Dynamic Patient Scheduling for a Diagnostic Resource

by

Jonathan Patrick

B.Arts & Sc. McMaster University 1997

M.Sc. The University of British Columbia 1999

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY in

THE FACULTY OF GRADUATE STUDIES

(Business Administration)

THE UNIVERSITY OF BRITISH COLUMBIA

October 2006

©Jonathan Patrick, 2006
Abstract

We take an in-depth look at the scheduling of patients for a diagnostic resource. Our aim is to schedule patients in order to maintain reasonable waiting times for minimal cost. We assume a fixed capacity with stochastic demand coming from multiple priority classes. We further assume that it is the lower priority patients that must be booked first, therefore requiring the resource manager to implement a booking policy to assure room for later arriving higher priority patients. If too much capacity is reserved for higher priority patients then there will inevitably be unused capacity resulting in longer waiting times than might otherwise be the case. If not enough capacity is reserved then higher priority patients will have to be served through overtime or forced to wait longer than recommended. We begin, in Chapter 1, with an international overview of efforts to reduce waiting times. Chapter 2 proposes a scheduling policy assuming only two priority classes and a fixed limit on expected overtime. The higher priority class are inpatients who must be served the day the demand for a scan is placed. The lower priority class consists of outpatients who may be booked weeks in advance. We present a model that gives the optimal reservation policy and examine the benefit of introducing some flexibility into the higher priority (inpatient) class. Chapter 3 then restricts the modeling to outpatients where demand is divided into multiple priority classes. We present a Markov Decision Process that we solve through approximate dynamic programming in order to derive an approximately optimal booking policy that maintains reasonable waiting times for minimal cost.
Chapter 4 presents some strong theoretical results as to the nature of the optimal policy from chapter 3 as well as providing bounds on the "deviation from optimality" associated with our approximation. Chapter 5 then adds inpatients to the model in chapter 3 and compares the results of the full model to those given in chapter 2. Finally, we conclude with possible enhancements to the model and policy insights for the resource manager.
Contents

Abstract ......................................................... ii

Contents ......................................................... iv

List of Tables ................................................... vii

List of Figures .................................................. viii

Acknowledgements ............................................... ix

Dedication ....................................................... x

Chapter 1. Meeting the Waiting Time Challenge ............... 1
  1. The Challenge ............................................ 1
  2. Limiting Demand ........................................ 5
  3. Increasing Capacity ..................................... 10
  4. Improved Efficiency ..................................... 12
  5. The Impact of Scheduling on Waiting Times .......... 16
  6. Diagnostic Imaging ...................................... 18

Chapter 2. Improving Resource Utilization through Inpatient Scheduling ............................................. 25
  1. Introduction ............................................. 25
  2. The Problem ............................................. 25
  3. The Proposed Scheduling Approach .................. 27
2. Policy Insights ............................................................ 127
Bibliography ............................................................... 130
List of Tables

2.1 Illustrative IP Demand Data for One Week ........................................ 30
2.2 Performance of the Proposed Policy \((H = 65, L = 10)\) ..................... 31
2.3 Performance for the Policy \((H = 75, L = 0)\) ................................. 31
2.4 Simulation Results (mean with 95% confidence limits) ...................... 41
2.5 Effect of OT Limit on Capacity Usage ............................................. 44
2.6 Multiple Exam Lengths Comparison ................................................. 48
2.7 Stochastic Service Times ................................................................. 51

3.1 A Comparison of Three Scheduling Policies ...................................... 78
3.2 Optimal Approximation Policy with Increasing Rejection Costs ........... 78
3.3 Impact of the Overtime Policy ......................................................... 80
3.4 Capacity of 30 Scans per Day ........................................................... 82

5.1 The Impact of LIP on a Small Sized Hospital ................................... 114
5.2 A Larger Hospital ............................................................................ 115
5.3 Mirroring Chapter 2 Simulation ....................................................... 116
List of Figures

2.1 Daily Weekday IP Demand for VGH (May-December 2003) 36
2.2 Constrained Policy vs Current Practice (Simulation standard error of 0.01) 43
2.3 Varying IP Arrival Rate (Simulation standard error of 0.01) 46
2.4 Varying Capacity (Simulation standard error of 0.01) 47
2.5 Varying Percentage of OP On Call (Simulation standard error of 0.01) 49
3.1 Waiting Time With Rejection by Priority Class 76
3.2 Waiting Time With Overtime by Priority Class 81
3.3 Average Cost for a Variety of Demand Streams 83
5.1 Average Number of Daily OT Scans as a Function of Capacity 117
Acknowledgements

First and foremost, let me thank my supervisor Dr. Martin Puterman whose unwavering encouragement and insightful direction have truly been all that I could have asked for. I only hope to one day give the same level of supervision to my own graduate students. Secondly, the two other members of my committee, Dr. Maurice Queyranne and Dr. Michael Friedlander, have both given of their time to insure that this thesis did, in fact, reach completion. Dr. Queyranne was especially helpful in insuring that the proofs in chapter 4 were complete and coherent. Thirdly, the timely visit of Dr. Dan Adelman to Vancouver was pivotal in giving this research momentum. I am very much indebted both for his initial help and his continued advice along the way. Fourthly, it would be remiss not to mention Elaine Cho who has guided this administratively challenged student through the numerous pitfalls of a university bureaucracy - all with unwavering good will. I am sure at times she was tempted to roll her eyes at my incompetence! Finally, no PhD student can underestimate the importance of his fellow students. In particular let me thank Jeffrey Colpitts, Benny Mantin, Frederik Odegaard and Anita Parkinson who helped assure me that my own thesis struggles were not unusual!
Dedication

I dedicate this work to my wife, Jennifer, and my children, Rhys and Corwin, who have endured with great patience, providing encouragement to my sometimes lagging spirits, while I have struggled to complete this PhD. Thank you somehow doesn’t seem adequate! Nonetheless, thank you for your patience and your support and may God continue to direct our steps together. "In His will, our peace".
CHAPTER 1

Meeting the Waiting Time Challenge

1. The Challenge

A middle-aged man reports to his doctor complaining of mild but persistent headaches and some hearing loss in one ear. His doctor recommends he visit a specialist and books him for an appointment in one month. On seeing the specialist, an MRI is recommended but cannot be performed for two weeks. Upon analysis, it is shown that he has a tumour growing on the right side of his brain requiring surgery. Two weeks later, the surgery is performed but turns out to be significantly more difficult than expected due to the size of the tumour. The patient survives the operation but is now paralyzed on one side of his face and has lost all hearing in one ear. The question that keeps recurring to him is this - what would have been the outcome if the surgery had been performed two months earlier?

The above story is, of course, anecdotal but not unusual. Nor is the challenge of potentially damaging wait times unique to brain surgery. A British article reported that an elderly patient waited 28 hours for treatment at an emergency center [29] while in Saskatchewan, waiting time for a magnetic resonance imaging (MRI) scan reportedly reached 22 months [19]. A recent case in Canada concerned an eighteen month girl who over several weeks suffered severe pain, her limbs swelling to the point that she could not walk. A bone scan was deemed necessary but there remained a three week wait for the MRI. Her parents drove her out of province where she was finally diagnosed with leukemia [42].
On top of this growing list of tragic stories there are now numerous studies that demonstrate alarmingly long and growing waiting lists in a variety of OECD countries. Australia, Canada, Denmark, Finland, Ireland, Italy, Netherlands, New Zealand, Norway, Spain, Sweden and the UK have all reported significant waiting time challenges within their health care systems [51]. A 2002 report stated that 38% of patients wait at least four months for elective surgery in the UK, 27% in Canada, 26% in New Zealand and 23% in Australia [27]. A 1995 study showed that 89% of patients waited more than 3 months for cardiovascular procedures in the UK, 47% in Canada and 18% in Sweden [8]. The excessive nature of these waiting times is highlighted by a Canadian study that surveyed physicians across the provinces and across specialties. Each physician was asked to give both the length of time a patient should expect to wait for treatment as well as their own assessment as to what would be the maximum recommended waiting time. In 88% of the cases (each case representing a different specialty and province), the expected waiting time was significantly longer [20]. The same study suggested that overall average waiting time was 92% higher in 2004 than it was in 1993. In a separate study, 19% of patients who were surveyed after having visited a specialist reported having been adversely affected by the length of the waiting time [47]. Both these studies face the obvious criticism that they are based on surveys and not concrete waiting time data but they nonetheless highlight the fact that, in the opinion of both the professionals in the system and the patients using it, waiting times within Canada at least, have reach unacceptable levels.

The erosion of public confidence in the Canadian health care system is highly correlated with this apparent increase in waiting time over the last decade and a half. In 1991, 61% of Canadians ranked the health care system as "excellent" or
“very good”. By 1995 this number had dropped to 52% and by 1999 it was as low as 24% [46]. Even more recently, a 2006 Statistics Canada survey reported that waiting times were the number one health care barrier for those have difficulty gaining access to the health care system [21]. The importance placed on the expected waiting time for treatment as a measure of the well-being of our health care system was raised by Romanow in his influential report on the future of health care in Canada in 2002. In his estimation, “long waiting times are the main, and in many cases, the only reason some Canadians say they would be willing to pay for treatments outside of the public health care system” [45].

While there is almost universal agreement on the necessity of reducing these waiting times, there is little consensus as to the most appropriate means of doing so. One of the obstacles to coming to any agreement is that data on waiting times for hospital procedures are notoriously difficult to obtain and analyze. One challenge is the paucity of reliable data making it difficult even to know the extent of the current challenge. A second is that there is no agreed upon standard of what constitutes the official definition of “waiting time” making comparisons difficult. While the end date of the waiting time period is clear (the day of the treatment or diagnostic test), there are at least three potential candidates for the start date:

1. The initial visit to the family doctor,
2. The initial visit to the specialist who recommends treatment or,
3. The date the request for treatment is placed.

While it may be reasonable to assume that the latter two are not significantly different, there is plenty of evidence that the time between the initial visit to the family doctor and the visit to the specialist can differ substantially [20]. From the patient’s
perspective, total waiting time clearly includes the initial wait between the family doctor visit and the specialist visit. However, this data is rarely available for analysis and thus much of the reported waiting times actually represent the third option - from the day the request for treatment is placed to the day of treatment. Moreover, there is some logic to breaking total waiting time down into the two parts (family doctor to specialist and specialist to treatment) as the causes and potential responses to the waiting time challenge may differ between the two.

Irrespective of this disagreement as to the definition of waiting time, there are clearly three options available to anyone seeking to address the issue. One can seek to limit demand, increase capacity and/or improve efficiency. Which alternative is preferred depends very much upon the assumed causes for the existing waiting time challenge. If one believes that a large portion of the challenge is the result of an abuse of the system by the user then it is a reasonable strategy to attempt to eliminate this unnecessary demand. On the other hand, if one believes that it is simply a question of too much legitimate demand for too little capacity, then increasing current capacity becomes the only viable option. Finally, if one believes that the major cause behind excessive waiting times has to do with the mismanagement of existing resources then improving current capacity usage is paramount. Of course, there is no reason why a situation might not warrant the implementation of more than one, if not all, of the above options simultaneously as a health care system may be dealing with too much legitimate demand for current capacity while still facing potentially unnecessary demand and attempting to improve the efficiency of current capacity usage.

Incidentally, it is worth noting that zero waiting time is generally considered not to be optimal [48]. Queuing theory clearly demonstrates that to have zero waiting
Demand for hospital resources can vary dramatically from one day to the next. Thus, any system that seeks to meet all demand immediately must build enough capacity to satisfy peak demand levels. However, on days when demand is less than expected, this inevitably results in unused capacity. The existence of waiting patients allows the resource manager to improve capacity utilization through prior scheduling, smoothing out some of this variability in demand. On the other hand, a growing waiting list brings its own challenges. First, and most importantly, there is the potential for deterioration in the patient's health ([47], [51]). Secondly, there is an increase in resources being used simply to maintain the waiting list [54]. All this suggests that there are optimal queue lengths, depending on the severity of the condition being treated, that lie above zero but undoubtedly below the current levels in most countries with universal health care. Below is a very brief analysis of various responses to the waiting time challenge divided into the three broad classes of limiting demand, increasing capacity and improving efficiency.

2. Limiting Demand

There are at least three methods for reducing demand - one involving rationing demand and the other two seeking to eliminate unnecessary demand. The first involves imposing a financial burden on the user. The most obvious example is the United States where there is little to no waiting time problem for most medical services - provided you have the means to pay for it [51]. This is due to an excess of capacity resulting from the high cost to the user in the private system. Of course, the more affordable public versions of health care within the United States do experience the
usual waiting time challenges. Less radical methods impose a co-payment on the user that, while failing to cover the entirety of the cost, nevertheless acts as a disincentive to users.

One variation on this theme involves subsidizing users who seek care outside of the public health care system. Thus, rather than forcing users to pay a fee for service in the public system, this scheme seeks to minimize the financial barriers associated with seeking help in a co-existing private health system so as to relieve the stress on the public one. Australia has recently gone this route by agreeing to pay 30% of private insurance costs as well as providing a tax incentive for taking out a private insurance package [18].

Secondly, one can reduce demand by giving the professional experts the opportunity to sift out potentially unnecessary demand. New Zealand has taken this option by giving hospital doctors the ability to send patients back to the referring physician if they deem the request to be insufficiently urgent [35].

Finally, there is a whole area of demand reduction through health promotion and prevention. Much like improving efficiency, there is little argument raised against the idea of prevention as a means of demand reduction. The debate comes simply in terms of its potential impact. Here there is legitimate disagreement as it is extremely difficult to assess what demand might have been had a method of prevention not been put into effect. Thus, assessing prevention methods depends very much on subjective assumptions [24].

2.1. Demand Rationing Through Pricing. Turning first then to demand rationing through co-payments, Siciliani and Hurst [51] demonstrated that there is a significant correlation between a country not reporting significant waiting times for
hospital procedures and the presence of some form of co-payment. However, they suggest that the co-payments are often so low as to bring into question whether price rationing is really playing a major role.

Two options are available to those seeking to impose a financial constraint. The first is to develop a secondary (invariably private) health system that requires the user to pay a significant cost while still maintaining a public system open to all. England is the most obvious example. A second method is to impose a more nominal co-payment for all users and maintain a single health system. Sweden and Ireland both fall under this category [51].

Critics of the first method are not hard to find. It is pointed out that while such methods may decrease the mean waiting time by allowing some patients to receive treatment much more promptly, it does so only at the expense of increasing the length of time to treatment for those who remain in the public system. The rationale is that though a private health care system will remove some demand from the public system, it will also inevitably siphon off doctors and nurses as well. Even if, as in the UK, doctors are required to work part-time in the public system in order to maintain their license, there still remains the fact that a certain portion of their working hours has been removed from the public system. If the total pool of resources (doctors and nurses) remains the same when broken into two separate systems of private and public health care then it is self-evident that any decrease in the waiting time for those in the private system must necessarily come at the expense of an increase in the waiting time for those in the public system. Queuing theory has long known that dividing a single queue, originally serviced by a single pool of servers, into multiple
queues with dedicated servers for each queue will inevitably increase waiting times provided the number of servers (and the average service rate) remains the same [58].

Two potential refutations of this line of reasoning are one, that the introduction of private health care increases the number of doctors and nurses by attracting more people to these fields and two, that a private system will increase throughput without having to increase resources. It would seem reasonable therefore to explore whether either of these two claims have in fact proved true in past instances where private health care has been introduced and whether the potential increase in throughput is sufficient to mitigate the negatives associated with dividing one queue into two. Only then would we have a true grasp on whether private health care can in fact help reduce waiting times for the majority of patients.

Finally, there is good reason for believing that the introduction of a concurrent private system may actually increase per capita spending on health care. The United States per capita spending on health care is significantly higher than any other OECD country [51]. Australia’s health care spending has increased from 8.5% of the GDP to 9.5%, perhaps due to the subsidies given to those taking the private health care option [18].

Imposing a more nominal co-payment fee on all users is most obviously open to the criticism that, no matter how nominal the fee, it necessarily marginalizes the poor. And herein lies the real challenge to these methods of demand reduction - is it legitimate to ration demand for health care services through pricing? If we accept the premise that health care should be available to all and if the aim, therefore, is to remove unnecessary demand then, unless we have reason to believe that it is the poor who are imposing this unnecessary burden on the system, surely pricing is a poor
method of rationing. Is it not possible to determine a better means of identifying and rejecting unnecessary demand - if it indeed exists?

2.2. Demand Rationing Through Prioritization. This is precisely what New Zealand has attempted to do in taking a somewhat different route to demand reduction. In 2001, New Zealand introduced a booking system relying on a priority score assigned to patients, coupled with a maximum six-month waiting time guarantee. Each hospital sets a priority threshold that corresponds to the level at which they can maintain a maximum six month wait. Any patient who receives a priority score below that threshold is referred back for clinical review (essentially delaying their surgery) and re-enters the system at a later date (eventually receiving a score that allows them to be booked). Patients who score below a second threshold are referred back to their family doctor [35]. While this does mean that no patient waits more than six months from the time of booking to the time of surgery, it does so only at the expense of leaving a portion of the demand in limbo. Thus, actual waiting times may significantly exceed the six-month limit. Again, such a policy only makes sense if one believes that there is a portion of the demand that is essentially unnecessary and thus can reasonably be put off without incurring a cost to patient health. The advantage of the New Zealand solution is that it places the decision of whether a medical procedure is unnecessary in the hands of the medical experts whereas a form of co-payment has the effect of placing that decision in the hands of the patient prior to consultation with a physician. The disadvantage is that even patients who are deemed not to need treatment still need to be prioritized and thus such a policy incurs a significant administrative cost that would not be incurred with a co-payment plan. Nor is a priority system free from controversy as doctors will inevitably disagree
on what is and what is not necessary. However, prioritizing patients is unavoidable so it is difficult to see how this subjectivity can be avoided by any system.

It seems reasonable to suppose that, in a system where the user pays nothing, there is potentially a tendency to overuse the service. However, it is also clear that placing too great a roadblock can often lead potential patients to ignore symptoms until they reach a certain level of severity. At that point, it may very well be more costly to treat than it would have been initially. Finding an intelligent means of removing unnecessary demand while at the same time not deterring legitimate demand is a delicate and complicated task that may in fact turn out to be the illusive pot at the end of the rainbow.

2.3. Demand Rationing Through Prevention. Finally, there is the option of reducing demand through prevention as opposed to rationing. Is it possible to find means of reducing the demand for a resource by providing other, less costly, alternatives? This option has been most extensively explored in the case of emergency departments. Some examples include ambulatory services that allow some patients to be treated without having to be seen at the hospital, short stay units and fast track systems [9]. There is, of course, little disagreement over whether or not one should reduce demand through prevention if possible. However there is plenty of disagreement over what methods are appropriate - meaning that they are effective without harming patient health.

3. Increasing Capacity

In most countries with universal health care attempts to limit demand are highly contentious and thus the focus often shifts to increasing throughput - either through
increased capacity and/or improved efficiency. Political pressure in Spain has led to a number of initiatives designed to increase capacity. From 1996 to 1998, Spain managed to reduce mean waiting time (from request to treatment) for surgery by five months; largely as the result of increased day surgery and improved hospital management. The size of the waiting list shrunk from 190,000 to 132,221. However, by 2000, about a third of that improvement had been lost putting the pressure back on the government. Initiatives were put in place to provide additional surgery time by having surgeries performed during the evening. As a last resort, the government promised to pay for private sector treatment for any patient whose waiting time had become excessive ([10], [11]). One of the arguments against increasing capacity has always been that it will lead to a resulting increase in demand. It is difficult to know whether this was the case in Spain or whether it was simply that hospitals reverted to old practices, leading to a rebound in waiting times.

The international experience of attempting to improve waiting time through increased capacity has not generally been very positive. Ireland [38], the Netherlands [50], the UK [29] and Canada [39] have all earmarked significant funds (ranging from $100 million to $4 billion) to initiatives designed to reduce waiting times but with little evidence of success (though, to be fair, for some it may be too early to judge). Whether this is a necessary outcome of attempts to increase capacity or simply the result of poor choices on the part of each country is much too detailed a question for this introductory survey. It does seem to demonstrate, however, that simply increasing funding does not necessarily lead to reduced waiting times. In fact, at least two studies have documented an increase in waiting time following an increase in
capacity. In Manitoba, the volume of cataract procedures increased considerably between 1992 and 1997. However, over the same period, the median waiting time also increased. A second study on surgical services in Britain found that as the number of hospital admissions increased, so too did the length of the waiting list [46]. The most common explanation for this rather surprising result is that demand is proportional to expected waiting time. Reduce the expected waiting time and demand increases bringing waiting times back up to what they originally were or even higher. The implication is that waiting times act as a form of rationing already deterring some demand or diverting it elsewhere. A specialist may, for instance, determine that an MRI scan would be helpful for diagnosis but that, due to the length of wait, it might be easier simply to treat and see whether the patient improves. Were waiting times less lengthy, he might choose instead to confirm the diagnosis first.

4. Improved Efficiency

While every resource manager would love to have greater capacity, the reality is that there are limits to how much capacity can be made available. Given the already high financial burden associated with running a public health care system, it is perhaps not surprising that the major policy for tackling the waiting time challenge is improving the use of available resources. Street and Duckett present the argument that a large factor in current waiting times is the result of poor funding methods rather than too much demand for too little capacity. “If governments exhibit a positive willingness to pay in reaction to waiting list increases and hospitals give greater weight to increased admissions than reducing waiting times, then waiting times will grow beyond what is optimal” [54]. Thus, this argument would suggest that current
funding policies give incentives to hospitals to maintain long waiting lists precisely to secure more funding. Implicit in this argument is the belief that there is the potential for improving efficiency if only the funding scheme gave hospitals the incentive to do so.

In Australia, attempts have been made to remove such negative incentives. The first measure was to link hospital funding to activity rather than to the size of the waiting list. Each treated case at a hospital attracts payment according to the patient's diagnostic type. Thus hospitals are rewarded for increased throughput but are not (if the weighting has been done correctly) penalized for taking on long-term patients. Approximately 50% of all hospital funding in Australia is now linked to the number and type of patients treated. Additional funding was also made available to those hospitals that reduced their waiting list. Specifically, hospitals were not permitted to claim this additional funding unless they had reduced their priority one patients' waiting times to acceptable levels by December 1993. Priority two patients' waiting times were to be within acceptable limits by April 1994. The results were dramatic for the top two priority levels with little change in the lower priority levels [54]. This would seem to suggest that the implicit assumption that there is room for improved efficiency given the right incentive is justified. However, it is possible that some demand was going unmet in order to keep waiting times at the levels required for the additional funding. The more recent Australian study mentioned earlier [18] also suggests that despite the initially good results, waiting times have not remained within "acceptable limits". The danger here is of a "Catch 22" situation where one can only receive additional funding if the queue size is reduced but one can only reduce the queue size if given additional funding. This is especially the case if it is
true that any reduction in waiting times leads to an increase in demand. Moreover, though removing the incentive for long waiting lists, the Australian funding scheme gives a decided incentive towards rushing patients through the system. Might this not be a case of substituting one poison for another?

A second method of improving efficiency, applying primarily to emergency departments, is to create a separate stream for patients whose treatment time is relatively short. A study in the UK looked at the implementation of a separate stream policy at an emergency department. They collected data for the five weeks prior to the introduction of the policy change as well as for the five weeks after and demonstrated significant improvements in the percentage of patients whose waiting time was less than an hour. These findings were duplicated in a Canadian study published in 2006 [9]. Again, queuing theory would have predicted as much. Intuitively, one can imagine the benefit of such a system by considering two people in a queue - one with a long service time and one with a short service time. It is clear that average waiting time is minimized by allowing the shorter service time patient to go first. This is essentially what a fast-track separate stream does. Such a policy would seem reasonable provided the separate stream was kept constantly busy. Queuing theory can easily determine how much slack in the separate stream would be necessary in order to negate the potential benefit.

Finally, a number of countries (including Sweden, New Zealand, Canada and the UK) have sought to improve efficiency by imposing maximum waiting time limits in an effort to force hospitals into greater efficiency ([26], [35], [59]). Again, the implicit assumption is that there is room for improved throughput if the appropriate incentive is provided. The Australian answer was to dangle the proverbial “carrot”
of extra funding. These countries chose instead to wield the “stick” of imposed waiting time guarantees. Sweden is perhaps a good model of the consequences. In 1992, Sweden introduced a maximum waiting time guarantee of 12 weeks for 12 key procedures. In the first year, an additional $70 million was set aside to help hospitals meet those targets but after that they were required to maintain the same targets without additional funding. At the beginning of the initiative there was 51 000 patients in the queue. By the end of the first year, this was reduced by 21%. Those waiting more than 13 weeks went from being 50% of patients to 10% of patients. At the same time, throughput increased from 161 300 to 180 000. Interestingly, waiting times began to improve as soon as the decision to introduce maximum limits was introduced but before the additional funding was provided. However, during the second year waiting times once again began to rise [26]. What was not reported, and what is surely of vital importance to the analysis, is whether or not the improvement in the 12 key procedures was achieved at the expense of other procedures not subject to the maximum waiting time guarantee. For instance, one study found that imposing a two week rule on the maximum waiting time for urgent cancer patients resulted in a doubling of the waiting time for routine cancer referrals [59]. “A tank of water under pressure, if plugged in one area will simply explode in another”. Likewise, unless the pressure to the health system is relieved (i.e. capacity made sufficient to meet demand), it is difficult to imagine maximum waiting time guarantees doing anything other than transferring the problem from one class of patients to another.

Is there enough inefficiency in the system for improved efficiency to relieve the pressure? That question can surely only be answered on a case by case basis. Certainly, it seems reasonable that, before increasing capacity, we seek to use the capacity
that we already have as efficiently as possible. As mentioned earlier, one of the major barriers to doing so has persistently been the lack of consistent, reliable data. Most medical databases are designed primarily for diagnosis and treatment and not policy. Thus, the data necessary for monitoring the efficiency of the system is often lacking. As one Canadian report states,

The waiting list “nonsystem” in Canada is a classic case of forced decision-making in the absence of good management information. There is a surfeit of non-standardized data and a dearth of usable policy-oriented information about waiting lists. The most serious consequence is that information and management defects are almost always prematurely diagnosed as financial shortages [33].

Most governments have realized this and are taking steps towards rectifying the lack of data ([36], [52]). In the mean time, it is difficult in the extreme to determine whether there is sufficient capacity, if managed efficiently, to meet demand or whether any initiative to improve efficiency must be coupled with increased capacity. Nevertheless, it stands to reason that any improvement in the efficient use of current resources can only help to mitigate the negative impact of the waiting time challenge.

5. The Impact of Scheduling on Waiting Times

It is not the intention, in this introductory chapter, to give an exhaustive survey of all that has been done to maintain reasonable waiting times. Rather, we seek to give the reader a sense of the urgency of the waiting time challenge as well as the incredible variety of possible responses. This urgency is heightened by the well-known fact that, for most OECD countries, the proportion of the total population above 65,
who are the largest users of the health care system, is increasing rapidly. The research presented in this thesis necessarily focuses much more narrowly on a specific aspect of the waiting time challenge - namely the optimal scheduling of patients to a hospital resource.

Scheduling for a hospital resource is most often complicated by the fact that demand for the resource comes from numerous sources with varying degrees of urgency. Since the lower the urgency of a patient, the further in advance treatment is booked, it is unavoidable that, for any given day, it is the lower priority patients who are booked first. Booking too many lower priority patients into a day may result in there being insufficient capacity left to meet the later arriving higher priority demand. Booking too few lower priority patients into a day may result in unused capacity which might otherwise have been used to serve these lower priority patients sooner. Thus, either way, sub-optimal scheduling policies may lead to excessive waiting times for certain priority classes. There are two angles to attacking the scheduling challenge. First, there is the issue of how to optimally schedule patients given there is a fixed available capacity. Chapter 2 takes this approach placing a limit on available capacity and then seeking to determine the effect on patient waiting time for the various priority classes when an optimal scheduling policy is put into practice. Secondly, and perhaps more importantly, there is the issue of determining how much capacity you need in order to meet the waiting time limits recommended for each patient priority class. Chapter 3 takes this approach presenting policies that seek to meet the waiting time limits with minimum capacity. Of course, how much capacity is needed depends heavily on how efficiently one uses the capacity available. For the purposes of meeting the scheduling challenge we will assume that there is a fixed number of patients who can be served in
a given day with some additional capacity available through overtime. We leave aside
the issue as to whether there might be efficiencies to be gained through procedural
improvements so that more patients might be served during regular hours.

6. Diagnostic Imaging

To be more concrete, the research focuses particularly at the scheduling of patients
for a diagnostic resource. This resource faces precisely the challenge mentioned above
as demand comes from three sources. Of highest priority is demand coming from
the emergency department that must be met immediately. Second, there is a highly
variable demand from the inpatient (IP) population within the hospital that must
either be served today or, in some cases, by tomorrow at the latest (though current
practice seems to serve all IP demand the day the request is received). The third
source consists of outpatient (OP) demand. This demand arrives in the form of faxed
requisitions coming from specialist physicians in the region serviced by the hospital.
Upon arrival, these requisitions are given to a radiologist who breaks the total OP
demand into numerous priority classes - each one with a maximum recommended
waiting time assigned to it. The requisitions are then returned to the booking clerk
who is responsible for booking these OP days or even weeks in advance. Thus,
diagnostic resources are a classic example where a scheduling policy is crucial as, for
any given day, lower priority demand must be booked prior to knowing the extent of
the higher priority demand that may arrive between now and the day in question.
Though the mathematical models developed in the following chapters will focus on
diagnostic resources, it is hoped that the above description of the scheduling challenge
makes it clear that the methods and insights developed here are applicable to any
hospital resource (such as surgery or cancer treatment) facing demand from multiple priority classes.

It is not difficult to show that diagnostic imaging faces, and will continue to face, the challenge of excessive waiting times. According to the World Health Organization, “about 30% of medical cases around the world require diagnostic imaging in order for referring physicians to make a proper diagnosis” [53]. This number is most likely to increase as more and more potential uses for CT (computed tomography) and MRI scans are discovered. In the last decade alone, the number of MRI scans performed in Canada has quadrupled [37] so that currently 35 million diagnostic scans are performed every year with an annual increase of 6% [53]. Two Canadian studies mentioned earlier ([20], [47] both reported a significant portion (more than 50% in the one, just under 50% in the other) of patients who waited more than a month for a CT scan. Considering the maximum recommended waiting time for the lowest priority class in both Alberta and British Columbia is 28 days, this would suggest that waiting times are in fact excessive. The reasons behind this massive increase in the use of diagnostic imaging are numerous and better documented elsewhere [31] but the salient fact is that the demand continues to outstrip the capacity exacerbating the need to use the available capacity as efficiently as possible. Answering the scheduling challenge is one necessary step towards doing so.

7. Related Literature in Operations Research

Our research, using mathematical models to manage queues, falls under the general operations research field of queueing theory. A helpful overview of research within
queueing theory applied directly to health care is provided by Preater from Keele University [40]. The specific issue of servicing demand from multiple priority classes has received extensive coverage within operations research. The two main applications have been airline revenue management ([6], [12] and [60]) and call center management ([4], [14] and [30]). Though useful for comparison, there are significant differences between these applications and the scheduling challenge within health care.

The airline revenue management has the advantage of generally looking at a small number of flights over a finite horizon. In our situation, each potential booking day could be viewed as a flight and, though the booking horizon (the maximum number of days in advance patients are booked) is finite, it is also continuously evolving, leading to an infinite, rolling horizon problem. Moreover, the objective in airline revenue management is, understandably, to maximize revenue which allows for a different formulation than for the case where the objective is to minimize excessive waiting times. Most notably, minimizing excessive waiting times forces one to take into consideration the whole horizon rather than truncating the horizon to a finite setting.

The call center management papers deal with similar issues of minimizing waiting times with limited capacity so that it is perhaps here that the comparisons are potentially most useful. Here too the trade-off is between service level (often given in terms of waiting time guarantees) and cost. Moreover, if there are multiple priority categories, they are generally differentiated according to a maximum waiting time guarantee that necessitates some kind of reservation policy for later arriving higher priority clients. However, there are some significant differences. For instance, in call center management, clients are not booked days in advance but held in a queue.
Essentially, this means that call centers are equivalent to hospitals that serve only inpatients as all "clients" are there and can be served immediately if space becomes available. Secondly, call centers do not deal with a booking horizon as all clients are either serviced within a specified time frame or lost to the system. Thus, each day the call center begins with an empty queue. This allows for a finite horizon model as opposed to the rolling, infinite horizon model necessary in the hospital setting. Nonetheless, there are clearly useful comparisons to be made.

When turning directly to health care applications, the scope of research narrows considerably. Two distinct branches of the literature emerge. In the first, the trade-off modeled is between unused capacity (resulting in longer waiting times) that results from under-scheduling a resource and excessive overtime that results from over-scheduling that resource. Gerchak, Gupta and Henig model this trade-off in the case of elective surgery scheduling [22]. The challenge is to schedule elective surgeries while leaving room for possible emergency surgeries. Strum et al. develop a minimal cost analysis model based on over-utilization and under-utilization applied to OR blocked time for subspecialties ([55], [56]). Liu and Liu analyse an OP clinic with multiple doctors with the complication that doctor arrival times are stochastic [34] and Ridge et al. model the effect of balancing the cost of unused beds in the ICU against the cost of insufficient capacity [43]. Finally, Rohleder and Klassen present a simulation based model that compares various potential methods for reducing waiting lists in the presence of uncertain demand with a fixed number of slots reserved for emergencies [44].

A second branch of research looks to choose the optimal schedule in order to maximize revenue ([23], [25]). Such models are only reasonable in the private health
care setting. For a resource manager in public health care, it is not revenue but accessibility and efficiency that are the objectives. Excessive waiting times address both these objectives. Both papers mentioned here present finite horizon models that do not capture the effect of the scheduling policy on the waiting times of the patients. As far as we know, no paper has modeled how to schedule more than two priority classes to a given resource in such a way as to minimize excessive waiting times.

Chapter 2 of the thesis mirrors a situation very similar to Gerchak, Gupta and Henig as we assume only one OP priority class and examine the optimal reservation policy for IP and emergency demand. Our analysis differs in that we directly examine the impact of the scheduling policy on OP waiting times as well as demonstrating the significant advantage of introducing some flexibility into the scheduling of the highest priority class. The end result is an optimal reservation policy for IP and emergency demand under a strict limit on expected overtime per day.

Chapter 3 turns the attention towards booking the remaining capacity amongst the various OP priority classes. The extension to more than two priority classes scheduled over a longer period of time makes for a much more involved scheduling challenge. Chapter 3 models the scheduling process as a Markov Decision process (MDP). However, the state space for the MDP is much too large for a direct solution so we derive an approximately optimal scheduling policy through approximate dynamic programming (ADP). ADP is an emerging and highly promising field of research. The two main textbooks in the area are *Neuro-Dynamic Programming* [5] and *Reinforcement Learning* [57]. More recent insights into the field have been provided by Adelman ([1], [2] and [3]) and by De Farias and Van Roy ([15], [16] and [17]). Both textbooks provide an array of potential approximation methods that will be
discussed briefly in Chapter 3. Our own work follows that of Adelman and De Farias and Van Roy in transforming the MDP into the equivalent linear program (LP) and then solving the LP by approximating the value function in the objective.

Chapter 4 of the thesis provides some strong theoretical results of the model presented in Chapter 3. We first establish analytically the form of the optimal linear approximation to the value function. As far as we know, this result is unique in the literature. The significance of these results is that it allows us, under mild constraints on the cost parameters, to derive the optimal value function approximation without having to solve the LP and, moreover, demonstrates the robustness of our results to changes in parameter values. Secondly, we present a discussion of potential bounds on the difference between what might have been achieved could we have solved the original MDP and the results we did in fact achieve through approximation. This remains a wide open area of research as practical bounds on this difference are not readily available.

Chapter 5 then seeks to compare the models presented in Chapters 2 and 3 suggesting some practical policy insights for the resource manager in terms of allocating available capacity to the various sources of demand.

Finally, Chapter 6 discusses some potential enhancements to the model as well as discussing the challenges that may arise in applying the methods and results presented in this thesis to other potential hospital resources outside of diagnostic imaging.

It is gratifying to note that the recent report by the Canadian federal advisor on Wait Times specifically encouraged the use of queueing theory to address the waiting time challenge [39]. We trust that the work presented in this thesis will help convince those working within the health care sector that operations research, and particularly
queueing theory and optimization, can play a significant role in improving resource utilization within hospitals.
CHAPTER 2

Improving Resource Utilization through Inpatient Scheduling

1. Introduction

Throughout the health system, both within Canada and elsewhere, excessive waiting times are severely hampering our ability to provide adequate health care. While there can be little argument that the primary reasons for excessive waiting times are limited resources and increasing demands, there are secondary factors which impede the efficient use of existing resources. Among these secondary factors, perhaps none are more important than the variability in the demand. One day a resource is stretched beyond its capacity, the next it stands idle for large portions of the day. We will address the challenge of excessive waiting times by developing an approach to improve resource utilization and then demonstrate via simulation the resultant effect on the growth rate of outpatient (OP) waiting time. We will concentrate on the growth rate in OP waiting time as the most relevant statistic since the issue of waiting time is of most import in those instances where demand exceeds capacity and therefore where queues are long and growing.

2. The Problem

At the most general level, this chapter addresses a situation where there is a fluctuating demand for a limited resource with two priority classes. Such a scenario occurs for instance in scheduling patients for operating room time (e.g. [22], [55] and [28]), or else booking MRI (magnetic resonance imaging) or CT (Computed...
Tomography) scans. The key challenge is that the low priority demand must be booked prior to knowing the high priority demand. Therefore, a significant portion of the total capacity must be reserved for this unknown high priority demand - leading inevitably to unused capacity on those days when the high priority demand is lower than expected. We develop a policy that reserves a certain amount of resource time to each priority level. Gerchak, Gupta and Henig demonstrated that such a cut-off policy (in the case of booking elective and emergency surgeries) is potentially sub-optimal but they do so under the assumption that there are no fixed limits on overtime [22]. In reality, most hospitals do work with a limited overtime budget. Gerchak et al. also acknowledge that the relative difference between the optimal policy and the cut-off policy is minor. Their main contention is that optimal cut-off policies are computationally difficult to determine. However, in the case of limited overtime there is, as we will show, a very simple means of determining an optimal cut-off policy. The potential for improving the ability to achieve reasonable waiting times for outpatients by following Gerchak, Gupta and Henig's approach of allowing a varying overtime limit will be explored in chapter 3 and chapter 5.

The motivation for this paper grew out of a detailed analysis of CT operations at the Vancouver General Hospital (VGH) where there is a large and fluctuating inpatient (IP) demand and three OP priority classes. OP priority levels are broken down according to the maximum recommended waiting time. What makes this a scarce resource problem is that waiting times for OP are, at present, significantly longer than recommended. One long-term goal of the project is to provide a reservation policy that will cover all priority levels. This larger model is addressed in chapters 3
and 5. For the purposes of this chapter, however, we will assume only one OP category. All IP are currently viewed as high priority - meaning that the demand must be met the day it arrives. Since OP demand is booked weeks in advance, it is clear that, for any given day, the low priority demand (OP) is booked prior to knowing the high priority demand (IP) for that day. In order not to overtax the system on days when IP demand is high, the department is forced to reserve a significant portion of scanning time for this unknown IP demand, leaving little room for OP. This results in unused capacity on days when IP demand is low and thus longer waiting times for OP than might be the case if this unused capacity could be utilized.

3. The Proposed Scheduling Approach

Our attempt to make use of the unused capacity in the system hinges on the fact that, contrary to current practice, not all IP need to be scanned the day the request for a scan is placed. Our proposal is as follows:

- Divide IP into two categories - high priority inpatients (HIP) who must be scanned the day the request is placed and low priority inpatients (LIP) whose scans can be delayed one day.
- Identify a group of OP who will be on call for scans with one-day notice.
- Reserve a specified number of slots for IP on each day.
- Reserve a specified number of slots for possible carry over of LIP to the next day.
- Limit overtime (OT) usage.

How to make this operational will be described in the remainder of this chapter. A division of inpatients as suggested above already exists at Vancouver General Hospital.
The prioritization scheme recently put in place contains a category for IP that can wait up to twenty-four hours. This group contains pre-operative investigation for head and neck, routine follow-up for intracranial disease, pre-operative oncology staging, numerous fractures and some chest infections. However, this distinction is currently being ignored since all IP are scanned the day the request arrives regardless of the priority classification.

The advantage of such a division is that there is a pool of patients who are available to be scanned on the current day but do not necessarily need to be. One might liken LIP to stand-by passengers on flights. Just as stand-by passengers provide the airlines with the ability to fill last minute unused capacity, so LIP would provide the hospital with the same flexibility. The major difference is that LIP must be scanned within the next 24 hours. As mentioned in chapter 1, there is potentially some useful crossover here as substantial research has gone into developing reservation policies for numerous priority levels within the airline industry. However, the airline industry is generally dealing with a small number of flights at a time over a finite horizon whereas a CT department is booking over an infinite rolling horizon where each day could be viewed as a separate flight.

We assume that the resource units are 15 minute scanning slots since the vast majority of scans are scheduled for 15, 30, 45 or 60 minutes. Thus a 60 minute long scan would equate to four units of resource time. The capacity of the system is the number of 15 minute scanning slots that can be scheduled for any given day. Initially, we will assume that all scans are scheduled for 15 minutes and later investigate, through simulation, the effect of introducing multiple exam lengths. Of course, actual exam lengths are stochastic and vary quite substantially from the scheduled length.
(see [56] for a comparison of normal vs. log normal exam lengths). However, from a booking point of view, it is the scheduled length of the exam rather than the actual length that is relevant. We will discuss the effect of introducing stochastic exam lengths later in the chapter.

We assume there are a total of $C$ regular hour slots available per day. Our proposed scheduling policy works as follows. Each day, $H$ slots are reserved for IP and an additional $L$ slots for any LIP carried over from the previous day. The remaining $C - H - L$ slots are available for booking OP demand. Note that the trade-off is clear - the more slots reserved for IP $(H + L)$, the less scanning time is available for OP resulting in longer OP waiting times. On any given day perform all HIP scans and as many LIP scans as possible in the $H$ reserved slots. At the end of the day, any additional LIP scans not performed are booked into the $L$ slots reserved for tomorrow. If any of the $L$ slots are not filled, then on call OP from the waiting list are called to fill the remaining slots - reducing the total number of reserved slots for tomorrow's IP to $H$. In other words, looking forward days in advance, the policy reserves a total of $H + L$ slots for IP but on the day before, it reduces this to $H$ reserved slots by booking any remaining LIP scans and/or additional OP scans into tomorrow's $L$ slots. (We impose the further restriction that no more than $L$ LIP scans are carried over to the following day in order to avoid any snowball effect. If more LIP scans arrive on any given day than can be dealt with through the $H$ slots reserved for IP today and the $L$ slots reserved for carry-over to tomorrow, then the excess is dealt with through overtime.)

A short numerical example further illustrates the proposed reservation policy. Table 2.1 presents hypothetical daily IP demand for a given week. (Note that although
the demand for the whole week is given, each day’s demand is not known until the day it arrives.)

Assume that the number of IP reserved slots, $H$, is set at 65 and the number of carryover reserved slots, $L$, is set at 10. On day 1, all 60 HIP and 5 LIP are scanned while the other 5 LIP are carried over to day 2. No overtime is required and there are no empty slots. Since only 5 of the 10 carryover slots are used by LIP, an additional 5 on-call OP can be scheduled for day 2. This reduces the total number of reserved slots on day 2 to $H$. On day 2 demand is low so all HIP and LIP patients can be accommodated ($45 + 12 < H$) without requiring overtime or holding any LIP over to day 3. However, there are 8 (65 - 57) unused slots. Since no LIP are held over, an additional 10 on-call OP can be booked for day 3. On day 3, total IP demand spikes and 10 overtime scans are required just to meet the HIP demand. In addition, since no more than 10 LIP can be held over from one day to the next, 5 of the third day’s LIP must be scanned using overtime with the other 10 being carried over to day 4. No additional OP can be booked for day 4 as all $L$ slots are filled with LIP. For the week, the number of unused slots, the overtime required, the number of LIP carried over and the number of additional OP booked are given in Table 2.2. The units again are the number of 15 minute scanning slots.

### Table 2.1. Illustrative IP Demand Data for One Week

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIP demand</td>
<td>60</td>
<td>45</td>
<td>75</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>LIP demand</td>
<td>10</td>
<td>12</td>
<td>15</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Total demand</td>
<td>70</td>
<td>57</td>
<td>90</td>
<td>68</td>
<td>67</td>
</tr>
</tbody>
</table>
### TABLE 2.2. Performance of the Proposed Policy ($H = 65, L = 10$)

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unused Capacity</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Overtime Scans</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>LIP held over to next day</td>
<td>5</td>
<td>0</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Additional OP booked</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>30</td>
</tr>
</tbody>
</table>

### TABLE 2.3. Performance for the Policy ($H = 75, L = 0$)

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unused Capacity</td>
<td>5</td>
<td>18</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>38</td>
</tr>
<tr>
<td>Overtime Scans</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Additional OP Booked</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Contrasting this with the policy that simply reserves 75 slots for IP (Table 2.3) gives an indication (despite the fictitious nature of the data) of the potential increase in resource utilization gained by the introduction of LIP. In particular, under the proposed policy, we have been able to fill 30 of the 38 unused slots that arose under the current policy. Note that this is done without increasing overtime.

### 4. Is The Proposed System Practical?

It is worth, at this point, pausing to consider the practicality of implementing the proposed policy. First, while it is known that several patients fall into the LIP category, no data is currently available to determine the proportion of the total IP demand that would fit such a classification. What we will show however is that the number of patients in the LIP category need not be a large percentage of the total IP demand in order to have a significant impact on OP waiting time. Second, there
is the practical issue of having a pool of on-call OP able to respond to a single day’s notice. Again, we will show that the proportion of OP willing to be on call need not be a large percentage of the total OP demand. It seems reasonable, at least for a reasonably large city, to assume the existence of a pool of OP within easy distance of the hospital who would gladly be on call in order to shorten their waiting time. Gerchak, Gupta and Henig assume, not unreasonably, that all OP can be called in to receive their surgery earlier than initially scheduled [22].

Third, there is a question of increased workload on the department staff. However, the only additional work placed on the staff is that of calling the on-call OP the night before. In our opinion, this added load is more than justified by the benefits of improved resource utilization. (In fact, one recommendation of the VGH study was to call OP the night before their scan in order to minimize no-shows so, in reality, there would be no additional work.) Fourth, there is a legitimate concern that such a system would increase costs due to increased IP hospitalization time. However, judicial use of the LIP category (i.e. not classifying as LIP those whose hospital stay might be shortened by a prompt scan) should alleviate this problem. Finally, we will show that such a system does not, as might be supposed, increase overtime.

The initial feedback that we have received both from hospital administration and the staff within the CT department has been favorable. Their major concern regards the size of the LIP category. No one could give us much insight as to what numbers to expect though they were cautiously optimistic that our proposed scheme could work - perhaps not at Vancouver General which is a tertiary care hospital, but at a community hospital where patients spend extended periods of time in hospital.
5. Choosing L and H

Of course, the obvious question concerns the choice of $L$ and $H$. From a practical standpoint, the most likely scenario is that a hospital will have a set overtime budget that should not be exceeded. Thus, one possible approach is to choose $L$ and $H$ to minimize the expected number of empty slots subject to a constraint on OT. The papers referred to above do not put a limit on OT but rather assign a penalty to OT and then seek a policy to minimize cost. This is mathematically much more difficult and the benefit is dubious. Assigning a "cost" to unused capacity is complex and thus one ends up adjusting the parameters to see what works regardless.

Our approach is to impose an OT limit and then adjust to see the effect. From a practical standpoint this is reasonable as hospitals do have a budget that they are trying not to exceed. The above scenario generates a simple optimization problem that can be solved for any given breakdown of IP demand into HIP and LIP. For instance, if we assume the daily arrival rates for HIP and LIP follow independent, stationary Poisson distributions (see later discussion for validation) with means $\lambda_H$ and $\lambda_L$ respectively then the optimization problem is to minimize

\begin{equation}
\text{Expected Number of Unused Slots} = \sum_{k=0}^{H} k \times \Pr(D_{HIP} + D_{LIP} = H - k)
\end{equation}

subject to an OT constraint. The random variables $D_{HIP}$ and $D_{LIP}$ are the demand for HIP and LIP respectively. Since they are independent, their sum is Poisson with parameter $\lambda_H + \lambda_L$. The above equation follows from the fact that there is only unused capacity if the sum of all IP demand is less than $H$. The OT constraint can
take one of two forms - a limit \( \alpha \) on the probability of OT or else a limit \( \beta \) on the expected number of units of OT. These lead to two optimization problems

\[
\text{minimize: Expected Number of Unused Slots} \\
\text{subject to: } \Pr(\text{Overtime}) \leq \alpha
\]

and

\[
\text{minimize: Expected Number of Unused Slots} \\
\text{subject to: } \mathbb{E}[\text{Overtime}] \leq \beta.
\]

The probability of overtime can be written as:

\[
(2.2) \quad \Pr(D_{HIP} > H) + \sum_{j=0}^{H} \Pr(D_{HIP} + D_{LIP} > H + L | D_{HIP} = j) \cdot \Pr(D_{HIP} = j)
\]

\[
= \sum_{j=H+1}^{\infty} \frac{(\lambda_H) e^{-\lambda_H}}{j!} + \left( \sum_{j=0}^{H} \sum_{i=H+L-j+1}^{\infty} \frac{(\lambda_L) e^{-\lambda_L}}{i!} \cdot \frac{(\lambda_H)^i e^{-\lambda_H}}{j!} \right).
\]

Note that the \( H \) in the above equation refers to today’s \( H \) reserved slots and the \( L \) refers to tomorrow’s \( L \) slots for carryover LIP demand, as it is these two numbers that determine today’s OT. This also holds true in the following equations. Computing expected overtime is more complicated as it depends on whether the OT is due to HIP, LIP or both. Thus, the expected amount of OT is given by

\[
(2.3) \quad \sum_{k=2}^{\infty} k \cdot \left( \sum_{0 < j < k} \Pr(D_{HIP} = H + j) \cdot \Pr(D_{LIP} = L + k - j) \right)
\]

\[
+ \sum_{k=0}^{\infty} k \cdot \Pr(D_{HIP} = k + H) \cdot \Pr(D_{LIP} \leq L)
\]

\[
+ \sum_{k=0}^{\infty} k \cdot \left( \sum_{j=1}^{H} \Pr(D_{HIP} = j) \cdot \Pr(D_{LIP} = H + L + k - j) \right)
\]

where the three sums refer to the overtime due to both HIP and LIP, HIP only and LIP only respectively. Again, if we assume independent, stationary Poisson
distributions for the arrival rates of both HIP and LIP we get the following equation for the expected amount of OT

\[
\sum_{k=2}^{\infty} k \cdot \left( \sum_{0<j<k} \frac{(\lambda_H)^j e^{-\lambda_H}}{(j+k)!} \cdot \frac{(\lambda_L)^{L+k-j} e^{-\lambda_L}}{(L+k-j)!} \right) + \sum_{k=0}^{\infty} k \cdot \frac{(\lambda_H)^{k+H} e^{-\lambda_H}}{j!} \cdot \frac{(\lambda_L)^{j} e^{-\lambda_L}}{j!} + \sum_{k=0}^{\infty} k \cdot \left( \sum_{j=1}^{H} \frac{(\lambda_H)^j e^{-\lambda_H}}{j!} \cdot \frac{(\lambda_L)^{L+H+k-j} e^{-\lambda_L}}{(L+H+k-j)!} \right)
\]

In our opinion, the second OT constraint is more appropriate since it is easier to place a dollar figure on the expectation than on the probability. Note that the OT constraint does not impose a strict limit on the number of OT scans performed per day but rather places a limit on the expected number of OT scans each day.

Our assumption of Poisson arrival rates is based on the data we collected for IP arrivals at Vancouver General Hospital (VGH) covering 8 months (see Figure 2.1). The larger than expected left-hand tail may plausibly be attributed to regular maintenance on the CT machines that causes a reduction in throughput on certain days. We also tested the data both for a 'day of the week' effect and for any autocorrelation with negative results in both cases. There is thus very little ability to predict tomorrow's IP demand in advance. There are, however, no figures on how much of the total IP demand could be classified as LIP. Thus our assumption that the two IP categories form two independent Poisson distributions, though reasonable, needs to be verified. VGH did not collect data on OP arrivals and thus the only data available to us was a two week period in which we collected the data manually. Thus our assumption of poisson arrivals rates for OP is based more on theory than data. Given that OP demand arises from numerous independent sources (GP and specialist
offices), it seems reasonable to assume that the sum total of daily demand would be Poisson. There is definitely the possibility (though we didn’t see it in the small data set we have) of a day-of-the-week effect that would be worth investigating.

**FIGURE 2.1. Daily Weekday IP Demand for VGH (May-December 2003)**

![Daily Weekday IP Demand for VGH](image)

The optimization problem with a constraint on expected overtime was solved by a simple directed search in MATLAB. The code searches for the smallest $H$ for which there exists an $L$ so that the pair $(H, L)$ satisfies the OT constraint. This will minimize unused capacity (subject to the OT constraint) since unused capacity depends only on $H$. (We are assuming here that there is enough OP on call to fill any of the $L$ slots not used by carryover LIP.) The algorithm starts with an $H$ large enough so that, together with $L = 0$, it satisfies the OT constraint. It proceeds by reducing $H$ until the OT constraint is violated and then, holding $H$ constant, increasing $L$ until the constraint is once more satisfied. While keeping $L$ constant, the program then reduces $H$ by one and attempts to find an $L$ so that the combination $(H, L)$ once again satisfies the OT constraint. It terminates when no such combination can be
found and returns the last viable combination. This combination minimizes expected unused capacity.

6. What Improvement Can We Expect?

System performance can be measured by two metrics - the expected throughput rate (expected number of scans per day) and the expected growth rate in OP waiting time. The first can be calculated analytically and the second will be analyzed through a simulation of the scheduling process.

Once $H$ and $L$ have been specified, it is a straightforward calculation to determine the expected throughput of the system. In such a scenario, the expected number of scans per day in the case where all IP are treated as HIP is given by:

$$ (C - H) + E[D_{HIP}] = C - H + \lambda_H $$

where $H$ is the number of slots reserved for IP and $C$ is the total capacity. The first term represents the number of scheduled OP scans and the second the expected number of IP scans performed per day. Here we are assuming that there is enough OP demand to fill however many scanning slots $(C - H)$ are made available. If a certain percentage of IP can be classified as LIP, the expected number of scans per day can be calculated as follows (provided we assume that we can fill all slots that are made available for on-call OP):

$$ (C - H) + E[D_{HIP} + \text{LIP not carried over}] $$

$$ = C - H + \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \left( i + \min[(H - i)^+, j] + [j - (H - i)^+ - L]^+ \right) \* \Pr(D_{HIP} = i)\Pr(D_{LIP} = j). $$

37
(Note that the $H$ here may be different from the $H$ generated for the initial case with zero LIP.) The three terms that constitute the coefficient multiplying the probabilities consist respectively of the number of HIP, the number of LIP scanned due to excess capacity (i.e. low number of HIP) and the number of LIP scanned due to having more than $L$ LIP carried over.

The above throughput rates as well as the resulting OP waiting times depend heavily on certain parameters - namely, the limit placed on OT, the IP demand rate, the percentage of IP classified as LIP and the capacity of the system. The rest of the paper will analyze system performance based on a simulation model of the scheduling process and study the sensitivity of results to variations in the parameters. Finally, we will consider the effect of introducing multiple exam lengths and stochastic service times as well as varying the percentage of OP willing to be on-call.

7. The Simulation Model

Our simulation model, created in ARENA, allows us to compare the effect of the model parameters on the growth rate in OP waiting time over any given period. We assume independent Poisson arrival rates for all priority levels. (Thus when we talk about LIP being 10% of total IP demand we mean that the mean arrival rate for HIP is nine times that of LIP.) A more contentious assumption is that all incoming IP demand for any given day arrives at the beginning of the day. This is unrealistic since, though most IP demand is known at the beginning of the day, emergency demand occurs throughout the day and is therefore not known at the outset. However, our results are a comparison of three possible reservation policies. If anything, the intra-day variation would be more detrimental (increasing OT and unused capacity) in the
case where all IP are assumed to be HIP than under our proposed policies since LIP patients would be identified early in the day (they would not be emergencies) and therefore would act as a pool of patients who could be scanned if and when space was available. As long as OP are generally scheduled earlier in the day (as at VGH) the intra-day variation will have little impact. Our own observation of the CT department revealed that for the majority of the day, the department had a pool of IP waiting for a scan that they called down when space was available. Thus unused capacity is unlikely to be affected by intra-day variation in IP demand. As for OT, it is possible that a late arriving emergency could cause additional OT but this scenario is equally likely under all reservation policies and therefore does not affect the comparison.

We assume that OP demand arrives at the beginning of the day and is scheduled for the earliest day available. Chapter 3 looks at more interesting scheduling policies for OP. For now, with only one priority class, there is no need to do anything other than book incoming demand in the first available slot. A certain percentage of the total OP demand is tentatively booked but also set aside as “on call”. These are the OP who can be called the day before to fill any of the \( L \) carry-over slots not filled by LIP for the next day. After each day, the simulation cycles through these “on-call” OP and places them into the available slots (on days when less than \( L \) LIP have been carried over to the next day) on a first in first out basis. As the simulation progresses, waiting times are recorded by priority level along with the arrival rates for the various categories, the amount of daily OT or unused capacity as well as the growth in OP waiting time from the previous day. We assume that an OP cannot be booked the day the request arrives.
Our initial testing of the simulation consisted of 100 replications of 1000 days each. We did not collect data for the first 500 days in order to mitigate the effect of initial conditions. We assumed a mean arrival rate for IP of 120 (15 minute scans) per day and a mean arrival rate for OP of 50 per day. This is clearly a large scale hospital (or multiple hospitals) with numerous CT machines. The regular-hour capacity is set at 165 which is five less than the average daily demand. We initially assume that all scans are 15 minutes and that an expected OT of five scans per day is acceptable. Thus the total capacity including OT is equal to the average demand. Actual practice at VGH tries to avoid OT completely which makes the use of LIP even more vital as shown below. Finally, we assume that 10% of OP are willing to be on-call and that 10% of IP can be classified as LIP. The results of the replications were very similar so we present the combined results with a 95% confidence interval.

The reason for doing 100 replications of 1000 days each rather than one longer replication is that, since we are dealing with the case where demand outstrips capacity, OP waiting time is increasing without bound. This is not realistic as, after a time, either overtime will be used to meet demand or else demand will slacken as people go elsewhere rather than face the lengthening waiting times. The purposes of this model is simply to show that the rate of increase in OP waiting time can be greatly reduced by introducing the LIP category meaning that the amount of OT required to meet demand is reduced. These findings are confirmed in the larger model presented in Chapter 5. We assess three reservation policies that we will call:

1. The constrained reservation policy
2. The unconstrained reservation policy
3. Current practice (i.e. treating all IP as HIP).
TABLE 2.4. Simulation Results (mean with 95% confidence limits)

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constrained Policy</td>
<td>Unconstrained Policy</td>
<td>Current Practice</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$H = 110$  $+/-$</td>
<td>$H = 107$  $+/-$</td>
<td>$H = 119$  $+/-$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$L = 9$</td>
<td>$L = 15$</td>
<td>$L = 0$</td>
<td></td>
</tr>
<tr>
<td>Daily IP Arrivals</td>
<td>119.98 0.09</td>
<td>119.98 0.09</td>
<td>120.01 0.08</td>
<td></td>
</tr>
<tr>
<td>Daily OP Arrivals</td>
<td>50.01 0.06</td>
<td>50.01 0.06</td>
<td>49.97 0.06</td>
<td></td>
</tr>
<tr>
<td>Regular Hour Scans</td>
<td>163.96 0.01</td>
<td>163.93 0.02</td>
<td>161.12 0.02</td>
<td></td>
</tr>
<tr>
<td>Overtime Scans</td>
<td>4.98 0.06</td>
<td>4.03 0.07</td>
<td>4.89 0.06</td>
<td></td>
</tr>
<tr>
<td>Unused Capacity</td>
<td>1.04 0.03</td>
<td>1.07 0.04</td>
<td>3.88 0.05</td>
<td></td>
</tr>
<tr>
<td>Throughput (Scans/day)</td>
<td>168.94 0.07</td>
<td>167.96 0.09</td>
<td>166.01 0.01</td>
<td></td>
</tr>
<tr>
<td>Weekly Growth in OP WT (days/week)</td>
<td>0.10 0.01</td>
<td>0.21 0.01</td>
<td>0.43 0.01</td>
<td></td>
</tr>
</tbody>
</table>

The constrained policy, $(H = 110, L = 9)$, is the optimum of the above optimization problem under the additional constraint that the total $H + L$ should be no more than would be reserved if all IP were HIP. This policy will perform at least as well as current practice since the same number of slots is reserved. This modified optimization problem can be solved in precisely the same manner as the previous one except that the viable combinations of $L$ and $H$ are restricted. The unconstrained policy, $(H = 107, L = 15)$, is the optimum of the above optimization problem without the additional constraint. Finally, we compare both of these to the current practice $(H = 119, L = 0)$ of treating all IP as HIP. Table 2.4 shows the average per day over all replications for the relevant statistics as well as the average weekly growth rate.

Both the constrained and unconstrained policies have significantly higher throughput rates and lower growth rates in OP waiting time than current practice. After 1000
days (starting from empty queues), the waiting time for OP under current practice grows to approximately 84 days while the waiting time for OP under the constrained optimal policy (column 1) only increases to 29 days. This significant reduction is achieved by identifying only 10% of total IP demand as LIP. It is accomplished with no additional expenditure (note that OT remained the same) and would seem to more than justify the procedural changes required. The fact that the more conservative constrained policy (column 1) outperforms (in reducing growth rate and increasing throughput) the optimal unconstrained policy (column 2) can be attributed (as will be shown later) to the fact that a 10% on call rate for OP is simply not large enough to take full advantage of the flexibility provide by the \((H = 107, L = 15)\) policy. Thus the higher \(L\) value is detrimental since it means that there is less scanning time available for those OP not on call.

The effect of classifying 10% of IP as LIP is to increase the number of scans per day by almost 3. This may not seem like a huge increase but, as Preater points out, it has been shown within queuing theory that “small changes to a heavily congested systems bring disproportionate changes in queue size” and hence waiting times [40]. Where demand outstrips throughput (under current practice) by just 4 scans per day, increasing the number of daily scans by almost 3 is a substantial step towards meeting total demand. Three additional scans per day over 1000 days results in 3000 more scans being performed than under the current practice. In that light it is perhaps understandable why there is such a significant difference in OP waiting time.

Clearly, none of the three policies satisfy demand in the above instance (average number of scans per day for all three is less than the combined arrival rate) thus waiting times are increasing in all cases. However, as Table 2.4 indicates, it is the
rate of increase in OP waiting time that is significantly reduced. Figure 2.2 shows the
growth in OP waiting time from a typical simulation run of 500 days in length for the
constrained reservation policy and for current practice demonstrating the significantly
lower growth rate in OP waiting time.

**Figure 2.2.** Constrained Policy vs Current Practice (Simulation standard error of 0.01)

The rest of the paper seeks to demonstrate the robustness of these results to
variations in the simulation parameters.

**7.1. Varying the OT Limit.** Table 2.5 presents the expected unused capacity
and expected number of scans per day, based on equations (2.5) and (2.6), as a
function of the limit on expected OT for the three reservation policies. As one would
expect, the more OT allowed, the less potential for unused capacity and therefore the
less difference between the three cases. It is worth noting that we are assuming here
that there is sufficient capacity to be able to reserve enough scanning time for IP so
as to keep OT as low as required. The trade-off is simply that the more scanning
table 2.5. effect of OT limit on capacity usage

<table>
<thead>
<tr>
<th>OT Limit</th>
<th>Reservation Policy</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constrained Policy</td>
<td>Unconstrained Policy</td>
<td>Current Practice</td>
<td></td>
</tr>
<tr>
<td></td>
<td>H L Unused Capacity</td>
<td>H L Unused Capacity</td>
<td>H Unused Capacity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scans per Day</td>
<td>Scans per Day</td>
<td>Scans per Day</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>121 10 4.89 161.09</td>
<td>118 17 3.43 162.56</td>
<td>131 11.94 154</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>116 11 2.63 164.28</td>
<td>114 15 1.98 165.00</td>
<td>127 8.77 158</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>113 11 1.70 166.21</td>
<td>111 16 1.23 166.71</td>
<td>124 6.68 161</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>112 9 1.45 167.53</td>
<td>109 15 0.86 168.01</td>
<td>121 4.89 164</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>110 9 1.03 168.96</td>
<td>107 15 0.59 169.3</td>
<td>119 3.88 166</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>107 11 0.59 170.25</td>
<td>105 16 0.39 170.56</td>
<td>118 3.43 167</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>106 10 0.48 171.43</td>
<td>104 14 0.32 171.55</td>
<td>116 2.63 169</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>108 6 0.72 172.26</td>
<td>102 16 0.20 172.26</td>
<td>114 1.98 171</td>
<td></td>
</tr>
</tbody>
</table>

slots reserved for IP, the less slots available for OP booking resulting in longer OP waiting times.

Of note is the fact that the overall number of reserved slots \((H + L)\) is less under the current practice of treating all IP as HIP than it is under the unconstrained policy. This is precisely the reason for including the constrained policy in the analysis, restricting the number of reserved slots to be the same as under current practice. The row with a limit of (an average of) 5 OT scans per day, demonstrates that the simulation results given earlier are a very good fit to the theoretical results presented here. The unconstrained policy under-performs in the simulation but this will be shown to be due to the lack of OP on call.

Significantly, the current policy must increase the OT limit to 8 before throughput matches demand. In contrast, both the constrained and unconstrained policies require
an OT limit of 6 in order for throughput to match demand. At Vancouver General Hospital, CT technicians are paid time and a half for OT giving an approximate cost of $60 per hour of OT. Thus, assuming two technicians per scan, an increase in OT by half an hour costs $60 in technician salaries per day yielding a $15,000 cost over the year. However, if OT exceeds two hours then technicians are paid double leading to a cost of $80 in technician salaries for 2 scans yielding a $20,000 cost over the year. CT technician salaries are clearly only the most obvious cost associated with OT. Increased OT would also imply longer hours for CT support staff, radiologists and porters. Thus the true cost of increased OT is undoubtedly significantly higher.

7.2. Varying IP demand. In order to determine the effect of the IP demand rate on the growth rate in OP waiting time, we ran the simulation with the IP arrival rate being 60 per day and 90 per day. Capacity was set at five under the total demand rate while all other parameters remained the same. Figure 2.3 compares the effect on the growth rate in OP waiting time for the three reservation policies, optimized for the different demand rates.

It is clear that reducing IP demand does not negate the benefit of the proposed reservation policy. Note also that the unconstrained policy is only outperformed by the constrained policy when the IP arrival rate reaches 120. This confirms the fact that the poorer results of this policy are due to having insufficient numbers of OP on call in order to take advantage of the flexibility provided by a policy with a large L value.

7.3. Varying the Percentage of LIP. As the percentage of LIP patients increases, the expectation is that the added flexibility would improve system throughput. To substantiate this we ran the simulation assuming that 20% of IP could be
classified as LIP and with 100% of OP on call to determine the maximum benefit that might arise from an additional 10% LIP. Again we ran the simulation for the optimal unconstrained policy ($H = 94, L = 35$) together with the optimal constrained policy ($H = 100, L = 19$) and compared them to the current practice. The growth rate in OP waiting time was dramatically reduced (from 0.432 days/week to 0.018 days/week) with throughput coming very close to matching demand. This was accomplished without increasing OT thus the improvement is due entirely to a reduction in unused capacity.

### 7.4. Varying Capacity

To determine the effect of increased capacity, we ran the simulation with the same initial conditions but increasing regular hour capacity.

It is clear from Figure 2.4 that as capacity approaches demand the effect of introducing LIP becomes less pronounced though percentage wise the reduction in the growth rate in OP waiting time increases. Moreover, as capacity approaches demand
some of the benefit derived from introducing a LIP category is transferred from increasing throughput to decreasing OT. With a capacity of 169, the average OT per day for the three policies was 3.39, 2.92 and 4.84 respectively. Thus, even in a system that is meeting demand through OT, there is some benefit to introducing LIP.

7.5. Effect of Multiple Exam Lengths. The above results are based on the unrealistic assumption that all exams are 15 minutes long. In reality, scan lengths can be scheduled for anywhere from 15 minutes to 90 minutes (though scan lengths greater than 60 are rare). Thus, though it is possible to simply view a 60 minute scan as four 15 minute scans, the result is that arrivals now come in batches which would, of course, affect the theoretical calculations. To determine the effect of multiple exam lengths we ran the simulation with exam lengths varying from 15 to 60 minutes. For both OP and IP, we assigned 50% of total demand to 15 minutes scans, 30% to 30 minutes scans, 10% to 45 minute scans and the remaining 10% to 60 minutes scans. We adjusted the overall arrival rates so that the total demand on the resource was
Table 2.6. Multiple Exam Lengths Comparison

<table>
<thead>
<tr>
<th></th>
<th>Constrained Policy</th>
<th>Unconstrained Policy</th>
<th>Current Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H = 110$</td>
<td>$H = 107$</td>
<td>$H = 119$</td>
</tr>
<tr>
<td></td>
<td>$L = 9$</td>
<td>$L = 15$</td>
<td>$L = 0$</td>
</tr>
<tr>
<td>Daily IP Arrivals</td>
<td>120.03 0.16</td>
<td>120.03 0.16</td>
<td>119.95 0.16</td>
</tr>
<tr>
<td>Daily OP Arrivals</td>
<td>49.98 0.09</td>
<td>49.98 0.09</td>
<td>49.98 0.09</td>
</tr>
<tr>
<td>Regular Hour Scans</td>
<td>162.24 0.15</td>
<td>161.99 0.16</td>
<td>158.82 0.16</td>
</tr>
<tr>
<td>Overtime Scans</td>
<td>7.70 0.07</td>
<td>6.22 0.08</td>
<td>7.13 0.08</td>
</tr>
<tr>
<td>Unused Capacity</td>
<td>2.77 0.1</td>
<td>3.01 0.11</td>
<td>6.17 0.1</td>
</tr>
<tr>
<td>Throughput (Scans/day)</td>
<td>169.94 0.15</td>
<td>168.21 0.16</td>
<td>165.95 0.16</td>
</tr>
<tr>
<td>Weekly Growth in OP WT (days/week)</td>
<td>0.0048 0.01</td>
<td>0.186 0.01</td>
<td>0.433 0.01</td>
</tr>
</tbody>
</table>

unchanged at 120 scanning slots (of 15 minutes each) per day for IP and 50 scanning slots per day for OP.

The results given in Table 2.6 show that the effect is to increase the advantage of introducing LIP to the system. However, once multiple exam lengths are added, the amount of OT increases beyond the allowable limit for all three policies considered. Again, the units are the number of 15 minute scanning slots.

**7.6. Increasing On Call Percentage.** One of the challenges facing this proposed reservation policy is the need to have a pool of OP who are on 24-hour call - that is, willing to be scanned if given a call the night before. In a city the size of Vancouver, this is unlikely to be a major problem. To test the range of results possible for the different reservation policies, we tested both the constrained policy ($H = 110, L = 9$) and the unconstrained policy ($H = 107, L = 15$) with the on call
percentage being 0, 10, 20, 30 and 100 percent of the total OP demand. The resultant effect on the weekly growth rate in OP waiting times is given in Figure 2.5.

**FIGURE 2.5. Varying Percentage of OP On Call (Simulation standard error of 0.01)**

As per expectation, increasing the percentage on OP on call increases the benefit of the proposed system up to a point, after which it makes little difference. (Note that the policy \((H = 110, L = 9)\) with 0% on call gives the same growth rate in waiting time for OP as the current practice of treating all IP as HIP.) Clearly, as the percentage of OP on call is reduced to zero, the effect on OP waiting times of introducing LIP is lost though there remains the advantage of reduced OT. In fact, for the policy \((H = 107, L = 15)\), the effect is detrimental since the sum of \(H + L\) is greater than would be reserved under current practice. Conversely, if the percentage of OP on call is sufficient to take advantage of the flexibility provided by LIP, then the policy \((H = 107, L = 15)\) is the most effective at reducing the growth rate in OP waiting time. However, the potential gain by using the unconstrained optimal policy is minimal and thus the more conservative policy \((H = 110, L = 9)\) that insures that
you do no worse than under the current practice of treating all IP as HIP would seem to be the wiser option. The salient message however is that the percentage of OP on call does not need to be a significant portion of the total OP demand in order for the introduction of LIP to have a significant impact on OP waiting times.

7.7. Stochastic Service Times. The above results assume that a 15 minute scan takes 15 minutes which is, of course, highly unrealistic. For instance, at VGH, the time required for a scan is systematically overestimated. Thus unused capacity would be much higher than reported in the simulation. However, our proposal looks at reducing the number of scheduling slots that stand idle in a given day. The efficiency within each scheduling slot is another issue entirely and largely depends on aligning the actual scanning time required for a given exam with the time scheduled. The overall impact on OT that would result from introducing stochastic scanning times is not, in our opinion, a significant factor - provided scheduled times are a fair reflection of the actual required scanning time and provided that regular hour capacity is sufficiently large. If these two criteria are met then the law of averages will mitigate the effect on OT of introducing stochastic scanning times. Moreover, whatever the effect produced, it would be the same for each policy and therefore would not affect the comparison.

To substantiate the above claim we ran the simulation with the same reservation policies but allowing for stochastic service times. We ran the simulation with the original parameters and allowed service times to follow a triangular distribution from 10 to 20 minutes. The results in Table 2.7 are almost identical to the initial results given in Table 2.4 with only a slight increase in OT.
TABLE 2.7. Stochastic Service Times

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constrained Policy</td>
<td>Unconstrained Policy</td>
<td>Current Practice</td>
</tr>
<tr>
<td></td>
<td>$H = 110$ $+/-$</td>
<td>$H = 107$ $+/-$</td>
<td>$H = 119$ $+/-$</td>
</tr>
<tr>
<td></td>
<td>$L = 9$</td>
<td>$L = 15$</td>
<td>$L = 0$</td>
</tr>
<tr>
<td>Daily IP Arrivals</td>
<td>119.97 0.09</td>
<td>119.97 0.09</td>
<td>119.98 0.1</td>
</tr>
<tr>
<td>Daily OP Arrivals</td>
<td>49.99 0.06</td>
<td>49.99 0.06</td>
<td>50 0.06</td>
</tr>
<tr>
<td>Regular Hour Scans</td>
<td>163.73 0.07</td>
<td>163.64 0.06</td>
<td>161.07 0.1</td>
</tr>
<tr>
<td>Overtime Scans</td>
<td>5.2 0.06</td>
<td>4.31 0.06</td>
<td>4.91 0.06</td>
</tr>
<tr>
<td>Unused Capacity</td>
<td>1.23 0.03</td>
<td>1.35 0.02</td>
<td>3.93 0.06</td>
</tr>
<tr>
<td>Throughput (Scans/day)</td>
<td>168.93 0.07</td>
<td>167.95 0.06</td>
<td>165.98 0.1</td>
</tr>
<tr>
<td>Weekly Growth in OP WT (days/week)</td>
<td>0.1 0.01</td>
<td>0.21 0.01</td>
<td>0.43 0.01</td>
</tr>
</tbody>
</table>

8. Implementation Challenges

There are, no doubt, a number of implementation challenges associated with any attempt to improve a system through process modification. We briefly discuss a few of the more obvious ones.

Firstly, our simulation assumes a five-day workweek whereas reality is somewhat different. VGH conducts OP scans Monday through Friday but performs IP scans 7 days a week. Since the weekends are only IP, the situation is actually better (favouring the proposed scheme over current practice) in that Monday starts off fresh with no carry over LIP and all of Friday's carry over LIP can be finished on the weekend (capacity is currently under-utilized on weekends). Our proposed scheduling scheme would therefore fit the current true situation with the only adjustment being that one
need only reserve $H$ slots on Monday (as you know in advance there will be no LIP) and hence can book more OP for Monday.

Secondly, though our own data collection suggests that a Poisson distribution is a reasonable fit to the data, this may not be true in other instances. It would be interesting to see the effect of running the simulation with different underlying arrival distributions. For instance, it could be argued that the Poisson does not represent enough variation. However, our simulation runs under the Poisson assumption with mean of 120 results in a standard deviation of about 11. Therefore three standard deviations to either side of the mean leads to a range of almost 70 slots per day. This seems to adequately reflect the fluctuation in demand we observed.

Finally, it is worth noting that the introduction of multiple OP classes does not affect the results given in this paper. One can view the overall problem as two sub-problems. First, how much capacity does one need to reserve for IP in order to meet the required limits on OT? Second, given that one has set aside a certain amount of overall capacity for IP demand, how does one proceed to schedule the various OP classes into the remaining capacity? The optimal answer to the first question is entirely independent of the second.

There are other implementation challenges (such as how no shows would affect the results) but none, in our mind, negate the fact that the introduction of the proposed reservation policy significantly increases throughput and reduces OP waiting times without increasing costs.
9. Conclusion

The introduction of a low priority IP class does not represent a huge procedural change on the part of a CT department. It requires some careful thought on the part of the radiologists to insure that the appropriate IP are classified as LIP and it requires some additional phone calling on the part of the staff. There are no additional financial costs involved. The results of the simulation show clearly that introducing a one day flexibility to only 10% of IP demand provides a significant reduction in the growth rate of OP waiting times - even if only 10% of OP are willing to be on call. Varying the parameters of the simulation, as well as introducing variability in the duration of each scan, did not affect the ability of the proposed policy to improve throughput and thereby reduce OP waiting times.

The proposal presented here applies readily to the CT operations that provided the impetus for this research. However, it could potentially be applied to any situation where there is a fluctuating demand and two or more priority levels (such as MRI-scheduling or operating room scheduling). The key is to determine whether there is any flexibility present in the high priority class and how to best exploit that flexibility. A potential application outside of health care might be the management and shipment of inventory. Inventory often involves multiple priority classes. The only question therefore is whether there is any flexibility in the scheduling of the higher priority class that is not currently being used efficiently. In the case of inventory management, Cattani and Souza compare a number of rationing policies in a scenario with two demand classes [13]. However, all of the policies only attempt to manipulate the lower demand class. Lawson and Porteus have demonstrated the advantage of allowing for the possible expediting of a shipment [32]. However, it might also be advantageous to
have a priority class that can potentially be delayed in order to insure that shipments leave with near full capacity. This would be analogous to the LIP category.

The research in this chapter could be viewed as clearing the ground for the more complex problem with multiple OP priority classes. Imposing an OT limit, as proposed above, allows for the separation of the total capacity between IP and emergency on the one hand and OP on the other. However, though a directed search was sufficient to determine the capacity required for IP demand, it does not answer the issue of when to schedule OP - particularly if there are numerous priority classes. Chapter 3 addresses the issue of scheduling OP into the remaining capacity as well as determining what action to take if the remaining capacity is not sufficient to maintain reasonable waiting times for the various OP priority classes. Chapter 5 will then expand the model in Chapter 3 to include IP demand and then compare the results with what has been presented here.

There is no doubt that limited resources are going to continue to challenge the provision of health care. It would seem reasonable therefore to insure that we are using the resources that we have as effectively as possible which is precisely what the proposal outlined in this chapter attempts to do.¹

¹A version of this Chapter has been accepted for publication in Journal of the Operations Research Society. Patrick J. and Puterman M. Improving Resource Utilization through Flexible Inpatient Scheduling
CHAPTER 3

Dynamic Multi-Priority Patient Scheduling with Uncertain Demand

1. Introduction

Chapter 2 simplified the scheduling process by assuming only one OP priority class. With only one priority class, there is essentially no scheduling issue for OP as one will simply allocate the first available slot within the capacity allocated to OP demand on a first come first served basis. However, if there are numerous priority classes then clearly there is reason for booking lower priority OP later than the first available slot in order to leave room for later arriving higher priority demand. It is this additional challenge that we seek to address in this chapter. We assume that total capacity has already been divided between IP and emergency demand on the one hand and OP demand on the other. We then seek to find the optimal scheduling policy for the capacity allocated to OP demand when there are multiple priority classes. Unfortunately, the optimization techniques used in chapter 2 are no longer useful in the setting where there are more than two priority classes. Thus, we resort to using (approximate) dynamic programming.

2. The Model

2.1. The State Space. We begin, therefore, by formulating the scheduling process as a discounted MDP. On any given day, the resource manager will have access to the schedule, as it currently stands from today to the end of the booking horizon,
as well as to the incoming demand waiting to be booked from each OP priority class.

Thus the state space, \( S \), takes the form:

\[(\bar{x}, \bar{y}) = (x_1, x_2, \ldots, x_N; y_1, y_2, \ldots, y_I)\]

where \( x_n \) = the number of OP already booked on day \( n \) (\( x_1 \) = tomorrow's bookings),

\( x_n \in \{0, 1, \ldots, C\} \) with \( C \) equal to the daily OP capacity,

\( y_i \) = the number of OP of priority \( i \) waiting to be booked,

\( N \) = the maximum number of days one will book in advance,

\( I \) = the number of OP priority levels.

It is worth noting that \( C \) is the result of an earlier decision that divides the total capacity between IP and emergency demand on the one hand and OP on the other.

### 2.2. The Action Set.

The resource manager's task is to decide on which day to schedule the incoming demand. In addition, there needs to be some mechanism for keeping the system viable even if there is not enough available capacity for waiting demand. There are at least three potential scenarios. In the first, the resource manager has the ability to reject demand if necessary. In the second, he can relieve the stress on the system by performing some scans through overtime (or through outsourcing). In the third, the booking of demand can be postponed to a future date so that any unbooked scans appear in tomorrow's demand. Since this last option only postpones the problem, we concentrate on the first two. Thus, a vector of possible
actions can be written as, \((a, z) = \{a_{in}, z_i\}\), where

\[a_{in} = \text{the number of priority } i \text{ OP to book on day } n,\]

\[z_i = \text{the number of priority } i \text{ OP to reject or serve through overtime.}\]

It is obviously undesirable to turn demand away or to perform extensive overtime so there needs to be a significant cost associated with these options. To be valid, any action must satisfy the following three constraints which insure that the capacity constraint is not violated, that all waiting OPs are booked and that actions are forced to be positive and integer.

\[
(3.1) \quad x_n + \sum_{i=1}^{I} a_{in} \leq C \quad \forall n, n = 1, ..., N
\]

\[
(3.2) \quad \sum_{n=1}^{N} a_{in} = y_i - z_i \quad \forall i, i = 1, ..., I
\]

\[
(3.3) \quad a, z \geq 0, \text{ integer } \forall i, i = 1, ..., I \quad \& \quad \forall n, n = 1, ..., N.
\]

Therefore, we can denote the action set for any given state by

\[
A_{x, i} = \{(a, z)|x_n + \sum_{i=1}^{I} a_{in} \leq C, \sum_{n=1}^{N} a_{in} = y_i - z_i, (a, z) \geq 0, \text{ integer}\}.
\]

2.3. The Transition Probabilities. Decisions are made at the end of each day. The model is complicated by the fact that the horizon is not static but rolling with day 2 becoming day 1 and so on after each decision epoch. We assume that no OP can be booked more than \(N\) days in advance so that, at the beginning of each decision epoch, the \(N^{th}\) day has no bookings in it. Once a decision is made, the only stochastic element in the transition to the new state consists of the next day's incoming arrivals.
Thus the transition can be written as:

\[(x_1, x_2, ..., x_N; y_1, y_2, ..., y_I) \rightarrow (x_2 + \sum_{i=1}^{I} a_i, ..., x_N + \sum_{i=1}^{I} a_i, 0; d_1, ..., d_I)\]

with probability \(p(\tilde{d}) = \prod_{i=1}^{I} p(d_i)\) where \(p(d_i)\) is the probability that \(d_i\) OP of priority \(i\) arrive on a given day. Here we are assuming that the demand for one priority level is independent of the demand for another. This assumption is not strictly necessary but does make the mathematics easier. Nor does there seem any reason to doubt that it would be true since OP demand arises from independent sources (the specialist doctors in the region serviced by the hospital). We are also assuming that demand from one day to the next is independent of each other. Again, though this does not seem unreasonable, it would be good to verify this in any application to real data.

2.4. The Cost Structure. The only remaining piece of information required for the optimality equation is the cost associated with a given state-action pair. The cost comes from two sources. First, a cost is incurred if an OP is booked later than the maximum recommended waiting time for the priority class in question. Second, there is a cost associated with rejecting demand or using overtime to serve excess demand. Thus, assuming a linear cost structure, we can write the costs as:

\[c(a, z) = \sum_{i,n} c(i, n) a_{i,n} + \sum_{i=1}^{I} h(i) z_i\]

where, for example, \(c(i, n) = k(i)[n - T(i)]^+\), and 

\[k(i)\] is the daily penalty for an OP of priority \(i\) who waits longer than the maximum recommended time, \(T(i)\),

\[h(i)\] is the penalty for rejecting or serving through overtime.
The cost function \( c(i, n) \) reflects the fact that the cost associated with a late booking is only incurred if a patient from priority class \( i \) is booked past \( T(i) \) and thereafter a fixed penalty cost, \( k(i) \) is paid for each day late. It is reasonable to assume that both \( k(i) \) and \( h(i) \) are non-increasing in \( i \). In fact, for the scenario where demand can be rejected, it seems reasonable to assume that \( h(i) \) is strictly decreasing in \( i \) and that for the scenario where overtime is used \( h(i) \) is constant since overtime costs are independent of the priority class.

### 2.5. The Value Function

The value function in the MDP is therefore the total discounted cost over the infinite horizon and satisfies the following optimality equation:

\[
(3.4) \quad v(\tilde{x}, \tilde{y}) = \min_{(\tilde{a}, \tilde{z}) \in \mathcal{A}_{\tilde{z}}} \left\{ \frac{c(\tilde{a}, \tilde{z}) + \gamma \sum_{d \in D} p(d) v \left( x_2 + \sum_{i=1}^{l} a_{i2}, ..., x_N + \sum_{i=1}^{l} a_{iN}, 0; d_1, ..., d_I \right)}{\mathcal{D}} \right\}
\]

where \( \gamma \) is the daily discount factor and \( D \) is the set of all possible incoming demand streams.

Of course, it is here that we run into dynamic programming's greatest nemesis - 'the curse of dimensionality' - since the state space is much too large to allow a direct numerical solution. In particular, the dimension of the state space is \( CNIM \), where \( M \) is the maximum number of patients from a single priority class who can arrive on a given day. This is, of course, extremely large for any reasonable values of \( C, N, I \) and \( M \) and therefore does not allow the above DP to be solved by standard methods.
2.6. An Introduction to Approximate Dynamic Programming. There now exists a whole field of potential methods for dealing with precisely this problem, all of which are grouped under the title of approximate dynamic programming. These potential methods begin by restricting the value function to a certain class of functions and then seeking to find the "optimal" value function within this class. The question is how do you determine what class of functions to choose and how best to find the "optimal" approximate value function within the chosen class? The first question is still wide open but there are a number of available answers to the second.

One possibility is to update the approximation via simulation. As an example of this tactic, consider approximate policy iteration using temporal difference methods. The process begins by assuming a certain form to the value function in the MDP that is dependent on a vector of parameters, \( \vec{r} \), and then fine-tunes the value of \( \vec{r} \) via a simulated run, \( (s_0, s_1, s_2, ..., s_n) \), of the MDP generated by a fixed policy. To see how this works in detail, let us assume that the approximate value function can be written as \( v(s, r) \) where \( s \) is a member of the state space. Further, let \( g(s_k, s_{k+1}) \) be the reward (or cost) for going from state \( s_k \) to state \( s_{k+1} \) and let \( \lambda \) be a number between 0 and 1. The update to the parameter vector, \( \vec{r} \), following the transition to state \( s_{k+1} \) is written as:

\[
(3.5) \quad r = r + \gamma d_k \sum_{m=0}^{k} \lambda^{k-m} \nabla v(s_m, r)
\]

where \( d_k \) is called the temporal difference (TD) and is equal to \( g(s_k, s_{k+1}) + v(s_{k+1}, r) - v(s_k, r) \), \( \gamma \) is the discount factor and where the gradient is taken with respect to \( \vec{r} \). Specifically, \( d_k \) gives you the difference between your current estimate of the value function at state \( s_k \) and the one-step realization of the value of being in state \( s_k \) based
on the simulation run. A positive temporal difference therefore suggests that you have underestimated the value of being in state $s_k$ and a negative temporal difference that you overestimated it. Intuitively, it makes sense that the value of $r_i$ should increase if and only if you have overestimated the value function ($d_k < 0$) and the value function is decreasing in $r_i$ ($\nabla_i v(s_k, r) < 0$) or else you have underestimated the value function ($d_k > 0$) and the value function is increasing in $r_i$ ($\nabla_i v(s_k, r) > 0$). Assuming $\lambda = 1$, this is precisely what equation (3.5) does except that the update is dependent not only on the transition $s_k$ to $s_{k+1}$ but on all transitions that have already occurred in the simulation run. The value of $\lambda$ allows one to discount the effect of a state the further away it is from the current transition.

There are two sources of error. First, the approximation may not be able to capture the true value function adequately and secondly, the update of the parameter vector is based on a simulation run and is therefore open to sampling error.

Once the approximate policy evaluation is complete and the approximate value function is available, a new policy is generated which is greedy with respect to $v(s, r)$. Sometimes this can be done exactly while at other times we can only obtain an approximation to a greedy policy which opens this method up to a third source of error. Thus, we have a an algorithm that alternates between policy evaluation steps and policy improvements steps.

A second possibility, and the one employed in this thesis, is to update the approximation through mathematical programming. This approach was first developed by Schweitzer and Seidmann [49] with more recent work done by both Adelman and De Farias and Van Roy. The general idea is to transform the above MDP into the equivalent linear program (LP) and then assume some form to the value function
that makes the LP tractable. As mentioned earlier, choosing the appropriate form for the approximation is still an open question in approximate dynamic programming and requires a certain amount of ingenuity and insight into the nature of the problem. One popular method is to pick "features" that best describe the problem as the basis for the approximation. The choice of "features" however remains a matter of judgement.

2.7. Approximating the Value Function. Dynamic Programming theory states that the above MDP can be transformed into the equivalent linear program of the following form:

(3.6) \[
\max_{\bar{x}, \bar{y} \in S} \sum_{\bar{x}, \bar{y} \in S} \alpha(\bar{x}, \bar{y})v(\bar{x}, \bar{y})
\]

subject to

(3.7) \[
c(\bar{a}, \bar{z}) + \gamma \sum_{\bar{d} \in D} \left( p(\bar{d}) v \left( x_2 + \sum_{i=1}^{I} a_{i2}, ..., x_N + \sum_{i=1}^{I} a_{iN}, 0; d_1, ..., d_I \right) \right) \\
\geq v(\bar{x}, \bar{y}) \quad \forall (\bar{a}, \bar{z}) \in A_{\bar{x}, \bar{y}} \text{ and } (\bar{x}, \bar{y}) \in S
\]

where \( \alpha \) is a vector of positive numbers. The solution to the above LP will equal the solution to the optimality equation given in (3.4) for any strictly positive \( \alpha \) [41]. Without loss of generality, we will assume that \( \alpha \) is a probability distribution (over the initial condition of the state space). The conversion to a linear program does not, however, avoid the curse of dimensionality as the above LP has a variable for every state and a constraint for every state-action pair in the MDP. A possible solution is to approximate the value function, \( v \), with a linear combination of basis functions.
(see [2] and [15]). As a first attempt, we will use the following affine approximation:

\[
\nu(\bar{x}, \bar{y}) = W_0 + \sum_{n=1}^{N} V_n x_n + \sum_{i=1}^{I} W_i y_i
\]

(3.8)

\(V_n\) can be interpreted as the marginal cost of booking a scan on day \(n\) and \(W_i\) can be interpreted as the marginal cost of having one more OP of priority class \(i\) waiting to be booked. We will impose the further restriction that \(V_n, n = 1, ..., N\) and \(W_i, i = 1, ..., I\) are non-negative while \(W_0\) is unconstrained. A working paper by Adelman [3] gives some reason for believing that for a MDP that consists of multiple sub-problems that are joined together by some linking constraints (as in our model), the choice of a linear approximation for the value function is reasonable. However, there is certainly no guarantee that this is the case and thus one resorts to determining the appropriateness of fit by analyzing the eventual outcome.

Once we have chosen a form for the approximate value function, we can reformulate the LP as the following:

\[
\begin{align*}
\max & \quad V, W, \sum_{x, y} \alpha(\bar{x}, \bar{y}) \left( W_0 + \sum_{n=1}^{N} V_n x_n + \sum_{i=1}^{I} W_i y_i \right) \\
\text{subject to} & \quad W_0 + \sum_{n=1}^{N} V_n x_n + \sum_{i=1}^{I} W_i y_i - \gamma \sum_{d \in D} \left[ p(\bar{d}) \left( W_0 + \sum_{n=1}^{N-1} V_n x_{n+1} + \sum_{i=1}^{I} a_{i,n+1} + \sum_{i=1}^{I} W_i d_i \right) \right] \\
& \quad \leq c(\bar{a}, \bar{x}) \quad \forall (\bar{a}, \bar{x}) \in A_{x,y} \text{ and } (\bar{x}, \bar{y}) \in S \\
& \quad \bar{V}, \bar{W} \geq 0.
\end{align*}
\]

\[63\]
Rearranging terms and using the assumption that \( \alpha \) is a probability distribution over the initial condition of the state space yields

\[
\max_{\bar{V}, \bar{W}} \left\{ W_0 + \sum_{n=1}^{N} E_\alpha[X_n]V_n + \sum_{i=1}^{I} E_\alpha[Y_i]W_i \right\}
\]

subject to

\[
(1 - \gamma)W_0 + \sum_{n=1}^{N} V_n \left( x_n - \gamma x_{n+1} - \gamma \sum_{i=1}^{I} a_{i,n+1} \right) + \sum_{i=1}^{I} W_i \left( y_i - \gamma E_\alpha[Y_i] \right) \leq c(\bar{a}, \bar{z}) \quad \forall (\bar{a}, \bar{z}) \in A_{\bar{x}, \bar{y}} \text{ and } (\bar{x}, \bar{y}) \in S'
\]

\[
\bar{V}, \bar{W} \geq 0
\]

with \( x_{N+1} = 0 \) and \( a_{i,N+1} = 0 \), for all \( i \in \{1, ..., I\} \) and where \( X_n \) and \( Y_i \) are random variables representing the number of OPs already booked on day \( n \) and the number from priority class \( i \) waiting to be booked respectively. By \( E_\alpha[X_n] \) we mean \( \sum_{x,y} \alpha(x,y) x_n \). However, it is reasonable to assume that \( \alpha(x,y) = \alpha(x)\alpha(y) \) and that \( \alpha(y) \) can be chosen to equal the demand distribution by priority class.

Though we now have a linear program with a reasonable number of variables, the number of constraints remains intractable. This suggests solving the dual through column generation. Solving the dual has the advantage of a reasonable number of constraints but at the expense of creating an intractable number of variables - one for each state-action pair. Column generation solves this problem by starting with a small set \( S' \) of feasible state-action pairs to the dual and then (using the dual prices as estimates for \( W_0, V_n \) and \( W_i \)) finding one or more violated constraints in the primal. It then adds the state-action pair(s) associated with these violated constraints into the set \( S' \) before re-solving the dual. The process iterates until either no primal
constraint is violated or one is "close enough" to optimality to quit. The dual of the above LP can be written as:

\[
\begin{align*}
(3.11) \quad & \min \sum_{(x,y) \in \mathcal{S}} X(x,y,a,z) \left( \sum_{i,n} c(i,n)a_{in} + \sum_{i=1}^{I} h(i)z_i \right) \\
& \text{subject to} \\
(3.12) \quad & (1 - \gamma) \sum_{(x,y) \in \mathcal{S}} X(x,y,a,z) = 1 \\
(3.13) \quad & \sum_{(x,y) \in \mathcal{S}} X(x,y,a,z) \left( x_n - \gamma x_{n+1} - \gamma \sum_{i=1}^{I} a_{i,n+1} \right) \geq E_{\alpha}[X_n] \quad \forall n = 1, \ldots, N \\
(3.14) \quad & \sum_{(x,y) \in \mathcal{S}} X(x,y,a,z) (y_i - \gamma E_y[Y_i]) \geq E_{\alpha}[Y_i] \quad \forall i = 1, \ldots, I \\
(3.15) \quad & \bar{X} \geq 0,
\end{align*}
\]

where \( X(x,y,a,z) \) is the dual variable associated with the state-action pair \((x,y,a,z)\).

In general, it may be difficult both to find an initial feasible set \( S' \) and to find a violated primal constraint. Fortunately, in this case, an initial feasible state-action pair for the dual consists of a state with no available slots and where all incoming demand is rejected or served through overtime. Finding a most violated primal constraint involves solving the following (tractable) integer program:
\[ z(\mathbf{V}, \mathbf{W}) = \min_{(x,y) \in \mathbb{S}} \left[ \sum_{i,n} c(i,n) a_{in} + \sum_{i=1}^{I} h(i) z_i + \gamma \sum_{n=1}^{N} \left( V_n \sum_{i=1}^{I} a_{i,n+1} \right) \right. \]

\[ \left. - \sum_{n=1}^{N} V_n (x_n - \gamma x_{n+1}) - \sum_{i=1}^{I} W_i \left( y_i - \gamma E[Y_i] \right) - (1 - \gamma) W_0 \right] \]

subject to equations (3.1) to (3.3) as well as the predictable positivity and integrality restrictions on the state vectors \( \mathbf{x} \) and \( \mathbf{y} \).

Due to equation (3.2), we can eliminate the \( y \) variables and decompose the above minimization by day:

\[ z(\mathbf{V}, \mathbf{W}) = \sum_{n=1}^{N} b_n + \gamma \sum_{i=1}^{I} W_i E[Y_i] - (1 - \gamma) W_0 + \sum_{i=1}^{I} (h(i) - W_i) z_i^* \]

where each \( b_n \) has the form

\[ b_n = \min_{\mathbf{a}_{*,n}, \mathbf{x}_n} \left\{ \sum_{i=1}^{I} \left[ c(i,n) + \gamma V_{n-1} - W_i \right] a_{i,n} + \left[ \gamma V_{n-1} - V_n \right] x_n \right\} \]

subject to

\[ x_n + \sum_{i=1}^{I} a_{i,n} \leq C \]

\( x_n \) and \( \mathbf{a}_{*,n} = (a_{1n}, ..., a_{In}) \geq 0 \), integer

with \( \gamma \) and the optimal \( z_i^* \) given by

\[ z_i^* = \begin{cases} 
0 & \text{if } W_i \leq h(i) \\
M & \text{otherwise.} 
\end{cases} \]

(Recall that \( M \) is the maximum number of arrivals for each priority class on any given day.) The form of each \( b_n \) has a nice intuitive interpretation. For each action, \( a_{in} \) there are two potential costs and one benefit - the cost, \( c(i,n) \), associated with a (possibly)
late scan, a cost, $\gamma V_{n-1}$, associated with a loss of available capacity tomorrow and a benefit, $W_i$, associated with removing some OP from the current demand. There is also a net loss or gain, $\gamma V_{n-1} - V_n$, in each day's value.

3. The Form of the Optimal Approximate Value Function

Once we have solved the dual, the prices associated with each constraint give us the optimal value function approximation. What is perhaps surprising is that we can, under mild conditions, give the exact form of the optimal value function approximation in both the scenario with rejected demand and the scenario with overtime without having to solve the approximate LP. Of course, the approximate value function outlined in the theorems below was not apparent initially and only suggested itself after extensive analysis of the optimal solutions to the approximate LP over a variety of parameter combinations. We present both theorems but defer the proofs to chapter 4. The first theorem gives the optimal value function approximation, over all approximations of the form given in equation (3.8), for the scenario where overtime is permitted.

**Theorem 3.1.** Assume that the cost of overtime, $h(i)$, is the same for all priority classes and $T(i)$ is strictly decreasing in $i$. If

\begin{align}
& k(i)[n - T(i)] + \gamma^{n-T(1)} h(1) > \gamma^{T(i)-T(1)} h(1) \\
& \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(1)}}{1 - \gamma} \lambda_i + \sum_{n=1}^{N} \gamma^{n-T(1)} E_a[X_n] > T(1)C + \frac{\gamma C}{1 - \gamma}
\end{align}

for all $n > T(i)$ and for all $i$, and

where $\lambda_i = E_a[Y_i]$ is the arrival rate for demand from priority class $i$, $C$ is equal to capacity and $\gamma$ is the discount rate, then the optimal approximate value function
amongst all linear approximations for the discounted MDP will have the following form:

- \( V_n = \gamma^{n-T(1)+} h(1) \) for all \( n \in \{1, \ldots, N - 1\} \)
- \( V_N = 0 \)
- \( W_i = \gamma^{T(i)-T(1)} h(1) \) for all \( i \)
- \( W_0 = h(1) \left( -T(1)C + \frac{\pi C}{1-\gamma} + \gamma \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(1)}}{1-\gamma} \lambda_i \right) \)

(The assumption that \( E_{\alpha}[Y_i] = \lambda_i \) is not strictly necessary but allows for some insightful interpretation below.) The above form of the optimal approximate value function has some intuitive appeal. From a cost standpoint, there is no difference between the days up to and including \( T(1) \), thus one would expect to value scanning slots on these days equally (\( V_n = V_{n-1} \) for all \( n \leq T(1) \)). Thereafter, the value of a slot tomorrow is equal to \( \gamma \) times the value of a slot today (\( V_n = \gamma V_{n-1} \) for all \( n > T(1) \)) since a spot tomorrow will become a spot today by the next decision epoch. Both these expectations hold true in the above formulation.

Equation (3.16) lends itself to numerous interpretations. If one sets \( i = 1 \) then the interpretation is that the cost of booking a patient on day \( n > T(1) \) and then performing an OT scan \( n - T(1) \) days into the future is greater than the cost of performing an OT scan now. In other words, you do not benefit from postponing the need for overtime by booking patients late. For \( i > 1 \), the interpretation is essentially the same except that the cost of performing an OT scan now is discounted reflecting the fact that the slots freed up by booking a lower priority patient late are less valuable and therefore there is less incentive to do so. Alternatively, equation (3.16) can be interpreted as stating that the cost of booking a priority \( i \) patient on day \( n > T(i) \) plus the value of a scanning slot on day \( n \) is greater than the value of
a scanning slot on day \( T(i) \). In other words, the cost to the system is greater if you book late. Note that the more you value the future, the more likely that equation (3.16) will be satisfied. For example, with \( \gamma = 0.9 \), it would only be violated if the cost of an OT scan is more than ten times the cost of a late booking. If \( \gamma = 0.99 \) then the cost of an OT scan needs to be approximately 100 times that of a late booking before equation (3.16) is violated. There is little reason for a resource manager of a diagnostic facility to discount future costs (especially considering the time periods are days) and thus in the current application, the above condition will undoubtedly be satisfied for any reasonable values of \( k(i) \) and \( h(i) \). This is a significant result as \( k(i) \) may be extremely difficult to quantify exactly. Thus to know that the above form of the value function is the optimal approximation over a wide variety of potential cost parameter values lends much more credence to the results.

Rearranging equation (3.16) yields the following condition on the ratio between the penalty for a late booking, \( k(i) \), and the penalty for OT, \( h(i) \).

\[
\frac{k(i)}{h(1)} > \frac{\gamma^{T(i)-T(1)} - \gamma^{n-T(i)}}{k(i)[n-T(i)]^+}
\]

Note that the right hand side is always positive and decreasing in \( n \). Thus if this condition is violated for day \( m \) then it is violated for all \( n \leq m \). If equation (3.16) is violated then the relative values of \( k(i) \) for each priority class \( i \) become important. Varying the relative values leads to any number of interesting policies that are willing to book patients from each priority class up to a certain number of days late before performing OT. However, once the late penalties get low enough then the emerging policy only uses overtime as a last resort. The closer the discount factor is to one, the quicker one reaches this "last resort" policy. For \( \gamma \) equal to 0.99, there is essentially
only the two potential optimal solutions - the one given in the above theorem and one that sets all $V_n$ and all $W_i$ equal to zero (and thus uses OT as a last resort).

Finally, in traditional DP theory, the solution to the LP is known to be independent of $\alpha$ provided $\alpha$ is strictly positive for all states [41]. However, in approximate DP, it is well known (see [1] and [15]) that this is not the case but the nature of the dependence of the optimal solution on $\alpha$ is not very well understood. Equation (3.17) provides the necessary bound on $\alpha$ in order for the optimal solution to have the desired form. The right-hand side can be viewed as the total available capacity in the infinite horizon setting. Since, capacity before day $T(1)$ is equally valuable, discounting need only start on day $T(1) + 1$. Capacity after day $T(1)$ forms a geometric sequence with the first term equal to $\gamma C$ and the ratio equal to the discount factor, $\gamma$. Thus, total capacity consists of the capacity available in the next $T(1)$ days, $T(1)C$, plus the capacity available thereafter, $\frac{\alpha C}{1-\gamma}$. Demand too forms a geometric series (with ratio $\gamma$) and thus the first term on the left-hand side can be viewed as the total expected demand over an infinite horizon. Finally, if one views $\alpha$ as a probability distribution on the number of initially booked slots per day, then the second term on the left-hand side consists of a discounted function of the total capacity already in use. Thus equation (3.17) states that total expected demand plus expected initial bookings should be greater than total capacity. In other words, there must be reason to believe that one will have to use OT at some point. If $\alpha$ is chosen so that this condition is violated, the dual solution will yield an approximate value function identical to zero. This will of course lead to a myopic booking policy that places every patient into the slot that minimizes the immediate cost.
A similar analysis, for the scenario where demand can be rejected, yields the following theorem.

**Theorem 3.2.** Assume that the cost of rejecting demand, $h(i)$, satisfies

\[ h(i) > \gamma^{T(i)-T(I)} h(I) \]

for all $i < I$ and that $T(i)$ is strictly decreasing in $i$. If

\[ k(i)[n-T(i)]^+ + \gamma^{n-T(I)} h(I) > \gamma^{T(i)-T(I)} h(I) \]

for all $n > T(i)$ and for all $i$ and

\[ \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(I)}}{1-\gamma} \lambda_i + \sum_{n=1}^{N} \gamma^{n-T(I)} E[X_n] > T(1)C + \frac{\gamma C}{1-\gamma} \]

then the optimal approximate value function amongst all linear approximations for the discounted MDP will have the following form:

- $V_n = \gamma^n \gamma^{T(1)-T(I)} h(I)$ for all $n \in \{1, \ldots, N-1\}$
- $V_N = 0$
- $W_i = \gamma^{T(i)-T(I)} h(I)$ for all $i \in \{1, \ldots, I\}$
- $W_0 = \gamma^{T(I)-T(I)} h(I) \left( - T(1)C - \frac{\gamma C}{1-\gamma} + \gamma \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(I)}}{1-\gamma} \lambda_i \right)$

where $n \sqrt{T(1)} = \max(n, T(1))$.

The intuition here is entirely analogous to the overtime scenario and thus will not be re-iterated. The difference between the two optimal approximate value functions stems directly from our assumptions on the cost structure for the two scenarios - namely that $h(i)$ is strictly decreasing in $i$ when demand is rejected but constant when overtime is used.
3.1. Determining the Scheduling Policy. The traditional LP method for directly solving an MDP provides the optimal policy by letting the probability of using action \( a \) in state \( s \) be equal to the value of the dual variable, \( X(s,a) \), divided by the sum of the dual variables over all possible actions in state \( s \). The viability of this method depends on the fact that a direct solution to the LP will have at least one positive variable, \( X(s,a) \), for all states \( s \). Since we solved the approximate LP through column generation, we only visited a very small percentage of all possible states and thus we cannot derive a policy through the above method. However, we can perhaps gain some insight into the optimal policy by examining the optimal dual variables. If we assume that the optimal approximate value function has the form outlined in Theorems 3.1 and 3.2 then one result that emerges from the proof is that there are strict conditions on what state-action pairs can possibly have a positive dual value. Namely, for the scenario involving overtime, the dual variables can only be positive for state-action pairs satisfying:

- \( x_n + a_{1n} = C \) for all \( n \leq T(1) \)
- \( z_i = 0 \) for all \( i > 1 \)
- \( a_{1n} = 0 \) for all \( n > T(1) \)
- \( a_{in} = 0 \) for all \( i > 1, n \neq T(i) \).

This suggests that the optimal policy will only ever use overtime for OP from the highest priority class (OP1). It also suggests that it will never book OP1 patients late and that it will book all other priority class patients on the last day that does not incur a cost. Similarly, for the scenario involving rejected demand, the dual variables can only be positive for state-action pairs satisfying:

- \( x_n + a_{1n} = C \) for all \( n \leq T(1) \)
Again, this suggests that OP1 patients should never be booked late and that the lower priority classes should be booked into the last day that does not incur a late penalty. It also suggests that only the lowest priority class should ever be rejected. However, since we did not visit every possible state, this falls short of providing a full policy. For instance, what should a resource manager do when faced with a situation where there is not enough OP1 demand to fill all available slots before day $T(1)$? Or what should the resource manager do if there is not enough space available on day $T(i)$ to meet all waiting demand from priority class $i$? Thus, we need an alternative method for deriving a booking policy from the optimal approximate value function. Fortunately, this can be done straightforwardly. Given the approximate value function, one can find the optimal action, $(\vec{a}, \vec{z})$, for any given state, $(\vec{x}, \vec{y})$, as needed, by solving the following tractable integer programming problem ($V_0 = 0$):

$$
(3.21) \min_{(\vec{a}, \vec{z}) \in A_{x,y}} \left\{ \sum_{i=1}^{I} \sum_{n=1}^{N} c(i, n) a_{i,n} + \sum_{i=1}^{I} h(i) z_i 
 + \gamma \sum_{d \in D} p(d) \left[ W_0 + \sum_{n=1}^{N} V_n (x_{n+1} + \sum_{i=1}^{I} a_{i,n+1}) + \sum_{i=1}^{I} W_i d_i \right] \right\}
$$

$$= \min_{(\vec{a}, \vec{z}) \in A_{x,y}} \left\{ \sum_{n=1}^{N} \sum_{i=1}^{I} (c(i, n) + \gamma V_{n-1}) a_{i,n} + \sum_{i=1}^{I} h(i) z_i + \text{constant} \right\}
$$

subject to equations (3.1) to (3.3). Thus, the logical means of testing the usefulness of the approximate value function is to allow equation (3.21) to determine the booking
policy in a simulation of the scheduling process. The results are given in the next section.

4. Simulation Results

The coding of the column generation required to solve the approximate LP was done in AMPL with CPLEX as the solver. Time to solution was less than five minutes and involved between 300 and 500 columns for any problem with a booking horizon of 30 days and the average number of arrivals less than 120. (Solving larger instances is quite feasible but we did not explore any bigger than this.) The simulation of the scheduling process was also done in AMPL/CPLEX as it involves solving the integer program given in equation (3.21) with constraints (3.1) to (3.3). We ran the simulation for 20,000 days with a warm-up period of 5000 days (that is statistics are not collected until day 5001).

4.1. A Small Hospital Setting with Rejected Demand. We initially present the scenario where the resource manager may choose to incur a cost in order to reject some demand. The results are based on a small hospital with a capacity for 10 OP scans per day and with 3 OP priority classes. The booking horizon is set at 30 days and demand is assumed to be Poisson with means 5, 3 and 2 for each priority class respectively. In reality, we need to truncate the Poisson distributions in order to keep the state space finite. However, one can set the maximum daily arrivals for each priority class large enough so that the probability contained in the truncated tail of the Poisson distribution is vanishingly small. Setting demand equal to capacity seemed the most reasonable choice given our objective. If demand is less than capacity then optimal scheduling is not as essential as it might be (though obviously it would
still be useful to schedule as optimally as possible due to variation in demand). If demand is set greater than capacity then rejection or overtime costs will simply be higher. In theory, since one does have some leeway in booking, a scenario where capacity and demand are equal might seem to be manageable, however variability in demand generally causes this not to be the case. The daily cost for a late scan is set at 3, 2 and 1 for each priority class and the cost of rejection at 10 more than the cost of booking on day 30. In other words, it is more costly to reject a higher priority OP than a lower one and it is always more costly to reject than to book (ignoring the cost associated with a loss of capacity). These “costs” are essentially costs to patient health rather than financial costs. The maximum recommended waiting time by priority class is 7, 14 and 21 days respectively. The discount rate is .99 as one does not want to sacrifice the future for the present.

Using equation (3.21) together with the action constraints (3.1) to (3.3), we can determine “optimal” booking actions for a simulated random demand. Perhaps surprisingly, the optimal solution results in a simple heuristic policy. First, fill any unused capacity available for tomorrow (since this capacity will be lost if not filled today). Second, if possible, book any remaining priority 1 OP some time in the first week (otherwise on the first available day) but book lower OP classes into the latest available day that does not incur a late penalty (i.e. for $i > 1$, book priority $i$ OP on day $T(i)$). Finally, periodically reject a certain percentage of demand from the lowest priority class. Note that this policy confirms our initial conjectures based on the optimal dual variables.

Under such a scenario, the above booking policy manages to keep waiting times within the recommended lengths but only at the expense of rejecting around 13% of all
priority 3 OP (see Table 3.1). This works out on average to roughly one rejection every 4 days (though of course rejections do not occur uniformly but rather come in spurts). Unfortunately, it does not seem possible to provide the conditions under which these rejections occur without solving the integer program. Thus, implementation of this policy depends on the availability of software capable of solving the integer program. Figure 3.1 shows the waiting times by priority class for the above policy.

**FIGURE 3.1. Waiting Time With Rejection by Priority Class**

![Waiting Time With Rejection by Priority Class](image)

If we run the same policy *but rejecting only if absolutely necessary*, then the results are dramatically worse with over 50% of priority 1 OP being late as well as significant portions of the lower priority classes (see Table 3.1). It is worth noting that, in almost all cases, this is the policy that emerges if the conditions given in Theorem 3.2 are not satisfied. It is also worth noting that the current practice of simply postponing the booking of any demand that cannot be met will in fact perform worse than the “reject as last resort” policy. The only reason costs do not tend to infinity when demand is postponed is that, in reality, demand is negatively correlated with expected waiting
time. Thus, in fact, there is implicit rejection of demand occurring under current practice, but the decision as to when to reject demand is being made not by the resource manager but by the specialist doctors who recommend the OP scans in the first place. Our current model does not capture the dependence of demand on expected waiting time but could be adapted at a later date to do so.

A possible alternative policy is to impose a strict booking limit on top of the above policy (again rejecting patients only if absolutely necessary). To that end, we ran the above booking policy but enforced the condition that an OP of lower priority would only be booked on a given day if the number already booked on that day did not exceed a pre-specified value. Clearly that number would decrease the lower the priority class.

As long as the capacity and the number of priority classes is not too large, the optimal booking limits to impose can be determined through simulation since there are only a relatively small number of options. The resulting optimal booking limits for the scenario outlined above is to place no restriction on the priority 2 class OP but to restrict priority 3 class OP to only one a day. (Of course, there is no reason to place a booking limit on tomorrow’s bookings as those scanning slots will be lost if not filled today.) While coming close to matching the results of the initial policy, the booking limit policy manages to maintain reasonable waiting times only by rejecting 21% of all priority 3 OP and booking half of those OP3 that it accepts late (see Table 3.1).

The fairly intuitive message is that the optimal solution is to closely restrict the lowest priority class in order to allow room for later arriving higher priority demand. Moreover, the above suggests that restricting the lowest priority class in a dynamic
### Table 3.1. A Comparison of Three Scheduling Policies

<table>
<thead>
<tr>
<th>Policy</th>
<th>% Late</th>
<th>% Rejected</th>
<th>Daily Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OP1</td>
<td>OP2</td>
<td>OP3</td>
</tr>
<tr>
<td>Optimal DP Approx.</td>
<td>0.10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Reject as Last Resort</td>
<td>53.17</td>
<td>35.77</td>
<td>24.85</td>
</tr>
<tr>
<td>Booking Limit</td>
<td>0</td>
<td>0.02</td>
<td>49.52</td>
</tr>
</tbody>
</table>

### Table 3.2. Optimal Approximation Policy with Increasing Rejection Costs

<table>
<thead>
<tr>
<th>Rejection Cost</th>
<th>% Late</th>
<th>% Rejected</th>
<th>Average Daily Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OP1</td>
<td>OP2</td>
<td>OP3</td>
</tr>
<tr>
<td>19</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>109</td>
<td>0.56</td>
<td>0.27</td>
<td>69.10</td>
</tr>
<tr>
<td>1009</td>
<td>49.06</td>
<td>88.84</td>
<td>95.40</td>
</tr>
</tbody>
</table>

fashion rather than placing a static booking limit can be more efficient in the long run.

As one would expect, increasing the cost associated with a rejection has the effect of hampering the ability of the booking policy to maintain reasonable waiting times. Table 3.2 presents the impact of increasing the cost of rejection to 100 and 1000 above the cost of booking a priority \( I \) patient on day \( N \).

Note that condition (4.14) in Theorem 3.2 is violated in the bottom two instances and thus a different policy emerges than the one mentioned earlier. In fact, in the second instance, condition (4.14) is violated for \( i = 3 \) but not for \( i < 3 \), thus essentially leading to a policy that is only willing to book the lowest priority class late. In the
third instance, condition (4.14) is violated for all \( i \) and thus we get the "reject as last resort" policy.

There are two obvious difficulties with implementing such a model in practice. First, how does one determine the "cost" of a late scan? Clearly, the direct cost is to patient health but that is next to impossible to accurately quantify. The second difficulty is that rejecting patients outright may not be a viable option.

4.2. A Small Hospital Setting with Overtime Available. The second scenario where \( z \) represents the number of OP served through overtime addresses both of these difficulties simultaneously. In this scenario, the cost of rejection can be set as the cost of an overtime scan. In essence, this is a means of deliberately overbooking. It would be the task of the policy makers to suggest a second date, \( T'(i) \), for each priority class, that would correspond to the point at which the late costs equal the overtime cost. (Incidentally, \( T'(I) \) would also give you the length of the booking horizon.) The daily "cost" of a late scan could then be determined by setting \( k(i) \) at such a level that by the time you are booking \( T'(i) \) days in advance (but not before), the late costs outweigh the overtime costs. Such an explicit formulation of the late penalty is not strictly necessary as one could still leave these costs arbitrary. The emerging optimal policy will remain the same provided the late penalty is large enough to satisfy condition (3.16) in Theorem 3.1. Recall that the cost structure in this case is significantly different as the overtime costs are independent of the priority class.

Once again we get the surprising result that the optimal solution results in a simple heuristic booking policy. First, fill any unused capacity available for tomorrow. Second, if possible, book any remaining priority 1 OP some time in the first week. If
no space is available, serve any remaining priority 1 OP through overtime. Continue to book lower priority OP on the last available day that does not incur a late penalty (i.e. for $i > 1$, book priority $i$ OP on day $T(i)$).

Initially, we assume a cost of thirty dollars for an overtime scan (the cost associated with the overtime salary of two CT technicians for a 15 minute scan) and that an additional wait beyond $T(i)$ of 5, 10 and 15 days for each priority class respectively equates to the cost of overtime. Thus the daily cost for late scans for each priority class is 6, 3 and 2 respectively. Note that unlike the the rejection scenario costs are now related to an objective dollar figure. Patient health is still taken into account by determining how many days passed the recommended limit equates to overtime cost.

The resulting effect on the performance of the linear approximation is given in Table 3.3 along with the performance when the cost of overtime is increased to 90 and 150. Note that all other parameters are identical to the scenario outlined initially.

Table 3.3 suggests that the above policy results in one and a half percent of priority one OP being served through overtime which, over 15 000 days, is roughly 1000 patients or one every 15 days. Thus, if overtime is available, one can maintain reasonable waiting times by removing a small proportion of the highest priority class. Note that the same policy emerged for a wide range of potential costs associated
with overtime. Thus, it is clear that performing a very small proportion of the high priority demand through overtime in order to remove some of the strain on a system can dramatically improve waiting times. Figure 3.2 shows the waiting times by priority class when overtime is allowed.

**Figure 3.2. Waiting Time With Overtime by Priority Class**

The next section seeks to show that these results extend to a larger hospital as well as to a variety of potential demand streams.

**4.3. A Large Hospital Setting.** A hospital such as the Vancouver General Hospital has a capacity of more than 30 OP scans per day. Data available from VGH only provided us with the number of OP scans by priority class per day. It did not give us the length of each exam requested. Thus, though most OP scans are only 15 minutes, the reality is that using 30 scans per day is an underestimate of the actual OP demand for 15 minute scanning slots. This will become more of an issue in Chapter 5 when we extend the model to include IP demand as the length of IP scans varies from 15 minutes to 90 minutes. Nonetheless, increasing the daily capacity to represent a larger hospital with 30 scans per day did not greatly affect
the results as shown in Table 3.4. ("Removed" simply refers to OP who were either rejected or served through overtime depending on the scenario.) The one advantage is that the increased capacity meant that the required percentage of OP served through overtime was significantly smaller. We tested the initial "reject" policy with rejection costs equal to 10 more than the cost of booking on day $N$ and the overtime policy with OT costs set at 90. Demand was still assumed to be Poisson but with mean arrival rates of 15, 10, and 5 for each priority class respectively.

These results would suggest that the policy using overtime is less sensitive to changes in capacity than one based on outright rejection. This robustness of the overtime policy is further underlined if one increases the percentage of OP demand coming from the lower priority classes. (The data we have from the Vancouver General Hospital suggests that the larger demand does actually come from the lower priority classes.) Figure 3.3 compares the average cost for various arrival streams in the case of outright rejection and overtime.

Clearly, the rejection policy fares worse the larger the percentage of the OP population coming from the lowest priority class. The overtime policy, on the other hand, fares reasonably well regardless of the relative frequency of each priority class.

<table>
<thead>
<tr>
<th>Rejection</th>
<th>% Late</th>
<th>% Removed</th>
<th>Daily Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP1</td>
<td>OP2</td>
<td>OP3</td>
<td>OP1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Overtime</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
5. Conclusion

The above analysis suggests that without some reasonable means of removing some portion of demand (be it through overtime or through outright rejection), a system where demand is equal to (or greater than) capacity will run into serious waiting time challenges. The questions facing a resource manager are how much demand must be removed (either through overtime or rejection) as well as what type of demand (which priority class) and when to remove it. Through the use of approximate dynamic programming, we have been able to answer these questions for any reasonably sized hospital - a feat that would have been impossible through traditional dynamic programming. Of course, the "cost" is that we had to approximate the original MDP and thus our solution may in fact be sub-optimal. However, our results suggest that even our approximate results are a substantial improvement over current practice. Moreover, the value functions derived through our approximation yield easily implementable policies that maintain reasonable waiting times while removing
only a small percentage of demand both in the case of overtime and of outright rejection. We further showed that if overtime is available then it is the highest priority class whose demand is occasionally removed while if one has to reject demand then it is the lowest priority class that is culled. Finally, from a policy point of view, the use of overtime has shown itself to be more robust than outright rejection (as a means of maintaining reasonable waiting times for minimal cost) over a variety of capacity constraints and demand streams.

To implement the policies presented here the only necessary data is the arrival distribution of patients by priority class as well as the available capacity, the length of the booking horizon and the maximum recommended waiting times for each class. The resource manager must also agree on the appropriate cost structure (overtime vs. rejection, late booking penalty). One would also need to ascertain that the assumption of independence between demand from various priority classes and between demand on successive days is verified. If it turns out that there is a “day-of-the-week” effect then this may alter the results. There is nothing in the results that depends on the Poisson assumption so alternative arrival distributions should lead to the same results.

Current practice seems to largely depend on postponing demand (by booking further and further in advance or else postponing booking altogether). However, this only compounds the problem. The research in this paper strongly suggests that it is possible to prevent the need for expanding wait lists by judiciously performing very minimal amounts of overtime in order to prevent congestion.

Since we used approximate dynamic programming, it is certainly possible that one might be able to do better if one knew the true value function in the MDP. The next
Chapter will provide, along with the proof for Theorems 3.1 and 3.2, a discussion of the potential gap between the what is achieved through the above approximation and what might have been achieved had the original MDP been tractable.¹

¹A version of this Chapter will be submitted for publication in OR. Patrick J. Puterman M. Queyranne M. Dynamic Multi-Priority Patient Scheduling with Uncertain Demand
CHAPTER 4

The Optimal Approximation and Bounding the Optimality Gap

1. Introduction

Presented in this chapter is a proof of the form of the optimal approximation to the value function (amongst the set of linear approximations):

\[ V(x, y) = W_0 + \sum_{n=1}^{N} V_n x_n + \sum_{i=1}^{I} W_i y_i \]

We analyze both the scenario where OT is an option and the scenario where demand can be rejected. The cost of an OT scan is assumed to be independent of the priority class while the cost of rejected demand is assumed to be decreasing in \( i \). We assume both OT/rejection costs and late booking costs are linear. In both cases, the maximum recommended waiting time, \( T(i) \), for each priority class increases with \( i \) as a high priority patient is a patient who must be served sooner. After presenting the proofs, we provide a discussion of the “cost” associated with having approximated the value function and present a more complicated approximation based on a convex formulation. For completeness, we restate the theorem before giving the proof.

2. The Form of the Optimal Policy with Overtime

THEOREM 4.1. Assume that the cost of overtime, \( h(i) \), is the same for all priority classes and \( T(i) \) is strictly decreasing in \( i \). If

\[ k(i)[n - T(i)]^+ + \gamma^{n-T(1)} h(1) > \gamma^{T(i)-T(1)} h(1) \]

Theorem 4.1.
for all \( n > T(i) \) and for all \( i \), and

\[
\sum_{i=1}^{I} \gamma^{T(i)-T(1)} \frac{\gamma^{[n-T(1)]^+}}{1-\gamma} \lambda_i + \sum_{n=1}^{N} \gamma^{[n-T(1)]^+} E_{\alpha}[X_n] > T(1)C + \frac{\gamma C}{1-\gamma}
\]

where \( \lambda_i \) is the arrival rate for demand from priority class \( i \), \( C \) is equal to capacity and \( \gamma \) is the discount rate, then the optimal approximate value function amongst all linear approximations for the discounted MDP will have the following form:

- \( V_n = \gamma^{[n-T(1)]^+} h(1) \) for all \( n \in \{1, \ldots, N-1\} \)
- \( V_N = 0 \)
- \( W_i = \gamma^{T(i)-T(1)} h(1) \) for all \( i \).
- \( W_0 = h(1) \left( -T(1)C - \frac{\gamma C}{1-\gamma} + \gamma \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(1)}}{1-\gamma} \lambda_i \right) \)

**Proof:** The outline of the proof is as follows. We first assume the above form to the value function and show that it gives a feasible solution to the primal LP. We then construct a dual solution that together with the above primal solution satisfies complementary slackness. We do this by determining what state-action pairs result in the primal constraint being tight and then constructing a dual solution that is only positive for this set of state-action pairs and for which all dual constraints are tight (since our primal solution is positive except for \( V_N \)). This construction is sufficient to prove the optimality of the above primal solution [7].

We begin therefore by proving the feasibility of the hypothesized primal solution. Clearly, it gives non-negative values for \( \bar{V} \) and \( \bar{W} \). Recall that the constraint for the primal LP can be written as:

87
\[(1 - \gamma)W_0 \leq \sum_{n=1}^{N} \left( \sum_{i=1}^{J} [c(i, n) + \gamma V_{n-1} - W_i] a_{i,n} + [\gamma V_{n-1} - V_n] x_n \right) + \gamma \sum_{i=1}^{J} W_i E_a[Y_i] + \sum_{i=1}^{J} (h(i) - W_i) z_i \]

Note that \(\bar{y}\) does not appear in the equation as we have made use of the fact that \(y_t = \sum_{n=1}^{N} a_{in} + z_i\) for all \(i \in \{1, \ldots, I\}\). Substituting the hypothesized solution into the primal constraint and assuming that \(E_a[Y_i] = \lambda_i\) (again not necessary but helps with the interpretation) yields:

\[
(1-\gamma)W_0 \leq -\sum_{i=1}^{J} \gamma^{T(i)-T(1)} h(1) a_{i1} - h(1)x_1 - \sum_{n=2}^{T(1)} (1 - \gamma)h(1)x_n + \sum_{i=1}^{J} \sum_{n=2}^{N} \left( k(i)[n - T(i)]^+ + \gamma^{[n-T(1)-1]^+} h(1) - \gamma^{T(i)-T(1)} h(1) \right) a_{in} + \sum_{i=1}^{J} \gamma^{T(i)-T(1)+1} h(1) \lambda_i + \sum_{i=1}^{J} (h(i) - \gamma^{T(i)-T(1)} h(1)) z_i
\]

for all \((\bar{a}, \bar{z}) \in A_{\bar{x}, \bar{y}}\) and \((\bar{x}, \bar{y}) \in S\). This follows from the fact that \(V_{-1} = 0\) and \(\gamma V_{n-1} - V_n = 0\) for all \(n > T(1)\). To prove primal feasibility, we simply need to show that the above definition for \(W_0\) satisfies equation (4.3) when the righthand side (RHS) is at its most negative. Fortunately, we can find a state-action pair that minimizes the RHS on a day to day basis. For each day the choice is between filling the day with previous bookings \((x_n)\) and/or with new arrivals \((a_{in})\) or simply leaving the day empty.
For \( n = 1 \), the coefficient in equation (4.3) for \( x_1 \) is at least as negative as the coefficient for \( a_{i1} \) with equality only if \( i = 1 \). Thus, any state-action pair minimizing the RHS will have \( x_1 + a_{11} = C \). In other words, we are indifferent between states where day 1 is fully booked with some combination of previous bookings and current priority 1 arrivals.

For \( n = 2, ..., T(1) \), we know that \( k(i)[n - T(i)]^+ = 0 \) for all \( i \). Thus, by inspecting the coefficient for \( x_n \), any state-action pair minimizing the RHS, will have \( x_n = C \) if

\[
(1 - \gamma) \geq \gamma^{T(i) - T(1)} - \gamma \quad \text{which is clearly true for all } i \text{ with equality for } i = 1.
\]

Again, in the case of \( i = 1 \), we are indifferent between any state where day \( n \) is fully booked with some combination of previous bookings and current priority 1 arrivals. Thus any state-action pair that minimizes the RHS will have \( x_n + a_{1n} = C \) for all \( n \leq T(1) \).

For \( n > T(1) \), the coefficient for \( a_{in} \) in equation (4.3) equals zero at \( n = T(i) \) for each \( i \geq 2 \). If \( n < T(i) \), the coefficient is clearly positive for each priority class since \( \gamma < 1 \). On the other hand, if \( n > T(i) \) then equation (4.1) insures that the coefficient is also positive. Thus, any state-action pair minimizing the RHS of equation (4.3) will have \( a_{in} = 0 \) for all \( n \neq T(i), i > 1 \). Moreover, it will have \( a_{1n} = 0 \) for all \( n > T(1) \).

The term in equation (4.3) involving the OT actions, \( z_i \), implies that any state-action pair minimizing the RHS will have \( z_i = 0 \) for all \( i > 1 \) and that, again, we are indifferent in the case of \( i = 1 \) (though \( z_1 \) can only be as large as the number of waiting priority 1 patients).

Using the above conditions on the set of state-action pairs minimizing the RHS of equation (4.3), one can show that \( W_0 \) has the proposed form by computing equation (4.3) with the appropriate values for \( \bar{a}, \bar{x} \) and \( \bar{z} \). Moreover, we have demonstrated that
the proposed primal solution will have tight constraints only for those state-action pairs satisfying the following conditions:

- \(x_n + a_{1n} = C\) for all \(n \leq T(1)\)
- \(z_i = 0\) for all \(i > 1\)
- \(a_{1n} = 0\) for all \(n > T(1)\)
- \(a_{in} = 0\) for all \(i > 1; n \neq T(i)\).

Thus, complementary slackness implies that the optimal dual solution must be zero for state-action pairs that do not satisfy the above conditions. The next step is to construct a feasible dual solution, \(X(\bar{x}, \bar{y}, \bar{a}, \bar{z})\), with the above properties so that we have a primal-dual pair satisfying complementary slackness. Since the primal variables given above are all non-zero (except for \(V_N\)), this implies that the dual constraints must all be tight (except for the one associated with day \(N\)). If we can find such a dual solution, we will have proved the optimality of the constructed primal-dual pair. In fact, we will impose the further restriction that a dual variable is positive only if, for all \(i\) and \(n\), \(x_n\) and \(a_{in}\) equal zero or \(C\).

Let \((s,a) = (\bar{x}, \bar{y}, \bar{a}, \bar{z})\) represent an arbitrary, feasible state-action pair and let \(B = \{(s,a) | X(s,a) > 0\}\). Recall that the dual constraints have the form:

\[
(1 - \gamma) \sum_{(s,a)\in B} X(s,a) = 1
\]
\[
\sum_{(s,a)\in B} X(s,a) \left( x_n - \gamma x_{n+1} - \gamma \sum_{i=1}^{I} a_{i,n+1} \right) \geq E_{\alpha}[X_n] \quad \forall n = 1, ..., N
\]
\[
\sum_{(s,a)\in B} X(s,a) (y_i - \gamma \lambda_i) \geq \lambda_i \quad \forall i = 1, ..., I
\]
\[
X \geq 0 \quad \forall (s,a) \in S \times A.
\]
Concentrating on equation (4.5), we know that \( E_\alpha[X_N] \) should be zero since at the beginning of each decision epoch day \( N \) stands empty. Thus, we need not put weight on state-action pairs with \( x_N \) non-zero. Thus, for \( n = N - 1 \), equation (4.5) yields

\[
\sum_{(s,a) \in B} X(s,a)(x_{N-1}) = E_\alpha[X_{N-1}] \Rightarrow \sum_{(s,a) \in B} X(s,a) = \frac{E_\alpha[X_{N-1}]}{C}.
\]

Further, for \( n = N - 2 \), equation (4.5) yields

\[
\sum_{(s,a) \in B} X(s,a)(x_{N-2} - \gamma x_{N-1}) = E_\alpha[X_{N-2}]
\Rightarrow \sum_{(s,a) \in B} X(s,a)(x_{N-2}) - \gamma \sum_{(s,a) \in B} X(s,a)(x_{N-1}) = E_\alpha[X_{N-2}]
\Rightarrow \sum_{(s,a) \in B} X(s,a) = \frac{E_\alpha[X_{N-2}]}{C} + \gamma \sum_{(s,a) \in B} X(s,a)
\Rightarrow \frac{1}{C} \left( E_\alpha[x_{N-2}] + \gamma E_\alpha[N-1] \right).
\]

Proceeding similarly, for arbitrary \( n \), we get

\[
(4.8) \quad \sum_{(s,a) \in B} X(s,a) = \frac{E_\alpha[x_n]}{C} + \gamma \sum_{(s,a) \in B} X(s,a) = \frac{1}{C} \sum_{m=n}^{N} \gamma^{m-n} E_\alpha[x_m].
\]

This would be for the case for all \( n > T(1) \) except that bookings can occur for lower priority classes on each \( T(i) \). Thus, for \( n = T(i) - 1 \), there is the added complication that bookings may be made on day \( n + 1 \). This leads to

\[
\sum_{(s,a) \in B} X(s,a)(x_{T(i)-1} - \gamma x_{T(i)} - \gamma a_tT(i)) = E_\alpha[X_{T(i)-1}]
\Rightarrow C \sum_{(s,a) \in B} X(s,a) - \gamma C \sum_{(s,a) \in B} X(s,a) - \gamma C \sum_{(s,a) \in B} X(s,a) = E_\alpha[X_{T(i)-1}].
\]

91
However, since \( z_i = 0 \) for all \( i > 1 \) and \( a_{in} = 0 \) for all \( n \neq T(i) \), equation (4.6) yields

\[
C \sum_{(s,a) \in B, a_1,T(i) > 0} X(s,a) = \sum_{(s,a) \in B} X(s,a)y_i = \frac{1}{1 - \gamma} \lambda_i
\]

\[
\Rightarrow \sum_{(s,a) \in B, a_1,T(i) > 0} X(s,a) = \left( \frac{1}{1 - \gamma} \right) \frac{\lambda_i}{C}.
\]

Thus,

\[
(4.9) \quad \sum_{(s,a) \in B, x_{T(i)-1} > 0} X(s,a) = \frac{1}{C} \left( E_\alpha[X_{T(i)-1}] + \frac{\gamma}{1 - \gamma} \lambda_i \right) + \gamma \sum_{(s,a) \in B, x_{T(i)} > 0} X(s,a)
\]

\[
= \frac{1}{C} \left( E_\alpha[X_{T(i)-1}] + \sum_{m=T(i)}^N \gamma^{m-T(i)+1} E_\alpha[x_m] + \frac{\gamma}{1 - \gamma} \lambda_i \right)
\]

and, in general, for \( n \geq T(1) \),

\[
(4.10) \quad \sum_{(s,a) \in B, x_n > 0} X(s,a) = \frac{1}{C} \left( \sum_{m=n}^N \gamma^{m-n} E_\alpha[x_m] + \frac{1}{1 - \gamma} \sum_{i=1}^I \gamma^{[T(i)-n]} I_{T(i)>n} \lambda_i \right)
\]

where \( I_{T(i)>n} \) equals one if \( T(i) > n \) and zero otherwise.

Finally, for \( n < T(1) \), \( a_{1n} \) may be non-zero. So, equation (4.5) yields:

\[
\sum_{(s,a) \in B} X(s,a)(x_n - \gamma x_{n+1} - \gamma a_{1,n+1}) = E_\alpha[X_n]
\]

\[
\Rightarrow \quad C \sum_{(s,a) \in B, x_{n+1} > 0} X(s,a) = E_\alpha[X_n] + \gamma C \left( \sum_{(s,a) \in B, x_{n+1} > 0} X(s,a) + \sum_{(s,a) \in B, a_{1,n+1} > 0} X(s,a) \right).
\]

However, not all of priority 1 patients are necessarily booked on the same day. This would make the system unsolvable but for the fact that we know that \( x_{n+1} + a_{1,n+1} = C \).
for all \( n \in \{0, ..., T(1) - 1\} \) and for all \((s, a) \in B\). Hence, equation (4.4) yields

\[
\sum_{(s, a) \in B} X(s, a) + \sum_{x_{n+1} > 0} X(s, a) = \frac{1}{1 - \gamma}.
\]

Therefore,

\[
\sum_{(s, a) \in B} X(s, a) = \frac{E_\alpha [X_n]}{C} + \frac{\gamma}{1 - \gamma} \quad \text{for } n < T(1).
\]

Thus equations (4.10) and (4.11) give the amount of dual weight that must be placed on each day in order to satisfy complementary slackness as well as dual feasibility. All weights are positive and none are greater than the total available weight \( \frac{1}{1 - \gamma} \).

To determine the dual weight to be placed on state-action pairs with positive \( z_1 \), we make use of the fact that

\[
\frac{\lambda_1}{1 - \gamma} = \sum_{(s, a) \in B} X(s, a) y_1 \quad \text{from equation 4.6}.
\]
Therefore,

\[ \sum_{(s,a) \in B} X(s,a)z_1 = \frac{\lambda_1}{1 - \gamma} - \sum_{n=1}^{T(1)} \left( \sum_{(s,a) \in B} X(s,a)a_{1,n} \right) \]

\[ = \frac{\lambda_1}{1 - \gamma} - C \sum_{n=1}^{T(1)} \left( \frac{1}{1 - \gamma} - \sum_{(s,a) \in B \setminus \{s_n - c\}} X(s,a) \right) \]

\[ = \frac{\lambda_1}{1 - \gamma} - CT(1) \frac{1}{1 - \gamma} + C \left( \sum_{n=1}^{T(1)-1} \left[ \frac{E_\alpha[X_n]}{C} + \frac{\gamma}{1 - \gamma} \right] \right) \]

\[ C \left( \sum_{n=T(1)}^{N} \gamma^{n-T(1)} \frac{E_\alpha[X_n]}{C} + \sum_{i=2}^{I} \gamma^{T(i)-T(1)} \left( \frac{1}{1 - \gamma} \right) \frac{\lambda_i}{C} \right) \]

\[ = \frac{\lambda_1}{1 - \gamma} - \frac{CT(1)}{1 - \gamma} + \frac{\gamma C(T(1) - 1)}{1 - \gamma} + \sum_{n=1}^{N} \gamma^{[n-T(1)]^+} E_\alpha[X_n] + \sum_{i=2}^{I} \frac{\gamma^{T(i)-T(1)}}{1 - \gamma} \lambda_i \]

\[ = \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(1)}}{1 - \gamma} \lambda_i - T(1)C - \frac{\gamma C}{1 - \gamma} + \sum_{n=1}^{N} \gamma^{[n-T(1)]^+} E_\alpha[X_n]. \]

Thus,

(4.12) \[ \sum_{(s,a) \in B} X(s,a)z_1 = \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(1)}}{1 - \gamma} \lambda_i - T(1)C - \frac{\gamma C}{1 - \gamma} + \sum_{n=1}^{N} \gamma^{[n-T(1)]^+} E_\alpha[X_n] \]

which is greater than zero by equation (4.2). It is worth noting that simply for dual feasibility, we would only need the above to be greater than or equal to zero. However, if equation (4.12) were to be identically zero then we would have degeneracy and the proposed primal solution would no longer be unique. In fact, the optimal objective function would be zero suggesting that an approximate value function identical to zero would be optimal.

The above argument suggests a weighting scheme for a dual feasible solution that, together with the proposed primal solution, satisfies complementary slackness. It remains to show that it is possible to construct positive dual variables satisfying the
above weighting scheme. However, since there are a near infinite number of state-action pairs (much greater than the length of the booking horizon plus the number of priority classes) and since one can choose to fill (or not fill) each day separately, finding a set of state-action pairs that will satisfy the above weighting scheme must, in fact, be possible - however difficult it may be in practice. A possible algorithm for determining such a dual solution would be the following:

- Begin with a state-action pair with days 1 to \( N - 1 \) full with previous bookings and with \( z_1 \) equal to the maximum number of arrivals for priority class 1.
- Assign weight to this state-action pair equal to the required weight for states with \( x_{N-1} \) positive.
- Create second state-action pair identical to the first except with day \( N - 1 \) empty.
- Give this new state-action pair weight equal to the required weight for states with \( x_{N-2} \) positive minus the weight already assigned to the first state-action pair. (If the assigned weight is negative simply make day \( x_{N-2} \) equal to zero in first state-action pair and assign weight to current state-action pair equal to the required weight for states with \( x_{N-2} \) positive.)
- Proceed similarly removing one day from the set of previous bookings each time.
- Whenever the required weight for states with positive \( z_1 \) is reached, split the current state into two with the only difference between the two being that one will have \( z_1 \) equal to the maximum arrivals for priority class 1 and the other will have \( z_1 = 0 \). Assign to the first the additional weight needed to
meet the required weight for states with positive $z_1$ and the remainder to the second (At what point this will happen cannot be known a priori).

- As soon as day $x_{i,T(i)}$ is empty in the new state-action paper add in the action $a_{i,T(i)}$ equal to $C$ and proceed as before.
- Whenever the required weight for states with positive $a_{i,T(i)}$ is met, repeat the process outlined above for $z_1$.
- Repeat all the way down to a state that has previous bookings only on day one.
- Add one final state-action pair that has no previous bookings but has $a_{1,n} = C$ for all $n \leq T(1)$. Assign weight to this state-action pair equal to $1/(1 - \gamma)$ minus the total weight already assigned.

This algorithm is not terribly illuminating but does provide assurance that a dual solution of the above form can be found which is enough to conclude the proof.

3. The Form of the Optimal Policy with Demand Rejection

Turning now to the scenario where some demand can be rejected and where it is assumed that it is more costly to reject higher priority patients, we restate the analogous theorem to the one given in the scenario involving OT.

**Theorem 4.2.** Assume that the cost of rejecting demand, $h(i)$, satisfies

\[(4.13) \quad h(i) > \gamma^{T(i) - T(I)} h(I)\]

for all $i < I$ and that $T(i)$ is strictly decreasing in $i$. If

\[(4.14) \quad k(i)[n - T(i)]^+ + \gamma^{n-T(I)} h(I) > \gamma^{T(i) - T(I)} h(I)\]
for all $n > T(i)$ and for all $i$ and

$$
(4.15) \quad \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(1)}}{1-\gamma} \lambda_i + \sum_{n=1}^{N} \gamma^{[n-T(1)]+} E_a[X_n] > T(1)C + \frac{\gamma C}{1-\gamma}
$$

then the optimal approximate value function amongst all linear approximations for

the discounted MDP will have the following form:

- $V_n = \gamma^n V^{T(1)-T(I)} h(I)$ for all $n \in \{1, ..., N-1\}$
- $V_N = 0$
- $W_i = \gamma^{T(i)-T(I)} h(I)$ for all $i \in \{1, ..., I\}$
- $W_0 = \gamma^{T(1)-T(I)} h(I) \left( - T(1)C - \frac{\gamma C}{1-\gamma} + \gamma \sum_{i=1}^{I} \frac{\gamma^{T(i)-T(1)}}{1-\gamma} \lambda_i \right)$

where $n \vee T(1) = \max(n, T(1))$.

**Proof:** Again we will begin by proving primal feasibility. If we assume the above form for the optimal approximate value function, the primal constraint yields:

$$
(4.16) \quad (1-\gamma)W_0 \leq -\gamma^{T(1)-T(I)} h(I) x_1 - \sum_{i=1}^{I} \gamma^{T(i)-T(I)} h(I) a_{i1}
$$

$$
+ \sum_{n=2}^{T(1)} \left[ k(i) [n - T(i)]^+ + \gamma^{T(1)-T(I)+1} h(I) - \gamma^{T(1)-T(I)} h(I) \right] a_{in}
$$

$$
- (1 - \gamma) \gamma^{T(1)-T(I)} h(I) x_n
$$

$$
+ \sum_{n=T(1)+1}^{N} \left( k(i) [n - T(i)]^+ + \gamma^{n-T(I)} h(I) - \gamma^{T(i)-T(I)} h(I) \right) a_{in}
$$

$$
+ \sum_{i=1}^{I} \left( \gamma^{T(i)-T(I)+1} h(I) \lambda_i + (h(i) - \gamma^{T(i)-T(I)} h(I)) z_i \right)
$$

As with the OT scenario, to prove primal feasibility we simply need to show that the above definition for $W_0$ satisfies equation (4.16) when the RHS is at its most negative. Again, we can find the state-action pair that minimizes the RHS on a day to day basis.
For $n = 1$, the coefficient in the above primal constraint for $x_1$ is at least as negative as the coefficient for $a_{i1}$ with equality only if $i = 1$. Thus, the state-action pair minimizing the RHS will have $x_1 + a_{11} = C$. In other words, we are indifferent between states that fill day $n$ with any combination of previous bookings and current priority 1 arrivals.

For $n = 2, \ldots, T(1)$, we know that $k(i)[n - T(i)]^+ = 0$ for all $i$. Moreover, the coefficient for $a_{in}, \left(\gamma^{T(1)} - T(I) + 1 - \gamma^{T(i) - T(I)}\right) h(I)$ is positive for all $i > 1$ and equals the coefficient of $x_n$ when $i = 1$. Thus, we are indifferent between states that fill day $n$ with any combination of previous bookings and current priority 1 arrivals. Therefore, any state-action pair that minimizes the RHS will have $x_n + a_{1n} = C$ for all $n \leq T(1)$.

For $n > T(1)$, the coefficient for $a_{in}$ equals zero if $n = T(i)$. If $n < T(i)$, the coefficient is clearly positive. On the other hand, if $n > T(i)$ it is also positive due to equation (4.14). Thus any state-action pair minimizing the RHS of equation (4.16) will have $a_{in} = 0$ for all $i > 1$ and $n \neq T(i)$. Moreover, $a_{1n} = 0$ for all $n > T(1)$.

The coefficient for $z_i$ is $h(i) - \gamma^{T(i) - T(I)} h(I)$ which is strictly positive for all $i < I$ by equation (4.13) and equal to zero for $i = I$. Note that if this condition is not satisfied then the coefficient for $z_i$ might be non-positive for other priority classes leading to a different optimal policy (one that might reject demand from more than one class). However, given equation (4.13) is satisfied, any state-action pair minimizing the RHS of equation (4.16) will have $z_i = 0$ for all $i < I$ and that we will be indifferent to the value of $z_i$ (although it is limited by the number of arrivals).

Using the above conditions on the set of state-action pairs minimizing the RHS of equation (4.16), one can demonstrate that $W_0$ has the form given initially simply by computing equation (4.16) with the appropriate values for $\bar{a}, \bar{x}$ and $\bar{z}$. Moreover,
we have demonstrated that the proposed primal solution will have tight constraints only for those state action pairs satisfying:

- \( x_n + a_{1n} = C \) for all \( n \leq T(1) \)
- \( z_i = 0 \) for all \( i < I \)
- \( a_{1n} = 0 \) for all \( n > T(1) \)
- \( a_{in} = 0 \) for all \( i > 1, n \neq T(i) \).

Again, complementary slackness implies that the optimal dual must be zero for state-action pairs not satisfying the above conditions. Moreover, since the primal variables given above are all non-zero (except for \( V_n \)), this implies that the dual constraints must all be tight. If we can construct such a dual solution, we will have proved the optimality of the constructed primal-dual pair. We first determine the amount of dual weight placed on states with positive \( z_i \) by equating the primal and dual objective functions:

\[
\sum_{(s,a) \in B} X(s,a)h(I)z_I = W_0 + \sum_{n=1}^N V_n E_\alpha[X_n] + \sum_{i=1}^I W_i E_\alpha[Y_i]
\]

\[
= \gamma^{T(1)-T(I)}h(I) \left( -T(1)C - \frac{\gamma C}{1-\gamma} + \gamma \sum_{i=1}^I \gamma^{T(i)-T(1)} \lambda_i \right)
\]

\[
+ \sum_{n=1}^N \gamma^{\max(T(1),n)-T(I)}h(I)E_\alpha[X_n] + \sum_{i=1}^I \gamma^{T(i)-T(I)}h(I)\lambda_i
\]

\[
= \gamma^{T(1)-T(I)}h(I) \left( \sum_{n=1}^N \gamma^{[n-T(1)]+} E_\alpha[X_n] + \sum_{i=1}^I \gamma^{T(i)-T(1)} \lambda_i \right)
\]

\[
- T(1)C - \frac{\gamma C}{1-\gamma}
\]
which is positive by equation (4.15). If the weighted sum of $E_\alpha[X_n]$ is not large enough (i.e. equation (4.15) is not satisfied), the dual problem may be infeasible or degenerate.

Turning to equation (4.5), we again know that $E_\alpha[X_N]$ should be set to zero since at the beginning of each decision epoch day $N$ stands empty. Thus, we need not put any weight onto state-action pairs with $x_N$ non-zero.

For $n = N - 1$, the dual constraint (4.5) yields

$$\sum_{(s,a) \in B} X(s,a)(x_{N-1}) = E_\alpha[X_{N-1}] \Rightarrow \sum_{(s,a) \in B} X(s,a) = \frac{E_\alpha[X_{N-1}]}{C}.$$  

For $n = N - 2$, the dual constraint (4.5) yields

$$\sum_{(s,a) \in B} X(s,a)(x_{N-2} - \gamma x_{N-1}) = E_\alpha[X_{N-2}]$$

$$\Rightarrow \sum_{(s,a) \in B} X(s,a)(x_{N-2}) - \gamma \sum_{(s,a) \in B} X(s,a)(x_{N-1}) = E_\alpha[X_{N-2}]$$

$$\Rightarrow \sum_{(s,a) \in B} X(s,a) = \frac{E_\alpha[x_{N-2}]}{C} + \gamma \sum_{(s,a) \in B, x_{N-1} > 0} X(s,a)$$

$$= \frac{1}{C} \left( E_\alpha[x_{N-2}] + \gamma E_\alpha[N - 1] \right).$$

Proceeding similarly, for arbitrary $n$, we get

$$\sum_{(s,a) \in B, x_n > 0} X(s,a) = \frac{E_\alpha[x_n]}{C} + \gamma \sum_{(s,a) \in B, x_{n+1} > 0} X(s,a) = \frac{1}{C} \sum_{m=n}^N \gamma^{m-n} E_\alpha[X_m].$$

This would be for the case for all $n > T(1)$ except that bookings can occur for lower priority classes on each $T(i)$. Thus, for $n = T(i) - 1$, there is the added
complication that bookings may be made on day $n + 1$. This leads to

$$\sum_{(s,a) \in B} X(s,a)(x_{T(i)} - \gamma x_{T(i)} - \gamma a_i, T(i)) = E_\alpha[X_{T(i)}]$$

$$\Rightarrow C \sum_{(s,a) \in B, x_{T(i)} > 0} X(s,a) = \gamma C \sum_{(s,a) \in B, x_{T(i)} > 0} X(s,a) - \gamma C \sum_{(s,a) \in B, a_i, T(i) > 0} X(s,a) = E_\alpha[X_{T(i)}].$$

However, equation (4.6) gives us that

$$C \sum_{(s,a) \in B, x_{T(i)} > 0} X(s,a) + \sum_{(s,a) \in B, z_i > 0} X(s,a)z_i = \sum_{(s,a) \in B} X(s,a)y_i = \frac{1}{1 - \gamma} \lambda_i$$

$$\Rightarrow \sum_{(s,a) \in B, z_i > 0} X(s,a) = \left( \frac{1}{1 - \gamma} \right) \lambda_i - \frac{1}{C} \sum_{(s,a) \in B, z_i > 0} X(s,a)z_i.$$

Thus, for $i = I$, we have

$$\sum_{(s,a) \in B, x_{T(i)} > 0} X(s,a)$$

$$= \frac{1}{C} \left( E_\alpha[X_{T(i)}] + \frac{\gamma}{1 - \gamma} \lambda_i - \gamma \sum_{(s,a) \in B, z_i > 0} X(s,a)z_i \right) + \gamma \sum_{(s,a) \in B, a_i, T(i) > 0} X(s,a)$$

$$= \frac{1}{C} \left( \sum_{m = T(i) - 1}^N \gamma^{m - T(i) + 1} E_\alpha[X_m] + \frac{\gamma}{1 - \gamma} \lambda_i - \gamma \sum_{(s,a) \in B, z_i > 0} X(s,a)z_i \right).$$

In general, for $n \geq T(1)$, we have

$$\sum_{(s,a) \in B, x_n > 0} X(s,a) = \frac{1}{C} \left( \sum_{m = n}^N \gamma^{m - n} E_\alpha[X_m] + \frac{1}{1 - \gamma} \sum_{i = 1}^I \gamma^{T(i) - n} I_{T(i) > n} \lambda_i \right.$$

$$- \gamma^{T(i) - n} I_{T(i) > n} \sum_{(s,a) \in B, z_i > 0} X(s,a)z_i \bigg).$$

since $z_i = 0$ for all $i \neq I$. Finally, for $n < T(1)$, the situation is again complicated by the fact that now $a_{1n}$ may be non-zero. However, we know that for all $n \leq T(1)$,
the day has to be full with either previous bookings or current bookings. Thus, for 

\( n < T(1) \), we have

\[
\sum_{(s,a) \in B} X(s,a)(x_n - \gamma x_{n+1} - \gamma a_{1,n+1}) = E_{a}[X_n]
\]

\[
\Rightarrow \sum_{(s,a) \in B} X(s,a) = E_{a}[X_n] + \gamma C \left( \sum_{(s,a) \in B} X(s,a) + \sum_{a_{1,n+1} > 0} X(s,a) \right).
\]

But since \( x_{n+1} + a_{1,n+1} = C \) for all \( (s,a) \in B \), equation (4.4) give us

\[
\sum_{(s,a) \in B} X(s,a) + \sum_{a_{1,n+1} = C} X(s,a) = \frac{1}{1 - \gamma}.
\]

Therefore,

\[
(4.19) \quad \sum_{(s,a) \in B} X(s,a) = \frac{E_{a}[X_n]}{C} + \frac{\gamma}{1 - \gamma}.
\]

Thus equations (4.18) and (4.19) give us the amount of dual weight that must be placed on each day in order to satisfy complementary slackness as well as feasibility.

The above argument suggests a weighting scheme for a dual feasible solution that, together with the proposed primal solution, satisfies complementary slackness. Again it remains to show that one can easily construct a feasible dual solution satisfying the above weighting scheme but the line of reasoning is no different and a similar algorithm for determining the dual solution can be applied. This concludes the proof.

4. How Good is the Approximation?

Perhaps the greatest challenge facing approximate dynamic programming is to determine the size of the gap between the cost associated with a policy derived from the approximate DP and the cost that would be incurred under the true optimal
policy. Unfortunately, current bounds are far from tight [15]. One alternative is to compare the results with that of current practice over a simulated data set [2]. However, in our case, current practice does not follow any discernible rule and thus defies simulation. Another option is to compare the outcome of the proposed booking policy with that of current practice over a known historical data set. This too is proving difficult in our current situation as hospitals have very little accurate data on OP waiting time. Though we have data giving the arrival rates for OP by priority class by day, there is no record of the state of the system at the beginning of data collection thus making comparison difficult. The simulation results given in Chapter 3 avoid this difficulty by providing a long enough “warm-up” period so that the effect of the initial conditions are minimized. The results substantiate the claim that our proposed policy outperforms current practice but it would be nice if we could quantify the gap between what we can achieve through approximation and what might have been achieved had we been able to solve the MDP directly.

4.1. Perfect Foresight. We can get an upper limit on how well we might have done by assuming perfect knowledge of the incoming demand. In such a case, the optimal policy can be solved as a straightforward assignment problem of the following form. Let $y_{it}$ be the number of OPs of priority $i$ that arrive on day $t$ of the simulation. Further, let $a_{itn}$ be the action that determines how many of priority $i$ patients that arrived on day $t$ to book on day $n$. Clearly, to be valid $n$ must be greater than $t$ but less than or equal to $t + N$. The assignment problem then becomes:

$$\min \left\{ \sum_{i,t,n} k(i) [n - t - T(i)]^+ a_{itn} + \sum_i h(i) z_i \right\}$$
subject to

\[
\sum_{i, t} a_{itn} \leq C \quad \forall n
\]

\[
\sum_{t, n} a_{itn} + z_i = \sum_{t} y_{it} \quad \forall i
\]

\[
a_{itn} \geq 0, \text{ integer}
\]

\[
a_{itn} = 0 \quad \forall n \leq t \text{ and } \forall n > N + t.
\]

Unfortunately, for a long simulation run or a large-sized hospital, this assignment problem becomes extremely memory intensive. For shorter simulation runs (1000 days) and a small-sized hospital, perfect foresight with an initially empty queue and the same parameters as the initial simulation run given in the results section of Chapter 3 seems to avoid all costs. Such perfect foresight is however impossibly optimistic and thus only provides a very loose lower bound on actual costs. Moreover, in a larger hospital with more variation in demand, the results are unlikely to be so good.

4.2. The Optimal Objective Value in the Approximate LP. One immediate lower bound on the total discounted cost is the optimal objective value for the approximate LP. Since we have restricted the value function to have a certain form, it is clear that the optimal primal objective value to the approximate LP will be a lower bound on the optimal objective in the original LP. By strong duality, we further get that the optimal dual objective in the approximation will also be a lower bound on the dual objective in the original LP. This dual version of the objective function can be interpreted as the long run total discounted cost and thus our optimal value for the objective in the approximate LP gives us a lower bound on the true discounted
total cost. In theory, we can therefore get a handle on the “cost” associated with having approximated the value function by comparing this lower bound with the total discounted cost incurred during the simulation (line of reasoning follows that of [3]). While this bound is better than “perfect foresight”, it suffers from being highly dependent on our choice of $\alpha$, the distribution on the initial state of the system, and on the demand stream early on in the simulation run. Thus, the total discounted cost incurred during the simulation varies dramatically with $\alpha$ and even varies significantly between simulation runs with the same $\alpha$. If we set $\alpha$ so that the initial booking horizon is full then we get the tightest bound with total discounted cost in the simulation being within 18% of the lower bound ($\pm 3.08\%$). However, if $\alpha$ is chosen to be slightly more realistic so that more capacity is available as one moves further out in the booking horizon then the optimality gap increases to 34% ($\pm 4.83\%$). The intuition is that the LP formulation does not fully capture the time dynamics so that it best mirrors the “true costs” when the costs are incurred sooner rather than later.

The above confidence intervals obviously give an upper bound on the optimality gap. If we were able to better capture the true value function in the original MDP one would see both the value of the objective function (the lower bound) increase and the simulated total discounted cost decrease. It is possible that even if one were able to solve the LP exactly, the gap would still be non-zero due to the fact mentioned earlier that the LP does not fully capture the time dynamics.

4.3. A Convex Approximation. The challenge of deriving practical bounds within approximate dynamic programming is an on-going one that we hope to pursue more in the future but there is one final piece of evidence to suggest that our approximation is, in fact, a reasonable facsimile of the true value function. It arises from
pursuing what was originally thought to be a defect in the model - namely, that a linear approximation does not take into account that on any given day, the marginal cost of booking a scanning slot should increase as the number of available slots decreases. Under a linear approximation, going from having two available slots to having one “costs” the same as going from having ten available slots to having nine. One would presume that this is not actually the case. Thus, a more reasonable conjecture of the value function might be non-decreasing and convex in the number already booked on each day \( (x_n) \). For simplicity we maintained a linear approximation in the demand variable.

In developing this new approximation, we took advantage of the fact that any non-decreasing, convex function \( f : \{0, 1, \ldots, C\} \rightarrow \mathbb{R} \) with \( f(0) = 0 \), can be written as:

\[
f = \sum_{v=0}^{C-1} \tau_v \phi_v
\]

with all \( \tau_v \geq 0 \) and where \( \phi_v(x) = \max\{0, x - v\}, v \in \{0, 1, \ldots, C - 1\} \). Thus our convex approximation can be written as

\[
u(\bar{x}, \bar{y}) = W_0 + \sum_{n=1}^{N} \sum_{v=0}^{C-1} \tau_{vn} \phi_v(x_n) + \sum_{i=1}^{I} W_i y_i.
\]

Precisely as before, we can then reformulate the LP in terms of this approximate value function:

\[
\max_{W_0, \tau, \tilde{\sigma}, \tilde{x}, \bar{y}} \left( W_0 + \sum_{n=1}^{N} \sum_{v=0}^{C-1} E_\alpha[\phi_v(X_n)] \tau_{vn} + \sum_{i=1}^{I} E_\alpha[Y_i] W_i \right)
\]
subject to

\[ (1 - \gamma)W_0 + \sum_{n=1}^{N-1} \tau_{vn} \left( \phi_v(x_n) - \gamma \phi_v(x_{n+1} + \sum_{i=1}^{I} a_{i,n+1}) \right) \]

\[ + \sum_{i=1}^{I} W_i \left( y_i - \gamma E_d[d_i] \right) \]

\[ \leq \sum_{i,n} c(i,n) a_{in} + \sum_{i=1}^{I} h(i) z_i \quad \forall (\bar{a}, \bar{z}) \in A_{\bar{z}, \bar{g}} \text{ and } (\bar{x}, \bar{y}) \in S \]

where \( \alpha \) is a probability distribution over the initial conditions. Again, we are still left with a very large number of constraints suggesting that we solve the dual by column generation.

The dual of the above LP can be written as

\[ \min_X \sum_{(\bar{x}, \bar{g}) \in S, (\bar{a}, \bar{z}) \in A_{\bar{z}, \bar{g}}} X(\bar{x}, \bar{g}, \bar{a}, \bar{z}) \left( \sum_{i,n} c(i,n) a_{in} + \sum_{i=1}^{I} h(i) z_i \right) \]

subject to

\[ (1 - \gamma) \sum_{(\bar{x}, \bar{g}) \in S, (\bar{a}, \bar{z}) \in A_{\bar{z}, \bar{g}}} X(\bar{x}, \bar{g}, \bar{a}, \bar{z}) = 1 \]

\[ \sum_{(\bar{x}, \bar{g}) \in S, (\bar{a}, \bar{z}) \in A_{\bar{z}, \bar{g}}} X(\bar{x}, \bar{g}, \bar{a}, \bar{z}) \left( \phi_v(x_n) - \gamma \phi_v \left( x_{n+1} + \sum_{i=1}^{I} a_{i,n+1} \right) \right) \geq E_\alpha[\phi_v(x_n)] \]

\[ \forall n = 1, \ldots, N \text{ and } \forall v = 0, \ldots, C - 1 \]

\[ \sum_{(\bar{x}, \bar{g}) \in S, (\bar{a}, \bar{z}) \in A_{\bar{z}, \bar{g}}} X(\bar{x}, \bar{g}, \bar{a}, \bar{z}) \left( y_i - \gamma E_d[d_i] \right) \geq E_\alpha[\psi_w(y_i)] \quad \forall i = 1, \ldots, I \]

\[ X \geq 0. \]
Again we need to start off with a feasible set, \( S' \), of state-action pairs to the dual
and then find a violated constraint for the primal by solving:

\[
(4.23) \quad z(\tilde{r}, \tilde{W}) = \min_{(x, \tilde{z}) \in \mathcal{X}_g, \tilde{z}} \left\{ c(\tilde{a}, \tilde{z}) - \sum_{i=1}^{I} W_i (y_i - \gamma E_i[d_i]) - (1 - \gamma)W_0 \right. \\
- \sum_{n=1}^{N} \sum_{v=0}^{C-1} \tau_{vn} \left[ \phi_v(x_n) - \gamma \phi_v \left( x_{n+1} + \sum_{i=1}^{I} a_{i,n+1} \right) \right] \right\}
\]

subject to equations 3.1 to 3.3.

Due to the equality constraint involving \( y_i \) (and again ignoring the arbitrary limit
on arrival rates), we can still eliminate the \( y \) variables to leave:

\[
(4.24) \quad z(\tilde{r}, \tilde{W}) = \min_{(a, \tilde{z}) \in \mathcal{A}_g, \tilde{z}} \left\{ (c(i, n) - W_i)a_{in} + \sum_{i=1}^{I} (h(i) - W(i))z_i \\
+ \sum_{n=1}^{N} \sum_{v=0}^{C-1} \left[ \gamma \tau_{vn-1} \phi_v \left( x_n + \sum_{i=1}^{I} a_{in} \right) - \tau_{vn} \phi_v (x_n) \right] \\
+ \sum_{i=1}^{I} W_i E_i[Y_i] - (1 - \gamma)W_0 \right\}
\]

subject to

\[
(4.25) \quad x_n + \sum_{i=1}^{I} a_{in} \leq C \quad \forall n, n = 1, \ldots, N
\]

\[
(4.26) \quad \tilde{a}, \tilde{z} \geq 0, \text{ integer .}
\]

As with the linear case, we can decompose the minimization by day to get:

\[
(4.27) \quad \sum_{n=1}^{N} b_n + \sum_{i=1}^{I} (h(i) - W_i)z_i^* + \gamma \sum_{i=1}^{I} W_i E[Y_i] - (1 - \gamma)W_0
\]
where each $b_n$ has the form:

$$
(4.28) \quad b_n = \min_{a*,n,x_n} \left\{ \sum_{i=1}^{I} (c(i,n) - W_i) a_{in} \right\}
$$

$$
(4.29) \quad + \sum_{v=0}^{C-1} \left[ \gamma \tau_{v,n-1} \phi_v(x_n + \sum_{i=1}^{I} a_{in}) - \tau_{vn} \phi_v(x_n) \right]
$$

subject to

$$
(4.30) \quad x_n + \sum_{i=1}^{I} a_{in} \leq C \quad \forall n, n = 1, ..., N
$$

$$
\bar{x}, \bar{a} \geq 0, \text{ integer}
$$

where $\tau_{v,n-1} = 0$ for all $v$ and $n = 1$.

We now have a non-linear objective with linear constraints. However, if we fix $x_n = x$ then each $b_n$ can be transformed into the equivalent IP given below:

$$
(4.31) \quad \min_{a*,n,x_n} \left\{ \sum_{i=1}^{I} (c(i,n) - W_i) a_{in} + \sum_{v=0}^{C-1} \left[ \gamma \tau_{v,n-1} \Psi_v - \tau_{vn} \phi_v(x) \right] \right\}
$$

subject to

$$
(4.32) \quad x + \sum_{i=1}^{I} a_{in} \leq C \quad \forall n, n = 1, ..., N
$$

$$
(4.33) \quad \Psi_v \geq 0 \quad \forall v, v = 0, ..., C - 1
$$

$$
(4.34) \quad \Psi_v \geq x + \sum_{i=1}^{I} a_{in} - v \quad \forall v, v = 0, ..., C - 1
$$

$$
(4.35) \quad \bar{a} \geq 0, \text{ integer}.
$$

Finding the value of $x \in \{0, ..., C\}$ that provides the smallest objective (4.31) is, of course, equivalent to solving (4.23). Once the optimal parameters, $W_0, \tau_{vn}$ and
$W_i$ have been determined the optimal policy, for a given $(x, y)$ vector, reduces to a
convex transportation problem:

$$
\min_{(a, z) \in A(x, y)} \sum_{i=1}^{I} h(i)z_i + \sum_{n=1}^{N} \left[ \sum_{i=1}^{I} c(i, n)a_{in} + \gamma \sum_{\tau=0}^{\tau_{n}} \phi_{\tau} \left( x_{n+1} + \sum_{i=1}^{I} a_{i,n+1} \right) \right] \\
\text{subject to}
$$

\begin{align}
(4.36) \quad & \sum_{i=1}^{I} a_{in} \leq C - x_n \quad \forall n, n = 1, \ldots, N \\
(4.37) \quad & \sum_{n=1}^{N} a_{in} = y_i - z_i \quad \forall i, i = 1, \ldots, I \\
(4.38) \quad & \bar{a}, \bar{z} \geq 0, \text{ integer}.
\end{align}

This is the "local" optimization problem that, once $\tau$ and $W$ are known, can be
solved for any given state $(x, y)$ and which yields the optimal booking policy based
on the approximation. It can be linearized in exactly the same manner as equation
(4.31).

The surprising result is that the added flexibility in this model makes no difference
as the optimal value function chosen by the model turns out to be the exact same linear
model given earlier - at least for the simulation runs we have attempted to date.
This would seem to support the conclusion that the linear model is an adequate
representation of the true value function of the MDP and that the "cost" associated
with having approximated the value function is not in fact too significant. Further
simulation runs would help to strengthen this finding but ideally we would like to
show this analytically in a manner similar to the proofs given above.
5. Conclusion

The implication of the results in this chapter is that one can in fact derive the optimal approximate value without having to perform the arduous task of solving the approximate LP. Moreover, we have shown that for any reasonable ratio between the cost associated with overtime or rejection and the cost of late booking, the optimal value function (and therefore the optimal policy) will have the form outlined in the theorems presented here. Thus, the optimal policies presented in Chapter 3 will in fact remain optimal for a wide range of values for the cost parameters - provided the future is not heavily discounted. Our ability to bound the "cost" associated with having approximated the value function is admittedly the weakest part of the analysis. However, both the bounds derived through the optimal dual objective value and the fact that the convex approximation returned the same linear optimal solution, suggest that our approximation is a reasonable representation of the true value function. To be able to say more than this is the object of future research.¹

¹A version of this Chapter may be submitted for publication. Patrick J. Puterman M. Queyranne M.
CHAPTER 5

Comparing Models

1. Introduction

An astute reader might at this point wonder why one needs the model from Chapter 2 as clearly one could extend the MDP model given in Chapter 3 to include inpatient demand and thus derive a complete scheduling policy. While this is true, and certainly we will present the results of extending the MDP model in this chapter, there is a more significant difference between the model in Chapter 2 and that given in Chapter 3. In Chapter 2, the model begins with a fixed overtime limit and then seeks to assess the impact of that OT limit on the waiting times of OP. In Chapter 3, we begin with recommended maximum waiting times and assess how much overtime is necessary in order to minimize the trade-off between late scans and OT (or rejection). We showed that the optimal scheduling policy can maintain reasonable waiting times with minimal OT (or rejection) and is robust to changes in the cost parameters. Thus, essentially, the two models start from opposite assumptions - the first assumes a limit on OT, the second a (non-strict) limit on maximum waiting time. The first therefore is assessing the impact of a given policy, the second is suggesting the necessary policy to meet given waiting time targets. Moreover, the analytical method is different in the two chapters with the model in Chapter 3 is significantly more general.
2. Extending the MDP to include Inpatients

Extending the MDP to include IP demand simply involves adding an extra priority class with a prohibitively high cost associated with a late booking and with a maximum recommended waiting time of one day. In other words, a priority class that is served the day the demand is placed. There is a slight change in the interpretation of the model as $x_1$ in the original MDP represented tomorrow's slots whereas now it represents today's slots. Thus, days are shifted over one and an additional restriction must be placed to insure that no OP can be booked into $x_1$. We present here the results for the scenario involving OT. The theorems presented in Chapter 3 and proved in Chapter 4 suggest that the optimal policy will involve booking all lower priority patients (in this case all OP priority classes) on the last available day that does not incur a cost and using OT to do any IP demand that cannot be met during regular hours. In fact, this is precisely what emerges when the full model is run in the simulation (see Table 5.1). Of note, is the fact that the cost associated with this policy is much higher than was the case when we did not include IP. This is due to the fact that before our highest priority class could be booked anywhere within the first week without incurring a penalty whereas now, our highest priority class can only be booked on day 1. Thus we no longer have the same flexibility to smooth out the variations in demand.

This would suggest that any flexibility that might be added to the scheduling of IP might very well have a significant impact on costs - as was seen in Chapter 2. Including a LIP category into the MDP model simply involves adding another priority class where the maximum recommended waiting time is 2 days with a prohibitively high cost associated with a late booking. The effect of taking some of the demand
from IP and placing it into this LIP category is to significantly increase throughput and therefore reduce the amount of OT required (see Table 5.1). The results presented in Table 5.1 represent the same model as the initial results in Chapter 3 with OP demand following three independent Poisson distributions with means 5, 3 and 2 respectively. We assume an IP demand that is Poisson with mean 20. In the model that includes a LIP category, we assume 10% of IP demand can be classified as LIP. OT costs are set at 90 with a discount rate of 0.99 and late penalty costs of 1000, 500, 18, 9, 6 for each priority class respectively. (Note that all we really needed was to make the late penalty cost for HIP and LIP greater than the OT cost to insure that no IP was made to waited longer than the recommended time.) We set capacity equal to 30.

In terms of the number of OT scans performed, assuming only 10% of IP can be classified as low priority, yields a drop from an average of 2.1 OT scans per day to 1.7 OT scans per day. Note also, that the advantage of this model is that, unlike the model in Chapter 2, it does not require a pool of on-call OP who can be rescheduled as space becomes available. This is a significant advantage in that the booking policy becomes much simpler.

If we seek to mirror a larger hospital such as VGH then the advantage of adding a LIP category becomes more pronounced (see Table 5.2). Again, recall that the VGH

### Table 5.1. The Impact of LIP on a Small Sized Hospital

<table>
<thead>
<tr>
<th>IP Mix</th>
<th>% Late/% Served through OT</th>
<th>Daily Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIP</td>
<td>LIP</td>
</tr>
<tr>
<td>100% HIP</td>
<td>0/10.69%</td>
<td>-</td>
</tr>
<tr>
<td>90% HIP</td>
<td>0/9.52%</td>
<td>0/0</td>
</tr>
</tbody>
</table>
Table 5.2. A Larger Hospital

<table>
<thead>
<tr>
<th>IP Mix</th>
<th>% Late/% Served through OT</th>
<th>Daily Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIP</td>
<td>LIP</td>
</tr>
<tr>
<td>100% HIP</td>
<td>0/4.75%</td>
<td>-</td>
</tr>
<tr>
<td>90% HIP</td>
<td>0/3.3%</td>
<td>0/0</td>
</tr>
</tbody>
</table>

Data available to us simply gives the number of requests per priority class per day but does not give us the actual number of scanning slots requested by priority class per day. This latter number may be significantly higher as some requests require more than one 15 minute scanning slot. The average number of requests for IP scans per day at VGH is 58.3 while OP demand from the three priority classes was 5.3, 10.9 and 13.8 respectively. While most OP scans are 15 minutes in length, IP scans are often much longer. Thus, it makes sense to increase the number of scans requested by IP to an average of 90 per day while letting OP average demand be 5, 10 and 15 per day respectively. Table 5.2 presents the results.

Again, in terms of the number of OT scans performed, the LIP category yields a drop from an average of 4.3 OT scans per day to 2.7 OT scans per day. Capacity usage has also gone up from 96% to 98%.

Finally, for reasons already mentioned, it is difficult to compare the models from Chapter 2 and Chapter 3. However, in Chapter 2 we did investigate the amount of OT required in order to meet demand for the case with 90% HIP and the case with 100% HIP. In the case of 90% HIP, the OT limit needed to be set at 6 OT scans per day (on average) while in the case of 100% HIP, 8 OT scans per day were needed. Thus, it is worth seeing whether these predictions are in fact born out in the full MDP model when the same parameter values are used as in the initial simulation.
TABLE 5.3. Mirroring Chapter 2 Simulation

<table>
<thead>
<tr>
<th>IP Mix</th>
<th>% Late/% Served through OT</th>
<th>Daily Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIP</td>
<td>LIP</td>
</tr>
<tr>
<td>100% HIP</td>
<td>0/6.48%</td>
<td>-</td>
</tr>
<tr>
<td>90% HIP</td>
<td>0/5.72%</td>
<td>0/0</td>
</tr>
</tbody>
</table>

from Chapter 2. To that end we assumed a capacity of 165 with Poisson demand rate of 120 for IP and 50 for OP. OP demand was broken into the three priority classes with arrival rates 15, 15 and 20. The results are given in Table 5.3.

Dividing the average daily cost by 90 we see in fact that the prediction from Chapter 2 does in fact reflect the results given with the full MDP model in terms of the amount of OT required in order to meet demand.

From a policy standpoint, the full MDP model would seem to suggest a very simple booking policy with an easily quantifiable cost. A further challenge to the resource manager is to decide whether or not the potential reduction in OT cost resulting from an increase in regular hour capacity is worth the financial input required. In fact, our model can easily help answer this question by quantifying the difference in OT cost between a model with capacity \( C \) and a model with capacity \( C + k, k > 0 \). For instance, taking the example from Table 5.3 and increasing regular hour capacity from 165 to 170 yields the results given in Figure 5.1 (taking only the case with 10% LIP).

Thus our model allows the resource manager to determine exactly how much regular hour capacity is optimal and how much OT costs to expect once that investment in regular hour capacity has been made - subject to the proviso that our solution is based on an approximation. It would seem reasonable to suppose that the optimal
solution will not be one involving zero OT as that would lead to significant unused capacity on days with low demand. Of course, how much OT is optimal depends very much on the actual cost of an OT scan and the cost associated with adding additional regular hour capacity.
CHAPTER 6

Further Enhancements and Policy Recommendations

1. Further Enhancements

There are certainly numerous potential enhancements to our optimization model that we hope to pursue in the future. We discuss briefly some of the more interesting ones. These do not represent attempts to "complicate" the model but rather seek to address significant challenges experienced by resource managers charged with scheduling multi-priority patients.

1.1. OP No Shows. A potentially significant source of unused capacity are OP who do not arrive in time for their appointment or simply do not show up at all. It is well-known within the airline industry that the optimal booking policy in the presence of no shows invariably involves overbooking. At present, our model does not explicitly penalize unused capacity. Of course, implicitly, by creating a late penalty, unused capacity is penalized simply because it increases waiting times. However, if we wanted to include a "no show" rate into our model then we would need to include in the reward function a penalty associated with unused capacity which would clearly be affected by the "no show" rate. It would be of interest to determine if such an addition to the reward function does in fact affect the optimal policy.

1.2. Random Service Times. In our current MDP model, each request for a scan is assumed to take exactly 15 minutes. Thus OT is incurred as soon as more patients are booked into a given day than there is capacity. In reality, scanning time
is stochastic and thus the amount of OT incurred is random. It is our belief that the inclusion of random service times would have little impact on the results provided that the expected scanning time is approximately equal to the allotted time. However, it would be worthwhile to verify this.

1.3. Multiple Exam Lengths. As mentioned in Chapter 2, scheduled exam lengths are, in reality, not all 15 minutes in length. Instead, they can be 15, 30, 45, 60 or even 90 minutes (all multiples of a 15 minute scanning slot) in duration. To deal with this added complication, the MDP model would need to be expanded so that each priority class is broken down by exam length. Thus, the state space would be

\[(\bar{x}, \bar{y}) = (x_1, \ldots, x_N, y_{11}, \ldots, y_{1J}, \ldots, y_{J1}, \ldots, Y_{JJ})\]

where \(y_{ij}\) represents the number of priority \(i\) OP waiting to be booked for a scan requiring \(j\) slots and \(J\) is the number of potential exam lengths. The action would consist of where to place incoming demand from each priority level and each exam length and would take the form:

\[(\bar{a}, \bar{z}) = (a_{ijn}, z_{ij})\]

where \(a_{ijn}\) represents the number of priority \(i\) OP requiring a scan of \(j\) slots to book on day \(n\) and \(z_{ij}\) is the number of priority \(i\) OP requiring a scan of \(j\) slots to reject or serve through overtime. The constraints on the action set would similarly need to
be modified to the following:

\[ x_n + \sum_{i=1}^{I} \sum_{j=1}^{J} j \cdot a_{ijn} \leq C \quad \forall n, n = 1, ..., N \]

\[ \sum_{n=1}^{N} a_{ijn} = y_{ij} - z_{ij} \quad \forall i, i = 1, ..., I \text{ and } \forall j, j = 1, ..., J \]

\[ \bar{a}, \bar{z} \geq 0, \text{ integer} \quad \forall i, i = 1, ..., I \quad \& \quad \forall n, n = 1, ..., N. \]

The transition probability would then become

\[(x_1, x_2, ..., x_N; y_1, y_2, ..., y_I) \rightarrow (x_2 + \sum_{i=1}^{I} \sum_{j=1}^{J} j \cdot a_{ij2}, ..., x_N + \sum_{i=1}^{I} \sum_{j=1}^{J} j \cdot a_{ijN}, 0; \bar{d}_1, ..., \bar{d}_I)\]

where \( \bar{d}_i \) now represents the incoming demand for priority \( i \) OP broken down into the exam length categories.

Finally, the cost to the system is given by:

\[ c(\bar{a}, \bar{z}) = \sum_{i,j,n} c(i,n) a_{i,j,n} + \sum_{i=1}^{I} \sum_{j=1}^{J} j \cdot h(i) z_i \]

where \( c(i,n) \) has precisely the same form as before. Of course, in order to implement such a model, one would need to know the arrivals rates not only for each priority class but for each priority class and exam length category. Nevertheless, given that such data is available, the above can be solved in precisely the same manner through approximate dynamic programming. The alternative is simply to view a 45 minute scan as three separate 15 minutes scan. However, this leads to batch arrivals and does not take into account that a 45 minute scan cannot be broken into one 15 minute scan today and two tomorrow. Our suspicion is that the simple heuristic booking policy would remain substantially unchanged but it would be useful to verify this.
1.4. Piece-wise Linear Costs. The reality for most hospitals is that staff salaries for OT are initially time and a half but jump to double time if the amount of OT reaches a certain limit, \( Z \). Introducing such piecewise OT costs would require adding a second rejection/OT variable \( z' \) such that \( z'_i \) equals to zero unless \( z_i \) equals \( Z \). Thus the cost structure would become

\[
c(\vec{a}, \vec{z}) = \sum_{i,n} c(i, n) a_{i,n} + \sum_{i=1}^{I} (h(i) z_i + h'(i) z'_i)
\]

where \( h'(i) \) represents the higher cost associated with double time.

1.5. Patient choice. Another reality facing hospitals is that patients often refuse the first date offered to them and thus may need to be offered a second choice. Adding this to the MDP model would greatly increase its complexity as the decision process would now require at least two decisions for each patient waiting to be booked. Moreover, the transition probabilities would have an additional stochastic element as a patient would refuse the initial date offered with some probability. More realistically, it might be worth incorporating patient choice into the simulation and seeing what potential effect it might have on the average daily cost.

1.6. Demand Dependent on Expected Waiting Time. Perhaps more interesting would be to incorporate the reality that demand for a hospital resource is most likely highly (negatively) correlated with expected waiting time. Thus, as waiting times decrease, demand increases. This could be incorporated into the MDP model simply by introducing balking to the model by making tomorrow's demand \( \vec{y} \) inversely proportional to the number of patients already in the system \( \sum_{n=1}^{N} x_n \).
Recall that in the original model the approximate LP could be written as:

\[
(6.1) \quad \max_{\tilde{V}, \tilde{W}} \left\{ W_0 + \sum_{n=1}^{N} E_\alpha[X_n]V_n + \sum_{i=1}^{I} E_\alpha[Y_i]W_i \right\}
\]

subject to

\[
(1 - \gamma)W_0 + \sum_{n=1}^{N} V_n \left( x_n - \gamma x_{n+1} - \gamma \sum_{i=1}^{I} a_{i,n+1} \right) + \sum_{i=1}^{I} W_i \left( y_i - \gamma E_\alpha[Y_i] \right) \\
\leq c(\bar{a}, \bar{z}) \quad \forall (\bar{a}, \bar{z}) \in A_{\bar{x}, \bar{y}} \text{ and } (\bar{x}, \bar{y}) \in S
\]

\[
\tilde{V}, \tilde{W} \geq 0
\]

with \( x_{N+1} = 0 \) and \( a_{i,N+1} = 0 \), for all \( i \in \{1, ..., I\} \) and where \( X_n \) and \( Y_i \) are random variables representing the number of OPs already booked on day \( n \) and the number from priority class \( i \) waiting to be booked respectively. The only difference emerging when demand is made inversely proportional to the number of patients already in the system is that the expectation over \( \alpha \) of \( Y_i \) will now be conditional upon \( \sum_{n=1}^{N} x_n \).

What effect this might have on the optimal policy is hard to conjecture.

1.7. Average Reward Criteria. We are very interested in developing an analogous model for the average reward MDP. The advantage would be that we avoid the ambiguities introduced by the arbitrary "cost" vector, \( \alpha \), in the discounted version. It will be of interest to see if we can in fact give as strong results in the average reward case as we have been able to in the discounted one. The complication is to determine whether or not the resulting average reward MDP model is unichain or not.

1.8. Multiple Hospitals. There is one last important practical complication that is worthy of consideration. Within Canada, and most likely elsewhere, the
situation of one hospital acting independently is becoming less and less common. Instead, regions are divided into health authorities, each of which may contain a number of hospitals with one centralized booking system. Thus, the challenge facing a resource manager is complicated by the fact that his resources may be spread out throughout the region at various hospitals. Thus, the state of the system at any given decision epoch includes the current booking schedule at each hospital as well as the current demand broken down not only by priority level but also by region. We are interested in the region of the OP simply because we will need to take into account a cost associated with traveling to various hospitals. (We need to be able avoid the absurd situation of sending an OP across the region in order to get a scan done one day earlier.) Thus the state space has the form:

$$(\bar{x}, \bar{y}) = (\bar{x}_1, ..., \bar{x}_n, \bar{y}_1, ..., \bar{y}_I)$$

where $\bar{x}_n = (x_{n1}, ..., x_{nM})$ is a vector consisting of the number of OP already booked on day $n$ for each hospital $m \in M = \{1, ..., M\}$, $\bar{y}_i = (y_{i1}, ..., y_{iJ})$ is a vector consisting of the number of OP of priority $i$ from each geographical region $j \in J = \{1, ..., J\}$ waiting to be booked, $N = $ the maximum number of days one will book in advance, $\tilde{N} = \{1, ..., N\}$, $I = $ the number of OP priority levels, $\tilde{I} = \{1, ..., I\}$. 
The actions are similarly more complicated as the resource manager needs to determine how many OP of each priority level and region to send to each hospital and to each day. Thus, a vector of potential actions can be written as:

\[(\bar{a}, \bar{z}) = \{a_{i,j,n,m}, z_{ij}\},\]

where

\[a_{i,j,n,m} = \text{the number of priority } i \text{ OP from region } j\]

\[\text{to book on day } n \text{ at hospital } m,\]

\[z_{ij} = \text{the number of OP of priority } i \text{ from region } j \text{ to reject.}\]

Of course, to be valid, any action must still satisfy the following two constraints, insuring that the capacity constraint is not violated and that all waiting OPs are booked, as well as constraints insuring that the actions are positive and integer-valued.

\[
(6.2) \quad x_{nm} + \sum_{i=1}^{I} \sum_{j=1}^{J} a_{i,j,n,m} \leq C_m \quad \forall(n,m) \in (\tilde{N}, \tilde{M})
\]

\[
(6.3) \quad \sum_{n=1}^{N} \sum_{m=1}^{M} a_{i,j,n,m} = y_{ij} - z_{ij} \quad \forall(i,j) \in (\tilde{I}, \tilde{J})
\]

\[
(6.4) \quad \bar{a}, \bar{z} \geq 0, \text{ integer.}
\]

The costs associated with such a scheme remain the same as before with an additional term that takes into account the cost of travel.

\[
c(\bar{a}, \bar{z}) = \sum_{i,j,n,m} c(i,j,n,m)a_{i,j,n,m} + \sum_{i=1}^{I} \sum_{j=1}^{J} h(i)z_{ij}
\]
where

\[ x = (x_1, \ldots, x_N), \bar{a} = \{a_{(i,j,n,m)}\} \text{ and } \vec{z} = \{z_{ij}\}, \]

\[ c(i, j, n, m) = k(i)[n - T(i)]^+ + D(j, m), \]

\[ k(i) \text{ is the daily penalty for a OP of priority } i \text{ who waits longer than} \]

the maximum recommended time, \( T(i) \),

\[ h(i) \text{ is the penalty for rejecting an OP of priority } i, \]

\[ D(j, m) \text{ is the cost of sending a OP from region } j \text{ to hospital } m. \]

Given the above formulation, we can give the transitions as:

\[
(\bar{x}_1, \ldots, \bar{x}_N, \bar{y}_1, \ldots, \bar{y}_l) \rightarrow \left( \bar{x}_2 + \sum_{i=1}^{I} \sum_{j=1}^{J} a_{(i,j,2,i)}, \ldots, \bar{x}_N + \sum_{i=1}^{I} \sum_{j=1}^{J} a_{(i,j,N,i)}, 0, \vec{d} \right)
\]

with probability \( p(\vec{d}) = \prod_{i=1}^{I} \left( \prod_{j=1}^{J} p_y(d_{ij}) \right) \) and where \( p(d_{ij}) \) is the probability that \( d_{ij} \) OP of priority \( i \) and from region \( j \) arrive on a given day. Here we are assuming that the demand for priority/region pair is independent of all others.

Finally, the optimality equation is written as:

\[
(6.6) \quad \min_{(\bar{a}, \vec{z}) \in \mathcal{A}_a, \mathcal{G}} \left\{ c(\bar{a}, \vec{z}) \right\} + \gamma \sum_{\vec{d} \in \mathcal{D}} \left[ p_y(\vec{d}) \ast v \left( \bar{x}_2 + \sum_{i=1}^{I} \sum_{j=1}^{J} a_{(i,j,2,i)}, \ldots, \bar{x}_N + \sum_{i=1}^{I} \sum_{j=1}^{J} a_{(i,j,N,i)}, 0, \vec{d} \right) \right]\}
\]

where \( \gamma \) is the discount factor.

From here, we can formulate the model much as we did before by deriving the related LP and approximating the value function by either a linear or convex approximation. Though it is clear that the state space is significantly larger, it seems
reasonable to hope that similar results can be obtained in this more general setting. Future work will seek to establish this.

1.9. Alternative Applications. Finally, we are interested in pursuing a number of possible alternative applications of the model. The most obvious is to surgery scheduling. The complication here is that most surgical scheduling works on a block design that assigns blocks of time to each surgeon who then books his/her patients into those blocks. Thus our model might be most useful to the surgeons themselves as they seek to minimize excessive waiting time for their patients based on a fixed capacity available to them. Moreover, surgeons do have access to OT as they can request more time if needed.

A second potential application would be to the scheduling of cancer radiation treatment. The added complication here is that when one books a patient for treatment, one is booking them for a number of treatment sessions rather than just one. Moreover, some treatments may require daily treatment for a period of days while others may require weekly treatment for a period of weeks. The length of treatment will vary by condition. Thus the patient classes will need to cover not only priority but also type and length of treatment. Given this information on the patients waiting to be booked, the update on the booking horizon, $\bar{\tau}$, though more complicated is nonetheless still entirely deterministic.

Finally, it would be interesting to see what purpose our model might serve in non-health care applications such as airline revenue management and call center management. Airline revenue management is not trading off waiting time and OT costs so, while the method may still be applicable, the objective would have to be modified significantly. Call centers on the other hand are dealing with the trade-off between
waiting time and capacity costs. Thus there is certainly the potential for useful cross-over.

2. Policy Insights

The ultimate goal of this research is, of course, to aid hospital resource managers to more efficiently schedule patients. Thus, it is fitting that we conclude this thesis with an overview of the policy insights to be gleaned from this work.

First and foremost, the salient message of this research is that a small amount of OT and intelligent scheduling can maintain reasonable waiting times. There will clearly be an optimal regular hour capacity that will be sufficient to keep OT costs at a reasonable level but not so large as to lead to excessive amounts of wasted resource time. Our model can help in determining the optimal choice for this trade-off. The intuition is that, in a system where demand and capacity are evenly matched, spikes in demand will cause the booking horizon to reach an undesirable state (i.e. one in which patients are booked later than recommended). If no means is available for relieving the pressure on the system then it is difficult for the resource manager to reduce waiting times back down to acceptable levels. Thus, when the next spike in demand arrives, it simply compounds the problem. However, if a small amount of OT is used judiciously, when a spike in demand occurs waiting times can be maintained within the recommended time frame.

The second policy insight is that the more flexibility present in the scheduling of your highest priority class, the greater your ability to manage waiting times for minimal cost. This was evident when we went from a model having only OP (where the highest priority class could be booked anywhere within the first week without
incurring a cost) to the full model involving IP (who had to be booked on the day the request was placed). Even adding a one day flexibility to only 10% of the IP class can have a significant impact on throughput and therefore on waiting times for all classes of patients. It is worth reminding the reader that this one day flexibility is already present in the priority categorization for CT scans at the site we studied but is currently not being used.

Thirdly, the simulation results in Chapter 3 clearly show that a strict reservation (booking limit) policy can be sub-optimal. Instead, one can do better by dynamically scheduling patients as described in Chapters 3 and 5. This confirms the work of Gerchak, Gupta and Henig who also showed that strict reservation policies can be sub-optimal in the case of two priority classes [22]. Here we demonstrate this result for multiple priority classes.

Fourthly, the research presented in this paper outlines simple booking policies that manage to restrict waiting times to the recommended levels. Due to the fact that we have employed approximations in order to solve the otherwise intractable MDP, we cannot categorically state that these booking policies are in fact optimal. However, we have provided bounds that suggest that these policies are, at least, close to optimal and certainly they would seem to be an improvement over current practice. Moreover, the simulation results suggest that there is very little unused capacity available for improving throughput. Thus, the potential for reducing the amount of OT required is limited. Finally, as health authorities move to using a centralized booking scheme for all hospitals in their region, it will become more and more necessary to have an automated scheduling process. This will require a more formal booking policy than
is currently used in practice. The simplicity of the booking policies that emerge from our research would lend themselves readily to an automated system.

Winston Churchill stated that “men often stumble on the truth but usually pick themselves up and carry on as though nothing had happened”. Our hope is that those resource managers who “stumble” across this research will not dismiss it as the work of academics in their ivory tower but will find it useful in aiding their attempts to more efficiently manage hospital resources.
Bibliography


