Lean beyond waste

Roemeling, Oskar-Paul

CHAPTER 4

Exploring variability and buffers in Lean initiatives in healthcare

4.1 INTRODUCTION

Adopting Lean concepts is one of the more popular strategies that healthcare organizations employ in an attempt to improve their performance (e.g. Serrano *et al.*, 2010; Hicks *et al.*, 2015), and improving patient flow is one of the main drivers of Lean initiatives. Improvements in flow can be obtained through reductions in variability and buffers, important Lean aspects (Hopp and Spearman, 2004).

According to Poksinska (2010), there is a need for more rigorous research in order to fully understand the underlying factors, which should include variability and buffers, that influence the impact and success of Lean initiatives in healthcare. In a review by Marodin and Saurin (2013) it becomes evident that Lean its theoretical underpinnings are still largely unclear. Andersen *et al.* (2014) similarly argued that the characteristics of Lean and its local application should be given more attention. More recently, Jasti and Kodali (2015) established that there is still need for more empirical research focused on Lean.

In response to these calls, and using case research, we investigate variability and buffers in Lean initiatives aimed at

improving patient flow. With this study we provide a stronger empirical basis for Lean and the underlying roles of variability and buffers. We study three different patient flows, in three different hospital departments. In this study, we adopt the definition of Hopp and Spearman (2004) by defining a Lean approach in services as the delivery of a service with the use of minimal buffers.

Buffers are seen as a consequence of variability, and variability is one of the main obstacles to a smooth flow (Schmenner and Swink, 1998). Further, alterations aimed at increasing flow will result in process changes that are likely to have a significant impact on performance (Fredendall *et al.*, 2009), and improvements in patient flow should lead to increased hospital productivity and greater patient satisfaction (Litvak, 2009; Villa *et al.*, 2009). Despite these positive assertions, the Lean-oriented literature is often lacking a focus on the variability and buffers that influence patient flow.

In this research, we will focus on the roles of buffers and reducing variability in patient flows, and elaborate on the interaction between the various forms of buffers. Through a multiple case study, we seek to answer the following research questions: (1) how do buffers and variability interact in a healthcare environment, and (2) how does the interchange, or 'trade-off', between buffers manifest itself in practice compared to theory? The remainder of this paper is structured as follows: first, we provide our theoretical background where we elaborate on the application of Lean ideas in healthcare and introduce the main concepts used - buffers and variability. Then we explain our

adopted methodology in more detail, coupled with a description of the cases studied. Following this, we present our analysis and results, showing the unique variability and buffer complexity in each case. The paper then ends with a discussion and draws preliminary conclusions based on our findings from which we present propositions for further testing.

4.2 THEORETICAL BACKGROUND

Krafcik, a researcher in the MIT international motor vehicle program in the 1960s and 1970s, coined the term Lean to label the new approach to manufacturing seen in Japanese industries (Krafcik, 1988; Womack *et al.,* 1990; Warnecke and Hüser,1995). The Lean concept originated at Toyota, in a production environment, and has become increasingly popular ever since (Bhamu and Sangwan, 2014).

Lean has since been applied not only in production settings but also in service environments including healthcare. The interest in applying Lean ideas to services began later than in production settings (Bowen and Youngdahl, 1998), and the transfer of Lean principles to healthcare environments is still seen as relatively novel (Burgess and Radnor, 2013). Nevertheless, Lean has been shown to improve healthcare quality, access, and efficiency (Mazzocato *et al*., 2010).

Hassell *et al.* (2010) report on the application of Lean principles in a medical laboratory, and show increased productivity alongside other improved performance outcomes. In a study by Chiodo *et al.* (2012), Lean ideas were successfully applied to improve the throughput of patients in a rehabilitation unit, with the result that

patients were now admitted and discharged on time. However, given that Lean ideas originated in a production setting, authors agree that Lean approaches cannot simply be copied from production to healthcare environments (Hines *et al.*, 2004).

Several studies have concluded that Lean applications in healthcare are inconsistent and fragmented (e.g. Ballé and Régnier, 2007; Proudlove *et al.*, 2008; Young and McClean, 2008). Additionally, Radnor and Osborne (2013) reported that studies on Lean in healthcare had mostly focused on the use of tools and practices. It would seem that Lean studies in the healthcare environment have focused on visible or practical aspects, and fail to address the underlying elements and characteristics (Andersen *et al.*, 2014; Hines *et al.*, 2008; Radnor, 2010).

In order to avoid the danger of selectively applying Lean practices, authors stress that a more holistic approach is necessary (Radnor *et al.*, 2012; Towill and Christopher, 2005; Waldman and Schargel, 2006). A holistic view requires organizations to take all the aspects of Lean into account, which means maximizing customer value through reducing waste and variability (Jayaram *et al.*, 2008; Shah and Ward, 2007).

Variability is an important factor in healthcare (Noon *et al.*, 2003; Allder *et al.*, 2010), and should have a central role in Lean applications. Given this importance, it is useful to distinguish clearly between natural and artificial variability (Litvak and Long, 2000). Natural variability is a given, and always present. Conversely, artificial variability exists because of a dysfunctional process (here within a healthcare organization) (Litvak and Long,

2000). In essence, artificial variability is a consequence of decisions made in an organization.

Alongside the fundamental distinction between natural and artificial variability, Litvak and Long (2000) propose a further distinction, identifying clinical, flow, and professional variability. Whereas these authors used this additional distinction only for natural variability, we would propose expanding their typology to include professional and flow variabilities as subtypes of artificial variability. However, we would expect clinical variability, the way patients respond to treatment, to be a purely natural affair, as the condition and recovery of patients can only be influenced by a healthcare provider to a very limited extent. We argue that flow and professional factors can cause artificial variability, for example when a patient is brought late into the operating room (flow variability) or a clinician's unfamiliarity with new technology slows the process (professional variability).

Variability influences the flow and results in the presence of buffers (Schmenner and Swink, 1998; Hopp and Spearman, 2004) in terms of either additional patient waiting time (patients waiting for treatment) or excess capacity (doctors waiting to treat), both leading to unnecessary healthcare costs. In a perfect world, these types of variability would be eliminated.

Buffers can be subdivided into capacity, time, and inventory buffers. An inventory buffer can relate to a stock of medicines, or medical attributes such as syringes. Inventory buffers cannot be used to absorb fluctuations in either flow variability or professional variability related to patient flow. This indicates that inventory buffers might be less relevant when addressing the core

healthcare process of providing patient care. Instead, healthcare providers are forced to manage variability through the use of time and capacity buffers.

In a hospital, capacity buffers could consist of an excess of nurses, physicians, or medical equipment, and would show itself in idle capacity or capacity waiting to deliver treatment. A time buffer on the other hand amounts to patients waiting to receive treatment, for example in the form of waiting lists. As such, patient flow problems result in either waiting patients or waiting staff (Mazzocato *et al.*, 2012).

To achieve a continuous flow, relevant stakeholders have to gain an understanding of the different processes, be able to identify waste, and investigate the source of the problems at hand (Poksinska, 2010). In particular, sources of variability that lead to buffers are crucial for improving flow performance and are integral to applying Lean ideas. In addressing these aspects, this research aims to create a better understanding of the link between variability and buffers, and the factors that add complexity to this relationship when focusing on patient flows.

Ultimately, our aim is to formulate propositions that can further Lean theory, especially where buffers and variability are concerned. In this study, we investigate the sources of artificial variability that hinder patient flows in a range of hospital departments, and we identify the relationship between the time and capacity buffers that are the consequence of artificial variability.

4.3 METHODOLOGY

We adopted a case study approach in this research because it is a suitable option for theory building, and for creating an understanding of a phenomenon in practice (Karlsson, 2009). In addition, case research is especially appropriate when, as in our healthcare environment, context knowledge is important (Voss *et al.*, 2002).

We have chosen to conduct a multiple case study since this should allow a thorough in-depth analysis of the relationship between variability and buffers in our selected cases (Voss *et al.*, 2002; Burgess and Radnor, 2013). Further, the investigation of multiple cases typically provides stronger support for theory than single case studies (Yin, 2009). Furthermore, multiple cases help prevent observer bias and strengthen validity (Karlsson, 2009), and we obtained data from different sources to enhance triangulation.

Through pattern matching we were able to compare empirically found patterns with patterns predicted by theory (Yin, 2009), and we can attempt to identify interchanges between buffers as described by Hopp and Spearman (2004).

Since we wished to focus on situations where improvements to patient flows were realized, our selected unit of analysis was the patient flow. Consequently, we searched for patterns to identify which buffers and variability sources were present, and how these affected patient flow. Our analysis started with a within-case analysis where we identified the main causes of variability in each flow, and looked for the buffering mechanism employed to cope with the specific variability. Here, we also investigated disruptive

factors since we are interested in issues that add complexity to the buffer - variability relationship.

Onsite visits were undertaken to gain deeper insight into the patient flows under study. As part of these site visits, we observed the daily routines of physical therapists and we took field notes during the onsite visits and observations. These observations were crucial in clarifying why certain variabilities and buffers existed and how they interacted, and this enhances the interpretive validity of our study (Karlsson, 2009). Data from the observations were analyzed by comparing the field notes collected. The case sites were visited several times during the research period to discuss and reflect on intermediate findings.

In addition to observations and case visits, we obtained secondary data on patient flows. These data consisted of internal documents, Excel data files, and presentations that reported on the steps that were undertaken as part of projects to improve patient flow. Hospital databases provided insight into access times, process structures, waiting times, and planning schedules.

To provide a richer story we conducted three semi-structured interviews with the unit managers responsible for patient flows in the various departments. The unit managers should be considered to be the most knowledgeable on the specific interventions and the related outcomes. These interviews were based on an interview protocol to enhance reliability and lasted between 45 and 60 minutes; they were recorded and transcribed within 24 hours as recommended by Eisenhardt (1989). These interviews provided insight into patient flows and aided the identification of flow

problems. The specific interview questions are included in appendix C.

The formal interviews were complemented by unstructured interviews with improvement project leaders, secretaries, a radiologist, and a surgeon. These interviews provided an understanding of the approaches that different disciplines took in coping with patient flow problems. The categorization of the interviews started by determining what type of variability or buffer was being discussed. Then, through deductive reasoning, we identified relevant quotes and irrelevant data. The combination of different types of collection methods and data sources facilitated construct validity, which is essential to theory building (Karlsson, 2009).

Cases

In our study, we focus on three different departments in a large teaching hospital in the Netherlands. This major clinical hospital has over 2900 employees and 643 beds at its disposal. Recently, the hospital adopted a Lean strategy as its main approach to continuous improvement.

For our study, we purposefully selected, on the basis of their flow orientation and perspectives, three cases out of a large number of improvement projects. This selection of these cases should ensure external validity, and should allow the results to be generalized (Karlsson, 2009).

All the patient flow improvement projects studied had been completed within the previous year. The projects focused respectively on improving flow in breast, physiotherapy, and

rheumatology units. The projects varied in their perspectives and we aimed to cover both a patient perspective (waiting for capacity) and a capacity resource perspective (waiting for a patient to be available). These distinct perspectives should also result in theoretical replication since we can expect contradictory outcomes between cases but for predictable reasons (Voss *et al.*, 2002).

At first we considered using only two cases, the rheumatology project (which had a patient perspective) and the physiotherapy project (a capacity resource perspective). However, we decided to include a hybrid form that was being adopted in the breast unit. Although the main focus in the breast project was reducing patient waiting times, this flow improvement project also identified situations where capacity was waiting for other capacity resources.

Overall, our selected cases provide a set of different perspectives on flow issues: patient-oriented, capacity-oriented, and hybrid. The differences among our selected cases meant that we had achieved theoretical sampling (Glaser and Strauss, 1967; McCutcheon and Meredith, 1993) in that we had covered the different conditions that we expected to influence the flow project outcomes. Additionally, the similarities and differences between the cases support the identification of patterns, and these patterns can form the basis for further theory development (McCutcheon and Meredith, 1993). Below, we further introduce the breast, physiotherapy, and the rheumatology units.

Breast: hybrid perspective

A breast clinic provides screening and treatment for women who are diagnosed with or suspected of having breast cancer. The main

goals of the flow improvement project were to decrease access time to 24 hours, to provide a final diagnosis within one day, and to ensure that 90 percent of those patients requiring treatment start this within four weeks.

During pre-screening, patients are categorized into six groups to determine urgency – by the so-called Breast Imaging-Reporting and Data System (BI-RADS) score. A patient visit starts with an intake interview and, depending on the situation, the patient undergoes a mammography or an echography. Mammograms and echograms are performed by an echo analyst and the outcomes later assessed by a radiologist. Patients have to wait while the results are produced, after which these are discussed with a surgeon. At this point, there are two possible outcomes: (1) the patient does not have a malignancy and exits the patient flow, or (2) the results are inconclusive or malignancy may be suspected, in which case the patient needs additional tests. These additional tests will require the same or maybe other medical disciplines to enter the care process.

Apart from new patients, a breast clinic also sees numerous patients returning for check-ups after treatment. Surgeons, echo analysts, and radiologists are involved in these appointments. The number of appointment slots per day is limited, and these are mostly at predetermined times. Five radiologists and five surgeons deal with these check-ups, and each staff member is available to the breast clinic one day a week.

Physiotherapy: capacity resource perspective

Physiotherapists help patients with their movement and aim to restore flexibility, strength, coordination, and balance. Central to the tasks of a physiotherapist is the 'mobilization' of patients, preferably within four hours of surgery in order to have them fit enough to be discharged from the hospital.

The main goal of the flow improvement project was to determine what caused physiotherapists to sometimes be waiting around for a patient to see. Physiotherapists would often encounter situations where they could not provide patient care because of a range of external disruptions. For example, patients could be unavailable because they were being treated by another medical professional.

The therapists can be classified as a shared resource because they provide their services in various departments throughout the hospital. During preliminary interviews, we were made aware that it was in the orthopedics ward and a general surgery ward that physiotherapists encountered the most disruptions. We therefore focused on these two departments, as these should provide the best opportunities to create new knowledge.

Rheumatology: patient perspective

The rheumatology department treats patients with arthritis and other rheumatic illnesses. These health problems affect muscles, bones, joints and other body parts, and are often highly complex.

The goal of the flow improvement project was to free up capacity in order to increase production and reduce access times. Previously, patients that needed to see a rheumatologist faced long

and fluctuating access times. The capacity of the rheumatology clinic was determined by its staffing of two rheumatologists, one specialized rheumatology nurse, one counsellor, and a front office.

Most of the patients were treated by the rheumatologists, one of whom had limited availability because of commitments in other departments. Patients considered to be relatively straightforward could be seen by the specialist nurse. The counsellor provided support activities, such as guidance on the course of treatment, and was not involved in medical procedures.

Patients come into contact with the rheumatology clinic through referral by their general practitioner (GP). One of the rheumatologists performs an initial screening during a triage process in which the information provided by the GP is assessed to determine subsequent actions. Then, in a first appointment, a patient is subjected to several tests by the rheumatologists and may receive a prescription for medication. In practice, as the illnesses are often chronic in nature, a large number of the new patients will return to the department several times for follow-up appointments.

4.4 ANALYSIS AND RESULTS

In this section, we analyze the data collected and present the results of our study. First, we focus on patient flow in the breast unit, then we shift our attention towards the patient flow in the physiotherapy unit, and conclude this section with the rheumatology patient flow.

Breast: hybrid perspective

In our analysis of Excel datasets on patient inflow, and based on earlier experiences of the clinic, we conclude that the variability in the inflow of patients appears predictable. Additionally, the number of patients that can be expected to show signs of a malignant growth is limited.

With low flow variability and low clinical variability, one would expect these issues to have no more than minor consequences on buffer sizes. However, patients have to wait before they can visit the clinic and encounter in-process waiting times in various appearances. Thus, contrary to our initial expectations, time buffers, especially in the form of patient waiting time, are prominent. Further, our analysis shows that patient waiting times, between intake, final diagnosis, and the start of the treatment, are mostly caused by the differences between patients, which appear to be highly unpredictable. Furthermore, it is the patient's degree of illness, reflected in the required number of tests and outcomes, that largely determines if a short wait for access is granted.

Regarding the daily differences between patients, the department head commented: *"when we observe an increase in new patients, then a percentage of those patients will receive bad news, and this means we will need additional time"*. According to the department head, the uncertainty linked to new patients makes it difficult to determine how much time per patient will be required: *"Since you do not always know exactly what your workload will be during the day, you might plan a Cobra examination* [a time-intensive scan] *and then receive a request*

for additional ultrasounds. If so, these ultrasounds have to wait, and work grinds to a halt".

In addition to a time buffer in the form of waiting patients, we also saw capacity buffers. Given the uncertain variability in the need for additional scans or tests, radiologists often have spare capacity. All the available appointment slots are not always filled but the radiologists have to be available in case a high number of patients with malignant symptoms enter the clinic. According to one of our interviewees, *"occasionally, not all the slots for scans will be used, there is a gap between scans. If this happens, a radiologist might wait until a batch of tasks is available before starting"*. We perceive this as the radiologists coming up with their own coping mechanism to avoid idle periods: sometimes they delay starting their work to ensure there is a build-up of mammograms or echogram to assess.

We encountered an interesting form of buffering behavior employed when there are a large number of patients requiring further treatment. At such busy times, personnel tend to work through their lunchbreaks or work beyond their normal finishing time. While this is not a capacity buffer in the strictest sense, it does show that a form of buffering is used, and one that is normally hidden in that it does not show itself in the form of either idle or excess capacity or of waiting patients.

The presence of both capacity and time buffers is difficult to explain given the limited variability. However, if we take the policies of the breast clinic into account, the underlying causes become much clearer. Achieving a patient access time of 24 hours and same-day-discharge can only be achieved with a specific mix

of patients. For example, it could perhaps be achieved if only five patients required tests, and only one of these patients required an additional scan. Given the unpredictable variability, it is impossible to ensure the required mix, and therefore the goals set by the clinic for patient times are only feasible under very specific conditions.

In essence, our study indicates that, while the desired specific patient mix might seem possible based on the overall annual inflow of patients, the uncontrollable variability between days ensures that this is unachievable on a day-to-day basis. Although the BI-RAD scores do offer some pre-screening information that can be used in planning, the test outcomes are not always definitive and further tests may be required. As such, there remains a relatively high level of uncertainty attached to each patient.

Even though most patients will be able to leave the clinic without requiring further testing or surgery, it is impossible to determine in advance whether a patient will require additional testing because of a suspected malignancy. Given the tight goals set, there then becomes no alternative but to accept the presence of a capacity buffer, for example in terms of the radiologists having excess capacity. Based on the current policies, the clinic has to accept that, at certain times, there will be idle capacity waiting for patients to treat. In essence, there is a trade-off between achieving a 24-hour access time and providing a conclusive diagnosis on the day of admittance.

Physiotherapy: capacity resource perspective

Physiotherapists have to find a balance between their own limited capacity and the limited availability of patients. For example, patients that are being treated by other disciplines cannot receive physiotherapy at the same time. Despite this variability in patient availability being expected to cause major issues, it does not translate itself into obvious capacity buffers. When we investigated further, we came across an interesting phenomenon that appears specific to healthcare environments.

During a patient's stay in a hospital, the unavoidable time it takes them to recover creates an 'invisible' time buffer. If a patient should receive physical therapy 'today', and is also scheduled to be discharged 'today', then there can be a time window of several hours in which the therapy can take place without delaying the patient's discharge. Apparently, there are no consequences related to patients receiving treatment either early or late in a day, provided it is given before the day ends. Rather than experiencing their stay in the ward as waiting time, patients see it as time spent recovering.

The physiotherapists are able to use this hidden time buffer to their advantage, and they use it to create flexibility in their capacity. It allows them to focus only on patients that are immediately available for treatment, and it enables them to work around other disciplines that also place demands on the patient. Intuitively, we would expect the hospital policy to result in a time buffer or, in other words, patients waiting for physiotherapy. Yet waiting time is 'absorbed' in the recovery time.

Moreover, we also failed to identify a capacity buffer in the sense of idle capacity. During observations and case visits, it seemed as if people were never idle, and that excess capacity did not exist. It seems that physiotherapists are able to find other activities to fill their capacity.

Nevertheless, we were able to identify the same 'capacity' buffer as we saw in the breast clinic, that is physical therapists continuing working during breaks or after regular working hours. The unit manager reported: *"If everything runs smoothly, you cannot say we need more physical therapists because, if you compare the number of patients with the time each patient needs, it should be possible. However, this is only feasible if you start therapy at exactly eight o'clock and finish at twelve o'clock, with two physical therapists available, and if you have access to the next patient immediately after completing an appointment. However, that is not how it works in practice"*. Here, the unit manager is alluding to the fact that capacity planning ignores the role of variability during the day. This implies that some form of buffering will be required. To an extent, this variability is absorbed through the flexibility that a patient's recovery time offers, and partly through physiotherapists working during breaks or after normal hours. However, we also observed an additional mechanism: physiotherapists adjust processing times in order to cope with variability. Especially at busy times, physiotherapists tend to increase their working speed or omit certain elements of their tasks.

Rheumatology: patient perspective

The rheumatology unit was struggling with long access times of up to 16 weeks for new patients that were referred to the department. In a response to this issue, the clinic decided to temporarily increase capacity, in order to reduce the backlog of patients and waiting times. As expected, the increase in capacity did shorten the waiting list, and patients could again be seen quickly by one of the rheumatologists.

Essentially, the rheumatology clinic had exchanged a time buffer for a capacity buffer. However, the temporary increase in capacity had another consequence in addition to shortening access times through reducing the reliance on time buffers. According to the medical professionals and the quantitative dataset, the reduction in access times resulted in an increased rate of referrals to the clinic. Hence, the increase in capacity buffer did not only substitute for the time buffer, it also functioned as a trigger for new demand.

Once access times had returned to an acceptable level, the capacity returned to the original level. This meant that the capacity that should have catered for the newly referred patients had been cut. In this overall process, the rheumatology clinic had introduced variability through a temporary increase in capacity. Although the decision to temporarily increase capacity looked a promising approach, the consequent inflow of new patients resulted in a lagged increase in demand. The new patients introduced into the system would require follow-up appointments, but the capacity to provide these follow-up sessions was no longer available. As a consequence, not only do access times start to increase again, it

also means that the additional capacity needed to provide treatment for follow-up patients came at the cost of capacity for accepting new patients.

This clinic provides an example of how underestimating the dynamics between buffers and self-induced variability can lead to problematic situations. The changes in the clinic were initiated with the best intentions, but the clinic was unaware of the external consequences of their actions. The improvement in access time led to more referrals, which led to inflow variability, which led to increased access times.

4.5 DISCUSSION

In carrying out this research, we were interested in the way buffers and variability interacted in a healthcare environment, and how the interchange or trade-off between buffers would manifest itself in practice. The interaction that one normally expects between variability and buffers was not so straightforward in our studied healthcare situations.

Moreover, on various occasions we observed what could be seen as buffering involving processing time, a mechanism described as a quality buffer by Hopp *et al.* (2007). To the best of our knowledge, this is the first study to identify this form of quality buffer in healthcare practice. The presence of this additional buffer has consequences for the interchange between buffers, and it could help explain situations where patients are not obviously waiting despite there being no excess capacity. We will expand on this quality buffer later in this section.

Contrary to earlier findings that reported limited attention being given to the underlying characteristics of Lean approaches (Andersen *et al.*, 2014; Hines *et al.*, 2008; Radnor, 2010), we would argue that, in the cases we studied, the attention given to flow does show that fundamental Lean issues are being considered. However, this apparently only seems to cover patient flows.

Our findings regarding patient flow in the breast unit underline the complexities of Lean initiatives in healthcare as already identified in the literature (Ballé and Régnier, 2007; Proudlove *et al.*, 2008; Young and McClean, 2008). Here, the need to address the desires of a range of stakeholders, such as the 24-hour access time with discharge on the same day, can result in waiting times. What we see here is that the goals set by the clinic conflict with each other.

When considering the variability in the inflow and mix of patients on an annual basis, this variability, and therefore the effects one might expect, appears to be limited. On what could be considered a macro-level, variability seems of minor importance because there is a relatively steady inflow and a low probability of malignancy in patients. This led us believe that the objectives set by the clinic were obtainable.

However, on a day-to-day basis, even the limited variability in patient inflow and malignancy rates can have major impacts. This makes it difficult to achieve the objectives set by the clinic in practice. These issues arise in an environment where organizational policies oppose buffering 'solutions' (additional time or capacity) based on the desire for quick access and

discharge. The findings from the first case in the breast unit lead to our first proposition: *Buffering 'solutions' should be implemented that reflect the period of time over which variability is targeted.*

The findings from our second case, involving the physiotherapists, showed that, in a healthcare organization, the interchange between buffers is not always obvious. We observed that a feature seemingly unique to healthcare, the unavoidable recovery time for patients, allowed a time buffer mechanism to be used during the recovery period without patients apparently waiting for treatment. In this situation, the patients acquire the characteristics normally associated with inventory buffers. For example, they seem to be 'stored' for future 'use'.

This goes against our original thinking and the ideas put forward in the literature (e.g. Jack and Powers, 2004) in that inventory buffers were not expected to have a role in the patient care process. It would therefore seem to be important to distinguish between occasions when patients can be considered as inventory, and others when waiting patients should be viewed as a time buffer. Additionally, despite patients not obviously waiting to receive treatment, they will often still require some form of treatment in due course. To illustrate this fluidity in interpretation, if, for example, the healthcare policy changed such that patients should be discharged as soon as the physiotherapists have delivered their treatment, the effect would be that, suddenly, there would be multiple patients waiting and an obvious use of a time buffer involving patients.

In addition to our observations related to this hidden time buffer, we came across another interesting phenomenon. At busy times, the physiotherapists were inclined to include fewer exercises in order to shorten consultations. In other words, they were reducing their processing time to cope with system load. This behavior could be considered as the use of a quality buffer, a special form of buffer for service environments proposed by Hopp *et al.* (2007). With a quality buffer, processing time is adjusted, influencing job quality, in response to system congestion. The study of the physiotherapy unit shows that the presence of buffers is partly dependent on organizational policies, and that an interchange between buffers (time and capacity) can be hidden behind these policies. Based on these findings, we formulate the following proposition: *In healthcare, a patient's recovery time before discharge can be used as a 'hidden' time buffer without resulting in obviously waiting patients.*

A study by Mazzocato *et al.* (2012) found that problems in patient flow result in increased waiting times. Our findings related to flow in the rheumatology unit underline these findings and reflect these earlier results. However, the way the problems in the rheumatology unit came into existence is both surprising and remarkable. Despite the best intentions of the medical professionals, the decrease in access times achieved through a temporary increase in capacity over several months, ended up in a worse situation.

Increasing the number of new patients seen led to an increase in the number of follow-up appointments which, without long-term added capacity, essentially blocked further new patients

entering the system. This mechanism resembles the service bullwhip effect (Lee *et al.*, 1997; Akkermans and Voss, 2013) in which limited information downstream of a process causes variability upstream of this process. The term bullwhip is applied because the further away the upstream process is, the greater the problem with variability (Geary *et al.*, 2006).

However, in our situation, rather than a lack of information being the cause, it was the availability of information, with referrers being kept appraised of current access times, that facilitated the effect. The temporary increase in capacity ultimately resulted in greater variability, and less stable access times than before. The lessons from our third case lead to our next proposition: *Buffering 'solutions' may create new variability and, as such, induce vicious cycles.*

The dynamics between patient recovery time and a time buffer, and between idle capacity and a quality buffer, add complexity to variability and buffer issues in healthcare. We had assumed that a capacity buffer in a hospital would manifest itself as idle capacity, in the form of under-occupied nurses, physicians, and medical equipment. However, we rarely found evidence of idle capacity resources. Maybe, and perhaps not surprisingly, professionals will find something to do if they are currently 'unneeded'.

Here, we come to the concept of the quality buffer (Hopp *et al.*, 2007). A quality buffer assumes that processing time can be adjusted in response to system load. If processing time can be lowered at busy times, we would expect the opposite to be also true: that, at times of low capacity demand (i.e. quiet periods), staff can increase their processing times to remain busy. As such,

the presence of a quality buffer could be interpreted either as a hidden capacity buffer, or as replacing it. This brings us to our final proposition: *A quality buffer, the ability to vary processing times to adapt to load levels, is an additional buffering mechanism that is able to manage variability that is neither buffered by explicit excess capacity nor by patients' waiting times.*

This research has its limitations and a major concern is that there is no standardized way to detect the causes of variability or buffers. The observations of buffers and their interchanges depend on the theoretical lens adopted. Our study is not exempt from the concerns regarding generalizability that are often associated with case studies. However, we have tried to mitigate generalizability issues by studying multiple, purposefully selected, cases, and by triangulating our data. Nevertheless, one should not assume that we have identified all the interactions between variability and buffers, or all the interchanges between buffers.

4.6 CONCLUSIONS

In this study we have taken a close look at the interaction between variability and buffers in a healthcare organization, investigated the 'buffer trade-off', and have uncovered additional complexities to the buffer and variability relationship in healthcare. We can now respond to our original research questions: (1) how do buffers and variability interact in a healthcare environment, and (2) how does the interchange, or 'trade-off', between buffers manifest itself in practice?

While the interaction between buffers and variability might initially appear straightforward, we observed various mechanisms in play. Hospital policies can hide or amplify time buffers. Capacity buffers are often hidden and we did not encounter capacity resources that were clearly idle. However, in times of high capacity demand, processing times are shortened, which we would see as the application of a quality buffer, and in a way that obscures the 'trade-off' between time and capacity buffers.

The outcomes of our current study lead to suggestions for further research. The mechanism we observed in which patient recovery times provide an additional time buffer appears unique to healthcare but it would be valuable to investigate whether a similar process occurs in other service environments. Additionally, while bullwhip effects have been studied over a considerable period, studies reporting on such effects in a service setting are far more limited. It would be interesting to revisit this concept and the related vicious cycles observed in this study. This is particularly relevant for healthcare settings where customers having information on access times appears to induce strong behavioral effects. Finally, we would like to suggest taking a closer look at a fourth conceptual buffer type that we identified: using processing time as a buffering mechanism. We saw this as a quality buffer but failed to find other studies that had investigated this phenomenon in practice.