How to deal with fluctuations in hospital processes to improve accessibility?

Joustra, P.E.

Citation for published version (APA):
The medical faculty is a collaborative effort of the Academic Medical Center and the University of Amsterdam.

"Even small healthcare institutions are complex, barely manageable places. Large healthcare institutions may be the most complex organisations in human history."

Peter Drucker
CHAPTER 1

General Introduction
1.1 The relevance of cost-effective ways for improving the accessibility of hospital care

Currently, hospitals have to deal with several external factors that require the efficient use of resources. Budget restrictions are forcing hospitals to work - and use their capacity - more efficiently, and at the same time, they are expected to deliver more care because of the aging of the population. In addition, the aging population will cause a structural shortage of health care professionals in the long term (see Figure 1.1). This situation requires hospitals to optimize their processes as much as possible and use their resources as efficiently as possible.

Moreover, hospitals are encountering more competition from new initiatives: private clinics that focus on specific patient groups (e.g., patients with cataract), and diagnostic centers (e.g., a specialized MRI center). Improved accessibility is known to be an important aspect of competition with other hospitals or specialized clinics. In addition to the distance involved and the recent experiences of either patients or their acquaintances with a hospital, low access times are important to general practitioners when deciding where to send their patients, and to patients when choosing a hospital [1]. This information is publicly available on the internet to make it easy for patients to compare the access times of various hospitals.

In addition, the Euro Health Consumer Index 2009 [2] reported that the Netherlands has the best health care system with the exception of accessibility, which, when compared to other European countries, could be improved.

From the patient’s perspective, low access times are preferable so that the period of uncertainty and inconvenience during their illness is kept as short as possible. More importantly, a patient’s condition may deteriorate while spending time on a waiting
list. For example, O’Rourke [3] concluded that lung cancer patients may become more seriously ill while waiting for radiotherapy treatment, therefore potentially requiring more complex treatment and/or decreasing the chance of full recovery. Furthermore, Huang et al. [4] concluded that for both breast cancer and head and neck cancer, the local recurrence rate was significantly higher for patients who had to wait longer for radiotherapy treatment.

Prentice and Pizer reported that the mortality rate within a 6-month follow-up period was significantly higher for patients who had to wait more than 31 days for an appointment at a geriatric outpatient clinic [5].

With respect to cardiac surgery, Plomp et al. [6] concluded that the causes and circumstances of death are waiting list-related for approximately 100 patients per year in the Netherlands. Moreover, Rexius et al. [7] concluded that long waiting lists for coronary artery bypass grafting are associated with considerable mortality, and the risk of death increases significantly with waiting time.

1.2 Why is it advisable to use quantitative decision support?

There are often multiple solutions available for improving the accessibility of hospitals. At the same time, the impact of a solution in a specific hospital setting cannot be known beforehand. To reduce the risk of implementing a solution that will not have the preferred impact on accessibility, it would thus be highly advisable to have quantitative decision support. With this quantitative support, the impact of each potential solution (or combination of solutions) could be predicted accurately before actually implementing the solution in practice. Moreover, it would provide an objective way of comparing alternative solutions before selecting the most cost-effective solution for meeting the performance target.

1.3 Why do waiting times occur?

There are different kinds of waiting times in hospitals: access times, throughput times, and in the waiting room. Waiting times are generally perceived as occurring due to a mismatch between demand and capacity. This is true to a certain extent. If demand structurally exceeds capacity, it is logical there will be waiting times. However, the mismatch between demand and capacity in hospitals is often only temporary and is caused by fluctuations - from day to day and week to week - in both demand as well as capacity. Although in most hospital departments the average capacity is sufficient for meeting the average demand, these temporary mismatches
do indeed result in waiting times for patients (see the example in Table 1.1). Note that because capacity that remained unused in the past cannot be used in the future, the average number of patients on the waiting list will keep growing.

To limit waiting times, management should add some slack capacity to compensate for the temporary mismatches due to fluctuations in demand and capacity. For example, if the capacity in Table 1.1 had been 7 instead of 6, the average number of patients on the waiting list would have been 0.7 instead of 2.7, though adding slack capacity results in a utilization rate of resources that is lower than 100%. The lower the preferred waiting times, the more slack capacity that should be added, and thus the lower the resulting utilization rate will be (see Figure 1.2). Therefore, urgent patients require more slack capacity than elective patients, and so an increase in the percentage of urgent patients in an outpatient department or operating room results in a lower utilization rate.

Although access times are often too high in practice, in most hospital settings they have been fairly stable over the past few years. This implies that capacity is sufficient for meeting demand (i.e., the utilization rate is below 100%). If the backlog could be reduced, often only a limited amount of slack capacity would have to be added to maintain these low access times.
To illustrate this, we used an outpatient department with an average of six new patients a day. Next, we used a standard formula from the literature on queuing theory [8] to determine the maximum utilization rate for various threshold values of the access time target (see Table 1.2). The results show that 1% (= 99.4% - 98.4%) slack should be added to maintain an average access time of two weeks compared to an average access time of six weeks (which is about one additional consultation in three weeks). Note that a structural reduction in access times from two weeks to only one week requires significantly more slack capacity than a reduction from six weeks to two weeks.

Table 1.2: Average access time versus utilization rate

<table>
<thead>
<tr>
<th>Average access time</th>
<th>Utilization rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 weeks</td>
<td>99.4%</td>
</tr>
<tr>
<td>5 weeks</td>
<td>99.3%</td>
</tr>
<tr>
<td>4 weeks</td>
<td>99.1%</td>
</tr>
<tr>
<td>3 weeks</td>
<td>98.9%</td>
</tr>
<tr>
<td>2 weeks</td>
<td>98.4%</td>
</tr>
<tr>
<td>1 week</td>
<td>96.8%</td>
</tr>
</tbody>
</table>

Table 1.2 also demonstrates that simply striving for a higher utilization rate will result in higher waiting times. Therefore, management should not aim to maximize the utilization rate exclusively, but should first decide upon the preferred waiting times in the waiting room and access times and then maximize the utilization rate according to these quality restrictions. Although adding slack capacity does indeed result in lower waiting times, there might be more efficient ways of reducing waiting times. In the remainder of this section, we will illustrate two ways for improving accessibility efficiently.

One way to efficiently reduce waiting times is to minimize the root cause of the waiting times, namely the fluctuations in demand and capacity. To illustrate this principal, we used the Pollaczek-Khintchine formula [9] to experiment with the degree of fluctuations in process times. The Pollaczek-Khintchine formula expresses that the average waiting time of a system with fluctuations in the process times \( W_{M|G|1} \) is a function of the variation coefficient of the process times \( c_s \) is the standard deviation divided by the mean process time) and the average waiting time with fixed process times \( W_{M|D|1} \). Formula (1) is scientifically correct for a single capacity unit.

\[
(1) \quad W_{M|G|1} = (1 + c_s^2) \cdot W_{M|D|1} \quad \text{with} \quad c_s^2 \text{ the variation coefficient of the process time distribution.}
\]
Table 1.3 contains four scenarios we used to illustrate the effect of a reduction in fluctuations in process times. The basic scenario has high fluctuations in process times (i.e., $c_s^2 = 1$) and the utilization rate is 95%, which results in an average waiting time of 19 (see Table 1.3). The first alternative scenario shows that fewer fluctuations in process times (i.e., $c_s^2 = 0.5$) with an equal utilization rate results in a lower average waiting time. The second alternative scenario demonstrates that the variation reduction in process times can also be used to increase the utilization rate while maintaining the same average waiting time as in the basic scenario. The final alternative scenario shows that a combination of both advantages – higher utilization rate and lower average waiting time – is also possible.

A second way to efficiently reduce waiting times is to pool various patient groups. The capacity of shared resources of central diagnostic departments in particular is often subdivided among many patient groups that differ from a medical perspective. It might be more efficient, though, to pool logistically similar patient groups. To illustrate the effect of increased economy of scope, we defined four scenarios. The basic scenario contains two patient groups with similar process times; each group has one dedicated capacity unit, and the utilization rate of both capacity units is 95%. Using a standard formula from the literature on queuing theory [8], this results in an average waiting time of 19 for both patient groups (see Table 1.4).

Table 1.4: Effect of increased economy of scope on waiting times

<table>
<thead>
<tr>
<th>Scenario</th>
<th>No. of capacity units</th>
<th>Utilization rate</th>
<th>Avg. waiting time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>2 x 1</td>
<td>95.0%</td>
<td>19</td>
</tr>
<tr>
<td>Alternative 1</td>
<td>1 x 2</td>
<td>95.0%</td>
<td>9</td>
</tr>
<tr>
<td>Alternative 2</td>
<td>1 x 2</td>
<td>97.5%</td>
<td>19</td>
</tr>
<tr>
<td>Alternative 3</td>
<td>1 x 2</td>
<td>96.6%</td>
<td>14</td>
</tr>
</tbody>
</table>

Alternative 1 shows that if we were to pool the two patient groups, the average waiting time for the pooled group would decrease to 9. Alternative 2 shows that the advantage of the increased economy of scope can also be used to increase the utilization rate to 97.5% while maintaining the same average waiting time as in the basic scenario. The last alternative scenario shows that a combination of both advantages – higher utilization rate and lower average waiting time – is also possible.
1.4 Settings

The research contained in this thesis was performed at the Academic Medical Center (AMC) in Amsterdam, the Netherlands. The AMC is an academic hospital with 1,002 beds, and 356,000 outpatient hospital visits and 55,000 surgical procedures in 2009. Nearly all departments use McKesson’s X/Care hospital appointment system to schedule the outpatient visits, and most surgical procedures are scheduled using the OKplus operating room information system. Elective patients are allowed to walk in without an appointment for only a small number of consultations and examinations.

To improve accessibility by reducing fluctuations in demand and capacity, we first need to know what types of fluctuations frequently occur in hospitals. An overview of these types is given in this section. We start by identifying the fluctuations that occur in all hospitals followed by the additional fluctuations in teaching hospitals and academic hospitals.

As is common in all hospitals, the number of patients referred to the various outpatient departments by general practitioners and other physicians fluctuate from day to day and week to week. Some outpatient departments may even have a seasonal pattern in the number of referrals. Also, the number of follow-up consultations may differ per patient, and the average number of follow-up consultations often differs significantly per subspecialty and/or per individual physician. As well as elective patients, physicians also have to see urgent patients. All these fluctuations in demand generally lead to access times for a first, a follow-up, or an urgent (or semi-urgent) consultation, even with a stable capacity. However, the capacity in terms of the number of available consultations also fluctuates heavily from day to day and week to week. In addition, all outpatient departments within the AMC experience a seasonal pattern in capacity due to holiday periods. On the operational level, a physician’s starting time and the arrival time of a patient with an appointment vary, which results in waiting time in the waiting room. At specific outpatient departments, patients need some preparation before a consultation (e.g., at an ophthalmology department), which causes even more fluctuations in the arrival time of patients in the consultation/examination room. Moreover, the actual duration of a consultation may differ per patient, and the average duration might differ per subspecialty and/or per individual physician.

There are more fluctuations in teaching hospitals than in non-teaching hospitals. For example, the average number of follow-up consultations done by a resident physician is likely to be higher than those done by an attending physician, and also the average number of follow-up consultations among the resident physicians probably differs more than those among the attending physicians. There are more fluctuations in teaching hospitals on an operational level as well. For example, due to a lack of experience, the duration of consultations performed by a resident physician...
is likely to fluctuate more than those performed by an attending physician. Moreover, if an attending physician is also seeing patients during the supervision period - which is often the case in the AMC - this supervision of a resident physician will cause interruptions in both their schedules and therefore more fluctuations in the duration of consultations.

Physicians in academic hospitals are likely to be absent more frequently due to other reasons (such as attending conferences and teaching medical students), which causes more fluctuations in the weekly number of available consultations. Also, academic hospitals have a more diverse and complex patient-mix compared to general hospitals. This often results in a higher degree of subspecialization and thus smaller groups of physicians per subspecialty, which reduces robustness and therefore often results in higher access times due to less flexibility in allocating the available capacity. On the operational level, the more diverse and complex diseases probably cause more fluctuations in the duration of consultations. In addition, research may require that specific patients are examined more extensively, which leads to more fluctuations in the duration of consultations.

Table 1.5 summarizes the types of fluctuations that occur in an outpatient department. The degree of fluctuations depends on the type of hospital; in general, academic hospitals show a higher degree of fluctuation than teaching hospitals, which in turn show more fluctuations than general hospitals.

The same types of fluctuations that occur in outpatient departments also occur in emergency departments, central diagnostic or therapeutic departments, and operating rooms.

Table 1.5: Types of fluctuations on tactical and operational levels

<table>
<thead>
<tr>
<th>Types of fluctuations in an outpatient department</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>On a tactical level</strong></td>
</tr>
<tr>
<td>Number of patients per day, per week, per month</td>
</tr>
<tr>
<td>Number of follow-up consultations per patient, per subspecialty, per individual physician</td>
</tr>
<tr>
<td>Various urgency levels</td>
</tr>
<tr>
<td>Weekly number of available consultations</td>
</tr>
<tr>
<td>Subspecialty required by patient</td>
</tr>
<tr>
<td><strong>On an operational level</strong></td>
</tr>
<tr>
<td>Physician’s starting time</td>
</tr>
<tr>
<td>Patient’s arrival time</td>
</tr>
<tr>
<td>Preparation time for a patient</td>
</tr>
<tr>
<td>Actual duration of a consultation per patient, per subspecialty, per individual physician</td>
</tr>
<tr>
<td>Attending physicians versus resident physicians</td>
</tr>
<tr>
<td>Supervision of a resident physician by an attending physician</td>
</tr>
</tbody>
</table>
1.5 Research question

The overall research objective of this thesis is to show the added value of quantitative methods such as queuing theory, discrete event simulation, and regression analysis for generally applied process improvement methods, in a complex hospital setting, in particular for decision support on a strategic level.

In addition to the overall research objective, this thesis aims to answer the following specific research questions:

1. How can waiting times, access times, and throughput times be reduced cost-effectively
   a. by decreasing the various fluctuations in health care processes and/or
   b. by pooling or separating various patient groups?
2. How can hospital processes be made more predictable so that management is able to anticipate future developments proactively, and structurally improve the hospital’s key performance indicators?

1.6 Methodology

Based on the different types of fluctuations that frequently occur in hospitals, we explored different methods for decreasing these fluctuations or minimizing the negative effect of these fluctuations on waiting and access times.

Generally applied process improvement methods such as Business Process Redesign [10], Total Quality Management [11], Theory of Constraints [12], Lean Management [13], and Six Sigma [14] give insights into which process steps to focus on to reduce waiting and access times or throughput times. Unfortunately, none of these methods quantitatively supports the expected impact of a potential solution on the waiting and access times before implementation in practice, nor do these methods quantitatively support the selection of alternative solutions.

Each of our research studies required quantitative decision support. Therefore, we have searched for quantitative methods that could be applied in situations where various fluctuations significantly influence a system’s performance, as is the case in hospitals. Queuing theory is an appropriate method for analyzing waiting times, and was frequently applied in health care settings [15, 16]. Discrete event simulation was often used for more complex hospital problems [17-19].

Though both of these methods were used to solve problems on an operational or tactical level, we did not find any articles where queuing theory or discrete event simulation were used for decision support on a strategic level.
1.7 Outline of this thesis

Chapters 2 through 8 are based on articles. Because this thesis should be seen as applied research, the order in which they are presented here is based primarily on the application area. Therefore, we have chosen to arrange the chapters according to hospital departments rather than the methodology. Subsequently, we arranged the articles so that sets of articles used to answer the same specific research question were combined wherever possible.

This reasoning led to the following order:
Chapter 3  To pool or not to pool in hospitals: A theoretical and practical comparison for a radiotherapy outpatient department.
Chapter 4  Reduce access times for an endoscopy department by an iterative combination of computer simulation and Linear Programming.
Chapter 5  Reducing MRI access times by tackling the appointment-scheduling strategy.
Chapter 6  How to reduce waiting times at the MRI of a University Radiology Department.
Chapter 7  Can statisticians beat surgeons at the planning of operations?
Chapter 8  How to juggle with priorities? An interactive tool to provide quantitative support for strategic patient-mix decisions.

Chapters 2 and 3 illustrate two alternative solutions for efficiently reducing access times in the AMC radiotherapy outpatient department to be able to meet the throughput time targets for radiotherapy treatment. The next three chapters all describe case studies of a central diagnostic department, namely the endoscopy department (Chapter 4) and the MRI department (Chapters 5 and 6). Chapters 4 and 5 demonstrate how to reduce access times for an endoscopic procedure and an MRI scan respectively. In Chapter 6, we evaluate various alternative scenarios to reduce waiting time in the waiting room of an MRI department. Finally, the last two chapters of this thesis focus mainly on the performance of the operating room. Chapter 7 explores how to improve the accuracy of the predicted duration of surgical procedures in order to limit the risk of overtime and the number of cancellations due to overruns of previous surgeries. Chapter 8 describes an interactive tool for supporting medical management with strategic patient-mix decisions, and takes the key performance indicators into account. In our case study of the AMC ophthalmology department we focused on the bottleneck, namely the operating room.
CHAPTER 1

Chapters 2, 4, and 6 were used to answer research question 1a. These studies demonstrate how decreasing the various fluctuations in health care processes result in lower throughput times, access times, and waiting times in the waiting room respectively. Chapter 2 explores the effect of a reduction in the fluctuations of capacity in the multi-step radiotherapy process in order to decrease the total throughput times. Chapter 4 also shows how a reduction in fluctuations in capacity could improve the accessibility of (in this case) an endoscopy department. Chapter 6 demonstrates two examples of variability reduction of MRI durations, namely reducing the supervision by radiologists, and scheduling patients at high risk of unexpected issues (e.g., patients with claustrophobia) for the open MRI scanner or at the end of the schedule for the regular MRI scanners. Chapters 3, 4, and 5 deal with the question of whether to pool or separate patient groups (research question 1b) for two different reasons and three different departments. More specifically, Chapter 3 explores the situations in which semi-urgent and elective patient groups should be pooled or separated. Furthermore, Chapters 4 and 5 both investigate whether pooling different groups of elective patients would reduce the access times for an endoscopic procedure and an MRI scan respectively.

To answer research question 2, we conducted the research described in Chapters 6, 7, and 8. In Chapter 6, we used variability reduction as well as improved predictability to reduce waiting times in the waiting room of an MRI department. Finally, Chapters 7 and 8 aim to improve the predictability of hospital processes so that management is able to anticipate future developments proactively instead of reactively.

With respect to methodology, in Chapters 2 and 3 we used a combination of queuing theory and computer simulation; in Chapters 4 through 6 we used only computer simulation. Finally, to improve predictability we applied regression analysis; in Chapter 8 we combined this with computer simulation.

Table 1.6 provides an overview of this thesis including the research question, hospital department, organizational level, research question and quantitative method(s) per chapter.
Table 1.6: Overview of this thesis

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Hospital department</th>
<th>Organizational level</th>
<th>Research question</th>
<th>Quantitative method(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Radiotherapy</td>
<td>Tactical</td>
<td>1a. Variability reduction</td>
<td>Queuing theory and computer simulation</td>
</tr>
<tr>
<td>3</td>
<td>Radiotherapy</td>
<td>Tactical</td>
<td>1b. To pool or separate</td>
<td>Queuing theory and computer simulation</td>
</tr>
<tr>
<td>4</td>
<td>Endoscopy</td>
<td>Tactical</td>
<td>1a. Variability reduction</td>
<td>Computer simulation and integer linear programming</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1b. To pool or separate</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MRI</td>
<td>Tactical</td>
<td>1b. To pool or separate</td>
<td>Computer simulation</td>
</tr>
<tr>
<td>6</td>
<td>MRI</td>
<td>Operational</td>
<td>1a. Variability reduction</td>
<td>Computer simulation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Improved predictability</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Operating room</td>
<td>Operational</td>
<td>2. Improved predictability</td>
<td>Regression analysis</td>
</tr>
<tr>
<td>8</td>
<td>Operating room</td>
<td>Strategic</td>
<td>2. Improved predictability</td>
<td>Regression analysis and computer simulation</td>
</tr>
</tbody>
</table>
CHAPTER 1

References


