2017

Flexible capacity in clinical pathways

RESEARCH PAPER BUSINESS ANALYTICS
VRIJE UNIVERSITEIT AMSTERDAM

JUDITH BOL

Preface

This paper is written as a part of the master's program Business Analytics. The goal of this course is to perform a research with a mathematical or computer science related background, with a link to a business case. The subject was chosen in consultation with my supervisor René Bekker, who I approached because I was interested in queueing theory and health care related research. I would like to thank René Bekker for the helpful support and advice.

Summary

The purpose of the research is to build a model for a clinical pathway where flexible capacity is taken into account, starting with a model for the outpatient department. The model should provide insight in the effects of adding capacity in hospital departments on the access time of patients.

From literature review it is derived that many factors influence hospital waiting lists and access times. To capture most of the elements influencing the access time, many studies use simulation models. In this paper, we use a stylized queueing model to study the impact of the extra flexible capacity.

In this research, every node in the clinical pathway is modelled as a separate stochastic birth-death process where capacity can be added if the waiting list becomes too long. According to Burke's theorem, in a stochastic birth-death process the output of the system is again a Poisson process with the same rate as the original arrival process and therefore the use of extra capacity does not influence the next stage the model of the clinical pathway. Consequently, the model of the clinical pathway can be decomposed to separate models of every node. A model of the first stage in the clinical pathway, the outpatient department, is created. Models for the other stages can be derived from this model.

The outpatient department increases the capacity of the clinic if the access time for a new patient becomes too long. We assume patients arrive according to a Poisson process and the service times of patients are exponentially distributed. Three different structures for increasing the capacity are implemented: abrupt increase, linear increase and "s-shaped" increase, all starting when the backlog exceeds a certain threshold, whereafter the expected waiting time of patients with respect to the load of the system is computed. When no extra capacity is used, the expected number of patients in the system tends to infinity if the load approaches 1. Using extra capacity reduces this effect. The expected number of patients in the system is compared to the relative amount of extra capacity that is used. The fluent increase structures lead to a lower amount of needed extra capacity and the expected access time for fluent increase is lower than in the case of abrupt increase until the load exceeds approximately 1.2. Thus, using only a small amount of extra capacity makes it possible to handle a load of 1. Therefore, we advise to increase the capacity in a fluent way. It is hard to determine which structure approaches the real situation because we did not have relevant data from the outpatient department available.

Furthermore, we found that a lower backlog threshold for increasing the capacity leads to a shorter waiting list. The models are compared to the case where the outpatient department does not increase the capacity but patients are assumed to abandon the queue if they have to wait too long, this also gives a lowering effect on the expected access time of patients. Interesting further research would combine increasing capacity and the effect of abandonments. Moreover, further research should focus on flexible capacity in departments without the assumption of exponential service times as this would approach the real situation better. However, analytical models are not sufficient for modelling this, simulation would be necessary.

Table of contents

1. Introduction

During the past decades, hospitals have become more interested in the use of clinical pathways. Clinical pathways are standardized routes of patients with a specific clinical problem. They are used both to increase efficiency and provide a higher level of care and transparency for patients. There are several definitions for clinical pathways. However, the most commonly used definition is from the website of the European Pathway Association (2017): “A care pathway is a complex intervention for the mutual decision making and organisation of care processes for a well-defined group of patients during a well-defined period”.

In 2012, Eline Bussing wrote a research paper about the use of the Queueing Network Analyzer (QNA) method in modelling clinical pathways to compute the lead time of the clinical pathways. This method gives an approach of the congestion measures of a (not necessarily Markovian) network of queues (Whitt 1983). The problem with this method is that it is not possible to make the used capacity variable based on the waiting times. Practise shows that the outpatient department works with a certain capacity, but if patients have to wait too long before an appointment is possible, the outpatient department will increase their capacity by planning more appointments on a day. The main objective of this paper is to build a realistic model of a clinical pathway where it is possible to increase capacity and investigate whether the use of extra capacity influences the access times of patients in every step of the clinical pathway and the expected total lead time of a patient through the clinical pathway.

To make the research more explicit, a particular clinical pathway is chosen. This example clinical pathway starts with referral from a general practitioner, whereafter the patient arrives at a particular outpatient department of the hospital. On the basis of medical examination, for example an X-ray, a CT scan and/or MRI scan the patient will be diagnosed. Subsequently, the patient returns to the outpatient department to discuss the results of the tests and determine the next steps. Often, the patient gets medication, some medical advice or referral to care outside the hospital (for example physiotherapy). If a patient has serious complaints, surgery is needed. After surgery, the patient must recover in the hospital for a couple of days. Finally, one or more clinic visits are planned that discuss what the next steps will be, followed by resignation from the hospital (Bussing, 2012). A schematic representation of this clinical pathway can be found in appendix A. The paper refers to this network of queues for a more intuitive understanding of the described models. However, the model could be applied to comparable clinical pathways as well.

To create a model of the clinical pathway, first a model of the outpatient department is created, comprising the option of using extra capacity. Different structures in increasing capacity (abruptly or fluent) are compared to determine which structure is reasonable. Computing both the expected number of patients in the system and relative used extra capacity gives us the possibility to give recommendations about which structure of increasing capacity is optimal. Furthermore, the influence of the moment from when extra capacity is used is investigated. The model is compared to a model with constant capacity where patients could abandon the queue if they have to wait too long, however, it is not possible to draw conclusions which model is plausible because we have no related data from the outpatient department available.

To get closed-form results it was necessary to assume exponential service times of patients in the various nodes of the clinical pathway. According to Burke’s theorem, this leads to a model with Poisson outflow and therefore the clinical pathway could be modelled as a standard Jackson Network where each node can be analysed separately. The model of the outpatient department could easily be adapted to suit the other nodes in the network. A brief description of how every node could be modelled is described in section 3.2.

This paper starts with a study of relevant literature about reducing waiting lists and access times of patients in hospitals and about patient flow modelling. Next, we describe the model for the outpatient department and briefly the model for the complete clinical pathway. Then we present the results of the expected number of patients and relative extra capacity for various situations. The conclusion gives recommendations about the structure of increasing capacity and some ideas for further research.

2. Review of the literature

One of the first papers concerning the application of queueing theory on hospital processes was written by Bailey and Welch in 1952 (Bailey & Welch, 1952). They mainly used statistics to determine the punctuality of both patients and medical staff and concluded that the unwillingness of medical staff to wait eventually a short moment fairly increases the waiting time of patients, whereas patients are usually on time and are likely to come earlier if the waiting time was known to be shorter. This paper was the basis of many studies related to hospital waiting times and waiting lists.

An interesting problem in trying to reduce hospital waiting lists, is the concept of 'feedback'. If the waiting lists for a hospital are successfully reduced, general practitioners will refer more patients to this hospital because of the short waiting list, and the waiting list will increase again (Culyer 1976). Worthington (1987) describes a queueing model that takes feedback into account by assuming the arrival rate decreases linearly until a certain number of customers on the waiting list, after which the arrival rate becomes zero. He used two options for modelling the feedback. The results of his research consist of the effects of various examples of management actions on reducing waiting lists, taking feedback into account, which leads to the conclusion that feedback is a very important phenomenon that has large influence on the length of hospital waiting lists.

Fomundan & Hermann (2007) provide an overview of papers that describe applications of queueing theory in health care. They state that hospitals want to minimize the waiting lists, but maximize the utilization of servers or resources, which leads to a goal conflict. Furthermore, they refer to Hall et al (2006) who stated that if the demand is larger than the capacity of a hospital department the only way the system will reach an equilibrium is reneging, patients abandoning the queue if it is too long.

Iversen (1992) describes several perspectives of patient arrivals in hospitals. He states two traditions in modelling queues, namely the assumption of either stochastic arrivals or arrivals based on patients deciding whether they will join a queue or not. He refers to a paper of Johansen (1987) who suggested a combination of the both traditions and stated that the arrival of patients follows a stochastic pattern established from rational behaviour of the patients that possibly join the queue. He adds a third aspect that influences waiting lists in hospitals, namely the fact that in many countries hospitals are paid by government instead of directly by the patients. This leads to long waiting lists as the government's willingness to pay for extra resources only increases if the queue becomes longer. As a result, hospitals prefer long queues since that leads to a larger contribution from government. To solve this, the institutional structure should be reformed.

A research with a more practical goal was performed and described by Elkhuizen et al. (Elkhuizen 2007). Their goal was to develop models for analysing the needed capacity in hospital outpatient departments. Therefore, they first modelled the department as an M/D/1 queue to gain global insight in the problem. This model was used to indicate the performance when using the actual capacity and to make a rough estimate of the needed capacity to reach a certain service level. Furthermore, they build a simulation model for a more detailed study of the outpatient department. This model needed less simplifications and provided more specific information. Both models were used to estimate the needed capacity and how much temporary extra capacity (and for how long) would be needed to eliminate a possible present backlog and how much capacity is needed to keep access time of patients within established norms. The model was originally build for the neurology department of the AMC, but was intended to be useful for other departments as well. To show their model was generic, they implemented the model as well for the gynecology outpatient department in the AMC. It was indeed easy to implement because of the relatively simple input. They were able to obtain directly applicable results about the performance of both departments and found out that the model could also be used to improve the efficiency in a department with sufficient capacity.

In the field of modelling clinical pathways, an interesting paper was written by Bhattacharjee & Ray (2014). They describe the complexity of patient flows through a hospital caused by many possible pathways, a large number of stages in a care pathway and possible repeating stages and various priority rules for the care of patients within a node, for example priority based on the urgency for care. They state that patient flow modelling makes it possible for hospitals to perform various types of analysis among which "*exploring the interrelationships between parameters (performance) at various stages*", which is the purpose of this paper as well. The authors provide an overview of all necessary steps to model patient flow in hospitals and an extended overview of papers written with respect to this subject, both based on Markovian models and non-Markovian models or discrete-event simulation. They conclude that for modelling patient flows, simulation is most-used whereas queueing models are mainly used in modelling a single stage of the clinical pathway. However, they see possibilities in using networks of queues in modelling clinical pathways to obtain numerical solutions with respect to patient flow characteristics in hospitals.

3. Model

We want to create a model of the clinical pathway. Therefore, we start by modelling the first node in the clinical pathway. As described in the introduction, the first node in our example clinical pathway is the outpatient department, where the patient meets the doctor for an anamnesis interview. Most hospitals have outpatient departments for several specialisms. In the outpatient department of a certain specialism, multiple doctors on that field are serving assigned patients. For the ease of this research we assume the clinic of a certain specialization had an overall capacity, where all doctors can serve all patients.

3.1 Model of the outpatient department

Two other important assumptions are made to model the outpatient department. First, the requests for an appointment are assumed to arrive according to a Poisson process with rate λ. This assumption is based on the fact that the superposition of many independent processes tends to be a Poisson process. The planning of patients is complicated and depends on the wishes of the patient and the schedules of the different doctors. However, it is assumed patients are served on a First Come First Served basis (Vis & Bekker, 2017).

Second, the service durations of the patients are assumed to be exponentially distributed. This is not completely realistic, as in practice the planned time for an appointment is set (for example ten minutes) and therefore deterministic. Having deterministic appointment durations leads to a pre-planned number of patients will be served on a day, which would be modelled by an M/D/1 queue. In practice, when the queue becomes longer, the outpatient department might decide to increase the capacity by shortening the appointment durations of patients and/or planning more appointments on a day. For the ease of modelling this flexible capacity we assume the service durations to be exponentially distributed. Therefore, the outpatient department will be modelled as an M/M/1 type queue, i.e. a stochastic birth-death process.

We assume the decision to increase the capacity of the outpatient department depends on the length of the waiting list. The outpatient department has a basis capacity m_{basis} and possible extra capacity m_* leading to a maximum capacity $m_{max} = m_{basis} + m_*$. Vis & Bekker (2017) describe the modelling of the appointment system with a queue in continuous time as an M/D/1 queue where the state is the number of slots at the waiting list for a patient arriving at time t:

$$X(t) = \text{backlog in slots for patients arriving at time } t.$$

We will model the appointment system in a similar way, assuming exponential service durations leading to an M/M/1 queue where the state is identical to the case of Vis & Bekker (2017). The flexible capacity can be modelled in two ways. The first option is to model the arrival rate λ to be dependent on the backlog x. This is analogous to the model described by Vis & Bekker (2017), measuring the arrival rate in terms of slots; taking $\mu = 1$ day. A day has a capacity of m_x slots, leading to an arrival rate:

$$\lambda_x = \frac{\lambda}{m_x}$$

The second option is to model the service rate μ to be dependent on the backlog x whereas the arrival rate λ does not depend on the queue length. Then the service rate becomes:

$$\mu_x = m_x \text{ slots per day} \cdot 1 \text{ day} = m_x$$

We are interested in the limiting distribution:

$$\pi_\infty(x) = \frac{\lambda_{x-1} \cdot \ldots \cdot \lambda_0}{\mu_x \cdot \ldots \cdot \mu_1} \pi_\infty(0)$$

We see that the limiting distribution only depends on the ratio $\frac{\lambda_x}{\mu}$ (first modelling option) or the ratio $\frac{\lambda}{\mu_x}$ (second modelling option). We can easily derive that these ratios are equal:

$$\frac{\lambda_x}{\mu} = \frac{\lambda/m_x}{\mu} = \frac{\lambda}{m_x \cdot \mu} = \frac{\lambda}{\mu_x}$$

Therefore, we can conclude both types of modelling are equivalent and will lead to the same results as each model could be rewritten in the other variant. From now on, we will model the increasing capacity by using λ_x, i.e. the arrival rate depends on the backlog x.

Basis scenario

We start by modelling the basis scenario where no extra capacity is available to see what happens to the queue length if the load of the system increases, i.e. the arrival rate of patients increases. The basis scenario can be modelled as a birth-death process with birth rate $\lambda' = \frac{\lambda}{m_{basis}}$ and death rate $\mu = 1$ day. The expected number of patients follows directly from the load:

$$\mathbb{E}(L) = \frac{\rho^2}{1-\rho}, \qquad \text{where } \rho = \frac{\lambda'}{\mu}.$$

In the basis scenario, we chose the capacity $m_{basis} = 4$, based on the paper of Bussing (2012). The load and expected queue length are computed using MS Excel for the range: $\lambda = [0, 4)$. The results are shown in Figure 1. We see that when the load increases, the expected number of patients increases almost exponentially, tending to infinity when the load gets close to 1.

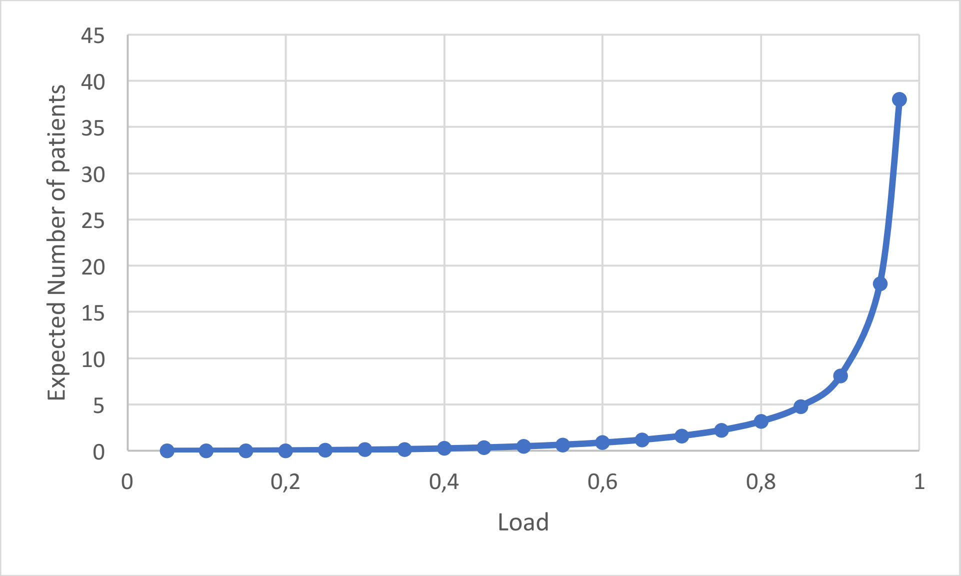


Figure 1: Expected number of patients for increasing load if no extra capacity is available

This tendency does not match practice. Even though the load in the outpatient department system often exceeds 1, the waiting list does not become infinitely long, as the outpatient department will undertake steps to avoid extremely long waiting lists. Extra capacity will be deployed by planning more appointment slots on a day, for example by the use of an extra doctor or by working longer days.

Three options of increasing capacity

We are interested in the effect of increasing capacity on the expected access time of arriving patients when the load increases. Therefore, the basis scenario where no extra capacity is available is compared to three different ways of increasing capacity.

The first option is to start with the basis capacity m_{basis} and increase the capacity to the maximum capacity m_{max} if the queue becomes longer than a certain threshold a_1. This abrupt increase of capacity gives the following structure of λ_x:

$$\lambda_x = \begin{cases} \dfrac{\lambda}{m_{basis}}, & x < a_1 \\ \dfrac{\lambda}{m_{max}}, & x \geq a_1 \end{cases}$$

However, it is more realistic that the capacity will be gradually increased. Therefore, the second option is to start with the basis capacity and gradually increase the capacity in a linear way if the queue exceeds the first threshold a_1. If the queue becomes more than a second threshold a_2, the maximum capacity is used. Of course, this gives a linear decrease in λ_x between the two thresholds a_1 and a_2, resulting in the following structure for λ_x:

$$\lambda_x = \begin{cases} \dfrac{\lambda}{m_{basis}}, & x < a_1 \\ \left(\dfrac{x - a_1}{a_2 - a_1}\right)\left(\dfrac{\lambda}{m_{max}} - \dfrac{\lambda}{m_{basis}}\right) + \dfrac{\lambda}{m_{basis}}, & a_1 \leq x \leq a_2 \\ \dfrac{\lambda}{m_{max}}, & x \geq a_2 \end{cases}$$

The third option again concerns a gradually increase of the capacity, but this time a smoother interpolation between the basis and maximum capacity. An appropriate function for this goal is the Cubic Hermite Spline (Cubic Hermite spline 2017). As λ_x should be decreasing between $\frac{\lambda}{m_{basis}}$ and $\frac{\lambda}{m_{max}}$, the Hermite basis function $h_{01}(t) = -2t^3 + 3t^2$ is used and transformed, resulting in the following structure for λ_x:

$$\lambda_x = \begin{cases} \dfrac{\lambda}{m_{basis}}, & x < a_1 \\ \left(-2\left(\dfrac{x - a_1}{a_2 - a_1}\right)^3 + 3\left(\dfrac{x - a_1}{a_2 - a_1}\right)^2\right)\left(\dfrac{\lambda}{m_{max}} - \dfrac{\lambda}{m_{basis}}\right) + \dfrac{\lambda}{m_{basis}}, & a_1 \leq x \leq a_2 \\ \dfrac{\lambda}{m_{max}}, & x \geq a_2 \end{cases}$$

Figure 2 shows the three different structures for $\lambda = 0.2$. For the first structure option $a_1 = 35$, for the second and third structure options $a_1 = 10$ and $a_2 = 60$.

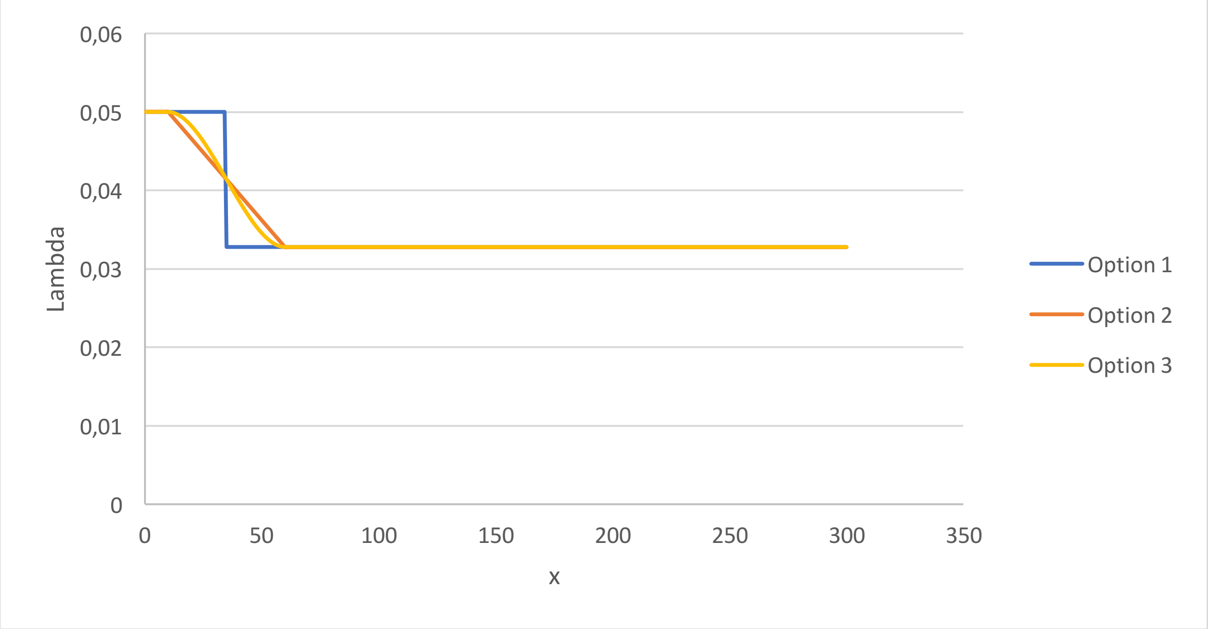


Figure 2: the structure of λ_x for three options of increasing capacity

Performance measures

We want to study the expected access time of patients when the load is increasing for all options of structures for λ_x. As our state x denotes the backlog, the expected number of patients in the system represents the number of slots a patient have to wait when he arrives. Thus, the expected number of patients equals the expected access time of a patient in slots.

The expected number of patients in the system can be computed using the following formula:

$$\mathbb{E}(L) = \sum_{0}^{\infty} x \cdot \pi_\infty(x)$$

where,

$$\pi_\infty(x) = \frac{\lambda_{x-1}}{\mu} \pi_\infty(x-1) = \frac{\lambda_{x-1} \cdot \ldots \cdot \lambda_0}{\mu_x \cdot \ldots \cdot \mu_1} \pi_\infty(0)$$

With λ_x as modelled according to the various structures and $\mu_x = \mu = 1$ day for all x.

The computational scheme for computing the expected number of patients works as follows. We set $\pi_\infty(0) = 1$. Then, $\pi_\infty(x)$ is computed for the first m values. We chose $m = 300$, large enough such that the thresholds $a_1, a_2 < m$. Thus, $\pi_\infty(x)$ is computed for $1 \leq x \leq 300$. For $x > 300$ we compute the sum:

$$\sum_{x=301}^{\infty} \pi_\infty(x)$$

To compute this sum, we need to assume λ_x is constant for $x > 300$, which is arranged by the choice of $m > a_1, a_2$. So, for $x > 300$ we have constant $\lambda_x = \frac{\lambda}{m_{max}}$. Then, for $x > 300$ we have:

$$\pi_\infty(x) = \frac{\lambda/m_{max}}{\mu} \pi_\infty(x-1)$$

Thus, for $x > 300$:

$$\pi_\infty(x) = \left(\frac{\lambda/m_{max}}{\mu}\right)^{x-300} \pi_\infty(300)$$

Where $\pi_\infty(300)$ is already directly computed in the model. Then:

$$\sum_{x=301}^{\infty} \pi_\infty(x) = \sum_{x=301}^{\infty} \left(\frac{\lambda/m_{max}}{\mu}\right)^{x-300} \pi_\infty(300) = \frac{\pi_\infty(300)}{\left(\frac{\lambda/m_{max}}{\mu}\right)^{300}} \cdot \sum_{x=301}^{\infty} \left(\frac{\lambda/m_{max}}{\mu}\right)^{x}$$

From series theory, we know if $|r| < 1$:

$$\sum_{x=n+1}^{\infty} r^x = \frac{1}{1-r} - \frac{1-r^{n+1}}{1-r} = \frac{r^{n+1}}{1-r}$$

Applying this for $r = \left(\frac{\lambda/m_{max}}{\mu}\right)$ and $n = 300$ gives:

$$\sum_{x=301}^{\infty} \pi_\infty(x) = \frac{\pi_\infty(300)}{\left(\frac{\lambda/m_{max}}{\mu}\right)^{300}} \cdot \sum_{x=301}^{\infty} \left(\frac{\lambda/m_{max}}{\mu}\right)^{x} = \frac{\pi_\infty(300)}{\left(\frac{\lambda/m_{max}}{\mu}\right)^{300}} \cdot \frac{\left(\frac{\lambda/m_{max}}{\mu}\right)^{301}}{1 - \frac{\lambda/m_{max}}{\mu}}$$

$$= \pi_\infty(300) \cdot \frac{\lambda/m_{max}}{\mu} \cdot \frac{1}{1 - \frac{\lambda/m_{max}}{\mu}} = \pi_\infty(300) \cdot \frac{\frac{\lambda}{m_{max}}}{\mu - \frac{\lambda}{m_{max}}}$$

This result is implemented in the model. Note that indeed $\left|\frac{\lambda/m_{max}}{\mu}\right| < 1$ as we set the capacity $m_{max} > \lambda$ for the whole range of λ and $\mu = 1$.

Now the summation of all $\pi_\infty(x)$ for the first 300 states and the sum to infinity is computed. All values of $\pi_\infty(x)$ are normalized by dividing by the total sum. The normalized values of $\pi_\infty(x)$ are used to compute the expected number of patients:

$$\mathbb{E}(L) = \sum_{0}^{\infty} x \cdot \pi_\infty(x)$$

This summation needs to be split as only the first 300 states are computed directly.

$$\mathbb{E}(L) = \sum_{0}^{\infty} x \cdot \pi_\infty(x) = \sum_{0}^{300} x \cdot \pi_\infty(x) + \sum_{301}^{\infty} x \cdot \pi_\infty(x)$$

For the first 300 states, the sum product of the state and $\pi_\infty(x)$ can be computed. For $x > 300$ the sum will be computed in a similar way as before, assuming constant $\lambda_x = \frac{\lambda}{m_{max}}$ for $x > 300$ and $\left|\frac{\lambda/m_{max}}{\mu}\right| < 1$. Similar to the previous summation we find

$$\sum_{301}^{\infty} x \cdot \pi_\infty(x) = \sum_{301}^{\infty} x \cdot \pi_\infty(x) = \sum_{x=301}^{\infty} x \cdot \left(\frac{\lambda/m_{max}}{\mu}\right)^{x-300} \pi_\infty(300)$$
$$= \frac{\pi_\infty(300)}{\left(\frac{\lambda/m_{max}}{\mu}\right)^{300}} \cdot \sum_{x=301}^{\infty} x \cdot \left(\frac{\lambda/m_{max}}{\mu}\right)^{x}$$

Using series theory, we know if $|r| < 1$:

$$\sum_{x=0}^{\infty} x \cdot r^x = \frac{r}{(r-1)^2}$$

and

$$\sum_{x=0}^{n} x \cdot r^x = \frac{(n \cdot r - n - 1)r^{n+1} + r}{(r-1)^2}$$

Combining these leads to:

$$\sum_{x=n+1}^{\infty} x \cdot r^x = \sum_{x=0}^{\infty} x \cdot r^x - \sum_{x=0}^{n} x \cdot r^x = \frac{r}{(r-1)^2} - \frac{(n \cdot r - n - 1)r^{n+1} + r}{(r-1)^2}$$
$$= \frac{(-n \cdot r + n + 1)r^{n+1}}{(r-1)^2}$$

Applying this for $r = \left(\frac{\lambda/m_{max}}{\mu}\right)$ and $n = 300$ gives:

$$\sum_{301}^{\infty} x \cdot \pi_{\infty}(x) = \frac{\pi_{\infty}(300)}{\left(\frac{\lambda/m_{max}}{\mu}\right)^{300}} \cdot \sum_{x=301}^{\infty} x \cdot \left(\frac{\lambda/m_{max}}{\mu}\right)^{x}$$

$$= \frac{\pi_{\infty}(300)}{\left(\frac{\lambda/m_{max}}{\mu}\right)^{300}} \cdot \frac{\left(-300 \cdot \left(\frac{\lambda/m_{max}}{\mu}\right) + 301\right)\left(\frac{\lambda/m_{max}}{\mu}\right)^{301}}{\left(\left(\frac{\lambda/m_{max}}{\mu}\right) - 1\right)^{2}}$$

$$= \pi_{\infty}(300) \cdot \frac{\left(-300 \cdot \left(\frac{\lambda/m_{max}}{\mu}\right) + 301\right)\left(\frac{\lambda/m_{max}}{\mu}\right)}{\left(\left(\frac{\lambda/m_{max}}{\mu}\right) - 1\right)^{2}}$$

This result is implemented in the model to complete the summation

$$\mathbb{E}(L) = \sum_{0}^{\infty} x \cdot \pi_{\infty}(x)$$

Now various results can be obtained about the access times of patients in the outpatient department and the influence of increasing capacity on the expected number of patients in the system, which are presented in the results section.

Furthermore, it is interesting to compare the expected number of patients in the system to the fraction of time extra capacity is used, i.e. the fraction of time that the backlog exceeded the lowest threshold:

$$\sum_{x>a_1}^{\infty} \pi_{\infty}(x)$$

However, this performance measure does not take the amount of extra used capacity into account. Therefore, the relative extra capacity used is more informative. The relative extra capacity used is the sum-product of the amount of extra capacity used and the fraction of time this amount is used:

$$\sum_{x=0}^{\infty} \left(\frac{\lambda}{\lambda_x} - m_{basis}\right) * \pi_{\infty}(x)$$

Abandonments

Finally, we want to compare the expected number of patients in the system when increasing capacity with the phenomenon that was described by Hall et al (2006) called reneging. In this situation patients leave the queue if their access time to the outpatient department is too long. This is modelled as an M/M/1 queue with abandonments, having a constant arrival rate λ, but the death rate consists of a summation of the service rate and a rate depending on the number of patients in the queue: $\mu + \theta \cdot i$, where μ represents the service rate $\left(\mu = \frac{1\,day}{m_{basis}}\right)$, θ represents the parameter of patients leaving the queue and i represents the state (number of patients in the system). Note that the birth- and death rates are no longer constant for $x > m = 300$, thus the summations needed to compute the expected number of patients in the system cannot be computed with the derivations above and only the first 300 states are taken into account. However, as the results of this summation is close to zero, the effect on the expected number of patients is negligible.

3.2 Clinical pathway

As explained in section 3.1, the model of the outpatient department can be rewritten as a birth-death process with constant birth rate λ and varying death rate depending on the capacity. According to Burke's theorem (Burke's Theorem 2016), in an M/M/1 queue with arrival rate λ, the departure process is also a Poisson process with parameter λ. If we interpret this in the clinical pathway, the arrival process of patients of the next node in the clinical pathway is again a Poisson process with parameter λ, independent of the use of extra capacity in the outpatient department.

Consequently, under the assumption of exponential service times, the clinical pathway can be modelled as a Jackson Network where every node can be modelled separately and every node does not influence the next node. When modelling the next node in the clinical pathway, in our example medical examination (for example a CT-scan), we can use the model of the outpatient department by changing some parameters. Again, we need to assume the service durations are exponentially distributed to obtain analytical results, although it is more likely the service durations are deterministic. Furthermore, In the case of medical examination it is less likely extra capacity is available as this does not only depend on the available time of the doctor's but also on the scanning equipment. Thus, for modelling medical examination we could even use the reduced model of the outpatient department, just having basis capacity.

Similarly, the model of the outpatient department could be rewritten to a model for surgery/ the operation room. In this case, it is unlikely extra capacity is available, as it depends on the availability of doctors, assistants, rooms and equipment. The assumption of exponential service times is more realistic for surgery as the time needed for surgery is more uncertain then at a scan or outpatient department visit.

The hospital stay is harder to model because the model of the outpatient department needs to be adapted to an M/M/S queue. The assumption of exponential service duration is plausible in this case, though the service duration is quite long compared to the outpatient department (often a few days instead of just a short appointment). Also in this situation, the use of extra capacity is not very likely, hospitals cannot easily add beds to a ward as an extra bed also comprises extra nurses and other care workers.

As all nodes in the clinical pathway could be modelled separately and do not influence each other, results could be obtained similar to those of the outpatient department. Therefore, only for the outpatient department results are obtained and described in the next section.

4. Results

Various results of performance measures are obtained for the outpatient department. First, the expected number of patients in the system $\mathbb{E}(L)$ is computed for all three structures of increasing capacity. In all scenarios, we take $m_{basis} = 4, m_{max} = 6.5$ and $\mu = 1$. These parameters are roughly based on the paper of Bussing (2012). To compare the effect of the three structures of λ_x, for both the second and third structure we choose $a_1 = 10$ and $a_2 = 60$. For the first structure (abrupt increase of the capacity) we choose $a_1 = 35$. Figure 3 shows the expected number of patients $\mathbb{E}(L)$ for the three options of increasing capacity in the range $\lambda = [0.2, 0.4, \dots, 6]$ and the basis scenario without increasing capacity for $\lambda = [0, 4)$, plotted against the basis load $\rho = \frac{\lambda/m_{basis}}{\mu}$.

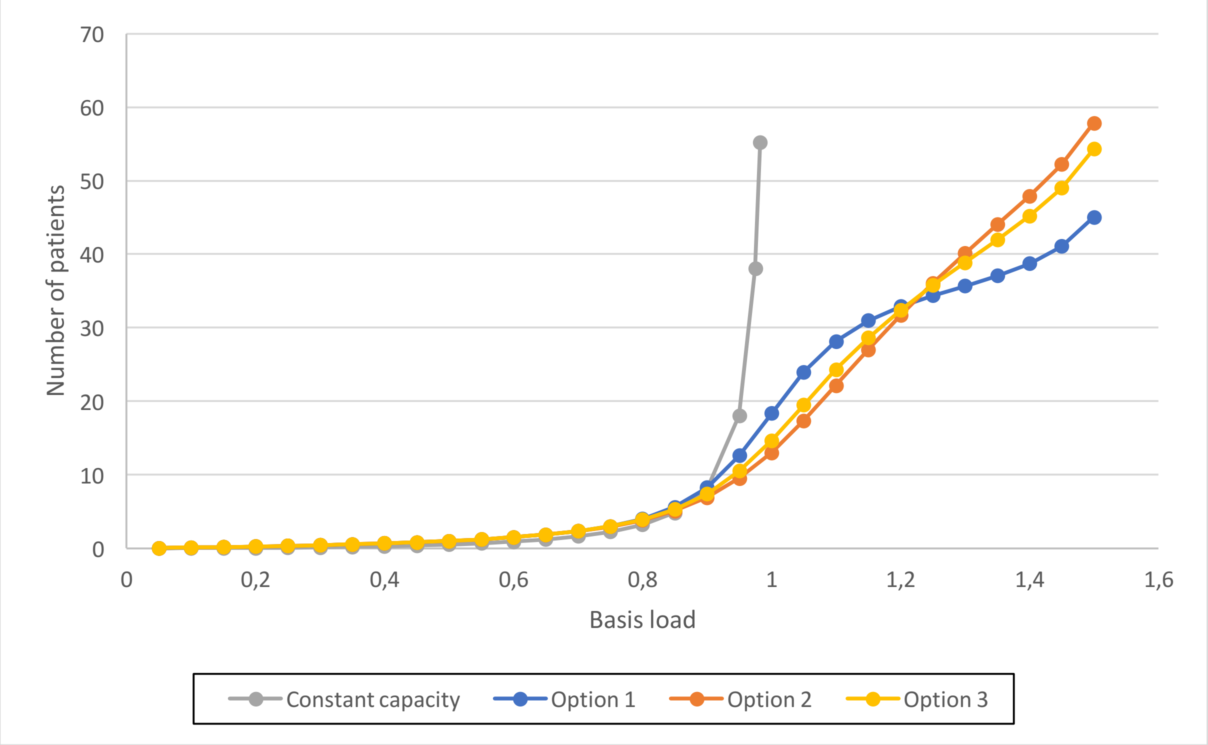


Figure 3: Expected number of patients in the system against load for basis capacity and three options of increasing capacity

We see that the expected number of patients in the system is lower in the cases where extra capacity is used from when the load approaches a value of 1. However, if the load becomes more than approximately 1.5, the system with extra capacity becomes instable and expected number of patients will tend to infinity again. Furthermore, the expected number of patients grows more fluently if the extra capacity is added fluently as in the second (linear) and third (Hermite spline) structures. The third structure does not have much influence on the expected number of patients in the system compared to the linear increase of capacity. It is not possible to estimate which structure is most realistic for the outpatient department, as we do not have data concerning the number of patients in the outpatient department system in relation to the used extra capacity available.

Figure 4 shows the corresponding relative extra capacity that is used for the three options, assuming the same parameters as before. We see the relative extra used capacity is almost equal for options 2 and 3, whereas more capacity is used when the first option for increasing capacity is applied.

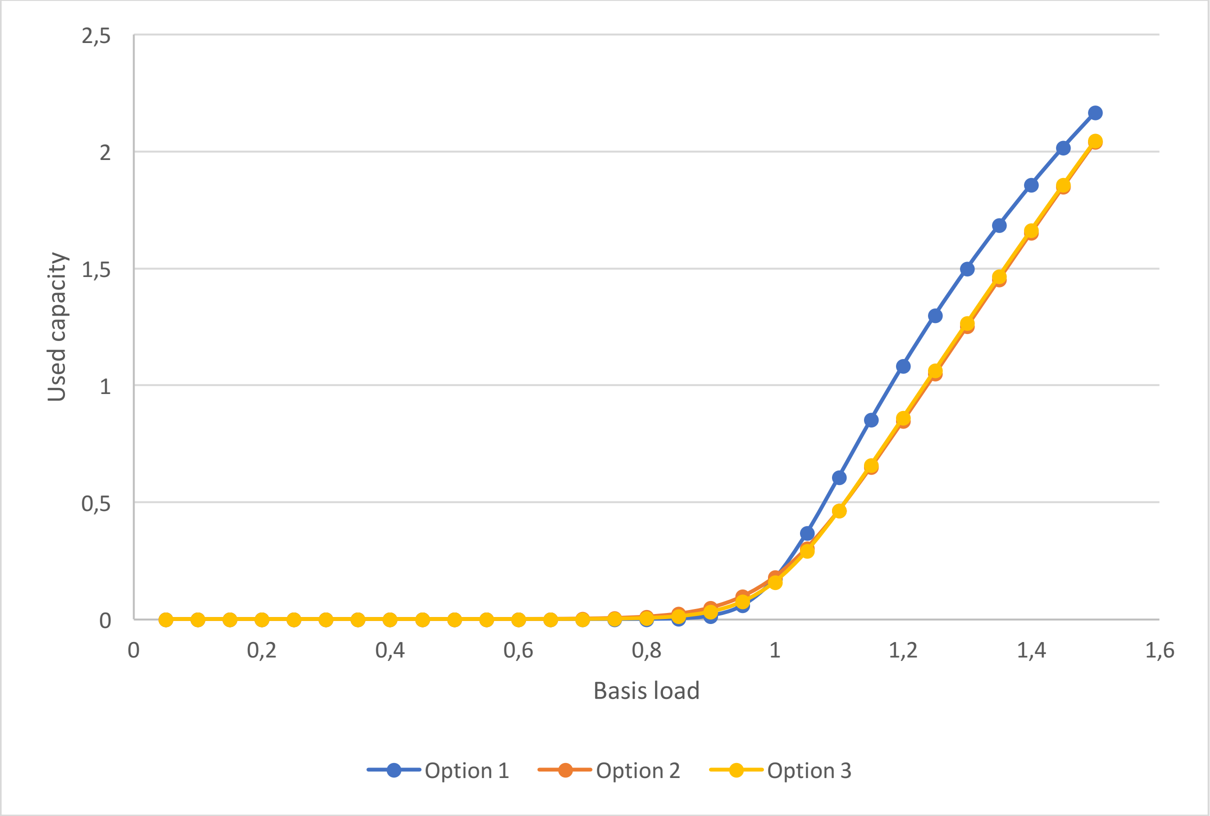


Figure 4: Relative extra capacity that is used against the load for three options of increasing capacity

Combining both figures 3 and 4, we see that if the load of the system is below 1.2, the second and third option, i.e. gradually increasing the capacity leads to both a lower expected number of patients in the system and a smaller amount of used extra capacity. Only if the load exceeds 1.2 the expected number of patients is smaller than when abrupt increase is used, as for the first option immediately the full available extra capacity is added. Nevertheless, in the scenario with parameters as above, we would advise to increase the capacity gradually rather than abrupt.

Influence of threshold parameters

To determine on which moment extra capacity should be used, we tried to gain some more insight in the influence of the parameters a_1 and a_2, the threshold for the backlog at which extra capacity is added. Figure 5 shows the expected number of patients in the system for the three different structures for various values of parameters a_1 and a_2. All other parameters are equal to the case above.

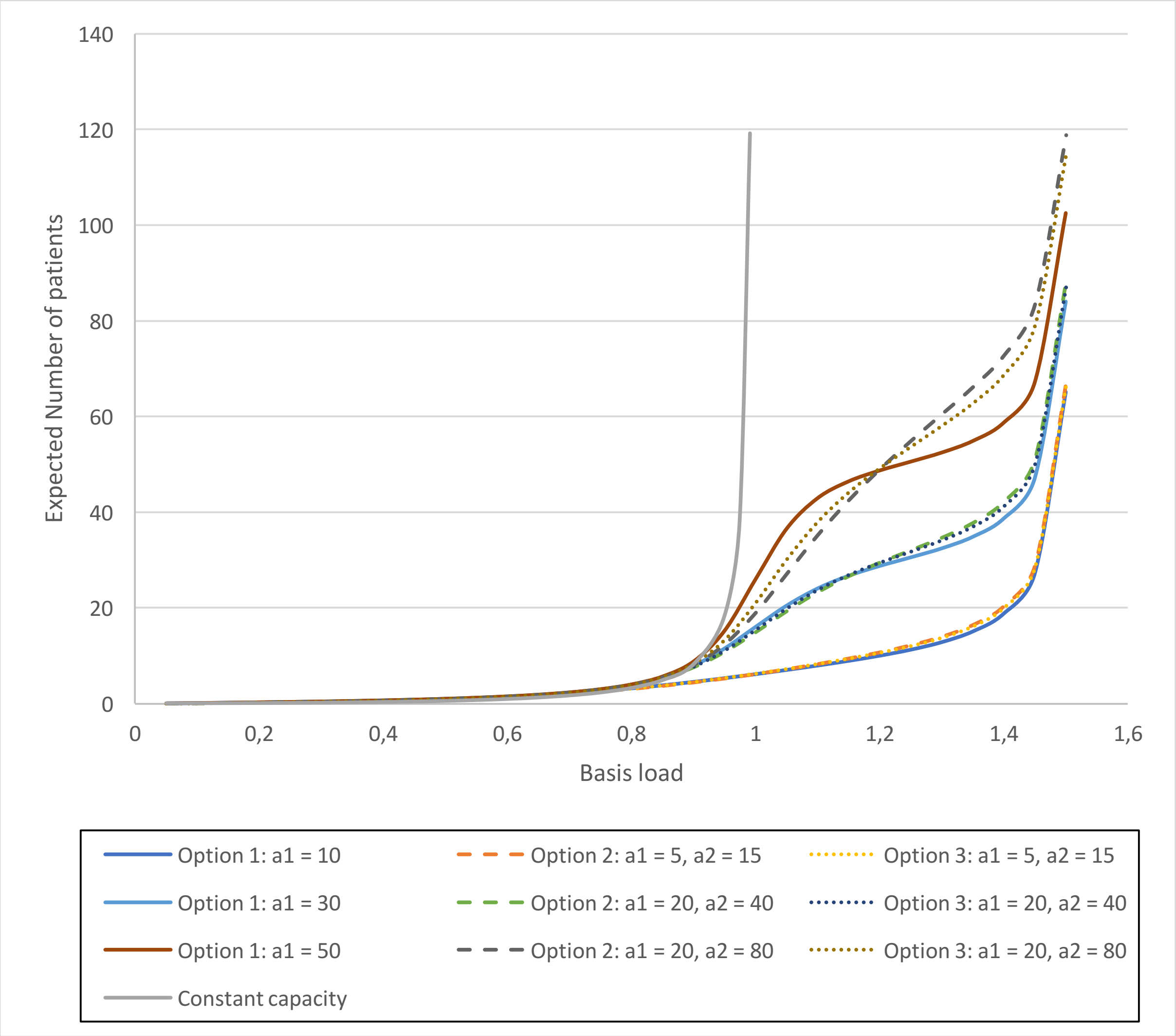


Figure 5: Expected number of patients in the system against load for various values of threshold parameters a_1 and a_2

Of course, lower threshold values a_1 and a_2 lead to a shorter expected queue, as extra capacity is added even when the backlog is relatively small. Note that if the thresholds are really small ($a_1 = 10/\ a_1 = 5$ and $a_2 = 15$) the shape of the graph of the expected number of patients looks like an elongated version of the graph of the expected number of patients without extra capacity, which is not realistic. It is more likely the queue will get longer for a while, after which extra capacity is made available which will shorten the waiting list. Apart from this, it is again hard to draw conclusions on which values for a_1 and a_2 are realistic and/or optimal because of a lack of data about the real situation in the outpatient department. The expected number of patients tends to infinity again if the basis load becomes approximately 1.5, for all structures and values of a_1 and a_2.

We see that the influence of the structure of adding capacity has most influence in the last row of parameters in the legend ($a_1 = 50/\ a_1 = 20$ and $a_2 = 80$). This is caused by the large spread in the values of parameters a_1 and a_2, which leads to a slow increase of the capacity in the second and third structure.

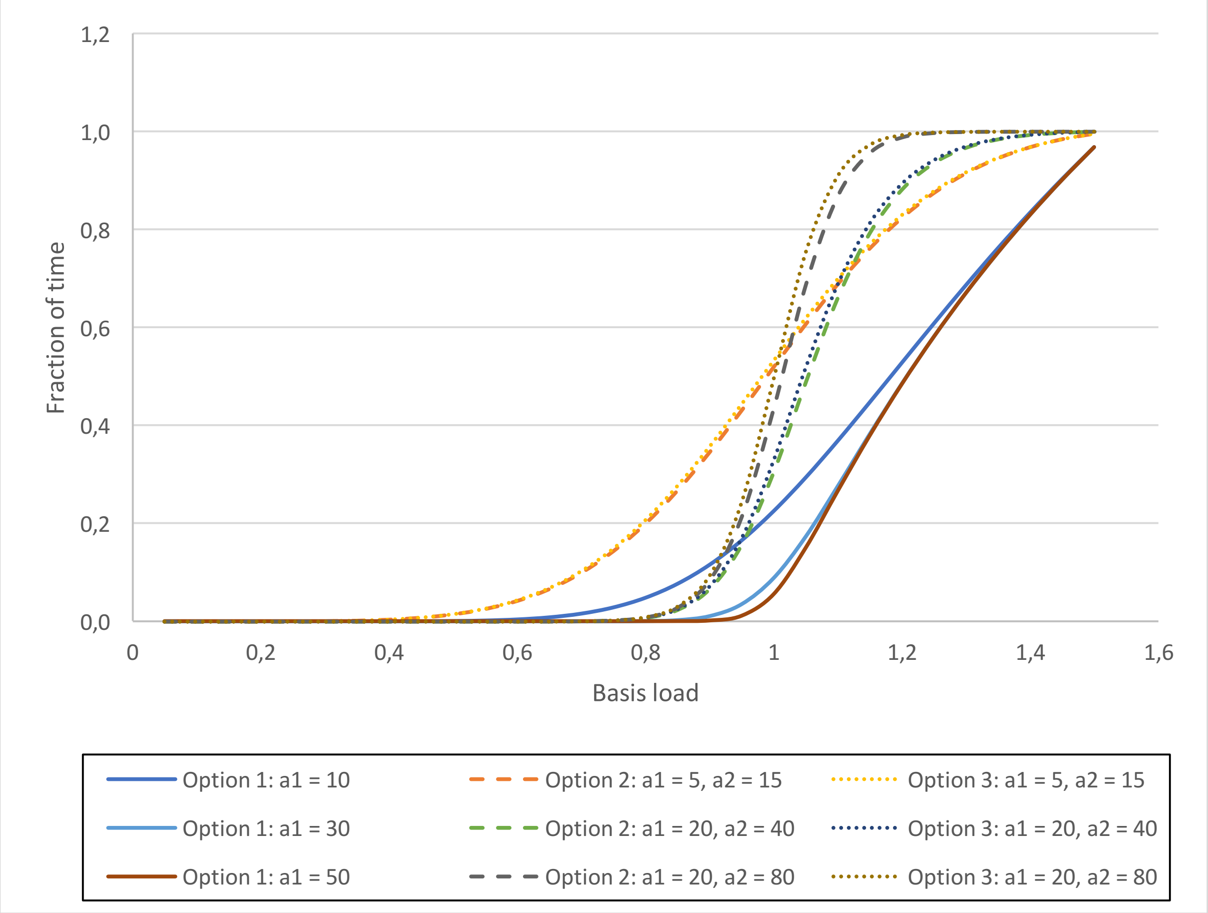


Figure 6: Fraction of time extra capacity is used against basis load for various values of threshold parameters a_1 and a_2

Figure 6 shows the fraction of time extra capacity is used belonging to the structures and parameters showed above. Clearly, the fraction of time extra capacity is used is larger if the parameter a_1 is smaller, as extra capacity is added from a lower state. There is not much difference in the fraction of time extra capacity is used between linear increase (option 2) and increasing capacity according to a Cubic Hermite Spline (option 3). We see that if the capacity is increased abruptly (option 1), the fraction of time extra capacity is used is lower than if the capacity is increased gradually, because the threshold for abrupt increase is higher (the mean of the parallel thresholds). In this figure, the fraction of used extra capacity is not taken into account.

Figure 7 shows the relative extra capacity used, we see that these graphs are all relatively close to each other. Of course, for lower threshold values a_1 and a_2 the relative extra used capacity is larger, as even when the backlog is small extra capacity is added. Furthermore, we see that abrupt increasing capacity (option 1) results in more used capacity. This can be explained by the fact that in option 1 immediately the maximum capacity is used if the threshold is exceeded.

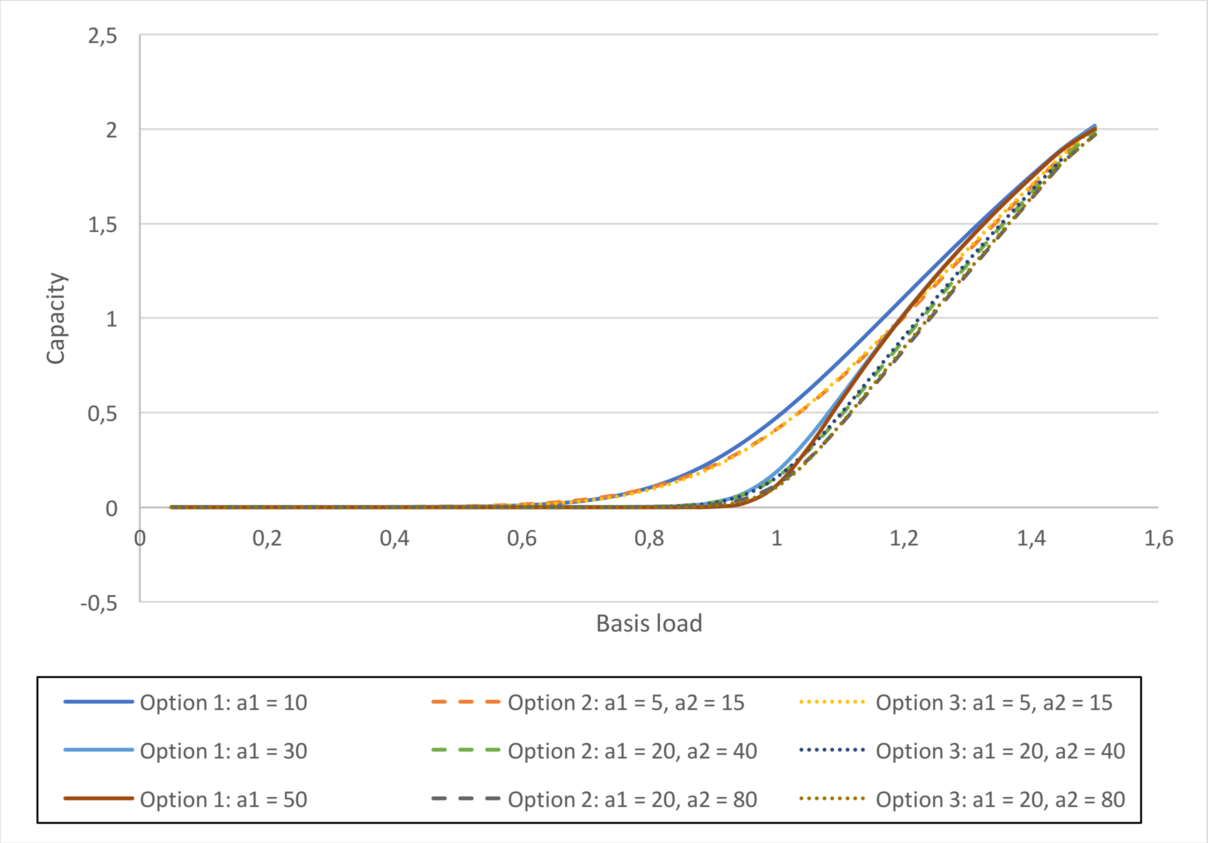


Figure 7: Relative extra capacity that is used against basis load for various values of threshold parameters a_1 and a_2

Abandonments

Now we compare the expected number of patients in the system of the situations in figure 3 with the situation where no extra capacity is available, but patients abandon the queue if the waiting list is too long. In figure 6 we see the expected number of patients as shown in figure 3 again, but the expected number of patients of the model with abandonments is added for two different values of the parameter θ. All other parameters are equal to the situation as shown in figure 3.

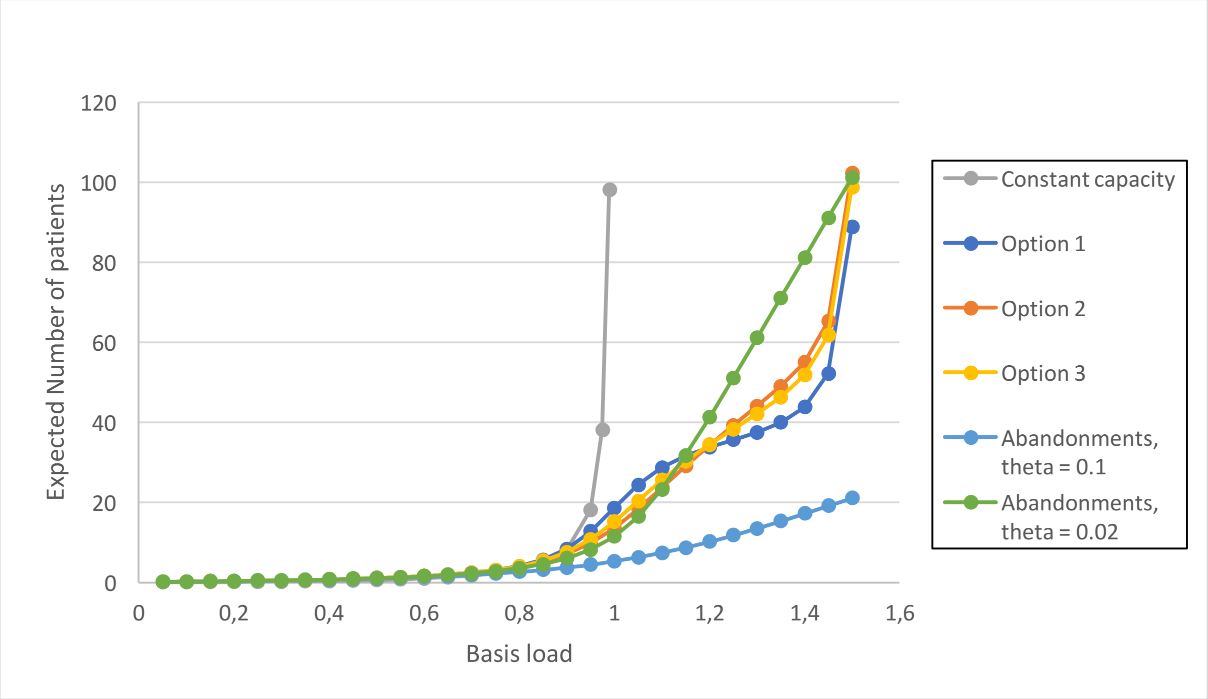


Figure 6: Expected number of patients in the system against load for the scenario of patients leaving the queue with rate θ, compared to figure 3

We see the expected number of patients strongly depends on the value of the parameter θ, which was not possible to estimate because we had no data available. If we expect a high abandoning rate θ, patients do not have much patience to wait and many patients leave the queue and for example choose to go to another hospital, which leads to a smaller expected number of patients. Note that the expected number of patients in the cases with abandonments remains stable, whereas the other situations tend to infinity when the load increases. Adding capacity leads to a stable system even when the load exceeds 1, but the system tends to infinity again when the load approaches 1.5.

5. Conclusion

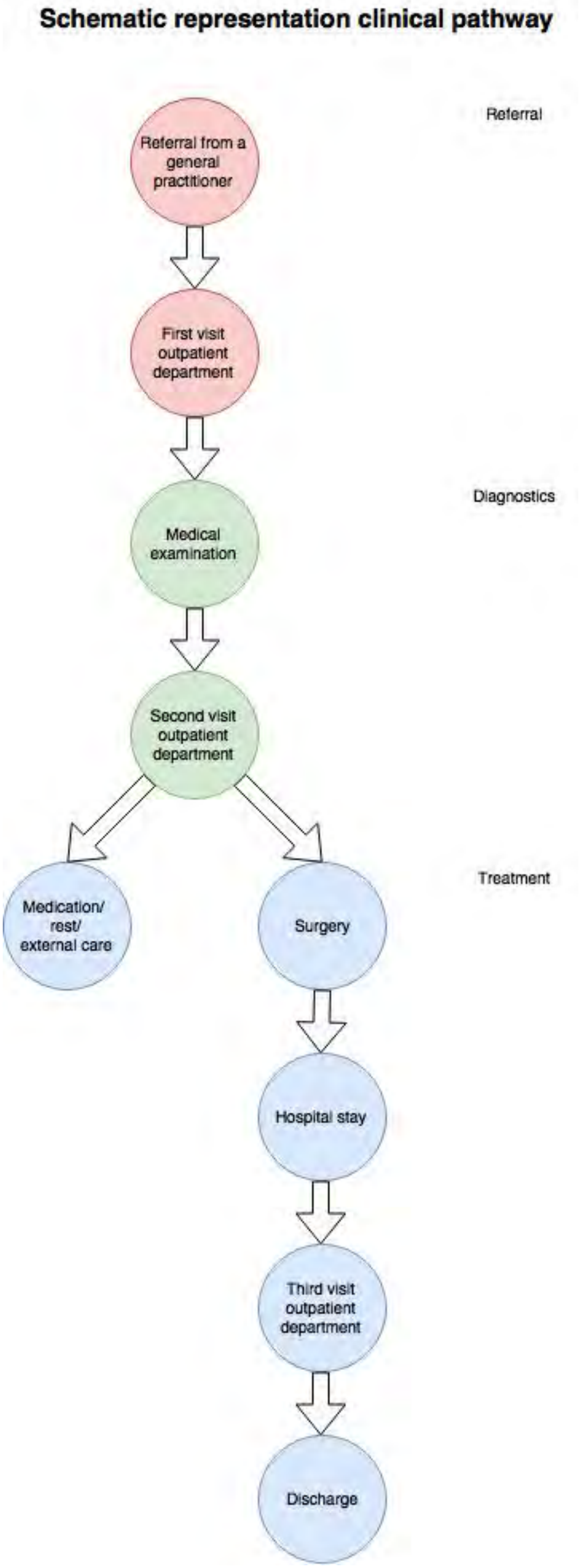
We derived a model for the clinical pathway assuming exponential service times, causing the model could be split in separate models for the various nodes in the network, as the nodes do not influence each other according to Burke's theorem. We expect increasing capacity in for example the first node (the outpatient department) will influence the arrival process of patients in the next node and the service times of appointments in the outpatient department are in real life deterministic. Therefore, we can conclude assuming exponential service duration is not appropriate in all nodes of the network. Interesting future research would comprise approaches to networks of queues where the service parameters depend on the state of the system without the exponential assumption, i.e. a network of M/G/1 queues. Based on the literature studied, we can expect it will be hard to model such a network without using simulation, as it is hard to obtain exact results for networks without the Markovian property.

Nevertheless, the model for the outpatient department was elaborated and leaded to interesting results. We see adding capacity significantly reduces the expected number of patients in the outpatient department system and therefore the access times of patients. We see the system remains stable until the load of the system is almost 1.5 if extra capacity is used, whereas the system without extra capacity becomes unstable if the load approaches 1. Furthermore, when taking both the expected number of patients and the relative used extra capacity in the system into account, we would advise to increase the capacity gradually rather than abrupt, as this leads to both a lower use of extra capacity and a lower expected number of patients until a load of approximately 1.2. However, we see it is hard to draw conclusions about which estimations for parameters and increasing capacity are realistic because we did not have data available. For further research, it would be interesting to collect data from outpatient departments and verify which of the increasing capacity structures described in this paper is most realistic and which threshold parameters are used in practice.

Additionally, it would be interesting to investigate whether outpatient departments perceive that patients leave the queue if they have to wait too long and in case they do, combine the effect of abandonments and using extra capacity and derive the influence on the expected number of patients in the system.

Appendices

Appendix A


Schematic representation clinical pathway
Referral
Referral from a general practitioner
First visit outpatient department
Diagnostics
Medical examination
Second visit outpatient department
Treatment
Medication/ rest/ external care
Surgery
Hospital stay
Third visit outpatient department
Discharge

References

Bhattacharjee, Papiya and Ray, Pradip Kumar. 2014. "Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections." *Elsevier Computers & Industrial Engineering Vol. 78.* pp. 299-312.

2016. *Burke's Theorem.* https://en.wikipedia.org/wiki/Burke%27s_theorem.

Bussing, Eline. 2012. "Capaciteitsverdeling zorgpaden." Vrije Universiteit Amsterdam.

2017. *Care Pathways.* http://e-p-a.org/care-pathways/.

2017. *Cubic Hermite spline.* August 1. https://en.wikipedia.org/wiki/Cubic_Hermite_spline.

Culyer, A.J. 1976. *Need and the National Health Service*. pp. 99.

Elkhuizen, S.G., Das, S.F., Bakker, P.J.M. and Hontelez, J.A.M. 2007. "Using computer simulation to reduce access time for outpatient departments." *Qual Saf Health Care Vol 5.* pp. 382–386.

Fomundam, Samuel and Herrman, Jeffrey. 2007. "A Survey of Queuing Theory Applications in Healthcare." *ISR TECHNICAL REPORT Vol.* 24 pp.1-22.

Hall, R., Belson, D., Murali, P. and Dessouky, M. 2006. "Modeling patient flows through the healthcare system." *Patient Flow: Reducing Delay in Healthcare Delivery, Hall, R.W.ed.* (Springer) pp. 1-44.

Iversen, Tor. 1993. "A theory of hospital waiting lists." *Journal of Health Economics Vol. 12* pp. 55-71.

Johansen, L. 1987. "Queues (and 'rent-seeking') as non-cooperative games, emphasizing mixed strategy solutions." In *Collective works of Leif Johansen Vol. 2*, by ed. F. Forsund, pp. 827-876. Amsterdam, North-Holland.

Roubos, Dennis. 2017. "Working paper: QNA with nodes in discrete time: A clinical pathway example."

Vis, Petra and Bekker, René. 2017. "Access times in appointment-driven systems and level-dependent MAP/G/1 queues."

Welch, J.D. and Bailey, Norman T.J. 1952. "Appointment systems in hospital outpatient departments." *The Lancet* pp. 1105-1108.

Whitt, W. 1983. "The Queueing Network Analyzer." *The Bell system technical journal.*

Worthington, D.J. 1987. "Queueing models for hospital waiting lists." *The Journal of the Operational Research Society, Vol. 38, No. 5* (Palgrave Macmillan Journals on behalf of the Operational Research Society) pp. 413-422.