriate, including neurologists. A patient's admitting physician decides when a neurologist should be called and then neurology consultations are provided by voluntary (i.e., not hospital-employed) private practice physicians in the community who travel to the hospital. After seeing an inpatient for an initial consultation, the neurologist returns for follow-up visits every subsequent day the inpatient remains admitted.

[1] Revised 27-Jan-2012; submitted to the 2012 Healthcare Best Paper Competition, POMS College of Healthcare Operations Management.

Reimbursement for these medical services is provided by the patient's medical insurance, billed privately by the neurologist for each day the patient is seen. We estimate that an initial consultation requires 45 minutes and reimburses \$150 on average; follow-ups need 15 minutes and reimburse \$50. Only one initial consultation is billed per patient admission with a follow-up billed every day the neurologist sees the patient thereafter.

Neurologists encounter a planning problem due to variability both in the number of daily new hospital referrals and each patient's length of stay (LOS) and therefore the number of follow-up visits required. The neurologist's dilemma is how much time to reserve for daily hospital rounds—if improperly managed there will be idle time or rounds late into the night.

We examined this planning problem using 395 days' empirical data collected June 1, 2009 to June 30, 2010 by the third author who leads an independent neurology office practice and responds to hospital referrals at our subject hospital. The daily consultation time history is summarized in Figure 1 in the appendix.

The neurologist's hospital work must be balanced by his outpatient office practice where excess appointment demand fluctuates between a one and two-week backlog (waiting-list) all year long. At \$200 to \$500 per hour, office revenue exceeds hospital referral insurance reimbursement rates and so one might ask why he serves the hospital rather than expanding office hours. Accepting hospital referrals benefits the local community through greater access to neurological services, assists the hospital in providing its patients access to care, furthers the reputation of the neurologist's practice, allows relationship-building with referring doctors, and provides extra income if planned correctly. If hospital care is done at times when office patients are not normally seen, such as early mornings or late evenings, total reimbursement is maximized and the greatest number of patients is served. Should hospital work become so time-consuming that it interferes with office visit time, however, it is in the neurologist's best interest to begin refusing hospital referrals because the opportunity cost is the loss of office patient time.

In short, the operational dilemma is how to handle hospital referrals given the finite available time and sources of variability. It is here that Operations Research (OR) modeling can improve the neurologist's allocation of time. The goal of this work is to build stochastic models that help the neurologist plan how he supports the hospital referral process in a manner that both maximizes patient access and is economically attractive. Without financial viability, supporting the hospital otherwise makes no sense.

Financial sustainment depends on happy customers. As is often the case in healthcare (e.g., Prahinski et al. 2003), the customers include the patients, insurance providers, hospital administrators, and referring physicians. Referring physicians are critical—if too many cases are sent simultaneously, the quality of care may be impacted. If rejected too often, though, a referrer may stop sending patients altogether; thus rejection is minimized, in general, by private specialists. In our analysis (and in practice), rejection is employed only during the busiest two or three days of the year. We demonstrate that this policy is also optimal financially yet sensitive to to certain exogenous parameters.

During our 2009-2010 study, the number of daily referrals fluctuated between 0 and 7 while the number of referring physicians remained constant. In this period, 18% of patients under the neurologist's care were discharged per day on average. In other words, if a patient is seen today, there is an 82% chance of a follow-up tomorrow, driven by a LOS that averages approximately 5 days for hospital inpatients receiving neurology care. Modeling each patient's need for a follow-up using the binomial distribution with probability of "success" 0.82 is the most tractable way to capture variability due to LOS fluctuations. All private-practice specialists who serve hospitals and bill privately presumably keep similar patient history information for billing purposes; therefore, the number and type of patient encounters per day are readily available in various specific scenarios, making this model easily reproducible and testable in a range of specialities and localities.

The majority of hospital admissions are in stable condition when referred, and so the neurologist may attend to his office's appointment schedule first and then make a single hospital trip (called *rounding*) for the referrals that accumulate during the day and follow-ups. Alternatively, rounding before office hours may be tried, but the time needed to see all the patients is not always enough to prevent a second trip after office hours.

To address the backlog of office practice demand and to expand patient access to care at the hospital, the neurologist partnered with a board-certified neurologist on a part-time basis. Thus all hospital referrals are "outsourced" to the part-time neurologist who is salaried at the equivalent of $100 per hour, and the practice keeps any the insurance reimbursement revenue less this salary and other overhead costs consistently mainly of the part-time physician's additional malpractice insurance—a separate annual policy required by the hospital with an annual premium for this specialty and region of $15,000 to $30,000 per physician covered.

Three decisions for managing variation. The employer-neurologist, as the practice owner, faces decisions that confront many businesses with variable demand. The first concerns the ideal number of "regular hours" to staff the part-time neurologist at the hospital. That is, how many hours R should the part-time neurologist plan for daily hospital rounds? The second decision is should the neurology practice reject new hospital referrals when the work is excessive. The delicate relationship between specialist and referring hospital physicians makes rejecting referrals unattractive in practice, as it may endanger the valuable relationship, so do models for the control of arrivals to queues add any new managerial insights? Third, in what order should new referrals and follow-up patients be seen during hospital rounds?

For the first decision of the ideal staffing time, the part-time physician would only agree to work with a regular guaranteed salary based on a set schedule at \$100 per hour, and agreed to serve overtime on busy days for \$150 per hour. The proper salary and time-availability for the part-time neurologist must be chosen to minimize the wasted fixed salary expense on low-activity hospital days, while maximizing the time the physician is available without overtime cost on busier days. That is, if the actual hospital workload is less than R then she is paid for R hours anyway. If the workload exceeds R hours, then it costs the practice \$100R for the first R hours, and \$150 for each hour thereafter. This presents a tradeoff: the optimal number of daily regular hours R^* (and corresponding salary) to agree upon at the beginning of the year so that overtime costs on busy days and opportunity costs on light days are properly balanced. We refer to R^* as the *optimal staffing solution.*

Recall that patient-contact time at the hospital reimburses at about \$200 per hour. For example, if $R = 3$ hours per day are planned and the actual hospital workload is 3 hours, then the practice's net revenue is about $3 \times (\$200 - \$100) = \$300$. However if the actual workload is 1 hour, then the net revenue is $1 \times \$200 - 3 \times \$100 = -\$100$. On the other hand if the actual workload is 4 hours, the net revenue is $4 \times \$200 - 3 \times \$100 - 1 \times \$150 = \$350$. In summary, the first decision is the ideal number of daily regular hours R^* on which to base the part-time neurologist's fixed salary.

This first decision concerns the daily hours to "outsource" hospital referrals to a part-time neurologist with different penalties when the actual workload is more or less than the "contracted" hours R. A convenient feature of our solution is that it applies to a single physician without the outsourcing option. Specifically, the foregoing costs may be viewed as opportunity costs for time not spent generating billable hours at one's office practice due to a planned number of hours R at the hospital. The \$100 per regular hour may be generalized as a cost c_r representing the average billable hourly rate that would have been earned with certainty had the physician not planned the hour at the hospital, assuming a backlog of office demand. Similarly, a \$150 overtime hour may be generalized as a cost c_o for the foregone billable hour (c_r) plus any additional costs incurred when the hours at the hospital exceed R and thus disrupt planned office appointments (e.g., such extra costs as delaying, rescheduling, and canceling appointments and related goodwill). We henceforth adhere strictly to the "outsourcing" interpretation of the problem but ask that the reader be mindful of the alternate "single-doctor" generalization; it is revisited in Section 6.

With these assumptions, our work makes the following contributions.

1. A formulation for the number of hours for daily hospital rounds that balances the cost of overtime with the cost of idleness, which we solve with real data.
2. A simple rule that allows a physician to rule out the rejection of referrals; this rule is often optimal and satisfied in practice, and requires only three parameters.
3. Simple heuristics based on the newsvendor model that give optimal or near-optimal staffing, and insights into the optimal staffing solution from the sensitivity analysis.
4. Observations on why the current United States healthcare reimbursement system puts this service planning process out of balance with incentives that are arguably unethical.

Relevant literature. One can view day-to-day fluctuations in the number of new referrals and follow-up visits as a Markov chain because the number of patients the neurologist serves tomorrow depends only on the number seen today. There is an expanding literature of Markov decision models for healthcare capacity reservation policies and medical treatment decision-making (see literature reviews in Fries 1976, Schaefer et al. 2005, Gupta and Denton 2008, Alagoz et al. 2010). Healthcare service delivery examples include the allocation of various patient priority classes to the limited capacity in an intensive care unit (Kao 1974), an operating room (Gerchak et al. 1996, Ayvaz and Huh 2010, Zonderland et al. 2010), nurse staffing at a hospital-based urology practice (Collart and Haurie 1980), a diagnostic service (Green et al. 2006, Patrick et al. 2008), and other hospital admissions (Kolesar 1970, Nunes et al. 2009). The dilemma is how much capacity to give elective/non-urgent patients given uncertain future emergency arrivals. In our case, the overwhelming majority of the patients—both in the office practice and at the hospital—are non-urgent and stable, and therefore segmenting patients into severity classes is not appropriate.

Other applications of the temporal aspect of healthcare capacity allocation include the control of the patient census size (Collart and Haurie 1976, Berger and Haurie 1981), the time to discharge patients (Kreke et al. 2008), and the slots to assign outpatients in an appointment scheduling system (Liu et al. 2010, Lin et al. 2011). We are unaware of prior work that helps independent physicians optimize their time while handling a random stream of hospital referrals.

The paper proceeds as follows. In Section 2 we formulate the problem, find the optimal number of hours R^* the part-time neurologist should plan for the daily hospital rounds, and develop effective heuristics that perform well for different distributions of the number of daily new referrals, assuming that all referrals must be accepted. In Section 3 we relax the accept-all-referrals assumption, and present effective heuristics and certain structural results that follow. We analyze the sensitivity of solutions to the actual parameters in Section 4, explore the optimal order to see patients in Section 5, and conclude in Section 6 with a summary of what happened when we implemented this paper's findings in practice.

2 Decision-making while accepting all referrals

We first concentrate on the problem of the optimal number of hours to assign the part-time neurologist to the hospital assuming a referral may never be rejected. We assume the number of hours per day to book appointments in the office is predetermined and are principally interested in how many additional hours per day to serve the hospital so that creating the access is financially viable for the practice and sustainable for the part-time neurologist. In Section 2.1, focusing on this hospital referral process in isolation, we formulate the problem as a discrete-time Markov decision process (MDP). In Section 2.2, we determine the optimal policy using empirical data. In Section 2.3, we provide some structural results, and in Section 2.4 develop two simple heuristics to determine the optimal R^* for various referral distributions.

2.1 Problem formulation

The number of hospital patients in the neurologist's care are observed the moment the part-time neurologist arrives at the hospital for daily rounds. We assume this number does not change until rounds are finished, at which time there is a transition due to new referrals and discharges over the next day. More specifically, on day $n \in \mathbb{N}^+$, we let X_1^n denote the number of new referrals who must be seen on day n, and X_2^n the number of patients seen on day $n-1$ who require a follow-up on day n with $X_2^n \leq X_1^{n-1} + X_2^{n-1}$. The stochastic process $X^n = (X_1^n, X_2^n)$ models the state of the system at the moment the part-time neurologist begins rounding on day n where the discrete-time Markov chain $\{X^n\}$ denotes the sequence of states visited over days $n = 0, 1, 2, \ldots$

The collection of all feasible states in the stochastic process is $\mathcal{S} \subset \mathbb{N}^2$. Given a system in state $x \in \mathcal{S}$ on day n, the neurologist must decide the number of *regular hours* R_x^n to assign the part-time neurologist to the hospital. Reasonable staffing choices occur in 30-minute increments, so R_x^n can be 0, 0.5, 1, 1.5, ..., 24 hours. We say the action space is $A_x = \{0, 0.5, 1, \dots, 24\}$ for all $x \in \mathcal{S}$ and define a decision rule d^n at day n as the collection of possible actions to be taken in each state, i.e., $d^n = \{R_x^n | x \in \mathcal{S}\}$. A staffing policy π consists of decision rules at each day, namely $\pi = (d^1, d^2, \dots)$.

The optimal number of regular hours R depends only on the state x, not on the day n, because $\mathcal{S}$ is discrete and A_x is finite for all x (by Puterman 1994, Theorem 6.2.10, there exists an optimal deterministic stationary policy). Without loss of generality, we assume that the decisions $R_x^0 = R_x^1 = R_x^2 = \cdots$ yield the same optimal action, and so we may denote the decision in any state x as R_x. A staffing solution R_x that changes dynamically for every state $x \in \mathcal{S}$ is undesirable in practice for two reasons. First, overtime hours cost more than regular hours, so if the action depends on the state, the solution is trivial: provide "regular hours" at the hospital long enough to cover all the patients that day and never pay overtime. Second, the part-time neurologist needs the schedule in advance, her annual salary is based on the daily regular hours, and indeed the agreement between the employer-neurologist and part-time neurologist is a constant number of hours R per day at the hospital, supplementing with overtime as needed. Our solution is to base the staffing decision on the overall pattern of patient referrals and follow-up visits (say, using the 2009-2010 data) so that the staffing decision does not depend on the current number of patients. Mathematically, we write that $R_x = R$ for all x with $R \in A$, where $A \in \{0, 0.5, 1, \dots, 24\}$ denotes the action space for all x. We seek an optimal decision rule $d \in \{R | x \in \mathcal{S}\}$ that results in the staffing policy using the same staffing decision in each state at each day (i.e., $\pi = (d)^\infty$) that maximizes the expected net present value (NPV) of total profit.

We subsequently denote the system state more simply as $x = (x_1, x_2)$ where x_i follows the definition of X_i^n, $i = 1, 2$. We assume (i) that all x_1 new referrals must be seen (no referrals may be rejected), and that the moment the part-time neurologist arrives at the hospital and (x_1, x_2) is revealed, (ii) no additional new referrals will arrive and (iii) none of the x_2 follow-ups will be discharged before being seen. We relax (i) in Section 3 and (iii) in Section 5.

A new referral's initial consultation requires $t_1 = 45$ minutes and reimburses $r_1 = \$150$; a follow-up requires $t_2 = 15$ minutes and reimburses $r_2 = \$50$. In reality, these vary from patient to patient, but fluctuations in individual consultation times and remuneration are trivial compared to the combined daily referral and LOS variability, and so for the purposes of this planning problem we consider r_i and t_i fixed for $i = 1, 2$. Therefore, the hospital reimburses at rate $r_1/t_1 = r_2/t_2 = \$200$ per hour. We let $Z = t_1 x_1 + t_2 x_2$ be the total daily hospital consultation time given (x_1, x_2); Figure 1 in the appendix gives the evolution of Z during the study period.

The cost of the part-time neurologist is $c_r = \$100$ per hour for each of the first R hours, and $c_o = \$150$ for each overtime hour thereafter. Overtime is prorated over fractions of an hour (e.g., if 15 minutes' overtime are worked, then a \$37.50 overtime charge applies), based on the total number of consultations and follow-ups seen for the day, and their respective time allotments. An example of the system dynamics using June 2009 may be useful.

	Wed 6/1	Thur 6/2	Fri 6/3	Sat 6/4	Sun 6/5	Mon 6/6	Tue 6/7	Wed 6/8	Thur 6/9
new referrals (x_1)	2	0	0	5	1	1	0	2	3
follow-ups (x_2)	7	7	5	4	9	8	8	8	9
time required (hours)	3.25	1.75	1.25	4.75	3.00	2.75	2.00	3.50	4.50
reimbursement	\$650	\$350	\$250	\$950	\$600	\$550	\$400	\$700	\$90

On Wednesday, 1 June 2009, there were $x_1 = 2$ new referrals and $x_2 = 7$ follow-ups. Seeing these 9 patients took (2)(0.75 hours) + (7)(0.25 hours) = 3.25 hours and reimbursed (2)(\$150) + (7)(\$50) = \$650. Of these 9 patients, 2 were discharged so there were $x_2 = 7$ follow-ups on 2 June 2009. There were no referrals on June 2 and so the system transitioned from (2,7) to (0,7) on June 1 to June 2.

With epochs of length one day, the discount factor is $\beta = 1/(i + i)$ with daily interest rate i (see, for example, Heyman and Sobel 1984, p. 25). Through other alternatives the practice can earn 5% APR with daily compounding (i.e., 5.13% APY), so $i = 0.05/365$ and thus $\beta = 7300/7301 \approx 0.99986$. We show in Section 4.1 that when all referrals are accepted the analysis is insensitive to the interest rate assumption.

If the system is in state (x_1, x_2) today, we now describe the probability of transitioning to state (y_1, y_2) tomorrow, $y_1 \in \{0, \ldots, N\}$, $y_2 \in \{0, \ldots, x_1 + x_2\}$. Let p_i be the probability of i new referrals with distribution observed empirically in 2009-2010 as follows:

i	0	1	2	3	4	5	6	7
p_i	0.208	0.248	0.218	0.167	0.094	0.030	0.020	0.015

Also $\alpha = 0.82$ is the empirically-obtained probability that a patient seen today needs a follow-up tomorrow. This implies if $x_1 + x_2$ patients are seen today, then y_2 follow-ups will be required tomorrow with probability

$$P_{x_1+x_2}^{y_2} = \binom{x_1 + x_2}{y_2} \alpha^{y_2} (1 - \alpha)^{x_1+x_2-y_2} \tag{1}$$

The referral probabilities p_i and the discharge proportion $1 - \alpha = 0.18$ make transitions to large follow-up numbers y_2 unlikely. For instance, $x_1 + x_2$ exceeded 20 only twice in 2009-2010 (see Figure 1), though (1) implies that the state space is theoretically unbounded.

We let $v_k(\cdot)$ be the expected NPV of discounted profits over the next $k \geq 0$ days, where

$$v_{k+1}(x_1, x_2) = r_1 x_1 + r_2 x_2 - c_r R - c_o(t_1 x_1 + t_2 x_2 - R)^+ + \beta \sum_{i=0}^{N} p_i \sum_{j=0}^{x_1+x_2} P^j_{x_1+x_2} v_k(i, j) \qquad (2)$$

with salvage value $v_0(x_1, x_2) \equiv 0$ for all $x \in \mathcal{S}$. Given R, standard MDP theory allows a solution to (2) via successive approximations (Bellman 1957). The neurologist reevaluates the optimal policy (i.e., the optimal R^*) every year, primarily to adjust for any shifts in hospital referral rates. This suggests that decisions should be based on the expected NPV of one year's profits. As we show in Section 4, the optimal policy holds over any horizon of $k = 91$ days or more, and we use a 365-day horizon for the following analysis.

2.2 Optimal solution at subject hospital

In this section, we determine the optimal number of regular hours from the empirical data collected while serving the subject hospital. In practice, the neurologists worry about the possibility of an overwhelming hospital workload and employ the following *skipping policy* that limits the state space: if the system transitions to (y_1, y_2) with $y_1 + y_2 > 20$, then the part-time neurologist sees all y_1 new referrals, $20 - y_1$ follow-ups, and skips the remaining $y_2 - (20 - y_1)$ follow-ups of lowest medical priority. Skipped patients will be seen on subsequent days if the total number in the part-time neurologist's care drops.

Formulating the skipping policy with Markovian transitions requires a convoluted state definition, and we therefore employ a simplification. When a transition occurs to $y_2 > 20$, we *artificially discharge* $y_2 - 20$ patients so that $y_2 = 20$. In other words, any transition to $y_2 > 20$ is approximated as a transition to $y_2 = 20$. This results in the finite state space $\mathcal{S} = \{(x_1, x_2) : 0 \leq x_1 \leq 7, 0 \leq x_2 \leq 20\}$ where the probability of transitioning from $x_1 + x_2$ patients today to y_2 follow-ups tomorrow is

$$P^{y_2}_{x_1+x_2} = \begin{cases} \binom{x_1+x_2}{y_2} \alpha^{y_2} (1-\alpha)^{x_1+x_2-y_2} & \text{if } y_2 \in \{0, \ldots, 19\} \\ \sum_{i=20}^{x_1+x_2} \binom{x_1+x_2}{i} \alpha^i (1-\alpha)^{x_1+x_2-i} & \text{if } y_2 = 20 \leq x_1 + x_2 \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

with $x_1 + x_2 \in \{0, \ldots, 27\}$, $y_2 \in \{0, \ldots, \min(x_1 + x_2, 20)\}$, and the probability of 20 follow-ups $P^{20}_{x_1+x_2}$ is in actuality the probability of 20 *or more* follow-ups.

Our skipping policy approximation results in a maximum daily workload of 27 patients which is 7 more than the 20 patients allowed by the actual policy. We chose 27 because had we set the limit at 20, then the artificial discharges could understate the true system workload in subsequent days. That said, the system transitions to these highly congested states so infrequently that the artificial discharge threshold (20 or 27 or any other number) exerts negligible influence on the optimal policy.

Taken together, the referral probabilities p and follow-up probabilities P give a 168×168 transition matrix where each entry denotes the probability of a transition from state (x_1, x_2) to (y_1, y_2) as $p_{y_1} P^{y_2}_{x_1+x_2}$. Using successive approximations to obtain $v_{365}(x_1, x_2)$ for all $x \in \mathcal{S}$ and for all $R \in \{0, 0.5, 1, \ldots, 23.5, 24\}$, we obtain the following results.

1. The optimal number of regular hours to staff the part-time neurologist at the hospital is 3 per day. That is, $R^* = 3$ hours maximizes $v_{365}(x_1, x_2)$ for all $x \in \mathcal{S}$. For example, consider the effect of R on the profits $v_{365}(0, 0)$:

Number of Regular Hours, R	Expected NPV of Profits, $v_{365}(0,0)$	
2.0	\$96,409	
2.5	\$100,561	
3.0	\$101,455	$\leftarrow R^* = 3$
3.5	\$98,699	
4.0	\$92,337	

Total profits are nearly maximized for values around R^* (e.g., $R^* \pm 1$ hour attenuates NPV of annual profits not more than 10%).

2. The maximum expected NPV of profits $v_{365}(x_1, x_2)$ for all $x \in \mathcal{S}$ due to $R^* = 3$ is shown in the appendix, Table 1, where it is seen that $v_{365}(x_1, x_2)$ increases monotonically in x_1 and x_2. This is true in general when hospital reimbursement rates (e.g., \$200/hr) exceed overtime rates (e.g., $c_o =$ \$150/hr).

2.3 Structural results

Next, we provide a structural result that simplifies the search algorithm for optimal number of regular hours R. Let $E_\pi(\cdot)$ denote the total expected discounted profit for a policy π, and let (X_1^n, X_2^n) be the random variable denoting the state observed in period n. The optimization problem is to maximize $E_\pi \left[\sum_{n=0}^{\infty} \beta^n u_{n,\pi} (X_1^n, X_2^n)\right]$, where $u_{n,\pi}(X_1^n, X_2^n)$ is the total profit obtained in period n when policy π is employed in state (X_1^n, X_2^n).

Proposition 1. *The total discounted profit is concave as a function of the number of regular hours, R.*

The proof is available in the appendix. Proposition 1 also holds for a finite decision horizon, and implies that the local optimal solution is a global one. Although we consider discrete R, Proposition 1 still applies. More specifically, to determine the optimal regular hours R^*, we can first calculate $v_n(x_1, x_2)$ for $R = 0$, and repeatedly increase R until $v_n(x_1, x_2)$ starts decreasing. At this point, we can stop the algorithm and conclude that optimal R^* is the one that maximizes $v_n(x_1, x_2)$ in the set of R values considered so far.

2.4 Heuristics

In this section are two simple heuristics that depend only on the mean and the standard deviation of the number of daily new referrals using ideas from corresponding single-period inventory ("newsvendor") problems (see, e.g., Arrow et al. 1951) where the total daily time Z required by the patients constitute the demand and the supply is R. Note, however, (3) implies that our problem is more difficult because demand in each period depends on the demand in the previous period, and different patients have different time demands. Hence, the optimal policy of the newsvendor problem is not optimal but provides guidelines for effective heuristics.

These heuristics estimate R^* for a single-period problem using approximations for the long-run probability distribution of the daily number of patients in the system, (X_1, X_2). Henceforth we omit the index n when referring to system states, and denote the expected daily profit $E\left[u\left(X_1, X_2\right)\right]$, where $u(\cdot)$ is the total daily profit given by (5) in the appendix. By Proposition 1, this expectation is concave in R, hence its partial derivative with respect to R (abbreviated $E'\left[u\left(X_1, X_2\right)\right]$) is zero at the optimal solution. Let $\mathbb{I}$ denote the indicator function. Then $E'\left[u\left(X_1, X_2\right)\right] = -c_r - E'\left[c_o(t_1X_1 + t_2X_2 - R)^+\right] = -c_r + c_oE\left[\mathbb{I}(t_1X_1 + t_2X_2 - R > 0)\right] = -c_r + c_o \Pr(t_1X_1 + t_2X_2 > R)$. Setting $E'[\cdot]$ to zero yields $\Pr(t_1X_1 + t_2X_2 > R) = c_r/c_o$ which implies that the hours R that maximize expected daily net profit are set so the probability of daily overtime equals the ratio of regular time cost to overtime cost. Hence, we conclude that the R^* that maximizes $E\left[u\left(X_1, X_2\right)\right]$ satisfies $F_Z(R) = 1 - \frac{c_r}{c_o}$, where $F_Z(\cdot)$ is the cumulative distribution function (c.d.f.) of the random variable $Z = t_1X_1 + t_2X_2$.

Let $\mu_{X_1} = \lambda$ denote the mean daily new patient referral rate. The long-run mean daily number of follow-up patients is therefore $\mu_{X_2} = \lambda\alpha + \lambda\alpha^2 + \lambda\alpha^3 + \ldots = \lambda\alpha/(1-\alpha)$.

Let random variable Y denote the total number of patients in the system (i.e., $Y = X_1 + X_2$) with c.d.f. $F_Y(\cdot)$ and expected value $\mu_Y = \mu_{X_1} + \mu_{X_2} = \lambda/(1-\alpha)$. Assuming that among X_1 patients that arrived $n \geq 0$ days ago, $X_1\alpha^n$ are still in the system, the standard deviation of Y can be approximated $\theta_Y = \sigma_{X_1}/(1-\alpha^4)^{0.5}$; this follows from summing over all $n \geq 0$

and assuming independence among random variables. Let t_Y be the average treatment time required each day per patient, where $t_Y = \left(t_1\lambda + t_2\frac{\lambda\alpha}{1-\alpha}\right) \Big/ \left(\frac{\lambda}{1-\alpha}\right) = (1-\alpha)t_1 + \alpha t_2$. Two simple heuristics to determine the optimal R^* follow.

- **Heuristic** $R1$**:** Approximate the distribution of Z by a normal distribution with mean $t_Y\mu_Y$ and standard deviation $t_Y\theta_Y$. Let R^* be $F_Z^{-1}(1-\frac{c_r}{c_o})$ rounded to the closest number in set of allowable actions A.
- **Heuristic** $R2$**:** Approximate the distribution of Z by the distribution of $t_Y Y$, where Y follows a Poisson distribution with mean rate μ_Y. Let R^* be $t_Y D_Y$ rounded to the closest number in set of allowable actions A, where D_Y minimizes $|F_Y(D_Y) - (1-\frac{c_r}{c_o})|$.

The optimal R^* satisfies $F_Z(R^*) = F_Y(\frac{R^*}{t_Y}) = 1 - \frac{c_r}{c_o}$. In Heuristic $R1$, we approximate R^* by rounding $F_Z^{-1}(1-\frac{c_r}{c_o})$ to the closest number in A because Z follows a continuous distribution. The observed empirical distribution of Z shown in Figure 1 and a Chi-squared test suggest a normal distribution is appropriate. However, in Heuristic $R2$, we first approximate $\frac{R^*}{t_Y}$ by D_Y because Poisson is a discrete distribution; then we round $t_Y D_Y$ to the closest number in A to approximate R^*. Heuristics $R1$ and $R2$ are simple to apply because they only require the mean and standard deviation of the distribution of daily new referrals.

Using the parameters for our problem, we can obtain $\mu_{X_1} = 1.936$, $\mu_{X_2} = 8.820$, $\mu_Y = 10.756$, $t_Y = 0.34$ hours, and $\theta_Y = 3.491$ hours. Both Heuristics $R1$ and $R2$ yield the result given in Section 2.2 ($R^* = 3$).

Now consider different distributions of the number of daily referrals to test the robustness of $R1$ and $R2$, specifically, when the number of new referrals follows a discrete uniform distribution with range $\{0, \ldots, k^U\}$ with $k^U \in \{1, \ldots, 10\}$; a Poisson distribution with rate $\lambda^P \in \{0.5, 1, \ldots, 5\}$; and a geometric distribution with success probability $p^G \in \{0.15, 0.2, \ldots, 0.5\}$. Using the parameter values given in Section 2.1, the following table provides the heuristic and optimal policies for each referral distribution, as well as the performance of the heuristic policies as the percentage of the profit $v_{365}(0,0)$ of the optimal policy. Moreover, in the following numerical results, we considered a maximum total workload of 35 patients (we used a larger state space compared to Section 2.2 due to the higher average referral rate in some cases).

	Discrete Uniform		
k^U	Optimal R	Heuristic $R1$	Heuristic $R2$
1	0.5	0.5 (100%)	0 (72.74%)
2	1.5	1.5 (100%)	1.5 (100%)
3	2.5	2.0 (99.65%)	2.0 (100%)
4	3.0	3.0 (100%)	3.0 (100%)
5	4.0	3.5 (99.39%)	4.0 (100%)
6	4.5	4.5 (100%)	5.0 (99.87%)
7	5.5	5.5 (100%)	6.0 (99.43%)
8	6.0	6.0 (100%)	7.0 (99.03%)
9	7.0	7.0 (100%)	7.5 (99.41%)
10	7.5	7.5 (100%)	8.5 (98.53%)

	Poisson		
λ^P	Optimal R	Heuristic $R1$	Heuristic $R2$
0.5	0.5	0.5 (100%)	0.0 (76.39%)
1.0	1.5	1.5 (100%)	1.5 (100%)
1.5	2.0	2.0 (100%)	2.0 (100%)
2.0	3.0	3.0 (100%)	3.0 (100%)
2.5	4.0	4.0 (100%)	4.0 (100%)
3.0	5.0	4.5 (99.67%)	5.0 (100%)
3.5	5.5	5.5 (100%)	6.0 (99.89%)
4.0	6.5	6.5 (100%)	7.0 (99.55%)
4.5	7.5	7.5 (100%)	7.5 (100%)
5.0	8.0	8.0 (100%)	8.5 (99.68%)

	Geometric		
p^G	Optimal R	Heuristic $R1$	Heuristic $R2$
0.20	4.5	5 (98.95%)	7.0 (83.10%)
0.25	3.5	3.5 (100%)	5.0 (94.41%)
0.30	3.0	3.0 (100%)	3.5 (98.68%)
0.35	2.5	2.0 (99.00%)	2.5 (100%)
0.40	2.0	1.5 (98.41%)	2.0 (100%)
0.45	1.5	1.5 (100%)	1.5 (100%)
0.50	1.0	1.0 (100%)	1.5 (98.73%)

Both heuristics are effective for all three referral distributions. Heuristic $R1$ attains the optimal profit more than 50% of the time with 98.41% of the optimal profit in the worst case. Heuristic $R1$ is also robust to the shape of the distribution, although it seems to perform better in more symmetric distributions, discrete uniform and Poisson. Heuristic $R2$ also performs well, although its performance slightly deteriorates for small and high values of the mean referral rate (mostly due to the discrete nature of the Poisson distribution). However, $R2$ also attains the optimal profit approximately 50% of the time, and performs well when the referral distribution is Poisson.

3 Optimal policies when rejecting referrals

In addition to the number of regular hours R to staff the part-time neurologist at the hospital, we add a second decision—the number of referrals to accept and reject, assuming such actions do not jeopardize relationships with referring physicians, namely, that the distribution of referrals is independent of the rejection decision. We call this the *two-decision problem.* In Section 3.1, we determine the optimal acceptance and rejection policy for a given R using the parameters and the referral distribution for the subject hospital. Next, in Section 3.2, we provide structural results and heuristics that can be implemented in systems with different parameters and referral distributions.

3.1 Optimal solution at subject hospital

Let a be the number of new referrals the neurologist is willing to accept given that there are x_2 follow up patients. We extend optimality equation (2) to the two-decision problem as

$$v_{k+1}(x_1, x_2) = r_2 x_2 - c_r R + \max_{a=0,\ldots,x_1} \Big\{ r_1 a - c_o(t_1 a + t_2 x_2 - R)^+ + \beta \sum_{i=0}^{N} p_i \sum_{j=0}^{a+x_2} P^j_{a+x_2} v_k(i,j) \Big\} \tag{4}$$

We first set R, and then find the acceptance/rejection policy that maximizes (4) for all $x \in \mathcal{S}$ over a 365-day horizon (as in Section 2.2, we employ artificial discharges when the number of patients in the system exceed 20). Then we increment R; after repeating over $R \in \{0, 0.5, 1, \ldots, 24\}$, we determine which value R^* maximizes these value functions.

For the parameters defined in Section 2.1, $R^* = 3$ hours remains optimal and rejection is not optimal in any state. In other words, the assumption that rejections do not influence future referral rates does not matter because rejection is not economically attractive. Moreover, the values $v_{365}(x_1, x_2)$ corresponding to $R^* = 3$ are identical to those shown in Table 1. The primary reason that rejection is not optimal is that the cost of overtime is $c_o = \$150$ per hour, but an overtime hour reimburses at \$200 hour, and so it remains profitable to accept patients and incur overtime. Sensitivity analysis in Section 4.2 reveals that rejection becomes economically attractive when the reimbursement is less than the expected labor cost, which is not unlikely in hospital systems that serve more uninsured/underinsured populations.

3.2 Structural results and heuristics

In this section, we first formalize the above discussion with a proposition that partially characterizes the optimal referral acceptance policy. More specifically, the policy that accepts all referrals is optimal when the problem parameters satisfy a simple condition.

Proposition 2. *Given $R \geq 0$, it is optimal to accept all patient referrals if $r_i \geq c_o t_i$ for $i \in \{1, 2\}$.*

The proof is in the appendix. The parameters in our problem satisfy the condition in Proposition 2 and we conclude that accepting all patients is optimal. Proposition 2 also holds regardless of whether we assume an infinitely-large state space or employ the skipping policy. Proposition 2 is a sufficient condition for optimality, and as we demonstrate numerically in Section 4.2, acceptance of all referrals remains optimal for values that fall slightly outside the range $r_i \geq c_o t_i$, $i \in \{1, 2\}$. Moreover, for a given acceptance policy, Proposition 1 still holds because the transition probabilities are not affected by R. Hence, we conclude that if the two-decision problem starts with a given acceptance control policy and tries to determine the optimal R^*, we can use the simple search algorithm described in Section 2.3.

When accepting all patients is not optimal, optimal acceptance control policy for a given R can be found more easily with the following static heuristics. Recall that t_Y is an arbitrary patient's daily expected treatment time. Similarly, we let r_Y be the long-run average reimbursement per patient per day. Hence, we let $r_Y = \left(r_1\lambda + r_2\frac{\lambda\alpha}{1-\alpha}\right) \Big/ \left(\frac{\lambda}{1-\alpha}\right) = (1-\alpha)r_1 + \alpha r_2$ and define Y, $F_Y(\cdot)$, t_Y as in Section 2.4. In the remainder of this section, we concentrate on the $r_Y < c_0 t_Y$ case. For the optimal rejection policy, the following heuristics prevent the number of patients from exceeding a certain threshold.

- **Heuristic** $T1$**:** Let M_1 the expected number of patients seen during regular hours, namely, $\frac{R}{t_Y}$ rounded to integer above. Accept a new patients, where $a = max\{0, M_1 - x_2\}$.
- **Heuristic** $T2$**:** Approximate the distribution of Y with a normal distribution of mean μ_Y and standard deviation θ_Y. Let $\kappa = 1 - [r_y - c_o t_Y(1 - F_Y(R/t_Y))]/(c_o t_Y - r_Y)$. If $0 \leq \kappa \leq 1$, let M_2 be $F_Y^{-1}(\kappa)$ rounded to the integer above, otherwise let $M_2 = M_1$. Only accept a new patients, where $a = max\{0, M_2 - x_2\}$.

Heuristic $T1$ myopically limits overtime cost and does not account for patients whose entire hospital stay may be profitable considering follow-ups. Heuristic $T2$ is based on the value of the threshold M_2 such that a marginal loss $c_o t_Y - r_Y$ is incurred with probability $\Pr(Y > M_2)$ and the expected profit for the same patient in the next period is approximated $r_Y - c_o t_Y(1 - F_Y(R/t_Y))$ because the revenue r_Y is obtained for all the patients, and the cost $c_o t_Y$ is incurred with probability $1 - F_Y(R/t_Y)$. These heuristics do not consider the cost term $c_r R$ because the cost for regular hours is incurred regardless of the number of patients. Moreover, $T1$ and $T2$ do not account for the dependency between two periods (in reality, a higher number of patients in a period is more likely to generate overtime cost in the next).

We tested the heuristics with the parameters in Section 2.1 and referral distributions in Section 2.4, except $c_0 = \$250$ per hour to assure that some rejection is optimal. The following table provides the heuristic and optimal policies as well as the performance of the heuristic policy as the percentage of the profit $v_{365}(0, 0)$ of the optimal threshold policy. We use the optimal R^* obtained for each case in Section 2.4 as the number of regular hours. Below, "optimal" denotes the optimal number of referrals to accept.

		Discrete Uniform				*Poisson*	
k^U	Optimal	Heuristic $T1$	Heuristic $T2$	λ^P	Optimal	Heuristic $T1$	Heuristic $T2$
1	2	2 (100.00%)	4 (84.47%)	0.5	3	2 (99.80%)	3 (100 %)
2	6	5 (95.72%)	5 (95.72%)	1.0	7	3 (72.45%)	4 (85.13%)
3	11	8 (96.92%)	8 (96.92%)	1.5	8	5 (91.94%)	9 (99.20%)
4	12	9 (96.71%)	11 (99.38%)	2.0	12	6 (83.11%)	11 (99.79%)
5	16	12 (96.64%)	12 (96.64%)	2.5	15	8 (82.70%)	12 (98.14%)
6	18	13 (96.35%)	16 (99.61%)	3.0	18	9 (77.13%)	14 (97.52%)
7	22	17 (97.94%)	17 (97.94%)	3.5	19	11 (86.10%)	19 (100.00%)
8	24	18 (97.86%)	21 (99.80%)	4.0	22	12 (81.59%)	21 (99.95%)
9	28	21 (97.71%)	23 (99.15%)	4.5	25	14 (81.95%)	23 (99.71%)
10	28	23 (98.36%)	27 (99.97%)	5.0	25	15 (84.45%)	28 (99.81%)
		Geometric					
p^G	Optimal	Heuristic $T1$	Heuristic $T2$				
0.20	18	12 (87.35%)	22 (98.90%)				
0.25	15	9 (82.30%)	16 (99.73%)				
0.30	14	7 (72.96%)	11 (96.72%)				
0.35	9	6 (86.28%)	10 (99.04%)				
0.40	8	5 (72.79%)	6 (84.73%)				
0.45	7	4 (71.54%)	5 (86.39%)				
0.50	5	3 (71.07%)	5 (100.00%)				

Both heuristics are effective for a discrete uniform referral distribution. The performance of $T1$ deteriorates for the more realistic Poisson and geometric distributions, however $T2$ performs relatively well (even reaching the optimum in some cases). Although its performance is slightly worse compared to $T2$, Heuristic $T1$ is desirable for its simplicity. Finally, we also approximated the distribution of Y using the Poisson (as in Heuristic $R2$ of Section 2.4) but this performed worse and is more difficult to apply than Heuristic $T2$, hence omitted here.

4 Sensitivity analysis

We now describe how the optimal policies described in Sections 2 and 3 change if we vary one parameter at a time, all else the same. When appropriate in Section 4.2 we compare ranges with those given in Proposition 2.

Sensitivity to the decision horizon, k. The foregoing analysis maximized profits over a 365-day horizon, and this solution is optimal for any horizon $k \geq 91$. The reason can be seen in the two-decision problem (Section 3). Recall that $v_k(x)$ is the expected NPV of profits over the next k days starting in state $x \in \mathcal{S}$. By Proposition 2, rejection is never optimal over any horizon $k > 0$ when the per-patient reimbursement exceeds the labor investment (i.e., $r_1 \geq c_o t_1$ and $r_2 \geq c_o t_2$). Second, R^* is sensitive to the number of periods remaining in the process only if $k \leq 90$ as follows.

R	No. of states x for which R maximizes $v_k(x)$ $k=1$	$k=5$	$k=25$	$k=50$	$k=75$	$k=100$
0.0	2	0	0	0	0	0
0.5	3	1	0	0	0	0
1.0	4	5	0	0	0	0
1.5	6	7	0	0	0	0
2.0	7	10	0	0	0	0
2.5	8	18	34	17	6	0
3.0	138	127	134	151	162	168
sum	168	168	168	168	168	168

Over very short time horizons there are uncongested states where $R^* < 3$ hours because one can minimize regular labor costs today and not pay extra future overtime charges due to the salvage values $v_0(\cdot)$. In other words, as the system approaches shut-down, it is optimal to scale back the regular hours only if the system load is light. Given that 91 days is such a short horizon we may reasonably assume that the solution is insensitive to the horizon k over which the policy must hold in reality and thus all subsequent analysis is over $k = 365$ days.

4.1 One-decision problem

Consider the sensitivity of R^* to other parameters under the assumption that all referrals must be accepted. Recall that feasible policies are the half-hour increments $R \in \{0, 0.5, 1, \ldots, 24\}$. In the following table, we obtained the range on each parameter—all else the same—for which $R^* = 3$ maximizes $v_{365}(x)$ for all 168 states of $x \in \mathcal{S}$. Parameters outside the following lower and upper bounds result in $R^* \neq 3$ in one or more states. The first six bounds are rounded to integers.

parameter	units	original value	range with $R^* = 3$ for all $x \in \mathcal{S}$: lower bound	upper bound
c_r	\$/hr	100	86	105
c_o	\$/hr	150	143	171
r_1	\$/referral	150	0	∞
r_2	\$/follow-up	50	0	∞
t_1	minutes	45	39	61
t_2	minutes	15	15	17
α		0.82	0.811	0.838
β		0.99986	0.98843	1

This has the following interpretations. If the regular hourly cost c_r decreases, it becomes less expensive to pay regular hours relative to overtime hours, and thus R^* may increase. For example, if $c_r = \$85$ per hour, then in certain states $v_{365}(\cdot)$ is maximized with $R^* = 3.5$ hours (with $R^* = 3$ in the majority of the states). Similarly, if $c_r = \$106$ then $R^* = 2.5$ hours in a few states. The cost of overtime c_o has a similar effect; if c_o increases, then again regular hours become relatively less expensive and there is an incentive to increase R^* (if $c_o > \$171$ per hour, then $R^* > 3$ in one or more states; if $c_o < \$143$, then $R^* < 3$ in some states).

The reimbursements r_1 and r_2 do not influence R^* because when all referrals must be accepted the expected reimbursement is independent of the allocation of regular hours. This assumes the probability distribution of referrals is stationary. See related comments in Section 6.

The consultation times t_1 and t_2, however, influence R^* because if either time decreases, less time is needed at the hospital and so R^* may decrease as well. For example, if $t_1 < 39$ minutes, then in certain states $R^* < 3$. The follow-up proportion α has a similar effect—as α decreases, the number of follow-ups decreases and therefore R^* ultimately decreases. The narrow range $0.811 \leq \alpha \leq 0.838$ over which $R^* = 3$ indicates that the solution is sensitive to this parameter. See further comments in Section 6.

Finally, the solution is insensitive to the discount factor β since $R^* = 3$ holds for $0.9884 \leq \beta \leq 1$, namely, when attractive alternatives pay 0% to 6900% APR. While impossible in practice, theoretically if $\beta < 0.9884$ (APR > 6900%), then the expected NPV of profits $v_{365}(0,0)$ (and other low-congestion states) is maximized with $R^* < 3$ because it is optimal to shrink the immediate regular-hours cost in state (0, 0) when future overtime charges are so highly discounted.

4.2 Two-decision problem

We inspect how robustly the jointly-optimal solution holds when individual parameters change, and discovered that $R^* = 3$ hours and "never reject" remains optimal over exactly the range of parameters c_r, c_o, t_1, t_2, α, β presented in Section 4.1. This solution also holds for $r_1 \geq \$102$ and $r_2 \geq \$25$. However, if either r_1 or r_2 is less then the optimal policy shifts to $R^* < 3$ with rejection of referrals in congested states. Let us demonstrate.

Sensitivity to r_1 and r_2. The following table shows how changes in r_1 and r_2 influence R^* and the optimal amount of rejection. Columns 3 and 7 indicate the number of the 168 states x for which R maximizes the expected NPV of discounted profits $v_{365}(x)$. Columns 4 and 8 refer to the number of states x where $v_{365}(x)$ is maximized by rejecting one or more referrals.

Sensitivity to r_1 *with* $r_2 = \$50$.				*Sensitivity to* r_2 *with* $r_1 = \$150$.			
r_1	R	no. states where profits maximized	no. states where rejection optimal	r_2	R	no. states where profits maximized	no. states where rejection optimal
\$0	1.0	0	142	\$0.0	0.0	168	147
	1.5	8	137		0.5	0	145
	2.0	160	124		1.0	0	143
	2.5	0	104		1.5	0	140
	3.0	0	77		2.0	0	133
50	2.0	0	49	12.5	0.0	168	147
	2.5	0	21		0.5	0	145
	3.0	168	15		1.0	0	142
	3.5	0	10		1.5	0	138
100	2.0	0	1	25.0	2.0	0	86
	2.5	0	1		2.5	0	55
	3.0	168	1		3.0	168	0
	3.5	0	0		3.5	0	0
150	2.0	0	0	37.5	2.0	0	0
	2.5	0	0		2.5	0	0
	3.0	168	0		3.0	168	0
	3.5	0	0		3.5	0	0

The policy $R^* = 3$ hours and "never reject" is optimal over such a large range of profits r_1, r_2 that the following discussion is mostly theoretical yet provides useful intuition for the problem. Specifically, if $r_1 \leq \$101$ (or $r_2 \leq \$24$), then expected profits $v_{365}(\cdot)$ are maximized if R^* is coupled with referral rejection in one or more states, starting with the most congested state $x = (7, 20)$ when $r_1 = \$101$ (or $r_2 = \$24$). Compare this result to those given by Proposition 2 which guarantees the optimality of accepting all referrals when $r_1 \geq \$112.50$ and $r_2 \geq \$37.50$ (all else the same).

There are additional considerations about rejections hidden in above columns 4 and 8. For example, the table indicates when $r_1 = \$50$ it is optimal to set $R^* = 3$ and reject patients in the 15 most congested states. These are the states $x \in \{(x_1, x_2) : x_1 + x_2 \geq 23\}$ where the actual number of patients to accept into these states are prescribed in the following table (see the 15 "boxed" numbers):

Optimal no. of acceptances, $r_1 = \$50,\ R = 3$

follow-ups, x_2	new referrals, x_1							
	0	1	2	3	4	5	6	7
0	0	1	2	3	4	5	6	7
1	0	1	2	3	4	5	6	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
15	0	1	2	3	4	5	6	7
16	0	1	2	3	4	5	6	6
17	0	1	2	3	4	5	5	5
18	0	1	2	3	4	4	4	4
19	0	1	2	3	3	3	3	3
20	0	1	2	2	2	2	2	2

When r_1 (or r_2) is sufficiently small, the optimal policy shifts to $R^* < 3$ hours per day (e.g., it is shown above that $r_1 = \$0$ implies $R^* = 2$ hours). Why does R^* decrease? So many referrals are rejected that there is simply less work to be done, and so the optimal staffing plan calls for fewer daily regular hours.

Why does rejection become more attractive with lower r_1 (or r_2)? Equation (4) implies that if there are x_1 referrals today, then we should accept as many as possible such that the sum of (a) the immediate profits of initial consultations plus (b) the expected NPV of future profits due to follow-ups is positive. The more that are accepted (and the more congested the state), the greater the chance that (a) or (b) includes expensive overtime charges. When r_1 (or r_2) is sufficiently small, the net profit of accepting a patient is negative, and the optimal action is the reject referrals until the sum of (a) and (b) is nonnegative.

Large changes in any parameter that undermines the above profit types (a) and (b), in theory, makes some rejection economically attractive. These include the time to serve a patient (t_1, t_2), the cost of labor (c_r, c_o), the discharge rate ($1 - \alpha$), and the referral probabilities p. Even the discount factor β, theoretically, contributes to rejection because more/less discounting skews future profits relative to the immediate profits. However, this is not possible using the current problem parameters. Let us address each parameter now.

Sensitivity to c_r and c_o. If the cost of labor becomes more expensive, it is possible that the sum of (a) and (b) becomes negative and thus rejection is financially attractive. For example, consider the following range of c_r and c_o, all else the same.

Sensitivity to c_r with $c_o = \$150$.			*Sensitivity to c_o with $c_r = \$100$.*		
c_r	R^*	no. states where rejection optimal	c_o	R^*	no. states where rejection optimal
\$25	5.0	0	\$100	0.0	0
50	4.0	0	125	2.5	0
75	3.5	0	150	3.0	0
100	3.0	0	200	3.5	0
125	2.5	0	250	4.0	28
150	0.0	0	300	4.0	68

We conclude that if c_r increases, then it is optimal to use fewer regular hours relative to overtime hours. Similarly, if the cost of overtime hours c_o increases then more regular hours should be used. Rejection is only desirable when the overtime cost c_o exceeds the \$200 per hour hospital reimbursement rate, which is consistent with Proposition 2.

Sensitivity to t_1 and t_2.

Sensitivity to t_1 with $t_2 = 15$ minutes			*Sensitivity to t_2 with $t_1 = 45$ minutes*		
t_1 (minutes)	R^* (hours)	no. states where rejection optimal	t_2 (minutes)	R^* (hours)	no. states where rejection optimal
30	2.5	0	10	2.5	0
45	3.0	0	15	3.0	0
60	3.0	0	20	3.5	0
90	3.5	21	30	4.0/4.5	98/94†
120	2.5	136	60	2.0	144

† (A comment about $t_1 = 45$, $t_2 = 30$, above: The optimal policy is $R^* = 4$ hours in the 96 states with $x_2 \leq 11$, and $R^* = 4.5$ hours in the 72 states with $x_2 \geq 12$. The table depicts that if the policy in 168 states is $R = 4$, then it is optimal to reject in 98 states. Alternatively, if $R = 4.5$ is used in all 168 states, then we reject in 94 states.)

If either consultation time t_1 or t_2 decreases sufficiently from its original value, then daily rounds require less time and $R^* < 3$ is optimal. If either increases above its original value, at first, the daily rounds take longer and so $R^* > 3$ is optimal, but reimbursement exceeds the expected cost of a consultation and so "no rejection" remains optimal. Indeed, by Proposition 2, no rejection is optimal when $t_1 \leq 60$ minutes (or $t_2 \leq 20$). However if either time increases beyond these limits, the labor cost of a consultation exceeds the reimbursement, and thus rejections become attractive. As rejection becomes more attractive, the number of hours R^* may also decrease due to the reduced expected workload.

Sensitivity to α. If α increases (decreases) sufficiently—corresponding to an increase (decrease) in average LOS—then the expected sum of regular labor hours plus overtime hours can be optimized with $R^* > 3$ ($R^* < 3$). Rejection is never optimal for all $0 \leq \alpha < 1$ because as $\alpha \to 1$ the average LOS $\to \infty$, but the workload will be limited by the skipping policy, and the cost of labor per unit time will not exceed the reimbursements. Theoretically, when α is at impossibly high levels (such as $\alpha = 0.95, 0.99$, below), the skipping policy is invoked in every period.

Rejection is never optimal over α by Proposition 2. To summarize for some sample values, the optimal policy as a function of α follows.

α	R^*	no. states where rejection optimal
0.99	5.5	0
0.95	5.5	0
0.88	4.0	0
0.82	3.0	0
0.75	2.0	0
0.50	1.0	0

Sensitivity to $p_1, p_2, \ldots, p_N$. While formal analysis is omitted from this discussion, the distribution of daily referral probabilities influences the optimal policy much like the consultation time t_1. If such probabilities p shift toward fewer (more) daily referrals, then $R^* < 3$ ($R^* > 3$) optimally balances the overtime costs.

5 Optimal consultation order

We assumed that no patient is discharged while the part-time neurologist makes her rounds. In reality, a new or follow-up patient may be discharged after the part-time neurologist starts rounding but before being seen on a given day. The order in which new referrals and follow-ups should be seen can be considered a scheduling problem in a queueing system with impatient customers, where abandonment rates differ by the consultation type. We assume there are no medical priorities, namely, that all patients to be seen are in stable condition.

Since the number of referrals and follow-up patients are known at the beginning of the rounds this process may be considered a "clearing" system that has no outside referrals during the rounds and all (x_1, x_2) patients that are in the hospital at the beginning of the rounds must be served, assuming it is optimal to accept all x_1 referrals. Let ν_1 (ν_2) denote the abandonment rate of a new (follow-up) patient. Also recall that $t_1 = 45$ ($t_2 = 15$) minutes is the service time of a new (follow-up) patient, and $r_1 = \$150$ ($r_2 = \$50$) denotes the revenue obtained from a new (follow-up) patient. Glazebrook et al. (2004) suggest that giving priority to the patient class with a higher value of $r_i\nu_i/t_i$ asymptotically maximizes expected profits as the abandonment rate approaches zero when preemption is allowed and is a near-optimal heuristic if preemption is not employed, which is the case of the part-time neurologist given stable patients. We call this the *$r\nu/t$ heuristic*. Argon et al. (2008) consider a similar single server clearing system with two customer classes. They partially characterize the optimal policy, propose near-optimal heuristics, and compare their results with the $r\nu/t$ heuristic. They observe that the $r\nu/t$ heuristic performs better when the abandonment rates are small, which is the case in our system.

When we apply the $r\nu/t$ heuristic, the identical ratios $r_1/t_1 = r_2/t_2$ imply that the patient class with the higher abandonment rate should be treated first. Abandonments are rare among the new referrals, a follow-up patient is more likely to be discharged, and hence $\nu_1 < \nu_2$ and we conclude that the current practice of consulting the follow-up patients first is a near-optimal policy and it should be continued. If for some reason future changes result in new referrals being discharged with greater frequency (say, due to a shift toward considerably shorter LOSs, as has been the case in other medical specialties), the "rectangular" and "triangular" heuristics proposed by Argon et al. (2008) should be considered as they outperform $r\nu/t$ heuristic when the abandonment rates are higher.

6 Conclusion

This study was conducted with a private practice neurologist, but the model and solution methods are generalizable to any physician specialist who accepts hospital inpatient referrals, or any private practice or hospital system that hires such physicians. The variability of inpatient hospital consultations creates days in which reimbursed activities are above-average on some days and below-average on others. Proper management strikes the right balance in minimizing fixed salary losses on slow days while providing adequate physician time to maximize reimbursement on busier days. Decision-making is more difficult when the physician's time opportunity cost alternative is a better-reimbursed office practice, such as in a community practice. This holds true whether one approaches the problem as a planning dilemma, an "outsourcing" problem, or a "single-doctor" problem (see Section 1).

Implementation. Supported by the analysis in Section 2, the part-time neurologist planned her schedule around $R^* = 3$ hours per day at our subject hospital. The analysis in Section 3 suggested that never rejecting a referring physician's request is the most economically-attractive policy. The neurologist had been accepting all referrals motivated by maintaining healthy relations with referring physicians; it was comforting to confirm that this policy is also best from a financial perspective when the number of hours R is first chosen properly. However, there are two important insights due to the sensitivity analysis in Section 4: if the number of hours R

is not chosen well, then rejections of referrals in certain congested states does limit losses, and the optimal R is most sensitive to the length of stay, as reflected in the proportion of return follow-ups α. To get an idea of how α fluctuates from month to month, we considered the 13 months June 2009-June 2010.

month	average proportion of daily return follow-ups
Jun 2009	82%
Jul 2009	76%
Aug 2009	80%
Sep 2009	81%
Oct 2009	81%
Nov 2009	83%
Dec 2009	82%
Jan 2010	76%
Feb 2010	82%
Mar 2010	86%
Apr 2010	87%
May 2010	83%
Jun 2010	83%

The monthly averages were outside the 81.1% to 83.8% sensitivity range (see Section 4) five times, but the mean $\alpha = 0.82$ appears stationary. This suggests the proportion α should be monitored and R^* adjusted if necessary.

Finally, in Section 5 the analysis suggests that follow-up patients should be seen before new referrals. Again, this confirmed what the doctor had been doing to maximize the chance of seeing a follow-up before that patient is discharged; it is interesting that queueing theory supports this decision.

After the implementation, it was discovered that referral probabilities are not necessarily stationary for three reasons. First, when referring physicians see a neurologist more in the hospital, more referrals are given. On the other hand, it is possible to upset a referring physician and so an entire referral stream may evaporate. Third, referral rates tend to drift during a year for other reasons, such as weather conditions, holiday and vacations by patients and doctors, times of greater rates of illness in the general population, general economic factors, changes in other hospitals. The conclusion is that the models presented here need to be rerun periodically to adjust for "referral drift."

Epilog and discussion. Some time after this analysis was conducted, the part-time neurologist went on leave, leaving the employer-neurologist alone to cover both the office practice and hospital referrals. We therefore revisited the analysis using the "single-doctor" generalization introduced in Section 1, rather than the "outsourcing" interpretation. Recall that this requires redefining c_r as the single doctor's foregone office revenue due to a planned hour at the hospital. Conservatively, the office practice generates approximately $c_r = \$300$ per hour for the average patient-contact hour.

At the hospital, the neurologist's time bills for \$200 per hour, and we conclude the office practice generates at least 50% more revenue per hour than the hospital in the long run so some of the time at the hospital can be used more profitably by expanding office practice hours. The clear implication is that office time should never be sacrificed for voluntary hospital time, as the former is more economically sustainable, and in both cases the physician is spending just as much time caring for patients. Given the demand backlog for the office practice, working day and night to care for patients at both office and hospital—while still never rejecting a hospital referral—is financially feasible only if seeing patients at the hospital takes place at times of day or night when office patients cannot be accommodated.

We propose that a good way for United States community hospitals to balance the process is by employing full-time specialists as "administrative doctors" with shared clinical and administrative duties. The administrative duties would buffer the slow inpatient demand periods and slip when inpatient demands are high.

Hospital-employed physician specialists give the hospital another benefit. At present the private neurologist benefits from a patient's long LOS while the hospital bears the cost. This strikes us as a system out of balance and perhaps in ethical conflict. On the other hand, hospitals that employ full-time physicians to serve admitted patients can, alternatively, balance the additional costs incurred from longer lengths of stay with additional revenues from follow-up visits by their employee-physicians. This would be an important counterbalance to an ethical dilemma that we see as much greater: in the United States' current DRG (diagnosis-related group) system in which hospitals are reimbursed a single fee for a patient's whole admission regardless of length of stay (Fetter et al. 1980), hospitals are incentivized to discharge patients quickly because greater variable costs are incurred without additional revenues the longer a patient remains admitted (for example, see Dougherty 1989). Employing physicians and following our model may improve the economic viability of providing the best hospital care for a patient, regardless of the length of stay.

References

Alagoz, O., H. Hsu, A.J. Schaefer, M.S. Roberts. 2010. Markov Decision Processes: A Tool for Sequential Decision Making under Uncertainty. *Medical Decision Making*, 30(4), 474-483.

Argon, N.T., S. Ziya, R. Righter. 2008. Scheduling Impatient Jobs in a Clearing System with Insights on Patient Triage in Mass Casualty Incidents. *Probability in the Engineering and Informational Sciences*, 22(3), 301-332.

Ayvaz, N., W.T. Huh. 2010. Allocation of hospital capacity to multiple types of patients. *Journal of Revenue and Pricing Management*, 9(5), 386-398.

Arrow, K., T. Harris, J. Marshack. 1951. Optimal inventory policy. *Econometrica*, 19(3), 250-272.

Bazaraa, M.S., H.D. Sherali, C.M. Shetty. 2006. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, New York.

Bellman, R. 1957. A Markovian Decision Problem. *Journal of Mathematics and Mechanics*, 6(5), 679-684.

Berger, C., A. Haurie. 1981. Efficient Aggregation of a Class of Large Scale Markov Decision Processes through the use of a Census Process. Ecole des Hautes Etudes Commerciales, Report G-81-03.

Collart, D., A. Haurie. 1976. On a suboptimal control of a hospital inpatient admission system. *IEEE Transactions on Automatic Control*, 21(2), 233-238.

Collart, D., A. Haurie. 1980. On the control of care supply and demand in a urology department. *European Journal of Operations Research*, 4(3), 160-172.

Dougherty, C.J. 1989. Ethical perspectives on prospective payment. *Hastings Center Report*, 19(1), 5-11.

Fetter, R.B., Y. Shin, J.L. Freeman, R.F. Averill, J.D. Thompson. 1980. Case mix definition by diagnosis-related groups. *Medical Care*, 18(2) Supplement, 1-53.

Fries, B.E. 1976. Bibliography of Operations Research in Health-Care Systems. *Operations Research*, 24(5), 801-814.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation Planning for Elective Surgery under Uncertain Demand for Elective Surgery. *Management Science*, 42(3), 321-334.

Glazebrook, K.D., P.S. Ansell, R.T. Dunn, R.R. Lumley. 2004. On the Optimal Allocation of Service to Impatient Tasks. *Journal of Applied Probability*, 41(1), 51-72.

Green, L.V., S. Savin, B. Wang. 2006. Managing Patient Service in a Diagnostic Medical Facility. *Operations Research*, 54(1), 11-25.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9), 800-819.

Heyman, D.P., M.J. Sobel. 1984. *Stochastic Models in Operations Research, Vol. II: Stochastic Optimization*, 2003 reprint, Dover Publications.

Kao, E.P.C. 1974. Study of Patient Admission Policies for Specialized Care Facilities. *IEEE Transactions on Systems, Man and Cybernetics*, 4(6), 505-512.

Kolesar, P. 1970. A Markovian Model for Hospital Admission Scheduling. *Management Science*, 16(6), 384-396.

Kreke, J.E., M.D. Bailey, A.J. Schaefer, D.C. Angus, M.S. Roberts. 2008. Modeling hospital discharge policies for patients with pneumonia-related sepsis. *IIE Transactions*, 40(9), 853-860.

Lin, J., K. Muthuraman, M. Lawley. 2011. Optimal and approximate algorithms for sequential clinical scheduling with no-shows. *IIE Transactions on Healthcare Systems Engineering*, 1(1), 20-36.

Liu, N., S. Ziya, V.G. Kulkarni. 2010. Dynamic scheduling of outpatient appointments under patient under patient no-shows and cancellations. *Manufacturing & Service Operations Management*, 12(2), 347-364.

Nunes, L. G. N., S.V. de Carvalho, R.C.M. Rodrigues. 2009. Markov decision process applied to the control of hospital elective admissions. *Artificial Intelligence in Medicine*, 47(2), 159-171.

Patrick, J., M.L. Puterman, M. Queyranne. 2008. Dynamic multi-priority patient scheduling for a diagnostic resource. *Operations Research*, 56(6), 1507-1525.

Prahinski, C., J.S. Haywood-Farmer, D. Wright, K. Saskiw. 2003. Quinte MRI. Ivey Business Publishing, case number 9-B02-D024.

Puterman, M.L. 1994 *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, New York.

Schaefer, A.J., M.D. Bailey, S.M. Shechter, M.S. Roberts. 2005. Modeling medical treatment using Markov decision processes. *Operations Research and Health Care*, 70(4), 593-612.

Zonderland, M.E., R.J. Boucherie, N. Litvak, C.L.A.M. Vleggeert-Lankamp. 2010. Planning and scheduling of semi-urgent surgeries. *Health Care Management Science*, 13(3), 256-267.

Appendix

Proposition 1

Proof. Assume that the system starts in state x^0. Let $Q^n_{x_0,x}$ denote the probability that the system is in state $x = (x_1, x_2)$ at time n, given that it is in state x^0 at time 0. We can write the total discounted profit under policy π as

$$E_\pi \left[\sum_{n=0}^{\infty} \beta^n u_{n,\pi} \left(X_1^n, X_2^n \right) \right] = \sum_{n=0}^{\infty} \beta^n \left[\sum_{x \in S} Q^n_{x_0,x} u_{n,\pi} \left(x_1, x_2 \right) \right].$$

Moreover, the daily net profit obtained at time n is equal to

$$u_{n,\pi} \left(x_1, x_2 \right) = r_1 x_1 + r_2 x_2 - c_r R_\pi - c_o (t_1 x_1 + t_2 x_2 - R_\pi)^+. \tag{5}$$

For any $R_1, R_2 \in \mathbb{R}$ and $\gamma \in [0,1]$, we have

$$\begin{aligned}
& r_1 x_1 + r_2 x_2 - c_r(\gamma R_1 + (1-\gamma) R_2) - c_o(t_1 x_1 + t_2 x_2 - (\gamma R_1 + (1-\gamma) R_2))^+ \\
\geq \;& \gamma \Big(r_1 x_1 + r_2 x_2 - c_r R_1 - c_o(t_1 x_1 + t_2 x_2 - R_1)^+ \Big) \\
& + (1-\gamma) \Big(r_1 x_1 + r_2 x_2 - c_r R_2 - c_o(t_1 x_1 + t_2 x_2 - R_2)^+ \Big) \\
\Leftrightarrow \;& \gamma (t_1 x_1 + t_2 x_2 - R_1)^+ + (1-\gamma)(t_1 x_1 + t_2 x_2 - R_2)^+ \geq (t_1 x_1 + t_2 x_2 - (\gamma R_1 + (1-\gamma) R_2))^+. \quad (6)
\end{aligned}$$

When $R_1, R_2 \leq t_1 x_1 + t_2 x_2$, (6) holds because both the left and right hand sides equal $t_1 x_1 + t_2 x_2 - \gamma R_1 - (1-\gamma) R_2$. When $R_1, R_2 > t_1 x_1 + t_2 x_2$, (6) holds because both the left and right hand sides equal zero. When $R_1 > t_1 x_1 + t_2 x_2$ and $R_2 \leq t_1 x_1 + t_2 x_2$, (6) holds because $(1-\gamma)(t_1 x_1 + t_2 x_2 - R_2) \geq (t_1 x_1 + t_2 x_2 - (\gamma R_1 + (1-\gamma) R_2))^+ \geq \gamma(t_1 x_1 + t_2 x_2 - R_1) + (1-\gamma)(t_1 x_1 + t_2 x_2 - R_2)$. The last inequality holds because $t_1 x_1 + t_2 x_2 - R_1 < 0$. Similarly, when $R_1 \leq t_1 x_1 + t_2 x_2$ and $R_2 > t_1 x_1 + t_2 x_2$, (6) holds because $\gamma(t_1 x_1 + t_2 x_2 - R_1) \geq (t_1 x_1 + t_2 x_2 - (\gamma R_1 + (1-\gamma) R_2))^+ \geq \gamma(t_1 x_1 + t_2 x_2 - R_1) + (1-\gamma)(t_1 x_1 + t_2 x_2 - R_2)$, where the last inequality holds because $t_1 x_1 + t_2 x_2 - R_2 < 0$. Hence, $u_n \left(x_1, x_2 \right)$ is concave in R. Since the sum of concave functions is concave (see, e.g., Bazaraa et al. 2006, p. 99) and $Q^n_{x_0,x}$ is independent of $u_n \left(x_1, x_2 \right)$ when all referrals are accepted, the total discounted profit is concave as a function of R. □

Proposition 2

Proof. Let π' be the policy that accepts all referrals in all states. Let π'' be another policy identical to π' except that referrals are rejected in an arbitrary state $x^0 \in \mathcal{S}$. The acceptance policy affects the number of new and follow-up patients in the system, hence the system state is a function of π in the arguments that follow. Let n^0 be the first time the Markov chain reaches state x^0. Since acceptance does not affect any other patient's referral acceptance, arrival, or LOS, we can conclude that $X^n_{i,\pi'} \geq X^{n^0}_{i,\pi''}$ for all $n \geq n^0$ (the sample paths of the two Markov chains will not be the same after time n^0, but this does not change that the number of patients in the system will never be less under policy π'). For any patient in the system at a certain time, the revenue obtained is r_1 or r_2 with cost $c_o t_1$ or $c_o t_2$, respectively. Hence, the profit associated with a patient is nonnegative if $r_i \geq c_o t_i$ for $i \in \{1, 2\}$ and we can conclude that $u_n(X^n_{1,\pi'}, X^n_{2,\pi'}) \geq u_n(X^n_{1,\pi''}, X^n_{2,\pi''})$ and so $E_{\pi'}\left[\sum_{n=0}^{\infty} \beta^n u_n\left(X^n_{1,\pi'}, X^n_{2,\pi'}\right)\right] \geq E_{\pi''}\left[\sum_{n=0}^{\infty} \beta^n u_n\left(X^n_{1,\pi''}, X^n_{2,\pi''}\right)\right]$. □

		new referrals, x_1							
		0	1	2	3	4	5	6	7
follow-ups, x_2	0	101,455	101,771	102,084	102,394	102,697	102,882	103,064	103,239
	1	101,671	101,984	102,294	102,597	102,857	103,039	103,214	103,384
	2	101,884	102,194	102,497	102,795	103,014	103,189	103,359	103,523
	3	102,094	102,397	102,695	102,989	103,164	103,334	103,498	103,656
	4	102,297	102,595	102,889	103,139	103,309	103,473	103,631	103,783
	5	102,495	102,789	103,077	103,284	103,448	103,606	103,758	103,905
	6	102,689	102,977	103,259	103,423	103,581	103,733	103,880	104,021
	7	102,877	103,159	103,398	103,556	103,708	103,855	103,996	104,131
	8	103,059	103,335	103,531	103,683	103,830	103,971	104,106	104,237
	9	103,235	103,506	103,658	103,805	103,946	104,081	104,212	104,337
	10	103,406	103,633	103,780	103,921	104,056	104,187	104,312	104,434
	11	103,571	103,755	103,896	104,031	104,162	104,287	104,409	104,527
	12	103,730	103,871	104,006	104,137	104,262	104,384	104,502	104,617
	13	103,846	103,981	104,112	104,237	104,359	104,477	104,592	104,705
	14	103,956	104,087	104,212	104,334	104,452	104,567	104,680	104,788
	15	104,062	104,187	104,309	104,427	104,542	104,655	104,763	104,864
	16	104,162	104,284	104,402	104,517	104,630	104,738	104,839	104,929
	17	104,259	104,377	104,492	104,605	104,713	104,814	104,904	104,983
	18	104,352	104,467	104,580	104,688	104,789	104,879	104,958	105,024
	19	104,442	104,555	104,663	104,764	104,854	104,933	104,999	105,055
	20	104,530	104,638	104,739	104,829	104,908	104,974	105,030	105,079

Table 1: Values of $v_{365}(x_1, x_2)$ when $R^* = 3$ in the one-decision problem described in Section 2.

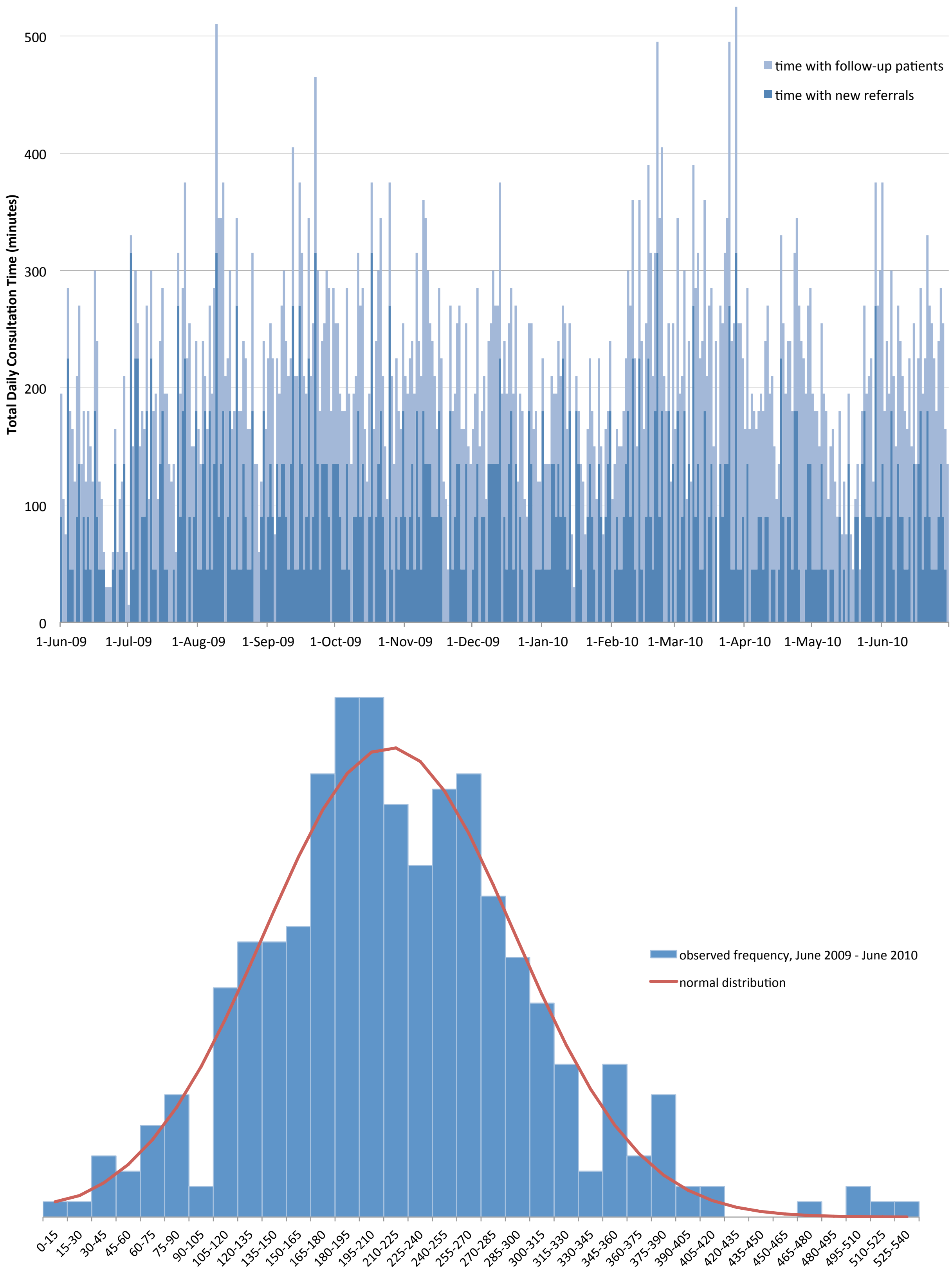

Figure 1: Total daily hospital consultation times, 1-June-2009 to 30-June-2010, shown as a time series and frequency distribution of 15-minute intervals. In the first, the height of each bar denotes the daily sum of times with new referrals and follow-up patients.