Reducing Wait Times through Operations Research: Optimizing the Use of Surge Capacity*

Jonathan Patrick, PhD and Martin L. Puterman, PhD

ABSTRACT

Widespread public demand for improved access, political pressure for shorter wait times, a stretched workforce, an aging population and overutilized equipment and facilities challenge healthcare leaders to adopt new management approaches. This paper highlights the significant benefits that can be achieved by applying operations research (OR) methods to healthcare management. It shows how queuing theory provides managers with insights into the causes for excessive wait times and the relationship between wait times and capacity. It provides a case study of the use of several OR methods, including Markov decision processes, linear programming and simulation, to optimize the scheduling of patients with multiple priorities. The study shows that by applying this approach, wait time targets can be attained with the judicious use of surge capacity in the form of overtime. It concludes with some policy insights.

Healthcare systems throughout the world face long and increasing wait times for medical services (Willcox et al. 2007; Siciliani and Hurst 2004; Hurst and Siciliani 2003; Blendon 2002). Sometimes these waits may have little medical impact, but excessive delays may be detrimental to patients' health (CIHR 2007). As a result, there is growing public and patient pressure on political leaders to reduce wait times to acceptable levels.

The First Ministers' Meeting on the Future of Health Care (2004) committed Canada to a program of determining, and then meeting, wait time benchmarks for cancer care, cardiac care, diagnostic imaging, joint replacement and sight restoration. These benchmarks provide "evidence based goals that express the amount of time that clinical evidence shows is appropriate to wait for a particular procedure or diagnostic test" (Postl 2006). Postl (2006), in his final report as Federal Advisor on Wait Times, noted that "we [in healthcare management] have not sufficiently exploited the academic resources available to us from business management schools or industrial engineering." In particular, he singled out operations research (OR) as especially relevant.

Operations research is the science of developing and applying mathematical models to provide decision-makers with better strategies to plan and operate systems. Through systems models and "what if?" analyses, it enables investigation of the impact of system changes prior to implementation. This paper uses OR methods to provide insight into the relationship between wait times and capacity. Through a case study, it also shows how our basic research on patient scheduling algorithms (Patrick et al. 2007) can reduce wait times by judiciously using surge capacity in the form of overtime.

Why Are There Waits for Access to Healthcare?
Wait times for health services arise because

- capacity does not match demand,
- capacity or demand is not well managed and
- there is significant variability over time in the demand for healthcare services.

By capacity, we mean the maximum rate at which a resource can deliver a service when operating at peak efficiency (Anupindi et al. 2005). Capacity is controlled through investment in and scheduling the use of people, physical plant and equipment. Setting capacity levels entails an unavoidable trade-off between wait times and resource utilization.

- When capacity significantly exceeds average demand, queues will be short and wait times minimal. Unfortunately,

* Reprinted from *Healthcare Policy* Volume 3, Number 3.

because of *variability* in the demand over time, resources will be idle a large portion of the time (Figure 1).

- When capacity is significantly below average demand, system resources will be fully utilized, but wait times will be excessive and will grow over time.
- Even when capacity equals or slightly exceeds average demand, queuing theory (Hillier and Lieberman 2001) implies that there will be long waits (in theory, they will be infinite in the limit). In this case, resources will be utilized most of the time.

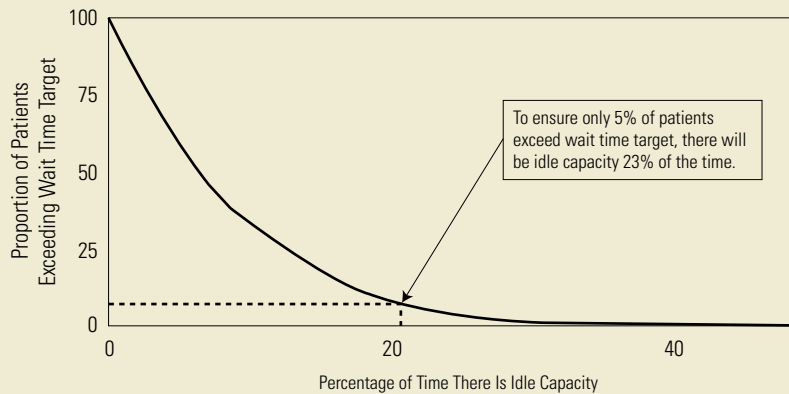Decisions regarding investment in capacity must explicitly account for the trade-off between capacity, idleness and wait times. The relation depicted in Figure 1 derives from fundamental queuing theory (Hillier and Lieberman 2001: 854–55). It shows that to ensure that a low proportion (percentage) of patients' wait times exceed specified targets, capacity must be set sufficiently high so that idle time is inevitable. The specific case indicated by the dotted line shows that to ensure that only 5% of patients exceed their wait time targets, capacity will be idle 23% of the time. Ingolfsson and Gallup (2003) developed the Queueing ToolPak, which is an easy-to-use Excel add-in that facilitates a wide range of queuing calculations that are useful for capacity planning.

Figure 1. Theoretical relationship between wait time targets and idle capacity

To ensure only 5% of patients exceed wait time target, there will be idle capacity 23% of the time.

Note:Calculations are based on a single server exponential queuing model with arrival rate 10 per week, service rate varied from 16 to 10 per week and a target of one-week service.

The Challenge of Measuring Wait Times

There are a number of complex issues that affect both the setting and achieving of wait time targets. They include the following:

- *Patients are not homogeneous.* Patients requiring urgent care must receive services more quickly than those requiring less urgent care. Hence, wait times must be assessed against appropriate benchmarks for each priority class. A report by the Health Council of Canada (2005) suggests using the terminology "urgency" instead of "priority" to avoid the negative connotations sometimes associated with the latter term. Further, it suggests using three patient urgency classes.
- *Wait times, as currently measured, do not tell the whole story.* Usually, wait times are measured and reported from the time at which a request for service (requisition) reaches the service provider until the service is provided. They do not account for upstream delays between the time at which a service is first needed and the points at which the

series of referring physicians can see and enter the patient into the appropriate queue.

- *Averages are not enough.* Wait times vary among patients, over time and among sites and measures. This variability must therefore be part of any performance measurement system. Further, wait time distributions tend to be skewed. We strongly recommend using metrics of the form: "What proportion of patients of a specific priority class receive the service within a specific, clinically desirable time?" The advantage of such metrics is that they provide meaningful guarantees to decision-makers and system users.
- *Accurate wait time data are not readily available.* Most data systems we have encountered do not provide complete wait time data. The biggest challenges are that time stamps are not accurate, data reside in different offices and databases are not linked. Further, relevant data are often not available in electronic form.

Levers for Managing Capacity: The Impact of Operations Research

Operations research methods can help health systems managers plan and manage capacity to meet wait time targets in the following ways:

- *Capacity planning* addresses the issue of how much capacity is needed to meet current and future wait time targets. Systemwide planning models based on linear and integer programming (Santibanez et al. 2007) can determine where and when to add system capacity.
- *Capacity management* addresses the question of how to assign demand to capacity to ensure that it is used as efficiently as possible. Levers to manage capacity efficiently

include improved patient scheduling, improved staff scheduling and modelling the entire system to account for and manage all resources a patient will need when scheduling services. Simulation, linear programming and Markov decision processes support such analyses.

- *Surge capacity management* provides a hybrid approach to planning and controlling capacity. Surge capacity refers to extra capacity (such as overtime) that can be sourced when needed to meet excess demand. Often, surge capacity is more expensive than base capacity, but its advantage is that it is not needed all the time. As noted above, when capacity is set sufficiently high so that wait time targets are achieved, capacity will be idle a significant fraction of the time. On the other hand, if capacity is set so that it is fully utilized, then a large fraction of patients will not achieve their wait time targets. A way around this trade-off is to have additional flexible or surge capacity that allows the system to function with less base capacity, and therefore less unused capacity, while still meeting demand. Managing surge capacity requires both determining an appropriate base capacity and developing rules that specify when and how much surge capacity should be used.

Postl (2006) stresses the need for research to explore the use and benefits of surge capacity. This paper answers that call by providing a case study that illustrates how operations research methods can help a manager develop rules to use surge capacity, in the form of overtime, to achieve wait time targets by developing efficient multi-priority patient scheduling rules. It does not address the issue of setting base capacity levels, which we shall address in future work.

The Need for Optimal Scheduling

In 2003, a team of investigators from the University of British Columbia (including the authors) began a study with the Vancouver Coastal Health Authority aimed at improving diagnostic imaging processes at several Vancouver hospitals. The team identified porter services and patient scheduling as promising areas for investigation. As a result, we reviewed porter services (Odegaard et al. 2007) and sought to develop new methods to improve patient scheduling (Patrick and Puterman 2007; Patrick et al. 2007). This paper translates the latter research into a decision-making context.

In most healthcare settings, patient scheduling is carried out by schedulers who must make complex trade-offs in the absence of intelligent software and precise decision rules to support their decisions. This activity becomes especially challenging and complex when

- patients are categorized into priority classes with different service time targets,

- there are multiple types of equipment with different capabilities on which a patient can be scheduled,
- patients must be booked for a course of treatment requiring several days or weeks or
- resources are spread across a wide geographic region.

Our research focused on the first issue and provides a foundation for investigating the other challenges. The specific problem our research investigated follows.

Each day, a random number of appointment requests arrive. A scheduler reviews these requests and assigns them to a pre-specified number of future appointment slots of constant length. Each request has a priority assigned to it. Patients with different priorities have different maximum recommended wait times. The challenge that the scheduler faces is that lower-priority patients must be booked "today" (for an appointment slot some time in the future) prior to knowing future demand. If patients are booked too far in the future, their maximum recommended wait times may be exceeded and staff and equipment may sit idle. If patients are booked too soon, then there may not be sufficient capacity to meet wait time targets for higher-priority patients arriving at a later date. Our research provides precise decision rules to enable schedulers to make these booking decisions and meet high-priority targets for all priority classes. We refer to such rules as *optimal schedules*. We will clarify what we mean by "optimal" below.

Problems of this type have received some attention in the operations research literature. Related papers within the healthcare field include the work of Gerchak et al. (1996) on allocating surgery time between elective and emergency surgeries, Gupta and Wang's (2008) paper on scheduling in a primary care clinic with multiple priority levels and the study by Green et al. (2006) on managing patient demand for a diagnostic facility. Our work differs from the first paper in that it considers more than two priority classes. In the two other papers, the objective is to maximize revenue, allowing the authors to focus on a single day rather than the entire planning horizon. Our interest in achieving wait time targets forces us to consider how actions taken on one day affect future decisions so that we cannot look at a single day in isolation. McGill and Van Ryzin (1999) summarize related research on customer scheduling in the airline industry, and Bassamboo et al. (2006) consider similar challenges facing call-centre managers when scheduling operators. As far as we know, there is no research on multi-class scheduling within healthcare, where the goal is to achieve wait time targets.

Methodology

The results and policy insights of our research are based on a Markov decision process (MDP) scheduling model. An MDP

models a system in which decisions are made sequentially over time, and future decisions and outcomes depend on current and past decisions (Puterman 1994). Applying an MDP provides an optimal policy that prescribes how best to manage the system in any contingency. It offers a systematic alternative to the "guess and check" approach that underlies using simulation on its own to determine good policies. In our setting, the system is described by the number of appointment slots available on each future day and the number of patients of each priority class waiting to be booked. A policy provides the scheduler with a set of rules specifying when in the future to schedule each waiting patient. Unfortunately, to determine optimal policies for realistic-sized systems, the MDP model becomes challenging, if not impossible, to apply.

Over the past decade, researchers in operations research, engineering and computer science (Bertsekas and Tsitsiklis 1996; Sutton and Barto 1998) have developed a new branch of operations research called *approximate dynamic programming* (ADP) that seeks to overcome such computational challenges. ADP methods produce good but not necessarily optimal solutions to the underlying problem. Policies obtained through ADP must be evaluated by testing them in a system simulation model. We use a simulation model to compare the optimal scheduling rules derived from the ADP with a range of alternatives, including current practice. As we will show below, the policy derived from our research outperforms both current practice and reasonable alternative policies. In this paper, we will use the expression "optimal" to refer to the policy obtained from the application of the ADP. However, how to quantify the proximity of this policy to the true optimal solution of the underlying problem remains a research challenge.

Optimal Scheduling Policy
The optimal scheduling (OS) policy assumes an externally determined fixed number of appointment slots of fixed length each day and the potential to use overtime as surge capacity. We find it by formulating an MDP model of the scheduling problem and using ADP methods to solve it. In this application, the policy has an intuitively appealing form, which we describe in Table 1. This policy can be easily integrated in a decision support tool, or communicated directly to schedulers as a set of easy-to-follow rules.

Our research shows that it is never advisable to book patients beyond their wait time targets. Doing so does not avoid the need for overtime; instead, it just delays when it is needed. For lower-priority patients, scheduling them as late as possible without exceeding the wait time target for their priority class gives the scheduler maximum flexibility to account for future demand variability.

Our research also shows that OS policy remains optimal

Table 1. Optimal scheduling policy

- Fill any unused capacity for the next day by assigning patients to that day in priority order.
- Schedule any remaining high-priority demand to the earliest available time slots before the maximum recommended wait time for this class.
- If there is outstanding high-priority demand that cannot be scheduled prior to its maximum recommended wait time, serve it through overtime.
- Schedule all other priority classes starting from the last available day that does not exceed the maximum recommended waiting time for that priority class, scheduling patients in priority class order.
- If there is insufficient capacity to schedule demand prior to its target date after all higher-priority classes have been allocated in the order described above, use overtime to serve this demand.

---

regardless of the number of priority classes, the specific wait time targets for each priority class and the length of the booking horizon. Also, it remains optimal for all reasonable overtime (OT) costs and as long as capacity is not significantly greater than average demand. In the unlikely circumstance that there is a large amount of excess capacity, then the optimal policy becomes instead a first come, first served policy. If capacity is significantly below average demand, the same policy remains optimal but OT costs may become excessive.

Comparison with Other Policies
To illustrate the benefits of using the OS policy, we focus on scheduling outpatient (OP) demand for a single diagnostic resource. We assume three OP priority classes; that regular hour capacity equals the average demand; that the three OP priority classes have maximum recommended waiting times of 7, 14 and 21 days, respectively; and that unlimited overtime capacity is available. Each day, the scheduler may assign patients to any day up to 30 days into the future (the booking horizon). We developed and used a simulation model of the patient arrival and scheduling process to compare the performance of different policies.

Figure 2 summarizes output for the simulation model and shows that the OS policy achieves all wait time targets. Additional output from the simulation (not shown here) indicates that fewer than 1.5% of patients are served through overtime. This percentage is equivalent to one overtime scan every 20 days (but the need for overtime comes in batches). Thus, the judicious use of surge capacity, in the form of overtime, achieves target wait times without adding costly base capacity. Though surge capacity may be more expensive to supply on a per case basis, using it in the manner described

above will prove less costly over the long term than maintaining a base capacity that is sufficiently high to achieve wait time targets. Of course, managers face the challenge of ensuring that a sufficient number of staff are available to work overtime.

We also used the simulation model to compare the OS policy to a booking limit (BL) policy and an overtime "only as a last resort" policy (OTLR), that is, a policy that uses overtime only when there is no available capacity in the booking horizon. BL policies have been widely used in the airline industry to reserve seating capacity for late-arriving, high-priority demand. A BL policy will book a patient on a given day only if the available unused capacity for that patient's priority class on that day exceeds a predetermined booking limit. The amount of reserved capacity varies with the priority class of the patient; the lower the priority class, the more space needs to be available before a booking will be made. Determining optimal booking limits requires solving another optimization problem, or enumerating all possibilities and evaluating them with simulation (Patrick et al. 2007). A BL policy will use overtime only if there is no available space for the priority class of the waiting patient over the planning horizon. The OTLR policy has appeared in some collective agreements. Comparisons are summarized in Table 2.

Clearly, the OS policy not only requires less overtime than the BL policy but also results in shorter wait times for all priority classes. The OTLR policy requires significantly less overtime but at the cost of a significant number of late bookings. Figure 3 provides the waiting time distribution by priority class for the OTLR policy. The dramatic improvement in wait times at the expense of low overtime costs clearly

Figure 2. Waiting times by priority class for the OS policy
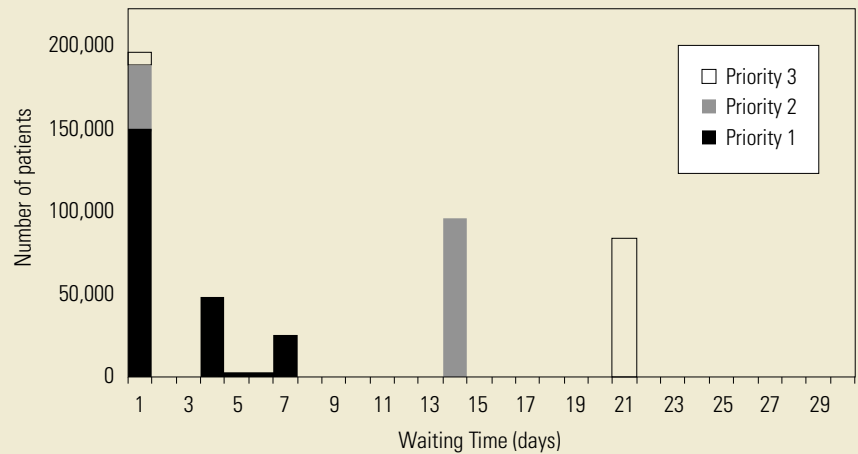
demonstrates the value of the OS policy over the OTLR policy from the perspective of meeting wait time targets.

Current practice does not seek to address outpatient demand through overtime. Instead, the booking horizon is pushed farther into the future. The OS policy provides a practical alternative to current practice without extending the booking horizon. As long as not scheduling a patient past a wait time target is an important criterion, the OS policy will significantly outperform current practice.
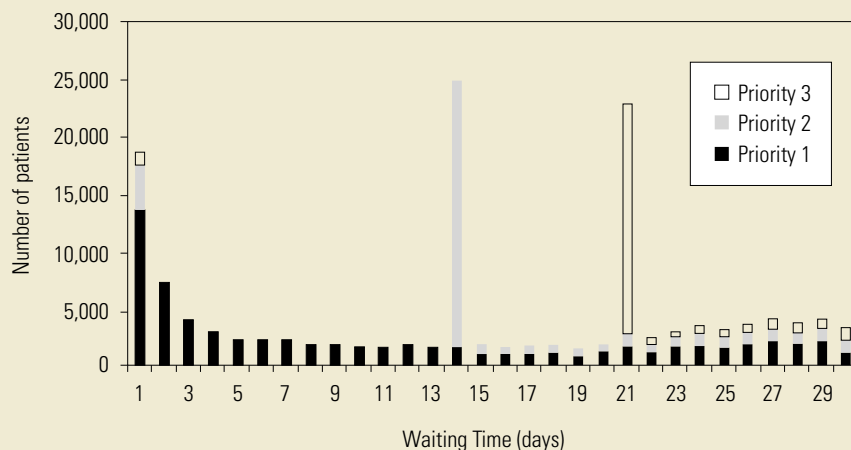
Including Inpatient Demand

Although the OS policy described in Table 1 remains "optimal" when inpatients are included in the model, there is a significant increase in overtime. This is because the highest-priority patient class must now be served the day the request is placed (current practice for inpatient demand), whereas in the setting described above, the highest-priority class could be served any time in the first week. This approach significantly affects a scheduler's ability to manage variability in demand. We therefore investigated the potential benefits of introducing flexibility into inpatient scheduling. In fact, in the setting we investigated, there was already flexibility in inpatient scheduling that was not being utilized. The priority scheme for inpatients implemented in Vancouver in 2003 designated a category of inpatients who can wait one day to receive

Table 2. Comparison of the OS, BL and OTLR policies based on simulation

Policy	% Late			% Served through OT			
	Priority 1	Priority 2	Priority 3	Priority1	Priority 2	Priority 3	Total
OS	0	0	0	1.44	0	0	0.72
BL	0	0.02	49.52	0	0	20.57	4.13
OTLR	53.17	35.77	24.85	0.08	0.43	0	0.17

Figure 3. Waiting time by priority class for the OTLR policy

a scan. Current practice ignores this flexibility and schedules all inpatient scans on the day of the request.

To apply our approach we used the OS policy with five priority classes, the three described above and one- and two-day targets for inpatients. Evaluating this policy through simulation showed that if only 10% of patients can wait one day for service, the average number of overtime scans per day was reduced from 4.27 to 2.67. Furthermore, overtime was required only for the highest-priority inpatients. Of course, if delaying a diagnostic procedure delayed inpatient release time, the resulting additional cost might offset any benefit.

These results run contrary to the current practice of seeking to improve wait times by pushing the wait time target of the lowest-priority class farther into the future. In fact, though there may be a temporary relief to the system from doing so, there is no long-term benefit to be gained from manipulating the wait time target of the lowest-priority class. Rather, it is the wait time target of the highest-priority class that is the driving factor in determining the required amount of overtime. Admittedly, there may be little flexibility in the scheduling of the highest-priority class, but the above results suggest that any amount of flexibility is worth pursuing.

When Is Optimal Scheduling Beneficial?

The OS policy proposed here manages the trade-off between unused capacity and overtime. When base capacity far exceeds average demand, the scheduler can use a first come, first served policy because capacity is unlikely to be fully utilized. However, in such a case, capacity will be underutilized, and resources will be idle a significant portion of the time. Conversely, if capacity is far below average demand, then a wide range of scheduling policies will probably do

equally well, as there is likely to be little unused capacity. In such a case, overtime costs will be large.

As Figure 1 suggests, providing sufficient base capacity to meet all demand is unlikely to be cost effective. Nor is setting base capacity so low that average demand significantly outstrips capacity. We suggest setting base capacity to minimize expected total costs of overtime and base capacity costs. Our limited exploration of this issue suggests that the base optimal capacity would be somewhat below average demand, the precise amount depending on overtime costs and demand rates.

We emphasize that two factors cause excessive waiting times: the amount of demand and its variability. Optimal scheduling can help mitigate the effect of variability. If average demand outstrips available capacity (regular and overtime), then no optimal schedule will help. However, even if base capacity is sufficient to meet average demand, there will still be increasing wait times owing to the variability in demand unless overtime is used appropriately. It is in such a scenario that the optimal scheduling policy will allow the resource manager to minimize unused capacity (which implicitly minimizes wait times) with the least amount of overtime.

Conclusions

When we began our study for the Vancouver Coastal Health Authority, the question posed to us was, "Where should a new CT scanner be located?" What our analysis revealed was that there was a significant amount of capacity that could be recovered by utilizing existing resources more efficiently. We recognized that one way to achieve this efficiency was through better patient scheduling. This finding led to our research on patient scheduling methods, which produced the OS policy, which in turn was shown through simulation to achieve wait time targets. Thus, even if more capacity is required, managers must first ensure that current capacity is used to its fullest potential.

In addition to the OS scheduling rule, we drew the following policy implications from our research:

- In the absence of surge capacity (in the form of overtime), there will either be significant wait times or significant idle capacity.
- With a small amount of overtime, wait times can be

maintained within the targets without significant excess capacity.

- The amount of overtime required depends heavily on the wait time target for the highest-priority class; if it is short, overtime needs are likely to be high. Classifying high-priority patients carefully and seeking subclasses with different targets could be useful levers for reducing overtime costs.

We hope that this work will prove useful to healthcare managers and policy makers and, as well, whet their appetite for further healthcare operations research studies.

Correspondence may be directed to: Martin L. Puterman, Research Director, UBC Centre for Health Care Management, 2053 Main Mall, Vancouver, BC, V6T 1Z2. Tel: 604-822-8388; e-mail: Martin.Puterman@sauder.ubc.ca.

Acknowledgement

References
Anupindi, R., S. Chopra, S. Deshmukh, J.A. Van Miegham and E. Zemel. 2005. *Managing Business Process Flows: Principles of Operations Management*. Englewood Cliffs, NJ: Prentice Hall.

Bassamboo, A., J.M. Harrison and A. Zeevi. 2006. "Design and Control of a Large Call Center: Asymptotic Analysis of an LP-Based Method." *Operations Research* 54: 419–35.

Bertsekas, D.P. and J. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Nashua, NH: Athena Scientific.

Blendon, R. 2002. "Inequities in Health Care: A Five-Country Survey." *Health Affairs* 21: 182–91.

Canadian Institute of Health Research (CIHR). 2007. "CIHR Releases Research Results to Inform the Development of Benchmarks for Wait Times. Backgrounder for Wait Times for Joint Replacement Surgery." Retrieved January 12, 2008. <http://www.cihr-irsc.gc.ca/e/29904.html>.

First Ministers' Meeting on the Future of Health Care. 2004. "A 10-Year Plan to Strengthen Health Care." Retrieved January 12, 2008. <http://www.hc-sc.gc.ca/hcs-sss/delivery-prestation/fptcollab/2004-fmm-rpm/index_e.html>.

Gerchak, Y., D. Gupta and M. Henig. 1996. "Reservation Planning for Elective Surgery under Uncertain Demand for Emergency Surgery." *Management Science* 42: 321–34.

Green, L., S. Savin and B. Wang. 2006. "Managing Patient Demand in a Diagnostic Medical Facility." *Operations Research* 54: 11–25.

Gupta, D. and L. Wang. 2008. In press. "Revenue Management for a Primary-Care Clinic in Presence of Patient Choice." *Operations Research*.

Health Council of Canada. 2005 (November). "Ten Steps to a Common Framework for Reporting Wait Times." Retrieved January 12, 2008. <http://www.healthcouncilcanada.ca/docs/papers/2005/WaitTimesEn.pdf>.

Hillier, F. and G. Lieberman. 2001. *Introduction to Operations Research* (7th ed.). New York: McGraw Hill.

Hurst, J. and L. Siciliani. 2006. *Tackling Excessive Waiting Times for Elective Surgery: A Comparison of Policies in Twelve OECD Countries*. OECD Health Working Papers 6. Paris: Organisation for Economic Co-operation and Development.

Ingolfsson, A. and F. Gallup. 2003. University of Alberta School of Business. "Queueing Toolpak 4.0." Retrieved January 12, 2008. <http://www.business.ualberta.ca/aingolfsson/qtp>.

McGill, J. and G.J. van Ryzin. 1999. "Revenue Management: Research Overview and Prospects." *Transportation Science* 33: 233–56.

Odegaard, F., L. Chen, R. Quee and M.L. Puterman. 2007. "Improving the Efficiency of Hospital Porter Services, Parts 1 and 2." *Journal for Health Care Quality* 29: 4–18.

Patrick, J. and M.L. Puterman. 2007. "Improving Resource Utilization for Diagnostic Services through Flexible Inpatient Scheduling: A Method for Improving Resource Utilization." *Journal of the Operational Research Society* 58: 235–45.

Patrick, J., M.L. Puterman and M. Queyranne. 2007. Under review. "Dynamic Multi-Priority Patient Scheduling." *Operations Research*.

Postl, B. 2006. "The Final Report of the Federal Advisor on Wait Times." Retrieved January 12, 2008. <http://www.hc-sc.gc.ca/hcs-sss/pubs/system-regime/2006-wait-attente/index_e.html>.

Puterman, M. 1994. *Markov Decision Processes*. New York: John Wiley and Sons.

Santibanez, P., M. Begen and D. Atkins. 2007. "Surgical Block Scheduling in a System of Hospitals: An Application to Resource and Wait List Management in a British Columbia Health Authority." *Health Care Management Science* 10: 269–82.

Siciliani, L. and J. Hurst. 2004. "Explaining Waiting-Time Variations for Elective Surgery across OECD Countries." *OECD Economic Studies* 38(1): 95–123.

Sutton, R.S. and A.G. Barto. 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Willcox, S., M. Seddon, S. Dunn, R. Tudor Edwards, J. Pearse and J.V. Tu. 2007. "Measuring and Reducing Wait Times: A Cross-National Comparison of Strategies." *Health Affairs (Millwood)* 26(4): 1078–87.