Frank van Rest

Advance appointment booking in chemotherapy

Master thesis, August 26, 2011

Thesis advisors:
Dr. F.M. Spieksma, M.L. Puterman, PhD

Mathematisch Instituut
Universiteit Leiden

Centre for Health Care Management
University of British Columbia

Preface

This thesis is the result of 8 months of research performed in order to obtain the M.Sc. degree in Applied mathematics from Leiden University. Eight years ago a new world opened for me when I was introduced to this university. In those eight years I saw all parts of this beautiful place. It started as it should start for all students, in classes for the introductory courses.

In my third year, I had the honour to be chosen board member of study society "De Leidsche Flesch". Here I got the opportunity to develop myself on a whole other dimension. This was also where I was introduced to the world of politics — together with Gonny Hauwert and the other board members we set up a campaign called "Lieve Maria" ("Dear Maria", `www.lievemaria.nl`). This campaign successfully convinced national politics of the necessity of enough and proper mathematical education for high school students. This brought me in contact with the political student party "Bewust en Progressief" for which I was a representative in the University Council in the following year.

All this time I was combining study with entrepreneurship, which might have been a cause for the delay of graduation. Not only had I to share my precious time between study and company, but I even started to doubt the use of finalizing my study at all. After all, why would I need a Master degree, convinced that I would never apply for a job anyway.

> Regret is insight that comes a day too late — H. N. Frye.[1]

I would like to thank my friends and family who helped me get insight before I had to regret. Special thanks to the Mathematical Institute in the person of Marcel de Jeu, who helped me get back on track.

This is when I started to work with Floske Spieksma, who was my bachelor thesis supervisor at the time. I would like to thank her for guiding me

[1] Herman Northrop Frye (1912 - 1991) was a Canadian literary critic and literary theorist, considered one of the most influential of the 20th century.

in the following years. She was the one who brought me in contact with Martin Puterman at the University of British Columbia. He gave me the opportunity to come to Vancouver, Canada and do research at the British Columbia Cancer Agency. I would like to thank Martin Puterman and the CIHR team, especially Pablo, Vincent, Ruben, Antoine and Travis, for the way they welcomed and helped me. This is an experience I greatly cherish.

Coming home after such an experience can be quite harsh. Thankfully my brother Jeroen van Rest, a fresh father, took me by the hand and helped me in the final phase of my research, together with Floske Spieksma, who was guiding me — again – in the final months of writing this thesis. Thank you.

I would like to conclude that I love to say that the last eight years, of which this thesis is the final result, were the best investment I ever did. Not only did I develop myself as a student in the mathematics, but also as a student in life. But first and foremost, I met people whom I can call friends for the rest of my journey.

Leiden, August 2011

Frank van Rest

Abstract

This thesis describes the research done at the BC Cancer Agency in Vancouver, Canada, as part of a CIHR-team, using operations research to solve health care problems. Appointment booking data is used to determine where problems occur in the Chemo unit. One of these problems is the mismatch of capacity and demand. Some days there is slack capacity and other days appointments need to be rescheduled due to lack of capacity. A Markov Decision Process of the process is introduced and via Markov Chain theory a simulation model is introduced. New booking methods and ways to add capacity are tested using this simulation model. We conclude that using a booking method that is a bit more advanced and uses the booking tolerance earlier in the process gives the best results, and is significantly better than adding extra capacity.

Contents

1 Introduction

1.1 Operations Research in Health Care

Healthcare is getting more challenging with more people hospitalized and less people available to deliver the care that is needed. Healthcare systems throughout the world face long and increasing wait times for medical services [19][17][8][3]. These don't always have medical impact, but longer delays might have negative consequences for patients' health. Therefore, there is growing pressure on political leaders to reduce wait times to acceptable levels.

> The First Ministers' Meeting on the Future of Health Care (2004) committed Canada to a program of determining, and then meeting, wait time benchmarks for cancer care, cardiac care, diagnostic imaging, joint replacement and sight restoration. These benchmarks provide "evidence based goals that express the amount of time that clinical evidence shows is appropriate to wait for a particular procedure or diagnostic test" (Postl 2006). Postl (2006), in his final report as Federal Advisor on Wait Times, noted that "we [in healthcare management] have not sufficiently exploited the academic resources available to us from business management schools or industrial engineering." In particular, he singled out operations research (OR) as especially relevant. [9]

1.2 British Columbia Cancer Agency - CIHR Team

The British Columbia Cancer Agency (BCCA) delivers cancer care to British Columbian and Yukon residents (total population is about 4.5 million). It is responsible for the whole spectrum of cancer care from prevention to treatment and rehabilitation. The agency has five centers across the province: one on Vancouver Island, one in Kelowna and three in the greater Vancouver area. The Vancouver Cancer Center (VCC) is the largest. All the data that is used in the research described in this thesis, is from the VCC.

The CIHR Team (Canadian Institutes of Health Research - Team in Operations Research for Improved Cancer Care) is a research team that brings together researchers, clinicians and managers from the University of British Columbia (UBC) and BCCA. The goal of the CIHR team is to develop, test, and implement modern management practices, especially from the field of Operations Research, to increase the efficiency of the cancer system and to enhance patient outcomes.

One of the many cases this team is working on is the VCC Ambulatory Chemotherapy Care Unit (ACCU). This often called 'chemo unit' provides more than 14,000 appointments to over 2,000 patients each year. The ACCU has 9 rooms and 33 chairs to deliver treatment. Each room is supervised by a nurse, who delivers treatment to approximately seven patients each day. During this treatment there is considerable interaction between the patient and the nurses. Each patient's chemotherapy treatment follows one or more of 200 different protocols. These protocols provide specific treatment guidelines including drugs to be used, drug delivery instructions, frequency of appointments, estimated nursing time required and appointment date tolerance. Typically, each protocol involves a series of chemotherapy appointments that are given on a weekly or monthly basis. Some protocols even require treatment on several consecutive days.

After an initial consult when the protocol and the chemo start date is specified, the patient visits his or her oncologist the day before the appointment to identify any complications that may require the need for cancellation or modification of the chemotherapy appointment or its characteristics.

The ACCU scheduling process seeks to provide patients with timely and reliable notification of their appointment so as to eliminate any additional stress for the patient and their family. Of late, due to an increase in patient numbers, this complex task has become more arduous and lead times for patient notification have been decreasing. In the context of the patient focus of the BCCA, this problem has come to the attention of management on several occasions. Further, with a projected further increase in case volumes in the future, the time was ripe for a review of the scheduling process.

The whole scheduling process involves two stages. The first stage of setting a appointment to a date (and not yet a time and nurse) is called 'booking'.

The second stage - where all booked appointments are set to a specific time of the day and appointed to a specific nurse - is called ‘scheduling’.

The CIHR team has developed the tool ‘Chemo Smartbook’ described in [14] to deal with this second stage ‘scheduling’ problem. This tool was implemented in June 2010. The research in this thesis is focused on the first stage ‘booking’ problem.

Since the new scheduling process the clerks at the ACCU schedule at most 60-65 patients per day.[2] The patients are put on a book list, but will not be given a specific time. The schedule for the day will be made a week in advance by another clerk, and the clerk will inform all patients about the appointment.

Info in this section and more is available at `www.orincancercare.org`.

My role in this team was to develop a mathematical approach to the problems they encounter. I was the first student with a more mathematical background in the team. Follow-up research on the work in this thesis is now being done by Post-doc Yasin Gocgun.

[2]When more place is needed a call to management needs to be made, so management can decide whether or not to add the additional personnel required.

2 Context and available data

2.1 Process

The chemo unit at the BC Cancer Agency is facing problems to meet the need to operate within budget. It does not have good insight in the capacity needs. By capacity, we mean the maximum rate at which a resource can deliver a service when operating at peak efficiency [1]. There is a feeling that the capacity is not used to its full potential, i.e. Fridays are perceived as being very busy time, while Mondays can be either very busy or very quiet. Moving appointments around is not easy, since this is constrained by medical guidelines. In general, setting capacity levels entails an unavoidable trade-off between wait times and resource utilization.

Requests for appointments are the result of a longer process:

1. symptoms appear;
2. visit to physician;
3. send to clinic;
4. request for appointment.

What follows is a brief description of the current situation. This information comes from interviews with the head nurse and the CIHR team (see Appendices).

As mentioned in the introduction, the chemo unit has 9 rooms and 33 chairs to deliver treatment. Each room is supervised by at most one nurse. On the first day of the month the head nurse makes the work schedule for all nurses for the following month (e.g. April 1 the schedule is made for May). Therefore the nurses know their work schedule at least a month in advance, which is requested by union rules. This schedule sets the capacity for that month. This schedule only provides the working hours for each nurse and not the specific patients they will treat, and is thus independent of the booked appointments.

The ACCU schedule is comparable to a regular office schedule, i.e. daily from 08:30 till 18:30. No work is done during the weekends.[3]

For this schedule they will try to get full capacity — 9 nurses for 8 hours a day — for each weekday. Already booked appointments - which may predict the load - are not considered during this phase. When we exclude holidays the average number of nursing hours available each day is 53. The standard deviation is 3 hours. Nurses are involved in other tasks (than delivering treatment) as well, which explains why the available amount of nursing hours is less than 9 times 8.

Requests for appointments come in every weekday. There are different types of appointments, namely

- regular appointments;
- clinical trials[4] (they do not have any tolerance);
- new patient appointments (for patients that have there first appointment ever);
- combo treatments (for patients having radiotherapy and chemotherapy simultaneously);
- multiple day appointments (these treatments are multiple days in a row, i.e. Monday, Tuesday and Wednesday).

These appointments have strict time windows: they have a target date and a tolerance (the number of days the appointment may be moved forward or backward in time). The tolerance is defined for the number of days before and the number of days after the target date. For both, it is at most 2 days.

[3]They are planning to add a Saturday morning shift though.

[4]Clinical trials are a set of procedures in research conducted to allow safety and efficacy data to be collected for health interventions

Clerks process the requests for appointments and book them to certain dates. They book it to the target date unless that date is full. When full, they look for other days given the tolerance (time window). A date is full when all appointment slots are taken[5].

Some slots are reserved for special appointments, such as clinical trials, or for new patients. This is to control availability for certain appointment types. E.g. new patients need more time for explanation and to see whether the patient is responding to the treatment as expected. The number of new patients is limited, so that nurses have enough time to give the needed guidance. The reservation for new patients also makes sure that new patients are able to enter the system, since recurring patients tend to overload the system. Recurring patients have appointments for the further future, and thereby filling appointment slots earlier than new patients.

2.2 Performance indicators

In this section we describe important performance indicators. These are partly based on the same performance indicators described in [14]. The research in [14] uses the performance indicator 'confirmation time', i.e. the amount of time in advance the appointment was confirmed with time and location, which is irrelevant to the research in this thesis. We introduce the 'slack capacity' to quantify the waste of resources.

The **slack capacity** (1) is the capacity that is available but not used. E.g. 9 nurses can treat 65 patients a day. When less patients have appointments that day, capacity is not used to its full potential.

The **wait list length** (2) is the number of appointment requests that are above capacity on treatment day.

The percentage of **days with appointment requests exceeding capacity** (3).

[5]For BC Cancer Agency the number of appointment slots per day is 65.

The **wait time for recurring patients** (4) is the time difference between the target date and the actual appointment date for patients that already started their treatment. This time difference should be minimal, since the target date is the date that the treatment has the best effect on the patient's health. When the first appointment in the treatment is set, the further (i.e. recurring) appointment dates follow from this given the treatment protocol.

The **wait time for new patients** (5) is the time difference between their target date and their actual appointment date. The target date is the first day that the patient can be treated and is provided by the physician. This wait time should be minimal as well, since longer wait time keeps the patients from getting their treatment.

These performance indicators are highly dependent. In this thesis we focus on the first three, but other choices could have been made.

2.2.1 Preconditions

The performance indicators are bound by the following preconditions. There are 9 rooms, each occupied by at most 1 nurse at a time. Each room has at most 4 treatment chairs. Pharmacy can do a limited amount of drug preparation each day.

Result of these preconditions is that at most 65 patients can be delivered treatment each day. See section 8.2 for more on this.

2.3 Data description

Four dates are relevant from the data. The 'arrival date' is the moment of the request for the appointment. The 'target date' is the optimal date for the most effective treatment. For the first appointment of a treatment series, this date is provided by the physician. Target dates for further appointments in the same treatment series are provided by the protocol.

The ‘appointment date’ is the date the appointment was realized (in the end). The ‘cancellation date’ is the date of the cancellation, and only available when the appointment is indeed cancelled. When an appointment is cancelled, and then later still continued, the appointment remains cancelled in the database and a new appointment for the same patient is made.

The following table shows the data scheme of the dates.

name	type	required	definition
arrival date	date	yes	The date the appointment request came in.
target date	date	yes	The optimal date for treatment.
appointment date	date	yes	The date the appointment was realized.
cancellation date	date	no	The date the appointment was cancelled.

The following figure shows how lead times can be found by using the above arrival, target, appointment and cancellation dates.

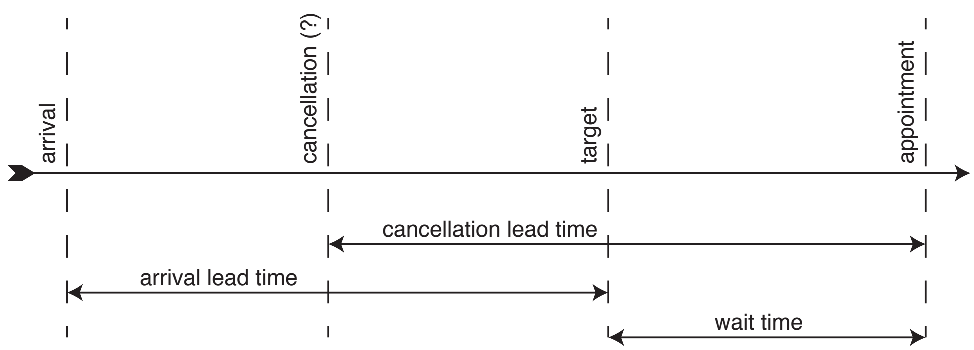

Figure 1: Lead times

The arrival lead time is the time difference between the arrival date and the target date. The cancellation lead time (if relevant) is the time difference between the cancellation date and the appointment date. The wait time is the time difference between the target date and the appointment date.

For the data example in section 2.5 the following dates and lead times are found:

input	
arrival date	September 16, 2008
target date	February 9, 2009
appointment date	February 26, 2009
cancellation date	February 4, 2009

output	
arrival lead time	146 days
cancellation lead time	22 days
wait time	17 days

2.4 Data set

The data set available consists of 225,433 appointments records. 47,729 distinct appointments are being described, for 4,185 different patients. So each appointment has on average 4.5 records. These records indicate changes to the appointment. The whole chemo treatment process of a patient can be reconstructed from these records.

The data covers a period from January 1st, 2009 until January 16, 2011. Not all records for all appointments are available, e.g. it is unknown what happens with unrealized appointments after the period we covered. For the data analysis in section 3 this is taken into account, in other words, sometimes some data is excluded to avoid censoring.

On June 9, 2010 a new scheduling system came into order. Some small irregularities are shown in the data from before and after this date. However these don't have a significant effect on the results described in this thesis.

The data is used for analysis as described in section 3. Other than that, the data is used to feed the parameters of the simulations described in section 5.5.

2.5 Example extract of the data

What follows is an example of data records from the data set. These records describe one single appointment. The different records indicate changes to this appointment. Only relevant columns for this research are present, for simplicity and to preserve patient anonymity.

Appt-ID	modified-on	appointment	item-status
50036	2008-09-16 14:56:48 (†)	2009-02-09 (‡)	
50036	2008-09-16 15:16:17	2009-02-09	
50036	2009-02-02 13:49:42	2009-02-26	
50036	2009-02-04 13:55:54 (⋆)	2009-02-26	C
50036	2009-02-04 14:39:48	2009-02-26 (∗)	C

We see that these records describe one single appointment since the Appt-ID is the same (50036) for all records. The modified-on column indicates the time that a record is added to the database. The latest record indicates the current status of the appointment and earlier records describe the history of this appointment. The appointment column indicates the date the appointment is booked. Item status indicates whether the appointment is cancelled (C) or not.

Records are present for all changes to the appointment: also for changes to fields that are irrelevant to this research, such as notes that clerks add to an appointment[6]. This explains why the first two records (and the last two records) show no relevant changes. In record number 3 the appointment date was changed and in record number 4 the appointment was cancelled.

See the following table for how the dates from section 2.3 are extracted from the data. Symbols are added to this table and the one above for clarity.

[6]Notes often seen are "pt aware", which stands for "patient is aware of appointment date"

record	column	symbol	interpretation
First row	modified-on	†	arrival date
First row	appointment	‡	target date
Last row	appointment	$*$	(effective) appointment date
First cancelled row[7]	modified-on	$\star$	cancellation date

We note that the database table structure is not sufficient for the purpose that it is used for. F.e. crucial information like the target date, tolerance and patient time preferences are now added in a 'notes' field, instead of separate (checked and restricted) fields. This makes the possibility of input mistakes a lot bigger. Also, the target date can be deducted from the first entry in the appointment column, but only because the clerks use it that way. We do a recommendation for another database table structure in section 8.3.

2.6 Computing performance indicators

Now that we have described the raw data, we can describe the underlying metrics in terms of data described in sections 2.3 and 2.4 for the most relevant performance indicators.

Let x_i be the number of appointments with appointment date i. The **slack capacity** (1) for date i is $\max\{C_i - x_i, 0\}$, where C_i is the total capacity for date i. The **wait list length** (2) for date i is $\max\{x_i - C_i, 0\}$.

The percentage of **days with appointment requests exceeding capacity** (3) is the percentage of days with a nonzero wait list length.

[7]The first row with a C in the item-status column

3 Data analysis

The data analysis concentrates on a few aspects including arrival rate of new patients, waiting time for new and recurring patients and indicators of demand and workload. On `http://fvr.me/thesis` all analyses are viewable. What follows are the data characteristics most relevant for this thesis.

3.1 Arrival lead time

In the following chart we have plotted the number of appointments having a certain arrival lead time. The periodicity can be seen: a series of weekly appointments tends to be booked simultaneously for many patients. 95 percent of the appointments have an arrival lead time of 85 days or less. This gives an upper bound for the time window we have to consider for arriving appointments.

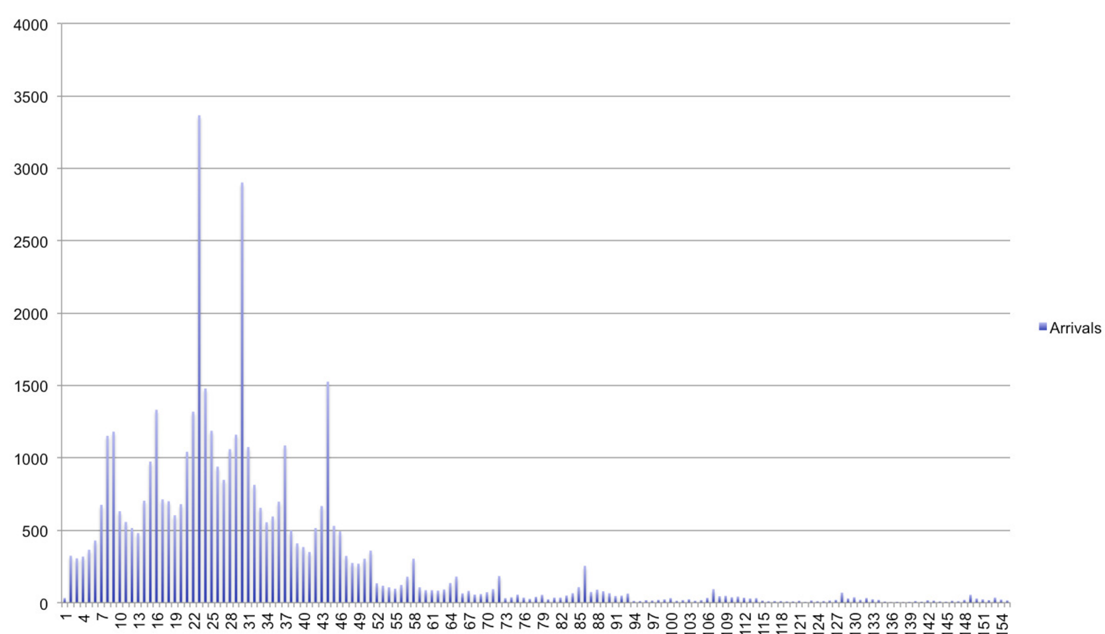

Figure 2: Number of arrivals by arrival lead time (days)

3.2 Cancellation lead time

In the following chart we have plotted the number of cancellations having a certain cancellation lead time. 25 percent of the cancellations occur in the last two days. Overbooking is therefore beneficial and being done. Periodicity in the cancellations can be seen as well. When one appointment is cancelled, the subsequent appointments in the same treatment are cancelled as well.

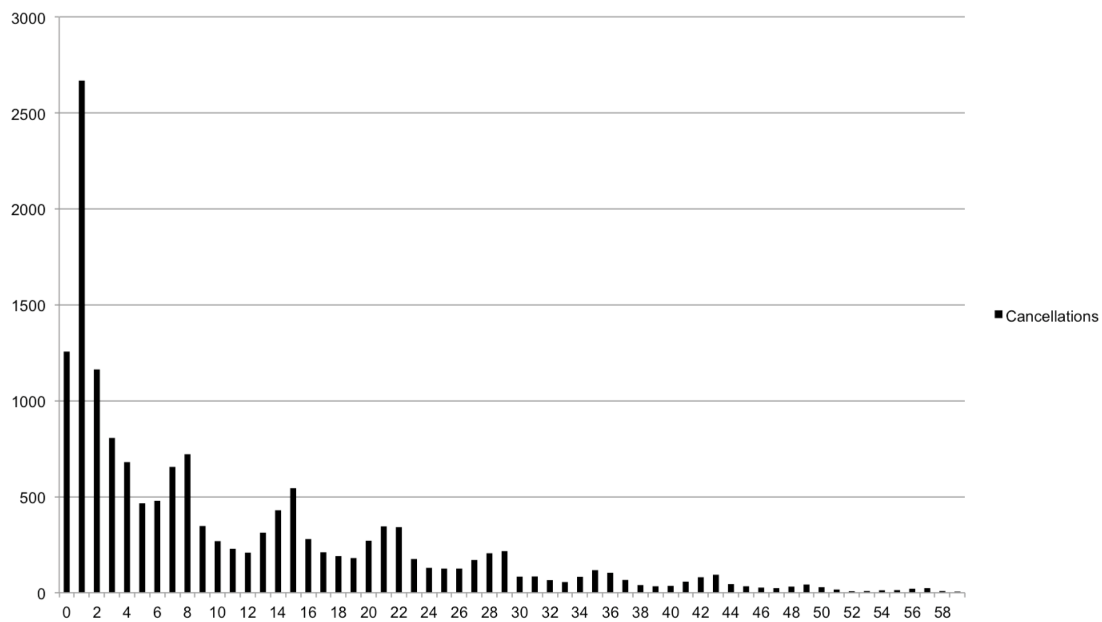

Figure 3: Number of cancellations by cancellation lead time (days)

3.3 Wait time

The following table shows the high percentiles for the wait time for new patients (first treatment) versus recurring patients.

Recurring patients:	
86.9% waits for	$\leq$ 0 days
90% waits for	$\leq$ 1 day
95% waits for	$\leq$ 2 days
99% waits for	$\leq$ 14 days

New patients:	
71.9% waits for	$\leq$ 0 days
90% waits for	$\leq$ 6 days
95% waits for	$\leq$ 10 days
99% waits for	$\leq$ 20 days

3.4 Nursing time utilization

All appointments have a protocol that defines the procedure for that appointment. In this protocol also the expected nursing time is included. This expected nursing time is the average amount of time a nurse has to deal with the patient for that protocol. In other words: the expected nurse time is the time that is stated for a treatment with that protocol. The total expected nursing time for a day is the sum of the expected nursing times of all realized appointments of the day.

For each day the total available nursing time can be calculated by summing the work times of all nurses for that day[8].

Nursing time utilization is the total expected nursing time divided by the total available nursing time. The following table shows the utilization percentage per day of the week.

Monday	77.6%
Tuesday	78.2%
Wednesday	78.9%
Thursday	84.7%
Friday	85.2%

[8]breaks are excluded

From communications with the chemotherapy unit manager we learn that 85% is an attainable goal for the nurse time utilization. We see that slack capacity is available on days other than Thursdays and Fridays and it supports the suggestion that there is room for improvement.

4 Problem description

The data analyses give insight into some problems at the chemo unit. Available capacity is not optimally used, but still some days are too busy causing appointments to be moved. This occurs when capacity and demand do not match. The less they match, the bigger the problems are.

Total available nursing time might seems a good indicator for capacity. However the BC Cancer Agency uses a different indicator, which is the number of appointment slots (65) per day, as described in section 2.2.1.

In reality some days are quiet and the total capacity (65) is not used, other days are busy and appointments need to be moved to another day, which is inconvenient for the patient and therefore lowers patient satisfaction scores [2].

The problem is therefore a mismatch between the capacity and the demand which is expressed in:

- exceeding capacity (postponing appointments), see section 3.3,
- slack capacity (waste of resources), see section 3.4.

The key performance indicators are slack capacity (1) and wait list size (2). The objective in this thesis is to find ways to minimize both. Increasing capacity will reduce the postponing of appointments, but will increase the slack capacity. However, changing the booking strategy will turn out not to have this 'conflicting' effect.

4.1 Approach

The approach to the problem above is to select and design a mathematical model for the process of appointment requests coming in, being booked and possibly being cancelled. The model is used to determine the performance indicators using the provided data on one hand and using different booking methods or methods for adding capacity on the other.

Because of conflicting objectives, we can minimize the slack capacity while putting a constraint on a fraction of wait listed appointments. Another approach is to minimize the product, a sum or a weighted sum of both performance indicators.

Tools available for application are Markov Chain theory, Poisson Processes and Monte Carlo Simulation.

4.2 Possible strategies

Increasing overall capacity (strategy 1) sounds like a good solution to make sure that postponing appointments occurs less frequently. However, this will increase the slack capacity as well. One way to possibly solve this is to add capacity based on demand predictions. Adding capacity can be done in various ways and on various times.

Decreasing overall capacity (strategy 2) might be a solution to decrease slack capacity. However, this will increase the number of appointments postponed, which is an unwanted effect. Because the purpose of the agency is to treat patients, this solution is not further researched in this thesis.

Using tolerance in another way than it's currently used (strategy 3), i.e. checking for tolerance when the target date is full[9], might decrease wait list size while not increasing the slack capacity.

We will investigate possible strategies 1 and 3.

[9] see section 2.1

5 Tool selection and design

5.1 Literature

In this section we describe the tools available to select the needed model. If the relationships that compose the model are simple enough, it may be possible to use mathematical methods to obtain exact information on questions of interest. This is called an analytical solution [10].

The goal is to make effective use of the resources by minimizing certain performance indicators. The mathematical research field which focuses on the effective use of resources is Operations Research [7].

Operations Research in health care often involves queueing theory and scheduling [18]. From queueing theory we learn that even when capacity equals or slightly exceeds average demand, there will be long waits [7]. Scheduling (i.e. job shop scheduling) is used for Chemo Smartbook — the second stage of the scheduling process — described in [14]. It is less relevant for the first stage problem, since we're dealing with an infinite horizon, not only for the jobs, i.e. appointments, but also for the resources, i.e. days. The same is true for queueing theory: we would need a rolling server model, which makes it too complex to analyse by the standard queueing theory.

When models get more and more complex, the standard theory in these research fields no longer apply. Therefore we need to use theory that is more fundamental, like stochastic processes and Markov theory [4]. The research field of Markov Decision Processes (MDP's) can be used to find optimal decisions for complex stochastic models [13].

Discounted Markov Decision Processes with an infinite horizon appears to be the best way to model the problem described, as seen in [12].

Input for these models can be found by statistical analysis of the provided data. Fitting methods are useful to find good estimators. For the appointment arrivals the best distribution has been argued to be a Poisson distribution in [16].

Techniques for using computers to imitate, or simulate, the operations of real-world facilities or processes is called simulation. In a simulation we use a computer to evaluate a model numerically, and data are gathered in order to estimate the desired true characteristics of the model [10].

Monte Carlo methods are a class of computational algorithms that rely on repeated random sampling to compute their results. Monte Carlo statistical methods, particularly those based on Markov chains, are now an essential component of the standard set of techniques used by statisticians [15].

A research field that has common grounds is load balancing, i.e. balancing a workload amongst multiple servers, a research field in computer science [5]. This theory might be applicable when using the tolerance to book appointments.

Another interesting research fields is that of 'yield' or 'revenue' management, which involves with concepts and the selective application of effective strategies and tactics for most hospitality operations [11].

5.2 The complete model

In this research we have used a discounted Markov Decision Process Model with an infinite horizon [12] to model the given problem. We use different priority classes for the different appointment types. These classes can be new versus recurring patients. Another option is to classify on cancer type. The way this classification is done is not relevant at this point. The results provided in this thesis don't depend on the classification.

5.2.1 Decision Epochs and the Booking Horizon

We consider a system that has a daily base capacity to perform

$$B_{total} = \sum_{i=1}^{I} B_i$$

variable-length procedures, where B_i is the number of slots reserved for priority class i and I is the number of priority classes.[10]

Total capacities for each day in the near future are given by

$$C_{ni} = B_i + c_{ni}$$
$$C_n = \sum_{i=1}^{I} C_{ni},$$

where c_{ni} is the capacity addition for day n and priority class i, C_{ni} is the total number of slots reserved for priority class i and I is again the number of priority classes.

At a specific time each day, referred to as the decision epoch, the scheduler observes the day of the month, the number of booked appointments on each future day over an N-day booking horizon to the booking list and to the waiting list, the capacity additions and the number of cases in each priority class to be scheduled. The booking horizon consists of the maximum number of days in advance that hospital management will allow appointments to be scheduled. In practice, this is usually not specified; however, we can set N sufficiently large (but still finite)[11]. Our model is complicated by the fact that the horizon is not static, but rolling. Thus, day n at the current decision epoch becomes day $n - 1$ at the subsequent decision epoch. Because no patient is scheduled more than N days in advance, at the beginning of each decision epoch, the Nth day has no appointments booked.

[10] At BCCA the slots are as follows on a regular Monday: 3 for clinical trials, 3 for new patients, 3 for new patients that also need radiation, 4 for regular patients that also need radiation, 3 for patients that need to have appointments in consecutive days, and 49 other slots

[11] In the BCCA data 95% of the appointments enter the system at most 85 days before their appointment date. Therefore N can be set to 85.

5.2.2 Definitions

n-day appointment: appointments booked n days from now.

n-day arrival: a request for an appointment for n days from now.

n-day cancellation: a cancellation of a appointment that is n days from now.

5.2.3 Notation

x	appointments booked
B	base capacity
c	capacity additions
C	total capacity
y	demand
q	cancellations
d	day of the month
m	number of days in a month (fixed to 20)
N	booking horizon
w	wait listed

5.2.4 The State Space

A typical state takes the form

$$s = (d, x, w, c, y) = (d; x_1, x_2, \dots, x_N; w_1, w_2, \dots, w_N; c_1, c_2, \dots, c_N; y_1, y_2, \dots, y_I),$$

where d is the current day of the month[12],

[12]Since capacity decisions are only made on the first of the month we keep track of the current day in the system

$$x_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nI} \end{pmatrix} \in \mathbb{N}_0^I$$

is a vector of appointments already booked on day n for each priority class i,

$$w_n = \begin{pmatrix} w_{n1} \\ w_{n2} \\ \vdots \\ w_{nI} \end{pmatrix} \in \mathbb{N}_0^I$$

is a vector of appointments on the wait list for day n for each priority class i,

$$c_n = \begin{pmatrix} c_{n1} \\ c_{n2} \\ \vdots \\ c_{nI} \end{pmatrix} \in \mathbb{Z}^I$$

is a vector of capacity additions on day n for each priority class and

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iP} \end{pmatrix},$$

where P is the number of priority i appointments waiting to be booked and

$$y_{ip} = (n, \min, \max), 1 \leq n \leq N, \min \geq 0, \max \geq 0$$

is a vector that holds the target date n and the tolerance for that target date given by min and max. To recall, the tolerance defines the time window given by the physician for this appointment.

The state space, S, is therefore

$$\begin{aligned} S \;=\; & \{(d,x,w,c,y)| \\ & 0 \le d \le m \\ & x_{ni} \le C_{ni}, 1 \le n \le N, 1 \le i \le I; \\ & w_{ni} \le C_{ni}, 1 \le n \le N, 1 \le i \le I; \\ & 0 \le \dim(y_i) \le Q_i, 1 \le i \le I; \\ & (d,x,w,c,y) \in \mathbb{Z}_m \times \mathbb{Z}_{NI} \times \mathbb{Z}_{NI} \times \mathbb{Z}_{NI} \times \mathbb{Z}_I\}, \end{aligned}$$

where m is the length of a month (fixed, 20) and Q_i is the maximum number of priority i arrivals in a given day. (Truncating arrival demand is necessary to keep the state space finite, but the maximum number of arrivals can be set sufficiently high as to be of little practical significance). Appointments in each priority class require one appointment slot. The nursing time required for an appointment is however dependent on the priority class.[13] Since the number of appointments each day (65) is quite high, the total nursing time of the appointments is stationary and not (yet) a big factor in optimizing the performance[14].

5.2.5 The Action Set

There are two categories of action sets. One involves setting the capacity and the other booking.

Nursing capacity Decisions involving the amount of nursing time are being made every m days. In the state space we keep track of the current day

[13] In our model we do not take into account the load an appointment has on the system. Different appointments require different amounts of “nursing time”. Further research is required to incorporate this into the model.

[14] as stated by the CIHR-team

of the month. Day 0 is an artificial day where decisions on nursing capacity are made.

Decisions made on day $d = 0$ are for the days $n \in \{d+m+1, d+m+2, \ldots, d+2m\}$.[15]

A matrix of possible actions can be written as

$$(c') = \{c'_{ni} | 1 \leq n \leq m, 1 \leq i \leq I, C_{lbound} \leq c'_{ni} \leq C_{ubound}\},$$

where c'_{ni} is the capacity addition for day n and priority class i, $C_{lbound} \in \mathbb{Z}$ is the lower bound of the capacity addition and $C_{ubound} \in \mathbb{Z}$ is the upper bound of the capacity addition.[16] These bounds may be negative. Negative capacity addition occurs when capacity is lower than the default capacity.

Booking appointments Decisions involving the scheduling of appointments are being made every day.[17] Thus, a matrix of possible actions can be written as $(a, b) = \{a_{ni}, b_{ni}\}$, where a_{ni} is the number of priority i appointments to book on day n and b_{ni} the number of priority i appointments to put on the wait list for day n.

To be valid, any action must satisfy the following constraints, insuring that the base capacity is not exceeded:

$$\begin{aligned} x_{ni} + a_{ni} \leq\ & C_{ni} \\ & \forall n \in \{1, \ldots, N\} \\ & \forall i \in \{1, \ldots, I\}, \end{aligned} \tag{1}$$

[15] At BCCA these decisions are made the first of each month for the whole next month, so m would be 20. An example to be clear: On January 1st, the capacity for every day of February is set.

[16] At BCCA the capacity is measured in the number of nursing minutes, but a translation to slots seems more intuitive.

[17] This is a generalization of the situation at BCCA, where decisions are made multiple times a day, i.e. whenever a request for an appointment comes in.

that the number of bookings equals the number waiting,

$$\sum_{n=1}^{N}(a_{ni}+b_{ni}) \;=\; |y_i| \qquad \forall i \in \{1,\ldots,I\}, \tag{2}$$

and that all actions are positive and integer,

$$(a,b) \in (\mathbb{Z}_N \times \mathbb{Z}_I, \mathbb{Z}_N \times \mathbb{Z}_I). \tag{3}$$

We denote the action set, A_s, for any given stage, $s=(d \neq 0, x, c, y)$, as the set of actions, (a,b), satisfying Equations 1 to 3.

5.2.6 Transition Probabilities

When $d=0$ the system is in a state where decisions need to be made about the capacity for the next few days. The transitions are deterministic. For action

$$c' = \begin{pmatrix} c'_{m+1} \\ c'_{m+2} \\ \vdots \\ c'_{2m} \end{pmatrix}$$

they are

$$\begin{aligned}
&(0; \\
&x_1, x_2, \ldots, x_N; \\
&w_1, w_2, \ldots, w_N; \\
&c_1, c_2, \ldots, c_N; \\
&y_1, y_2, \ldots, y_I) \;\rightarrow\; (1; \\
&\qquad x_1, x_2, \ldots, x_N; \\
&\qquad w_1, w_2, \ldots, w_N; \\
&\qquad c_1, c_2, \ldots, c_m + c'_m, c_{m+1} + c'_{m+1}, \ldots, c_{2m} + c'_{2m}, c_{2m+1}, \ldots, c_N; \\
&\qquad y_1, y_2, \ldots, y_I).
\end{aligned}$$

Once a booking decision for action (a, b) is made, the stochastic elements in the transition to the next state consist of the number of new arrivals and the number of cancellations in each priority class.

If the number of new arrivals is represented by y' and the number of cancellations by q', then the state transition,

$$\begin{aligned}
(d;&\\
x_1, x_2, \ldots, x_N;&\\
w_1, w_2, \ldots, w_N;&\\
c_1, c_2, \ldots, c_N;&\\
y_1, y_2, \ldots, y_I) &\rightarrow (d+1 \mod m;\\
&\quad x_2 + a_2 - q'_2 + w'_2, \ldots, x_N + a_N - q'_N + w'_N, 0;\\
&\quad w_2 - w'_2 + b_2, w_3 - w'_3 + b_3, \ldots, w_N - w'_N + b_N, 0;\\
&\quad c_2, \ldots, c_N, B;\\
&\quad y'_1, \ldots, y'_I),
\end{aligned}$$

occurs with probability

$$p(y') = \prod_{i=1}^{I} p(y'_i) \prod_{k=1}^{I} \prod_{l=2}^{N} p(q'_{kl}),$$

where w'_i is the number of appointments that is being moved from the wait list to the book list[18],

$$w'_{ni} = \min(q_{ni} + c_{ni} - x_{ni} - a_{ni}, w_{ni}),$$

$p(y'_i)$ is the probability that y'_i priority i appointments arrive on a given day and $p(q'_{kl})$ is the probability that q'_{kl} priority k appointments are being cancelled l days in advance. We assume that the demand for each priority class is independent and that each day's demand is independent as well. We do assume the same for the cancellations.

Because the demand arises from multiple independent sources, independence between classes seems a reasonable assumption.

[18]All appointments on the wait list are moved to the book list as long as there is capacity on the book list

5.2.7 Costs

The costs associated with a given state-action pair derives from four sources: a cost associated with booking a appointment out of the time window, a cost associated with an appointment booked on the wait list, a cost associated with appointments that are still on the wait list on the current day and a cost associated with using (additional) nurse capacity. For action $a = (a, b)$ in state $s = (d, x, w, c, y)$ we represent the costs r as follows:

$$r(a, s) = \sum_{i=1}^{I} \sum_{n=1}^{N} r_{ni}(y_{ni}, a_{ni}, b_{ni}) + \sum_{i=1}^{I} w_{0i} r_{WL} + \sum_{i=1}^{I} c_{0i} r_{cap}$$

where $r_{ni}(y_{ni}, a_{ni}, b_{ni})$ represents a cost function for booking appointments out of their time window and booking appointments to the wait list, r_{WL} is a penalty for appointments that are still on the wait list on the day of the appointment and r_{cap} is the cost for adding one slot of extra capacity.

5.2.8 Not taken into account

This goal of this model has been to describe the process as precise as possible. Some characteristics are however not included in the model.

Nurse time In the model all appointments require one appointment slot. Available nurse time and the required nurse time of the appointments is not taken into account. This choice is made since the BC Cancer Agency itself is not looking at the total available nurse time and the required nurse time per appointment. See section 8 for more on this. Adding nurse time to the model however is quite a simple adjustment and can be done easily.

Periodicity of appointments One patient might have multiple (recurring) appointments (e.g. every week). So in reality, groups of appointments are booked at once and cancelled at once. In the model each appointment is seen individually. This is done because concepts like the arrival lead time and the cancellation lead time are not clear for groups of appointments. The consequence of this is that individual appointments have higher arrival and cancellation rates than the rates for groups of appointments.

Wait time Individual appointments are not tracked throughout this model. Instead counters are used to describe the system. Therefore wait time for individual appointments can not be measured in this model. Therefore performance indicators (4) and (5) can not be measured with this model.

Moving appointments Appointments on the wait list are moved to another day in the (near) future if no capacity is available to treat them. This movement of appointments is not included in the model. This has a decreasing effect on the load in the system.

5.2.9 Further remarks

Unreliability Remarkably the casual staff (i.e. temporary staff) which is used to add capacity when needed, is cheaper than the standard staff. However, they are less reliable, in the sense that it is unclear whether they are available for specific days or hours. Therefore reliability should be added to the cost function. How and to which extent is to this point unclear.

Base level priority slots In the current situation base level priority slots are different for each day of the week. This is not yet included in the model, we now treat every day the same. It is however possible to change the base capacity B_i into a day of the week dependent B_i^{dw} where dw is a 'day of the week' counter, just like d is a day of the month counter.

Arrival rates At BC Cancer Agency the data shows that the arrival of new appointments is not evenly distributed over the week. Therefore arrival rates are different for each day of the week. This is not included in the model, but just like above, this can be added by introducing a 'day of the week' counter.

Opening of slots In practise, one day in advance all priority slots are ignored and waitlisted appointments can go to any slot. F.e. if there are still five slots open for clinical trials, and the wait list still has some 'non-clinical trials'-appointments, these appointments can fill those slots. This is not included in the model. For the results of this thesis this is not relevant, since we're not using different priority classes.

5.3 Simplified Markov Chain Model

The curse of dimensionality of the Markov Decision Process makes the model described in the previous section too hard to solve analytically.

We will therefore consider — as a next step — a new Markov process that is associated with the process of arrivals and cancellations. We create a continuous time Markov Process approximation with parameters for capacity, arrival rates and cancellations rates. For this process we will analyze the embedded Markov Chain and look for the stationary distribution. In this way we can see f.e. the long term effect of increasing capacity or the effect of increasing arrival rates due to ageing of the population. However this approach ends up to be computational intractable as well.

The choice for a continuous time Markov Process makes the transitions a lot simpler. The drawback is that we are bound to use the Poisson process, but as remarked earlier, the Poisson process has been argued to be the best process for describing the arrivals [16].

5.3.1 Parameters

We introduce a continuous time Markov Chain Process with parameters c, y and q. c is a capacity vector,

$$c = \begin{pmatrix} c_1 \\ \vdots \\ c_I \end{pmatrix}$$

where the elements represent the capacity per priority type per day. y is a demand matrix,

$$y = \begin{pmatrix} y_{11} & \cdots & y_{N1} \\ \vdots & \ddots & \vdots \\ y_{1I} & \cdots & y_{NI} \end{pmatrix},$$

where the element y_{ni} represents the rate for requests for a priority i appointment with target date n coming in.

q is a cancellation matrix,

$$q = \begin{pmatrix} q_{11} & \dots & q_{N1} \\ \vdots & \ddots & \vdots \\ q_{1I} & \dots & q_{NI} \end{pmatrix},$$

where the element q_{ni} represents the rate of cancellations for a priority i appointment with target date n.

5.3.2 State space

The state space is defined as S, the collection of booking matrices x, where

$$x = \begin{pmatrix} x_{11} & \dots & x_{N1} \\ \vdots & x_{ni} & \vdots \\ x_{1I} & \dots & x_{NI} \end{pmatrix},$$

with x_{ni} the number of booked priority i appointments for day n. We keep track of wait listed patients simply by checking whether the bookings exceed the capacity for a given priority class.

5.3.3 Transition rates

For the MCP with parameters (c, y, q) the "shift-to-the-next-day" transition rates are

$$\begin{pmatrix} x_{11} & \dots & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1I} & \dots & x_{NI} \end{pmatrix} \xrightarrow{1} \begin{pmatrix} x_{21} & \dots & 0 \\ \vdots & \ddots & \vdots \\ x_{2I} & \dots & 0 \end{pmatrix}, \tag{4}$$

which describes transitions to the next day.

For $1 \leq n \leq N$ and $1 \leq i \leq I$ "the arrival rates"

$$\begin{pmatrix} x_{11} & \dots & x_{N1} \\ \vdots & x_{ni} & \vdots \\ x_{1I} & \dots & x_{NI} \end{pmatrix} \xrightarrow{y_{ni}} \begin{pmatrix} x_{11} & \dots & x_{N1} \\ \vdots & x_{ni}+1 & \vdots \\ x_{1I} & \dots & x_{NI} \end{pmatrix}, \tag{5}$$

which describes the arrival of an appointment request.

For $1 \leq n \leq N$ and $1 \leq i \leq I$, "cancellations"

$$\begin{pmatrix} x_{11} & \dots & x_{N1} \\ \vdots & x_{ni} & \vdots \\ x_{1I} & \dots & x_{NI} \end{pmatrix} \xrightarrow{x_{ni}q_{ni}} \begin{pmatrix} x_{11} & \dots & x_{N1} \\ \vdots & x_{ni}-1 & \vdots \\ x_{1I} & \dots & x_{NI} \end{pmatrix}, \tag{6}$$

which describes the cancellation of a appointment. All other rates are 0.

5.3.4 Embedded Markov Chain

The embedded Markov Chain is obtained by looking at the Markov process only at transition instances. I.e., we ignore how long we are in a state and only concern ourselves with the probabilities of making transitions from each state to all of the other states when a transition actually occurs.

Transition probabilities For transition (4) the probability is

$$p^1 \equiv \frac{1}{1 + \sum_{\substack{1\leq n\leq N \\ 1\leq i\leq I}} y_{ni} + \sum_{\substack{1\leq n\leq N \\ 1\leq i\leq I}} x_{ni}q_{ni}}.$$

For transition (5) $(1 \leq n \leq N, 1 \leq i \leq I)$ the probability is

$$p^y_{ni} \equiv \frac{y_{ni}}{1 + \sum_{\substack{1\leq n\leq N \\ 1\leq i\leq I}} y_{ni} + \sum_{\substack{1\leq n\leq N \\ 1\leq i\leq I}} x_{ni}q_{ni}}.$$

Finally, for transitions (6) $(1 \leq n \leq N, 1 \leq i \leq I)$ the probability is

$$p^{xq}_{ni} \equiv \frac{x_{ni}q_{ni}}{1 + \sum_{\substack{1\leq n\leq N \\ 1\leq i\leq I}} y_{ni} + \sum_{\substack{1\leq n\leq N \\ 1\leq i\leq I}} x_{ni}q_{ni}}.$$

Balance equations The stationary distribution of the embedded Markov chain satisfies the global balance equations:

$$\sum_{j \in S \setminus \{k\}} \pi(j)p(j,k) = \sum_{j \in S \setminus \{k\}} \pi(k)p(k,j) \ \ \forall k.$$

For state k we define $k - d_{ni}$ as the state where 1 appointment of class i for day n was cancelled and we define $k + d_{ni}$ for the state with a corresponding arrival. We define k_{shift} for the state of the next day: values are shifted one to the left, and the last column is filled with zeroes. We define L_k as the set of states that are possible previous day states. So if

$$k = \begin{pmatrix} x_{11} & \dots & x_{(N-1)1} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ x_{1I} & \dots & x_{(N-1)I} & 0 \end{pmatrix},$$

then

$$\begin{pmatrix} * & x_{11} & \dots & x_{(N-1)1} \\ * & \vdots & \ddots & \vdots \\ * & x_{1I} & \dots & x_{(N-1)I} \end{pmatrix} \in L_k,$$

where the $*$'s can have any value.

We have the following global balance equation:

$$\sum_{\substack{1 \le n \le N \\ 1 \le i \le I}} (\pi(k+d_{ni})p_{ni}^{xq} + \pi(k-d_{ni})p_{ni}^{y}) + p^1 \sum_{i \in L_k} \pi_i = \pi(k) \sum_{\substack{1 \le n \le N \\ 1 \le i \le I}} (p_{ni}^{y} + p_{ni}^{xq}) + p^1 \ \ \forall k,$$

where

$$\pi(k + d_{ni})p_{ni}^{xq}$$

is the flow in a state through cancellations,

$$\pi(k - d_{ni})p_{ni}^{y}$$

is the flow in through arrival of new appointment requests,

$$p^1 \sum_{i \in L_k} \pi_i$$

is the flow in through next day transitions,

$$\pi(k)p_{ni}^{y}$$

is the flow out of a state by arrival of new appointment requests,

$$p_{ni}^{xq}$$

the flow out of a state by cancellations and p^1 is the flow out by next day transitions.

In general it is computationally intractable to solve this system of equations for most Markov chain models [6].

Detailed balance does not hold since the process is not reversible. This transition has positive probability but the reverse transition is impossible:

$$\begin{pmatrix} * & x_{11} & \dots & x_{(N-1)1} \\ * & \vdots & \ddots & \vdots \\ * & x_{1I} & \dots & x_{(N-1)I} \end{pmatrix} \rightarrow \begin{pmatrix} x_{11} & \dots & x_{(N-1)1} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ x_{1I} & \dots & x_{(N-1)I} & 0 \end{pmatrix}. \tag{7}$$

Note that for a time window of size N the stationary distribution of this Markov chain actually equals P^N, where P is the transition matrix. So this relative big P-matrix does have a relative short convergence time.

5.4 Markov chain - basis for simulation

We introduce a new (discrete-time) Markov chain inspired by the previous continuous Markov chain:

$$X(t) = (X_j^n(t)), \quad 1 \le j \le I,\ 1 \le n \le N,$$

where $X_j^n(t)$ is the number of booked class j appointments, n days before the target date at the end of day t and again I is the number of priority classes and N the time window.

In this model the appointments are booked to their target date.

The recursive representation of the Markov chain is

$$X_j^{n-1}(t) = (X_j^n(t-1) - C_j^{n-1}(t))^+ + A_j^{n-1}(t),$$

where $A_j^{n-1}(t)$ is the number of arrivals of appointments during day t for class j, $n-1$ days before target and $C_j^{n-1}(t)$ the number of cancellations during day t for class j, $n-1$ days before target. The meaning of this equation in words, is that the number of appointments booked for a day n in the future is the number of appointments booked yesterday for that same day plus todays arrivals of new appointment requests minus todays cancellations.

The arrivals $A_j^{n-1}(t)$ are Poisson distributed with parameter λ_j^{n-1},

$$A_j^{n-1}(t) \sim \text{Pois}(\lambda_j^{n-1}),$$

where λ_j^{n-1} is the arrival rate for $(n-1)$-appointments with priority class j.

Since the number of booked appointments can't be negative, we constrain it by only considering greater or equal than zero values. In practice this will not be an issue, since the number of cancellations is dependent on the number of booked appointments (the number of cancellations can not be higher than the number of appointments booked). Let q_j^{n-1} be the cancellation rate of a class j appointment $n-1$ days before the target date. Then

$$\mathbb{E}(C_j^{n-1}(t)|X_j^n(t-1)) = X_j^n(t-1)q_j^{n-1}.$$

The cancellations $C_j^{n-1}(t)$ have a binomial distribution with parameters $n = X_j^n(t-1)$ and $p = q_j^{n-1}$,

$$C_j^{n-1}(t) \sim \text{Bin}(X_j^n(t-1), q_j^{n-1}).$$

For any initial state $X_j^k(0)$, letting $t \to \infty$ we get

$$X_j^k(t) \to X_j^k(\infty),$$

where $X_j^k(\infty)$ is distributed according to the stationary distribution of the Markov chain, provided it exists. We will assume so.

For simulation we can start with any initial distribution $X_j^k(0)$.

5.4.1 Sample performance metrics

Let

$$X^1(t) = \sum_{j=1}^{M} X_j^1(t)$$

the total number of appointments on time t for the next day.

If T_C is the total capacity of the system, sample performance metrics are:

$$\mathbb{P}(X^1(\infty) > T_C),$$

the probability that the number of appointments exceeds the capacity.

$$\mathbb{E}_{slack} = \mathbb{E}((T_C - X^1(\infty))^+),$$

the expectation of the slack capacity.

$$\mathbb{E}_{pp} = \mathbb{E}((X^1(\infty) - T_C)^+),$$

the expectation of the number of appointments to be postponed.

$$\operatorname*{Var}_{pp} = \mathrm{Var}((X^1(\infty) - T_C)^+),$$

the variance of the number of appointments postponed.

5.5 Simulation

The Markov chain described in the previous section is the basis for the simulation. The standard simulation where appointments get booked to their target date is based on the recursive formulation. More complex simulations — f.e. the one where tolerance is used — are still based on the same principles, but don't have a nice mathematical representation. These are however small adjustments to the simulation code.

For the simulation we don't take priority classes into account. So with the general arrival rates (λ^n) and cancellation rates (q^n) available the simulation can be executed. These rates can be found by fitting probability distributions to the data set already provided and analyzed.

The following image describes the process from the real world to the simulations, and where the configurations have their effect.

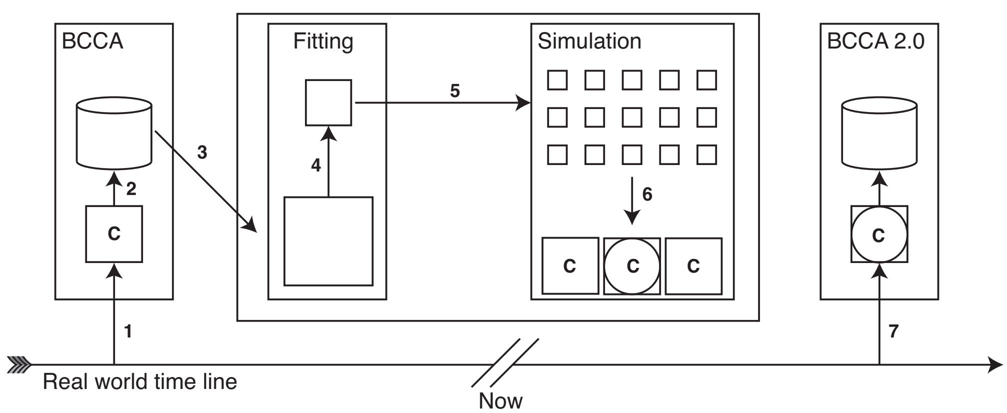

Figure 4: Description of approach

The arriving appointment requests in the real world are via the current configuration (1) saved in the database (2) available in the BC Cancer Agency. The performance indicators for this configuration can be calculated from the data. This database is interpreted and prepared for fitting (3). This data is fitted to a probability distribution (4). This distribution is used as input for the simulation (5), where Monte Carlo methods are used to generate numerous instances of a dataset. These datasets are used to test different configurations (6), so that the best performing (using the performance indicators) can be selected. The best performing configuration can be used in the future at the BC Cancer Agency (7).

The simulation is written in the free available software Octave, which can be found at `www.gnu.org/software/octave/`. The code can be found in Appendix C.

5.5.1 Configurations

For the simulation it's important to make sure that decision rules — configurations — can be implemented easily. Two important decision rules concern:

- add capacity (by a certain minimal amount),
- how to use the tolerance.

Configurations are used to test what is the impact on the performance metrics. Configurations differ in whether they use appointment tolerance or whether they add capacity a week in advance when the number of appointments booked y days in advance is greater than a certain threshold x.

5.5.2 Input

Arrival rates We find for all arrival lead times (1 to 85) the arrival rates by fitting the data to 85 Poisson distributions. The following table shows some examples.

arrival lead time	mean	variance	λ	r^2
1	0.63	0.9	0.5	0.99587
2	0.59	0.94	0.41	0.97616
3	0.6	0.94	0.41	0.97854
22	6.5	4.6	8.59	0.03912
85	4.4	4.3	3.97	0.34453

In this table λ is the mean (and variance) of the fitted Poisson distribution and r^2 the proportion of linear covariance with the observed versus the expected frequencies. As seen the Poisson distribution is not a good fit for the days with relative high arrival rates. Therefore not the fitted distributions are used for the simulation, but the distributions based on the mean values. This to guarantee that the average load on the system is like in reality. See section 8 for more on this.

Cancellation probabilities We find the n-day cancellation probabilities (85 in total) in the data by counting the number of n-day cancellations and dividing by the number of non-cancelled appointments $n+1$ days in advance. See the following table for an example of the calculations.

n	# n-day cancellations		# $n+1$-day appointments		# n-day cancellation probability
1	2669	/	33755	=	0.079069767
2	1164	/	34919	=	0.033334288
3	807	/	35719	=	0.022593018
8	722	/	38691	=	0.01866067
15	545	/	40987	=	0.013296899

In the simulation we used a binomial distribution (n, p) with n the number of booked appointments and p the cancellation probability.

5.5.3 Iterations

The simulation has a given time window of N days[19]. Therefore the state of the system is dependent on the last 85 days. To get an instance of a non-transient system the simulation has to run for at least 85 days. The initial state of the system is independent of the state after 85 days[20]. The simulation therefore stops after 85 iterations.

[19] $N = 85$ in this case

[20] Only when moving appointments forward (which is not included in the basic simulation) the simulation has to run longer

Simulation for all configurations are repeated 100 times to retrieve the results provided in section 6, i.e. averages and variations of the performance indicators.

6 Results

In the following table the results of four different configurations are shown.

The 'Standard' configuration is closest to the current practise. Appointment tolerance is not used, and no capacity is added. This one is completely based on the Markov chain of section 5.4.

The 'Tolerance' configuration uses a tolerance of 1 day for all appointments. Appointments are booked to the least busy day of the time window (plus and minus one from the target date).

The 'Add cap. (x, y)' configurations don't use tolerance, but they add capacity for a day when the number of booked appointments for that day is higher than x, y days before.

Model	$\mathbb{P}(X^1(\infty) > T_C)$	$\mathbb{E}_{pp}$	$\mathbb{E}_{slack}$	$\mathbb{E}_{pp} \times \mathbb{E}_{slack}$	$\mathbb{E}_{pp} + \mathbb{E}_{slack}$
Standard	0.35	2.32 ± 0.76 (90%)	5.12	11.88	7.44
Tolerance	0.24	0.56 ± 0.21 (90%)	2.99	1.67	3.55
Add cap. (50,7)	0.15	0.82 ± 0.39 (90%)	6.86	5.63	7.68
Add cap. (60,5)	0.13	0.85 ± 0.42 (90%)	8.2	6.97	9.05

As stated in section 4 the objective is to minimize the following properties:

- exceeding capacity (postponing appointments). The key metric for this property is $\mathbb{E}_{pp}$,
- slack capacity (waste of resources). The key metric for this property is $\mathbb{E}_{slack}$.

As already stated in section 4.1 a way to find the optimal configuration is to minimize the product, a sum or a weighted sum of both performance indicators. In the table above the product and the sum are added in the last two columns.

6.1 Visual results

The following pictures show the impact of using tolerance on the booking process.

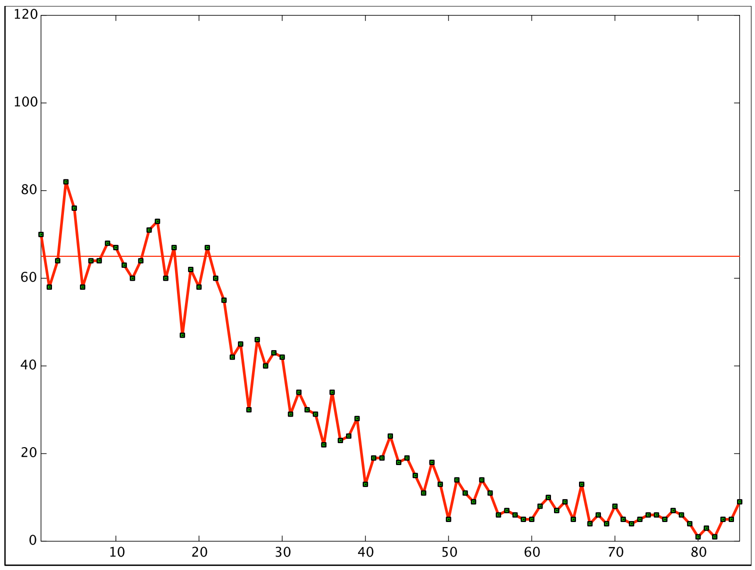

Figure 5: Number of appointments on booking list for next 85 days - not using tolerance.

Figure 6: Number of appointments on booking list for next 85 days - using tolerance.

7 Conclusions

As the results above show, in configuration 'Tolerance' the best results are retrieved. We see that Var_{pp} is smallest for this configuration as well[21] and there seems to be the key of the success.

In Figure 6 it can be seen that tolerance makes the whole process a lot smoother. Using the tolerance has the effect that the rather wild Poisson process that is the basis of Figure 5 is damped down. This 'bounding' or 'controlling' the process has a very positive effect on the outcome.

7.1 Tolerance

The results of the research show clearly that using the tolerance in a basic way can be very beneficial to decrease the wait list while decreasing the

[21]can be seen by the small confidence interval

waste of resources at the same time. There may be some disadvantages to using tolerance. It makes the process of booking appointments a bit more complicated. Mistakes in that process are life threatening, and should therefore be minimized.

8 Discussion

8.1 Automating booking process

Automating the booking process may be helpful to minimize (life threatening) mistakes while implementing a more advanced booking rule. However automating this process introduces new complexity to the booking system that needs to be tested and maintained.

8.2 Capacity

Capacity at BCCA is now given by number of patients per day (65). This is a result of different constraints such as available nurses, number of rooms, number of chairs and preparation time for drugs needed by the pharmacy. This number however doesn't reflect the difference in protocols for each patient. Therefore the recommendation is to make capacity dependent on the constraints itself. It might be possible to treat 70 patients on a day if the appointments all require a small amount of nursing time (and the other constraints are satisfied as well).

8.3 Database structure

The database structure the BC Cancer Agency is using is not optimized for data analysis. Of course the database is primarily used by the clerks and data analysis is just secondary. However the database is not optimized for the current use by the clerks as well. For example, there are no specified fields in the database for the target date, the tolerance or time preferences of the patient. And when the process of booking appointments is getting more complex, it would be beneficial to use a database which enables adding a smart input form, f.e. an input form that picks the best date given the patients target date using the provided tolerance. The CIHR-team would greatly benefit from a better organized database structure, in terms of being able to do better analysis and being able to improve the way the clerks have to deal with the system.

8.4 Validating tolerance on provided data

When a better database structure would be implemented, it would be a lot easier to validate the results provided in this thesis with real historical data. Now it involves writing a whole new program. This has not been done in this research project due to lack of time.

8.5 Medical and ethical implications of prioritizing

In the current situation at BCCA recurring patients have priority over new patients. So patients that receive a treatment (periodically) will have a place on the booking list before new patients have so. This is because the interval time is part of the protocol and thus the medical guidelines. Consequence is that new patients have to wait (a bit) longer for their first treatment. During the research it was and remained unclear whether this effect is wanted or acceptable. The following question needs to get answered: Is a delay for the first treatment worse than a suboptimal interval period during a periodical treatment?

8.6 Cancellation rate by patient type

Cancellation rates depend on patient characteristics. Therefore the simulation and prediction can be done in greater detail with these characteristics in mind. However this introduces ethical questions. What if we know that older patients tend to cancel their appointment more often than others? This would imply that we can (over) book more older patients, knowing that they might cancel anyway. Is this fair to younger patients? Implementing patient characteristics therefore needs to be done with ethics in mind.

8.7 Arrival process distribution

The Poisson distribution is not a very good fit for busy arrival lead times. As the reader can see in Appendix A for various reasons the Poisson distribution might not be the best distribution to use in the models. Better insight in the process before arrival would be greatly beneficial for this research.

8.8 Possible extensions

Decision variables can be added, like adding capacity at multiple time points in advance. I.e. if more than 40 total on list 2 days in advance add 1 unit capacity.

We can use another dimension. Use nursing hours instead of the appointment slots.

Deterministic forecast from current state is possible when using the current state as a initial state and start iterating.

Time can be scaled (i.e. let each time step be a week instead of a day).

References

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

[2] Ruben Aristizabal, Vincent S Chow, John French, Fareeza Khurshed, Colleen E McGahan, Martin L Puterman, Nancy Runzer, Pablo Santibáñez, and Scott Tyldesley. Impact evaluation of booking process redesign on patient experience. *CIHR Team Working Paper*, 2011.

[3] R. Blendon. Inequities in health care: A five-country survey. *Health Affairs*, 21:182–91, 2002.

[4] K.L. Chung. *Markov chains with stationary transition probabilities.* Springer (Berlin and New York), 1967.

[5] George Cybenko. Dynamic load balancing for distributed memory multiprocessors. *Journal of Parallel and Distributed Computing*, 7(2):279–301, 10 1989.

[6] Winfried K. Grassman. *Computational Probability.* Springer, 2000.

[7] F. Hillier and G. Lieberman. *Introduction to Operations Research.* New York: McGraw Hill., 2001.

[8] J. Hurst and L. Siciliani. Tackling excessive waiting times for elective surgery: A comparison of policies in twelve oecd countries. *OECD Health Working Papers 6. Paris: Organisation for Economic Co-operation and Development.*, 2006.

[9] PhD Jonathan Patrick and Martin L. Puterman. Reducing wait times through operations research: Optimizing the use of surge capacity. *Longwoods Review*, 6(2):77–83, 2008.

[10] Averill M. Law and W. David Kelton. *Simulation Modeling and Analysis.* McGraw-Hill Science/Engineering/Math, 1999.

[11] J. I. McGill and G. J. van Ryzin. Revenue management: Research overview and prospects. *Transportation Science*, 33(2), 5 1999.

[12] J. Patrick, M. L. Puterman, and M. Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *OPERATIONS RESEARCH*, 56(6):1507–1525, December 2008.

[13] M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley and Sons, 1994.

[14] Aristizabal R., Chow V.S., Huang K., Puterman M.L., Runzer N., and Santibáñez P. Chemo smartbook: An innovative redesign of the bc cancer agency chemotherapy booking system. *CIHR Team Working Paper.*, 2011.

[15] Christian Robert and George Casella. *Monte Carlo Statistical Methods.* Springer, 2004.

[16] Sheldon M. Ross. *Introduction to Probability Models.* Elsevier, 2003.

[17] L. Siciliani and J. Hurst. Explaining waiting-time variations for elective surgery across oecd countries. *OECD Economic Studies*, 38(1):95–123, 2004.

[18] Kiok Liang Teo. Practical operations research applications for healthcare manager. *Annals Academy of Medicin*, 38(6), 2009.

[19] S. Willcox, M. Seddon, S. Dunn, R. Tudor Edwards, J. Pearse, and J.V. Tu. Measuring and reducing wait times: A cross-national comparison of strategies. *Health Affairs (Millwood)*, 26(4):1078–87, 2007.

A Data analysis

A.1 Target date

For all appointments the cancer type is available. In the following chart the appointments are shown by day of the week. Once for their target date and once for their realized appointment date. We see that less appointments have a target date on Tuesdays compared to the other days of the week. Inquiry at the BCCA reveals that this is the result of less “breast cancer” physicians working on Mondays, referring less patients for Tuesdays. Since the cancer type “breast” is the biggest group, such effects have a big influence on the total system.

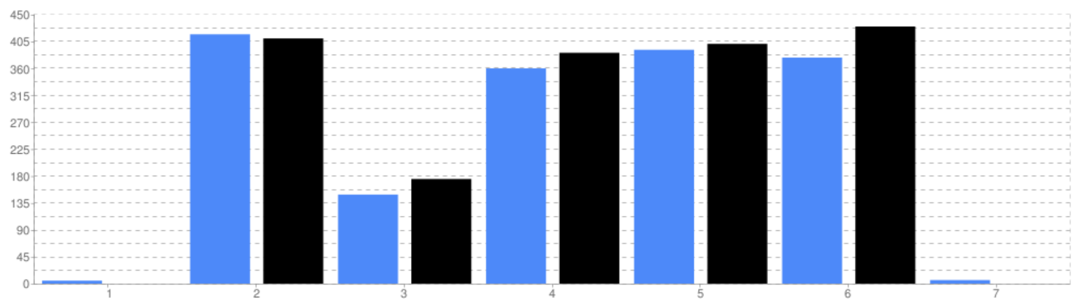

Figure 7: Number of breast-type appointments per day of the week by target date (blue) and realized appointment date (black) (1 = Sunday)

The key insight here is that the process of incoming appointments is not smooth. Requests for appointments are the result of a longer process from the point of view of the patient. First the symptoms appear; then they visit to physician; he sends the patient to a clinic and in the clinic a request for the appointment is made. This pre-process creates unnatural delays which effect the final process.

Therefore the arrival process is not what you may assume (Poisson distributed).

A.2 Friday effect

In the following graph the so called ‘Friday effect’ is seen. Fridays tend to be busier than the other days of the week. An often heard explanation for this effect deals with shifting appointments. Appointments are shifted to the next day when there is no capacity left. This can be done only when the tolerance of the appointment is satisfied. Tolerance often only provides flexibility for one or two days, which is not enough to shift the appointment over the weekend.

However when we look at the target dates we see that those are high on Fridays as well. Therefore, the explanation for the busy Fridays might have something to do with the process that defines the target dates, as seen in the figure above. This is confirmed in Appendix B.

The periodicity of appointments (i.e. every week) enlarges this ‘Friday effect’.

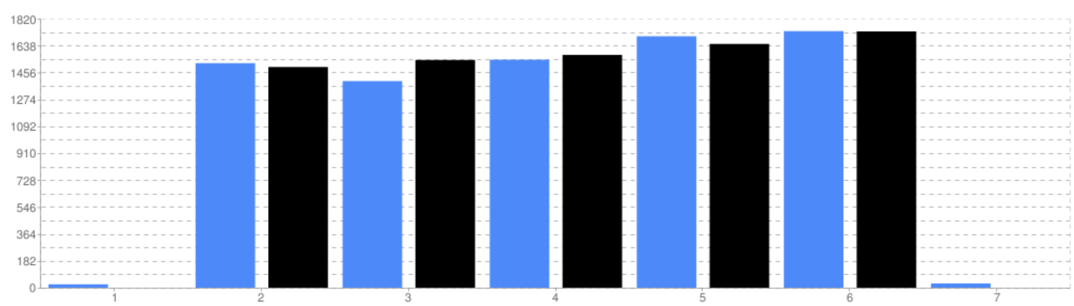

Figure 8: Number of appointments per day of the week by target date (blue) and realized appointment date (black) (1 = Sunday)

A.3 Arrivals

In the following graph the appointments are shown by the day of the week of the arrival. We see that appointments don't enter the system evenly distributed over the days of the week. The arrivals in the weekend indicate people working in the weekend. Flaws in the data are highly unlikely since this information is extracted from the computer system itself. In other words, these records in the database are indeed entered on a weekend day.

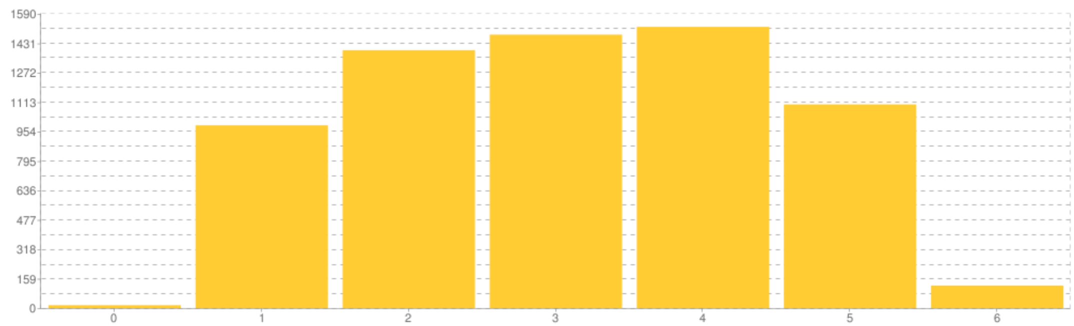

Figure 9: Number of appointments entering the system per day of the week (0 = Sunday)

A.4 Holidays

Capacity levels do change over the year, since in summer it is harder to find nurses (holidays). Demand is also down in summer, probably because potential patients are on holidays as well. Patients may have less symptoms during holidays or link the symptoms to the holiday, instead to an illness.

B Interview Nancy

Q: What may be reasons to change min/max from protocol?

A: This should not be happening. So these are errors by the clerks. If an appointment has in the notes ”moved by Nancy”, this is an indication that the appointment is booked out of it's time window. The flexibility (min/max) is only used for the first appointment and not for the following in the total treatment. This is because Nancy is afraid that the intervals will be messed up otherwise, or it'll costs the clerks a lot more time to manage that correctly, and this time is not available.

Q: When do you add capacity? How does it work?

A: At the first day of the month Nancy will make the nurse schedule for the next month. So on February 1st, she will know who is working when in March. She tries to get 9 nurses every day for 8 hour shifts, independent of the current state of the book list. Sometimes this is hard, because of holidays and vacations. When she is not able to get the nine nurses, she will start blocking some spots on the book list. She will only block new patient slots and clinical trial slots. So now Nancy isn't looking at demands for specific days. Forecasting may be useful to have more control on when to add shifts. However, trying to spread the load over the week given tolerance on the target date will probably improve results more and faster.

Every other Monday she will add 1.5 nurses. On Tuesdays, Wednesdays and Thursdays half a shift is added regularly. On Fridays a full shift is added and half a shift is removed. On average half a shift (4 hours) are added very day. This is in sync with our data analysis. Some clinical trial appointments do not need the special nurses in the chemo unit. These patients do not get a chemo treatment, but receive oral medication and only have blood drawing. We should find out which protocols are like that, since these patients don't have to be treated by the chemo unit. Every nurse in the building is able to treat those patients. This can decrease the load on ACCU.

Q: Is it correct that clinical trials don't have a target date?

A: Yes, they don't have tolerance, so the appointment date is the target date.

Nancy's question to us: What percentage of appointments is a clinical trial? This is load that other units don't have, and they require more time on average than non clinical trial appointments. This may be useful information to share with management. Ruben: We can find out what is the number of patients that is booked on Friday but have high tolerance. These patients can be moved to next week and lower the pressure on Fridays.

Q: Can you tell us more about the casual staff?

A: Casual staff is cheaper than standard staff, because of the benefits for regular staff. Casual staff has to be available for 4 shifts a month and they need to maintain the calendars, so that Nancy knows when they are available. Casual staff now consists of 5 nurses. She doesn't want too many, because than she can't guarantee enough work for them, and they'll find somewhere else to work. She doesn't want less too, because than it would be harder to meet the demand. The casuals normally want to work 2-3 days a week. The reason that not all nurses are casuals is because of reliability. The current situation can be seen as a balance between flexibility, price and reliability.

Q: What is the target date for new patients? A: The target date is the last date the physician says it can be scheduled. When there is info in the notes about CRT (Chemo ready to treat), the target date is in 14 days of CRT. This info is missing in around 90% of the cases.

Q: Why is Friday such a busy day?

A: The heavy treatment clinics are on Thursdays, so the patients needs treatments on Fridays. These appointments may have tolerance (like max 2), but because of the weekend they cannot be moved to the Monday.

Q: What happened on October 18, 2010 and July 26, 2010, since the utilization is so low?

A: Mondays are always a problem. Sometimes it's very busy, sometimes it's not. October 18 and July 26 are indeed Mondays.

Q: What happened on August 12, 2010 since capacity was low, but it wasn't a vacation day?

A: In August it's hard to find nurses, since they all want vacation then. It may therefore happen that not enough nurses can be found.

C Octave code

C.1 Simulate.m

```
function s = Simulate(tolerance)

  # time window
  s.N = 85;

  # number of classes
  s.M = 1;

  # parameters for adding capacity
  s.add_capacity_at_day = 5;
  s.add_capacity_if_above = 6000;
  s.add_capacity_set_to = 70;


  # creating state space
  s.state = zeros(s.N, s.M);

  # creating capacity vector
  s.capacity = 65*ones(s.N);

  # number of repetitions of the simulation
  s.repetitions = 100;

  # using tolerance? (1 = yes, 0 = no)
  s.useTolerance = tolerance;

  # how to deal with overtime
  s.overtime = 1;

  # creating the class
  s = class (s, "Simulate");

  # creating vector for total appointments on day 1 for
  # each repetition
  histogram_data = zeros(1,s.repetitions);

  # start timing
  tic;
```

```
# probability of going over capacity
PoverCapacity = 0;

# slack capacity
slackCapacity = 0;

# number postponed
numberPostponed = 0;

#
varNumberPostponed = zeros(1,s.repetitions);
increasedCapacity = 0;
numberInSystem = 0;

for p=1:s.repetitions

  # initializing state space
  s.state = setInitialState(s);

  for i=0:(s.N+5) # + 5 to get state images
    s.state = increaseT(s);

    # increasing capacity
    if((s.N-i) == s.add_capacity_at_day)
      if(sum(s.state(s.add_capacity_at_day,:)) > \
        s.add_capacity_if_above)
        s.capacity = s.add_capacity_set_to*ones(s.N);
        # counting days with increasedCapacity
        increasedCapacity = increasedCapacity + 1;
      endif;
    endif;


    # if(i > s.N)
    # h = figure(i);
    # showState(s);
    # print(h,'-dpng',i);
    # endif;
  endfor
  if (sum(s.state(1,:)) > s.capacity(1))
    PoverCapacity = PoverCapacity + 1;
    numberPostponed = numberPostponed + \
    sum(s.state(1,:)) - s.capacity(1);
    varNumberPostponed(p) = sum(s.state(1,:)) - \
```

```
      s.capacity(1);
    else
      slackCapacity = slackCapacity - sum(s.state(1,:)) \
      + s.capacity(1);
    endif;
    numberInSystem = numberInSystem+sum(s.state(:,1));
    histogram_data(p) = sum(s.state(1,:));
    #showState(s);
  endfor;

  #figure(1);
  #hist(histogram_data,40:1:85);

  disp('P over capacity');
  PoverCapacity/s.repetitions
  disp('increased capacity');
  increasedCapacity/s.repetitions
  disp('number postponed');
  numberPostponed/s.repetitions
  disp('number postponed when postponed');
  numberPostponed/  \
  size(varNumberPostponed(varNumberPostponed≠0),2)
  disp('slack capacity');
  slackCapacity/s.repetitions
  disp('variance number postponed');
  var(varNumberPostponed)
  disp('average appointments in system');
  numberInSystem/s.repetitions

  toc

endfunction
```

C.2 setInitialState.m

```
function state = setInitialState(s)
               state = ones(s.N, s.M);
endfunction
```

C.3 increaseT.m

```
function nextState = increaseT(s)
  nextState = zeros(s.N, s.M);
  increments = zeros(s.N, s.M);
  for n=2:(s.N+1)
    for j=1:s.M
      arrival = a(s,n−1,j);
      if(s.overtime == 0)
        arrival = arrival;  + max(0, s.state(1,j) − 65);
      endif;

      if(s.useTolerance == 1)
        if( n == (s.N+1) )
          arrival = arrival − binornd(arrival,q(s,n−1,j));
        endif;
        while(arrival > 0)
          if(n > 2 && n < (s.N+1) && (s.state(n−1,j) /
          +increments(n−2,j)) < (s.state(n,j) /
          +increments(n−1,j) ))
            increments(n−2,j) = increments(n−2,j)+1;
            arrival = arrival − 1;
            #disp('divide');
          elseif(n < (s.N−1) && (s.state(n+1,j) /
          +increments(n,j)) < s.state(n,j)+increments(n+1,j))
            increments(n,j) = increments(n,j)+1;
            arrival = arrival − 1;
            #disp('divide');
          else
            #disp('no divide');
            increments(n−1,j) = increments(n−1,j) + 1;
            arrival = arrival − 1;
          endif;

        endwhile;
        if(n ≤ s.N)
          nextState(n−1,j) = s.state(n,j) − /
          binornd(s.state(n,j),q(s,n−1,j));
        else
          # arrivals will be added by increments;
          nextState(n−1,j) = 0;
        endif;
      else
        if(n ≤ s.N)
```

```
        nextState(n-1,j) = s.state(n,j) - /
        binornd(s.state(n,j),q(s,n-1,j)) + arrival;
      else
        nextState(n-1,j) = arrival - /
        binornd(arrival,q(s,n-1,j));
      endif;
    endif;
  endfor;
endfor;

if(s.useTolerance == 1)
  increments;
  nextState = nextState + increments;
endif;
nextState;
norm = norm(s.state-nextState);
endfunction
```

C.4 a.m

```
function r = a(s, n, j)
  arrival_rates_per_day = [
    0.626204239 0.591522158 0.599229287 0.689788054 0.80539499  \
    1.289017341 2.221579961 2.275529865 1.192678227 1.050096339 \
    0.95761079  0.89017341  1.317919075 1.878612717 2.566473988 \
    1.342967245 1.325626204 1.088631985 1.196531792 1.971098266 \
    2.543352601 6.487475915 2.816955684 2.246628131 1.757225434 \
    1.560693642 2.02504817  2.23699422  5.591522158 2.061657033 \
    1.545279383 1.235067437 1.04238921  1.134874759 1.344894027 \
    2.094412331 0.944123314 0.757225434 0.707129094 0.651252408 \
    0.986512524 1.287090559 2.94026975  1.013487476 0.93256262  \
    0.610789981 0.516377649 0.514450867 0.585741811 0.693641618 \
    0.252408478 0.217726397 0.200385356 0.181117534 0.23699422  \
    0.346820809 0.585741811 0.202312139 0.157996146 0.163776493 \
    0.159922929 0.177263969 0.262042389 0.348747592 0.125240848 \
    0.157996146 0.104046243 0.113680154 0.140655106 0.181117534 \
    0.354527938 0.05973025  0.071290944 0.107899807 0.069364162 \
    0.052023121 0.078998073 0.105973025 0.04238921  0.063583815 \
    0.06743738  0.094412331 0.123314066 0.208092486 4.445086705 \
  ];
  r = poissrnd(arrival_rates_per_day(n));
```

```
endfunction
```

C.5 q.m

```
function r = q(s,n,j)
  rates_vector = [
    0.040436209 0.079069767 0.033334288 0.022593018 0.018711876 \
    0.012643803 0.012831847 0.017274978 0.01866067  0.008915076 \
    0.006844784 0.005794241 0.005261171 0.007819136 0.010630671 \
    0.013296899 0.006786234 0.005089611 0.004587156 0.00432859  \
    0.006440267 0.008156915 0.007999252 0.004099793 0.003019394 \
    0.002918356 0.002910065 0.003934834 0.00471925  0.004947222 \
    0.001911611 0.001930984 0.001497278 0.001268806 0.001877191 \
    0.002662034 0.002340972 0.001505956 0.000898372 0.000763085 \
    0.000807338 0.001299196 0.00181127  0.00209798  0.001003412 \
    0.00075761  0.000601296 0.0005342   0.000711775 0.000955598 \
    0.00064413  0.000377451 0.000177592 0.000199751 0.000288453 \
    0.000310559 0.000465642 0.000531915 0.000199428 0.000132935 \
    0.000199366 0.000243611 0.000243551 0.000553244 0.00064138  \
    0.000243228 0.000287369 0.000154713 0.000176788 0.000154665 \
    0.000154641 0.000287121 0.00015458  4.41638E-05 0.000110397 \
    6.62339E-05 4.4154E-05  0.000176585 0.000132424 0.000176534 \
    6.61959E-05 2.20648E-05 8.82515E-05 0.000132366 0.000154406 \
    0.004232316
  ];
  r = rates_vector(n+1);
endfunction
```

C.6 showState.m

```
# Histogram of the number of slots on day 1 based on 1000
# iterations of the simulation with different booking rules.

function showState(s)
  hold off;
  for i=1:s.M
    plot(s.state(:,i));
```

```
    hold on;
  endfor;

  plot(s.capacity,'r');

  sum_vector = zeros(1,s.N);
  for i=1:s.N
    sum_vector(i) = sum(s.state(i,:));
  endfor;
  axis([1 s.N 0 120])
  plot(sum_vector,'--rs','LineWidth', 2, \
  'MarkerEdgeColor','k','MarkerFaceColor','g','MarkerSize',3)
  #display(sum(sum_vector));
endfunction
```