Abstract number: 015-0732

Title: Applying manufacturing flow theory to health and social care management

Roy Stratton

Nottingham Business School, Nottingham Trent University, Burton Street, Nottingham,

NG1 4BU (E-mail: roy.stratton@ntu.ac.uk) (Tel: +44 1159 418418)

Alex Knight

QFI Consulting, POBox 953, Tring, Herts, HP23 4ZX, UK

POMS 21st Annual Conference
Vancouver, Canada
May 7 to May 10, 2010

Abstract

The concepts of flow and continuous improvement are central to production management theory, and there is growing evidence of their successful application in health and social care. In manufacturing, this development has been led by lean and the theory of constraints (TOC), and there are now similar parallels emerging in health and social care management. However, there is little research evidence that investigates the nature of the improvements claimed and how the respective approaches theoretically relate. The purpose of this paper is to clarify the role of lean and TOC in improving flow in health and social care management. The focus of the primary research in this paper is on the application and implementation of time buffer management by QFI consulting. Four UK hospital implementations of the QFI Jonah software and methodology were investigated to establish how buffer management was being applied and why the reported benefits were being achieved. This involved collecting service delivery data together with semi-structured interviews. To support this evaluation, four control functions of time buffer management have been identified as a basis for evaluation of the application designs and their implementation. Case research evidence shows significant and rapid improvement in length of stay following implementation of the approach, amounting to a reduction in length of stay of around 20% and significantly improved Accident & Emergency performance. Sustainability issues were evident however and were traced, at least in part, to lack of adherence to one or more of the functional elements of the system. This case research has considered four applications that represent some of the more successful implementations. This research is, therefore, limited by the range of applications considered and is also limited in its ability to evaluate the sustainability of the implementations in the longer term. This paper provides clearer theoretical understanding of the improvements experienced by

these hospitals, which is critical to expanding the use of traditional manufacturing approaches into complex service environments. This research has helped conceptualise how time buffer management control functions can be structured to evaluate the design and implementation of novel applications. The paper concludes by relating the TPS kanban rules and functions to the time buffer management control functions in the context of patient flow management and discussing the theoretical boundaries to their use.

Keywords: Theory of constraints, buffer management, health and social care, patient flow, case research

1.0 The operations challenge in healthcare and social care

With an aging population it is increasingly important to view health care delivery systems holistically, acknowledging that the wider delivery system encompasses both health and social care. Therefore, effective and efficient acute heath care must also ensure strategic alignment of the social care system. This fact further demonstrates the complexity of this wider system, as many different organisations need to be synchronised in order to ensure improvements in efficiency and timeliness associated with improved patient flow. However, the silo mentality, driven by separate organisational targets, often results in dysfunctional behaviour. In England the government drive to improve performance in the National Health Service (NHS) has had to address many examples of such behaviour, especially at the silo interface, between social and medical care services.

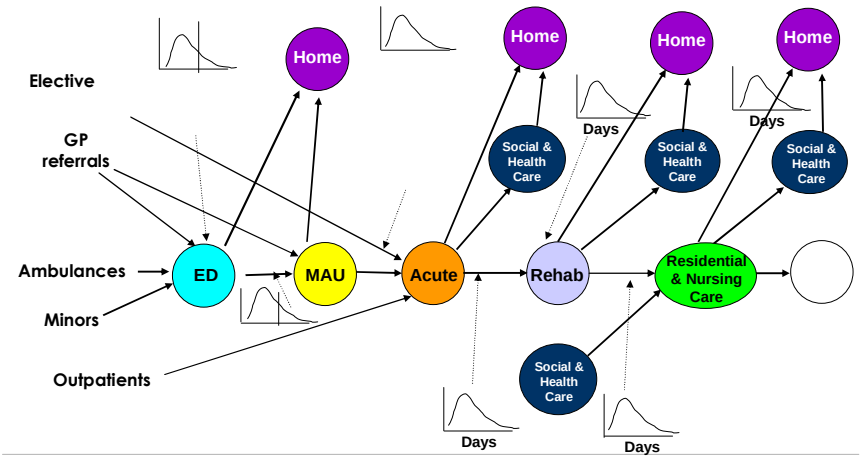

Figure 1: Health and Social Care:a systems perspective

Figure 1 illustrates the macro flow routes into and through the health care system, together with the interface between health and social care. This Figure, in addition to the major flows, illustrates the variation and uncertainty associated with demand, together with process and recovery times. In the UK, such times, together with budgetary control, are used to drive improvement and there are now emergency and planned treatment targets. The Emergency Department (ED) performance is measured against a government target for England that states 98% of patients are to be treated and discharged or admitted within 4 hours. The more recent referral to treatment target states that 95% of patients are to have started treatment within 18 weeks of referral by a General Practitioner (GP).

In England the government sets the waiting targets and the local hospital management are empowered to find appropriate means of achieving these targets, whist remaining within budget. Agencies, such as the NHS Institute for Innovation and Improvement support this process and over the years a wide mix of approaches have been used that

originated in manufacturing.

These targets have driven much improvement activity that now tends to centre on lean thinking but there is growing awareness of the merits of TOC and particularly buffer management which is the focus of this paper. Although the UK NHS started from a low level of access performance 10 years ago their current performance compares well with international standards. For example, the Emergency Department (ED) waiting time average across the US is 240 minutes (4 hours), up 18 minutes on 2005, with State averages ranging from 158 to 381 minutes (Press Ganey Associates, Inc, 2007). If we acknowledge the skewed nature of such distributions an average of 4 hours will result in many people waiting over 8 hours.

2.0 Meeting the operations challenge

Operations theory has established the merits of standardisation and the importance of reducing variation and uncertainty in enabling flow (Schmenner and Swink, 1997), but not all environments lend themselves to this. Although there is clear opportunity to standardise patient flows, much of the health care environment can be allied to non-standard manufacture where the flow complexity is characterised by uncertainty in routing and processing/recovery times. Over the years manufacturing based theory has been increasingly applied to health care. This includes the adoption of clinical pathways (Herck et al., 2004) to help standardise procedures and the use of Shewhart's (1931;39) continuous improvement cycles (PDSA) (Walley and Growland, 2004). The use of run charts and control charts are now common and this has more recently been expanded to explore the Six-Sigma methodology (Proudlove et al., 2008).

Over recent years, with the UK Government's emphasis on service delivery, the focus has moved to improving access times, with significant interest in manufacturing

parallels. The two most prominent approaches being kanban control, which is closely associated with lean manufacturing (Bhasin and Burcher, 2006) and time buffer management, which is closely associated with the Theory of Constraints. Although both these approaches are concerned with managing flow they come from very different environments.

TPS kanban

Kanban originated as a central concept and approach within the Toyota Production System (TPS) (Ohno, 1988) and continues to be a central feature of lean applications within manufacturing and service sectors. The Japanese word Kanban may be translated as signal in English but it has specific meaning in its use within the TPS. A kanban, typically in the form of a card, acts as a signal to produce or transfer a specified quantity of a product to the next stage of production. Between any two stages in the overall process there will be a number of kanbans at different points in this replenishment cycle. Ohno (1988), the originator, very usefully described the system in terms of six rules (See Figure 2).

Kanban Rules of Use
1. Later process picks up the number of items indicated by the kanban at the earlier process.
2. Earlier process produces items in the quantity and sequence indicated by the kanban.
3. No items are made or transported without a kanban.
4. Always attached a kanban to the goods.
5. Defective products are not sent on to the subsequent process. The result is 100% defect free goods.
6. Reducing the number of kanban increases their sesitivity.

Figure 2 The Rules of Kanban (source: Ohno, 1988: 30)

Ohno (1988) clearly identified these rules as being central to the TPS system.

> *In reality practicing these rules [the six rules of kanban] means nothing less than adopting the Toyota Production System as the management system of the whole company.* (Ohno, 1988:41)

Although kanban is not the focus of this research these rules and and associated functions will be subsequently used in trying to relate the derived time buffer management control functions.

TOC Time buffers

TOC based time buffer management originated at the other end of the operations spectrum to that of TPS kanban. Whereas kanban is closely associated with make to

availability, time buffer management is associated with make to order. The variable and uncertain nature of patients and their treatments has much in common with the complexity of a make to order environment and, therefore, it is not so surprising that the time buffer management approach has proved effective in certain environments (Umble and Umble, 2006). However, the question we are attempting to address here concerns how and why it has been effective. To address this question it is important to outline the traditional TOC applications of time buffer management and identify the associated control functions they satisfied, before relating this to the health and social care applications. In understanding the boundaries of applying time buffer management it is useful to understand the environment best suited to kanban control and the paper will finally discuss these distinctions.

3.0 Time buffer management – manufacturing origins and traditional applications

TOC based time buffer management is applied differently in manufacturing and project environments. The manufacturing approach, entitled Drum Buffer Rope, emerging in the late 1980s with the project management approach, entitled Critical Chain, emerging in the later 1990s. These traditional applications are briefly outlined in turn, before identifying common functional control elements. These functional elements will subsequently be used to both explain and evaluate the healthcare applications.

3.1 *Manufacturing based time buffer management*

The manufacturing application is termed Drum - Buffer - Rope (DBR). This approach is well documented and has been developed as a practical approach in the manufacturing environment (Goldratt, 1990; Umble and Srikanth, 1997; Schragenheim and Detmer,

2001; Stratton et al., 2008) as outlined below.

- Drum: The drum can be set by a resource constraint or market demand and we will assume here that it is the market demand. It may be interesting to note that DBR applications are now rarely associated with bottleneck management as any growth strategy will look to ensure market demand remains the constraint, as is the case in healthcare.
- Rope: This is the planning mechanism for releasing work and effectively choking material release in line with customer demand. This is the DBR mechanism for preventing over production where the rope represents the time offset that releases material in line with demand.
- Buffer: Once material is released the time remaining is termed buffer time, which is based on the assumption that the actual processing time (touch time) is negligible. To support the management of this buffer it is conceptually divided into three equal time zones (See Figure 3).

3.1.1 *The four control functions of DBR buffer management*

The associated control functions are related below.

- *Prioritise the flow of work according to buffer consumption.*

Work is release based on the choked release date set by the rope. Work is prioritised with reference to the buffer consumption, indicated coarsely by the colour coding or more precisely, percentage buffer penetration..

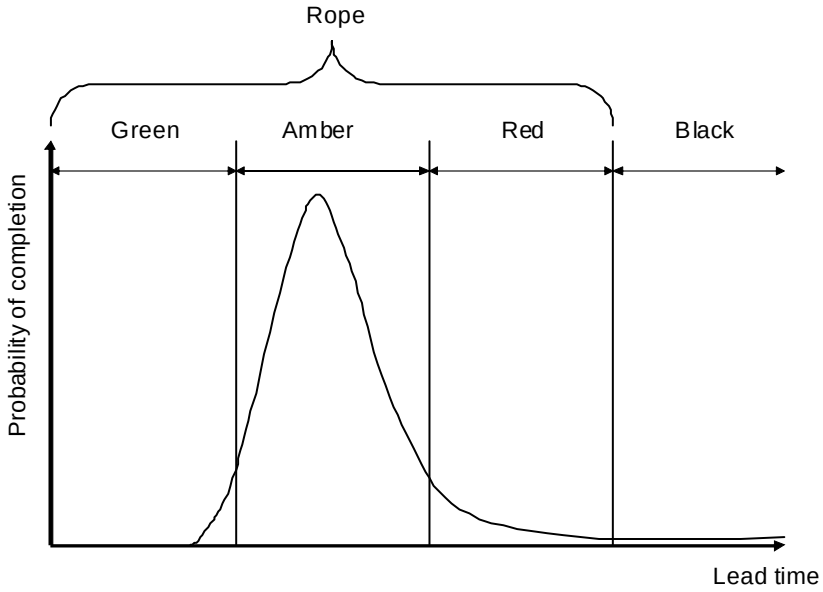

Figure 3:Buffer management in an operations environment showing buffer zoning of lead time distribution

- *Identify when to expedite potentially late arrivals.*

The red zone should effectively represent the tail of the distribution with relatively few orders entering this zone (see Figure 3). Having parts in the red zone indicates the need to act, i.e. expedite where necessary to ensure it is delivered within the remaining buffer time. As has already been stated, even though the red zone represents only one third of the lead time, there is still time to complete the order, even if processing has not started. This, however, assumes there are few orders in the red zone at any one time.

- *Signals when the whole production system is starting to become unstable.*

The system is acknowledged to have the capability to expedite only a limited number of parts but if this rises above a certain level (Goldratt (1990) suggests over 10%) the system is at risk of going out of control. In such cases, urgent action is required to bring the system back into control, typically by limiting order intake or increasing capacity in some way.

- *Identify main sources of delay in order to target improvement.*

Continuous improvement is about reducing variation (Deming, 1986) and in the flow context, targeting sources of variation that threaten delivery. This is illustrated well by the exposure of rocks in the river and rocks analogy commonly associated with lean manufacturing. The equivalent, in buffer management terms, is associated with the reasons for red zone penetration of the buffer, therefore, it is necessary to record what is the source of delay at this time.. This information is periodically analysed using Pareto analysis and used to target improvement..

3.2 Project based buffer management

In the project environment, where the processing time (touch time) is a significant part of the overall lead time, the DBR assumption that process time is insignificant is no longer valid. In these situations the time buffer element of the lead time needs separating form the process time.

If we consider traditional project management, a network of tasks are logically linked by dependencies and the longest path is referred to as the critical path (see Figure 3 – left side). The TOC approach to buffer management in projects is called Critical Chain Project Management (CCPM) (Goldratt, 1997; Steyn, 2000; Stratton 2009) and, as with DBR, it emphasises the need to manage the time buffer.

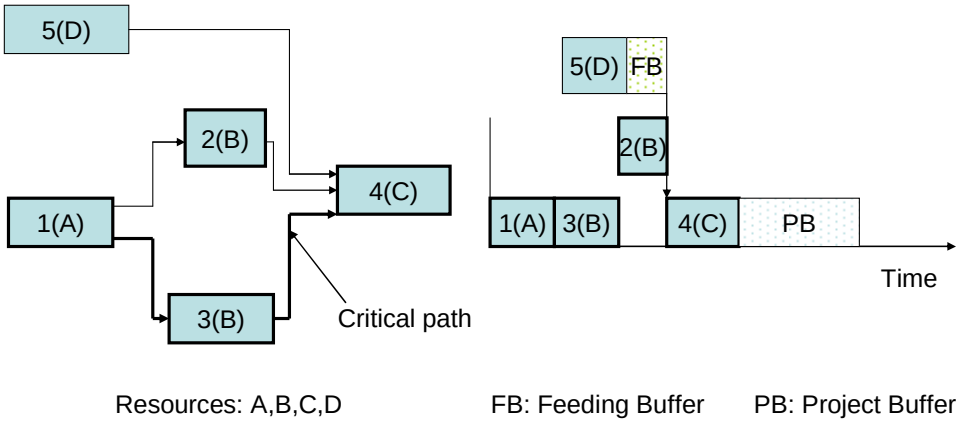

Figure 4 Traditional Network diagram (left)
Critical Chain showing schedule with time buffers (right)

In CCPM the time is extracted to make it visible simply halving the task time and placing half of the time removed into a time buffer which is conceptually located at the end. As can be seen in Figure 4, resource dependency is also acknowledged (the scheduling of resource B is staggered) and the project buffer protects the critical chain, with feeder buffers protecting activities with float from becoming critical. Goldratt (2007) advocates the project buffer and associated feeder buffers should be set at one third of the lead time.

3.2.1 *The four control functions of CCPM buffer management*

Again we can identify four prime functions of CCPM buffer management.

- *Provide a priority mechanism*

Priority is not just set by the time remaining, as in DBR, but by a simple ratio of project buffer remaining to critical chain remaining. This enables a resource provider to

prioritise different tasks within and across projects.

- *Provide a mechanism for expediting tasks consuming the project buffer.*

Unlike with DBR, where action is only taken in the red zone, it is important to reduce buffer consumption that can occur in any activity. As there is effectively only one activity consuming the project buffer at any time the improved visibility of buffer consumption creates awareness and opportunity to support the resource concerned and minimise wastage of the buffer at the activity level.

- *Escalate action when buffer consumption threatens delivery*

At the project and programme level the monitoring of the ratio of buffer consumption to completion of the critical chain is used to trigger escalation procedures.

- *Target ongoing improvement activity by tracking causes of buffer consumption.*

As with DBR there is a need to track the reasons for delay to focus improvement activity.

3.3 Summary

Traditional time buffer management has been shown to embrace four control functions that may be summarised as Prioritise, Expedite, Escalate and Improve. These functions will be used in the case study research that follows to provide a framework for comparison and evaluation of these health and social care applications of time buffer management.

4.0 Research method

This research has been initiated to help explain the basis of the improvement in

patient flow claimed by QFI consulting and numerous hospitals that have adopted their approach and software. Two buffer management applications were chosen for investigation due to the fact that they were in common use and embraced both planned and emergency care:

1) Emergency Department Buffer Management (termed A&E Jonah)

2) Discharge Buffer Management (termed Discharge Jonah)

It should also be noted that although these two buffer management applications operate at different points in the health and social care system they are not independent of each other. For example, one of the main reasons for breaching the Emergency Department (A&E) 4 hour access target is lack of beds due to delayed discharge. Hence, reducing length of stay impacts bed availability and this in turn impacts ED performance.

The primary research question set is:

How and why has time buffer management contributed to improved patient flow?

The inductive nature of this research question clearly fits the case method, but it was necessary to develop a conceptual framework on which to base the comparison and evaluation. This has been provided through identifying the underlying control functions represented in the two generic time buffer management applications DBR and Critical Chain.

This research involved access to the buffer management application software design, the implementation process and users of the system. Four hospitals were visited where the systems were seen in operation, archival data and reports were accessed and semi-structured interviews were conducted, in line with the good practice advocated by Yin (1994) and Eisenhardt (1989). In one case the unit of analysis was extended to include social care and cross buffer meetings at a regional level were attended. This research was conducted over a period of 20 months. In all four cases the discharge buffer

management application was implemented and in three cases the Emergency Department buffer management application. In all cases the implementations were less than 3 years old.

The research findings are presented below, with section 5 describing the applications and how they are designed to operate. Specific reference has been made to the conceptual framework, in the form of the four underlying control functions. Section 6 discusses the evidence concerning the merits of the functional elements of time buffer management before relating these functions to those associated with kanban in the context of healthcare management. The paper concludes by exploring the wider significance of these control functions and an attempt is made to identify under what conditions kanban and time buffer management are more suitably applied.

5.0 Emergency and planned care buffer management application design

The two health care buffer management systems investigated are described below in relation to the previously identified control functions.

5.1 *Emergency Department buffer management application*

The Emergency Department application is closely allied to the DBR buffer management approach as outlined below.

- Drum: The pace of the system is set by the drum and in this environment it is determined by demand with no prior warning apart from the ability to forecast as the Emergency Department is expected to manage capacity in line with demand.
- Rope: The rope would normally be used to control release of work into the system, but in this environment there is no attempt to control release as all patients are

considered to be in the 4 hour buffer on arrival.

- Buffer: The buffer time of 4 hours commences on patient entry and as with standard buffer management practice this time is equally divided into three zones (see Figure 3) of 80 minutes each. The application is supported by software that tracks the patient buffer consumption with the priority being set by the time of entry. The buffer changes colour as time progresses and at each transition (80,120 and 240 minutes) the system is designed to record what the patient is waiting for. This is selected from a drop down menu and used in Pareto analysis. This typically includes waiting for - consultant, waiting for X-ray, etc. The protocol is that this drop down menu is updated by the nurse responsible for the patient and on discharge, or admission, the patient is removed from the screen.

5.1.2 *Buffer management functions*

The practical repesentation of the four control functions are as follows:

- *Prioritise* - the order of priority regarding the discharge target is simply the order of presentation, however, this clearly does not acknowledge clinical priorities that are expected to override this default priority sequence.
- *Expedite* - patients entering the red zone are within 80 minutes of breaching the 4 hour target and the computer networked screen is used across the hospital to communicate the need to expedite action by the resource currently causing the delay. The resource causing the delay of a patient is also informed separately from this screen.
- *Escalate* - on a regular basis (hourly is recommended) a buffer meeting analyses delay reasons and escalates as necessary. This escalation is deemed to be an

essential part of the synchronisation of resources to enable flow. The overall stability of the Emergency Department performance is monitored by the number of red zone penetrations and middle and senior management have visibility of this screen. This effectively acts as an escalation signal and protocols can be established to assess the situation. Additional resource may need to be provided to avoid the Emergency Department becoming unstable, resulting in breaches of the 4 hours which have to be reported by the hospital centrally.

- *Improve* - the reasons for delay are subsequently analysed by charting the main causes of delay on entering the yellow, red and back zones. This data (produced in Pareto format) is used in buffer management meetings (typically weekly) aimed at focused continuous improvement effort. Although the Pareto graph is produced to present the main reasons for delay the data is also used to identify causal links with any common causes becoming a focus for rapid improvement.

5.2 *Patient discharge buffer management application*

The need for timely patient discharge relates to acute and community hospitals, as well as social care homes, and can be viewed as key to managing patient flow through the health and social care system. This application aims to shorten the length of stay (LOS) by subordinating all support activities to the medical needs of the patient, with the aim that the patient's position within the health and social care pathway is determined by medical needs alone.

A first step in this discharge planning process is to set a Planned Discharge Date (PDD) when the patient first arrives. This is followed by identifying the care plan and associated activities that need to be completed before this date. In reality the PDD is

initially indicative but can be updated as knowledge and circumstances dictate. The concept of a PDD is new to many hospitals but is an essential part of any planning and control system. As with CCPM task management, the emphasis is on updating the PDD if it significantly changes, to ensure priorities reflect the medical need.

With the time to discharge based on medical judgement alone it is necessary to ensure subordination (alignment) of the support activities. In this environment activity durations are more characteristic of a project where the touch time is significant, as illustrated in Figure 5. Therefore, the projected buffer time remaining comprises the time to the planned discharge date (PDD) minus the longest activity duration yet to be completed.

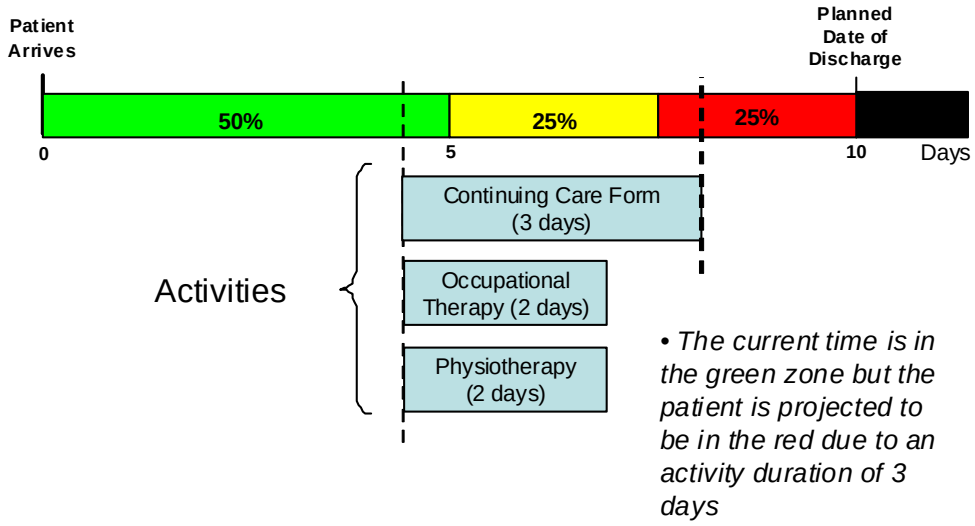

Figure 5 Discharge Buffer Management

Accounting for the duration of activities has been simplified with the assumption that there is no activity dependency. Therefore, the longest activity duration, together with the time now and PDD, enables the projected buffer penetration to be determined (see

Figure 5). This application is effectively a hybrid of the Critical Chain and DBR buffering approaches. That is, the duration to PDD effectively acts as a buffer as in DBR, but the longest activity duration is acknowledged in determining the buffer penetration. Finally, the buffer regions have been adjusted (see Figure 5) to make allowance for the activity duration offset. The software can provide priority lists for the different specialists (eg occupational therapists) showing buffer penetration.

5.2.1 *Buffer management control functions*

The practical representation of these four control functions are as follows:

- *Prioritise* - the priority status for the patient is determined by the time to planned discharge minus the longest remaining activity to complete. Unlike in the Emergency Department this may change. As the planned discharge dates can change in both directions the relative priority of different patient activities may change. Each patient is displayed on a computer screen, a line for each patient, in priority order, coded green, yellow, red and black. The software enables a subset of patients waiting for a resource, such as an occupational therapy, to be listed as a priority coded work to list.
- *Expedite* - patients that are projected to stay beyond the planned discharge date are shown in the black, therefore urgent, but as this predicts a future situation there is still time for recovery. Reasons for delay are recorderd on entry into the yellow/red/black zone and where it remains in the black zone the entry is repeated every three days to capture data for analysis purposes. The delay reasons are typically: waiting for – medical review, care package, community hospital, continuing care decision, etc. The ‘reasons for delay’ data is circulated daily at a ward based buffer meeting and the main reason for delay

communicated as part of the escalation process.

- *Escalation* - the stability of the overall system is signified by the overall number of red and black buffer penetrations which provide advanced warning of the pending instability and the need to escalate recovery action. Lack of beds is a major cause of Emergency Department breaches and increases in delayed patient days is a leading indicator that bed shortages across the hospital will follow. Typical delays can be due to - a need for medical review, shortage of a particular resource provider, lack of community hospital beds, etc. To be effective this escalation process needs to be activated before the system destabilises, which in the final instance may mean temporarily opening additional beds.
- *Improve* – with the reasons for delay recorded it is important to regularly identify and act on common causes, Regular buffer management meetings are encouraged at different levels of frequency locally and regionally. Figure 6 illustrates the Pareto format of the delay reasons analysis used at the regional level with other care providers.

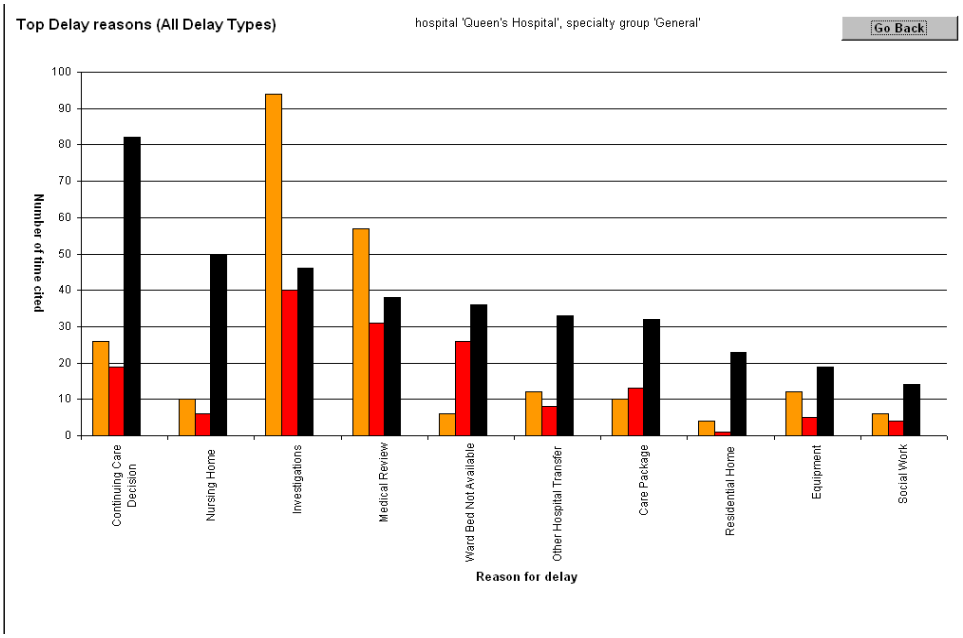

Figure 6 Discharge QFI Jonah top delay reasons by region presented to cross-buffer meetings

6.0 **Implementation Findings and Discussion**

This section focuses on the implementation of the applications. Firstly, establishing the overall benefit, before identifying observations and issues concerning the control functions, and then finally addressing the research question directly with reference to kanban functions.

Interviews with senior management regarding both applications were very positive as evidenced by the following quotes.

> *'TOC has been applied to improved patient flow in A&E, Assessment Units, and discharge planning. This has resulted in a sustained reduction in medical length of stay from 8.6 to 6.3 days (>25%). Released bed capacity supported the achievement of the 18 week GP referral to treatment target, a year ahead of schedule.'*
>
> (Director of Governance and Nursing)

> *'With the help of Theory of Constraint we have been able to move Barnet & Chase Farm Hospitals NHS Trust from one of the worst performing trusts in England to one of the top performing. In Q4 (2007-2008) we were the top performing trust in London for the 4 hour target and 6th across England. Also, by applying the Theory of Constraints to our discharge process we have been able to reduce our length of stay by 27% and we know we can improve further*

on this.'

(Chief Executive Officer)

'The application of TOC has helped us to reduce our length of stay by up to 23% in one of our hospitals, but the real benefits from QFI Jonah are around improving how we deliver care to our patients through better planning and co-ordination of their care - ultimately it's about recognising that patients should go home as quickly and safely as possible.'

(Chief Executive Officer)

The numerical data supports the wider claims of the CEOs, particularly in the early stages of implementation as illustrated in Figure 7. In this case the improvement in ED performance was rapid (4 months) following the implementation of both the Discharge and ED buffer management systems.

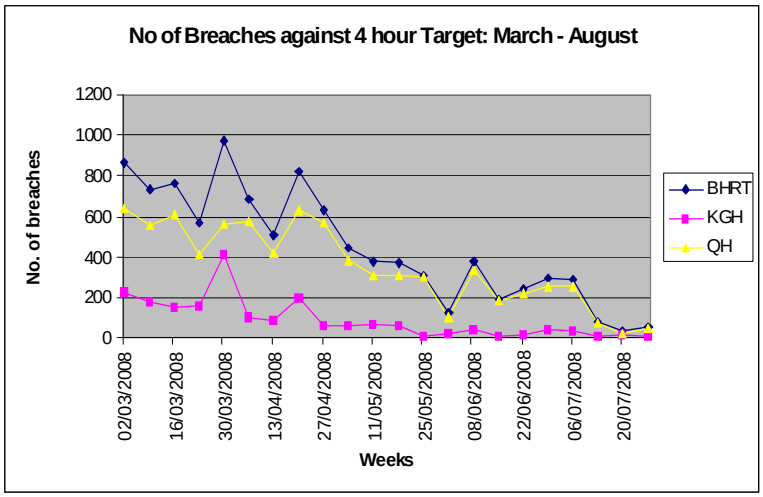

Figure 7: Improvement in emergency breaches following discharge and emergency buffer management implementations introduced from March to July 08.

In all the implementations there was a clear link between discharge management and delayed admissions in the Emergency Department as indicated in Figure 7. Therefore, the normal practice is to introduce discharge buffer management first, before implementing the emergency buffer management system. The marked improvement in ED performance resulting from the implementation of Discharge Jonah, clearly indicates that the main reason for breaching is bed availability. In one hospital the improvement in ED performance was such that they delayed the introduction of A&E Jonah. That was until they realised that the ED performance profile clearly showed the poor stability of the system as a large proportion of the patients were there over 3.5 hours.

The ED is where chaos reins when the system is out of control and in another hospital the Operations Director explained how the introduction of A&E Jonah had completely transformed their Emergency Department from a place of chaos where no one wanted to work to a relatively stress free stable environment.

What follows are specific observations and issues that illustrate the four control functions used in the evaluation.

6.1 *Observations and Issues Allied to the Four Control Functions*

Prioritise

Introducing Discharge Jonah requires key care plan activities to be promptly set together with a planned discharge date for each patient. In all cases these dates were not formally kept prior to the implementation but this practice is critical to providing priority control base on patient need. Adherence to the discipline of promptly setting these dates and plans varied between hospitals and across wards, especially when under pressure. It was evident that unless they were effectively maintained the system would

quickly lose its value to the organisation.

Expedite

It was also evident that the discipline of holding daily and weekly buffer meetings on the wards was necessary to both expedite specific patients and deal with common causes of delayed patient discharge. Practice varied between wards and it was evident there was correlation between reduced length of stay and the ward discipline in updating the system and holding buffer meetings. However, these buffer management meetings need to avoid any sense of blame, as there was evidence that this had resulted in discharge dates being extended in some wards to avoid the perceived pressure in such meetings.

Escalate

In one hospital the improvement resulted in a ward closure which was negatively viewed by some ward staff as the remaining beds were more intensively utilised and management were reluctant to open extra beds when the system was under pressure. Some hospitals had started to formalise the escalation procedure and one had recently set up a fast response ward of 20 beds. The ward was fully kitted out and ready to use at less than two hours notice. A team of staff were trained to open the ward in response to previously agreed escalation signals, which include ED buffer signals and other leading indicators. Prior to this, there was no formal preparation for opening extra beds and management only reluctantly approved the opening of a ward because of the cost, and particularly the difficulty in closing it back down again. It was, therefore, necessary to establish clear procedures that manage the closing of wards as well as the opening. This contrasted with the previous practice of cobbling together a mothballed and ill equipped

ward and staffing it at short notice with all the patient risks that entailed. The consequence of not escalating in a timely way typically resulted in patient outliers (patients not being placed on the correct wards) which makes the ongoing length of stay worse and, therefore, the additional beds are needed for longer.

Another illustration of less formal escalation is where an Operations Manager kept the A&E Jonah screen constantly running in the background on his computer and if he saw the red zone growing he would go down without being requested to lend a hand and escalate if necessary.

Improve

When the PDD is exceeded the system captures reasons for delayed discharge every three days and these are analysed periodically in buffer meetings held at the hospital and regional level with other health and social care partners. The regional cross buffer meetings were not in regular use in all the hospitals, but they provide a regional forum to discuss the external causes of delay. In one instance these included waiting for clinical review in an outlying community hospital, but more commonly waiting for care packages. In another situation at a cross buffer meeting the main cause of delayed discharge was ‘awaiting the continuing care decision’ which often meant the family had yet to decide on a nursing home. One of the actions that followed resulted in a communication campaign aimed at encouraging family members to start planning ahead and keeping them informed of the planned discharge date.

A more fundamental result of the cross buffer meetings was a change in attitude of a Social Services manager. Following several of these meetings he admitted to previously not viewing the wider system needs but focusing on protecting his budget.

Summary

These four control functions are clearly evident in the application of time buffer management to health and social care and provide an answer to the research question at one level. However, the nature of these control functions would appear to be distinctly different to the rules associated with kanban control identified earlier. Therefore, to address the research question more fully the following section compares the associated functions in the context of managing patient flow.

6.2 *Relating the kanban control rules and functions*

The six kanban rules and and associated functions (Figure 8) are outlined below and subsequently discussed in relation to the derived functions of time buffer management.

Functions of kanban	Kanban rules of use
1. Provides pick-up or transmission information.	1. Later process picks up the number of items indicated by the kanban at the earlier process.
2. Provides production information.	2. Earlier process produces items in the quantity and sequence indicated by the kanban.
3. Prevents over production and excessive transport.	3. No items are made or transported without a kanban.
4. Serves as a work order attached to goods.	4. Always attached a kanban to the goods.
5. Prevents defective products by identifying the process making the defectives.	5. Defective products are not sent on to the subsequent process. The result is 100% defect free goods.
6. Reveals existing problems and maintains inventory control.	6. Reducing the number of kanban increases their sesitivity.

Figure 8 The functions and rules of kanban (source: Ohno, 1988: 30)

Functions/rules 1, 2 and 4 are concerned with the transfer and production information associated with standard predefined specifications, routings and transfer data.

Function 3 is vital to the lean focus on Just in Time production and ensuring inventory between each work centre is kept to a predefined level.

Function 5 ensures the source of defects is made immediately visible, therefore ensuring rapid problem identification and resolution.

Function 6 enforces continuous improvement. The number of kanbans in the replenishment cycle represents the inventory currently needed to ensure reliable supply. Reducing the number of kanbans reduces the buffer inventory and therefore time, so making the system more sensitive to problems in the drive towards perfection.

Interpreting the kanban functions

Below the four buffer management control functions have been related to these kanban rules/functions with specific reference to patient flow.

Prioritise

Kanban function 1 (F1) effectively pulls work through the system to meet upstream replenishment demands. This in turn results in the need to provide a production instruction (F2) which will be prioritised by the order in which instructions are received with F4 ensuring the routing information is always available These functions are clearly allied to having predefined process routes and intermediate stock that is pulled through the system in line with a level schedule.

In the healthcare environment the flow path may be predefined, especially in elective surgery, and there is clearly potential to emulate the rules and functions of kanban where the flow route is standardized and predictable. This precondition, however, does not apply to time buffer management, although there is a need to instill the discipline of setting planned discharge dates, even if they subsequently move. Synchronization is

central to both approaches but achieved in different ways. In the case of kanban this is achieved through tight control of predefined process steps, whereas time buffer management allows for the dynamic alignment of changing and variable processes. Therefore, time buffer management can be more widely applied and is particularly appropriate where the healthcare process requirements are uncertain.

Expedite

In a kanban system the need to expedite may arise due to delayed replenishment or qualiy problems (F5). Again, due to the centralised nature of the production plan demand variation is not normally an issue and a reactive response (F5) is acceptable due to such disruptions being relatively rare and diminishing over time (Ohno, 1988:41).

In healthcare, demand variation is often not under control and the impact of the variation is unknown due to the varied and changing routes and process times. In this environment it is more important to have a means of proactively expediting specific patients by providing a suitable signal as in the case of red zone penetration.

Escallate

Traditional kanban applications require level production (Ohno, 1988: 39) with significant demand changes typically being centrally planned, therefore, the kanban functions do not explicitly address this need.

In many healthcare environments this assumption is not valid and the system often has to accommodate incremental and sudden increases in demand as in the case of winter influenza. In this situation the system needs to be able to escalate capacity at short notice in response to the changing situation. Therefore, a means of signalling the emerging instability is desirable together with protocols to enable a proactive response.

The segmented use of the time buffers provides a simple means of achieving this and this can be closely related to the management of instability associated with statistical process control.

Improve

A combination of F5 and F6 drives the kanban continuous improvement functionality. Again, the emphasis is on encouraging visibility of problems in combination with fast response to remedy the associated disruption. The removal of kanbans (F6) is indicative of a reduced need for buffering, which is associated with reduced disruptions and/or faster response. In a traditional kanban environment the causes of disruption are visible. However, in health and social care the timescales are longer and due to complex flowpaths there is less visibility of common causes of delay. Therefore, a means of identifying and recording the causes of delay is needed to enable subsequent analysis and targeted improvement as provided in time buffer management.

6.3 *How and why has time buffer management contributed to improved patient flow?*

Having illustrated and discussed how these time buffer management applications have worked in theory and practice let us consider the wider theoretical basis for its success in the context of the kanban functions.

The case evidence showed how the simplified hybrid applications of DBR and CCPM used in these hospitals embody the four control functions. In this environment, as in MTO manufacturing, the need to provide a time based priority system is evident, together with the need to expedite individual requirements and escalate system requirements in the face demand changes. This clearly contrasts with the kanban

functions which assume a standardised planning and stability. Both systems aim to reduce system fluctuations and instability but in the more complex environments capturing this data areguably needs to be more formalised in order that it can be aggregated and used to focus improvement activity. Whereas kanban is traditionally based on physical intermediate stock buffering the use of time is a natural extension of the buffering concept to more complex and dynamic make to order environments.

The essence of the time buffer control functions (prioritise, expedite, escalate and improve) can be conceptually related to statistical process control (Shewhart, 1939; Deming, 1986). That is, the buffer zoning, as with SPC, provides a simple means of interpreting the instability signals and in a flow environment focusing timely expediting, escalation and continuous improvement effort.

It is interesting to note that the growing adoption of lean healthcare has resulted in improvements through activities such as waste walking and the adoption of 5S. However, understanding how to manage patient flow is proving more elusive (Jones, 2008) and the direct kanban equivalent is proving difficult to establish (.Jones and Mitchell, 2006; Zidel, 2006; Fillingham, 2008; Graban, 2009; Baker et al., 2009). It is therefore proposed that time buffer management provides a natural development to kanban control suited to these complex and unstable environments.

7.0 Conclusions

There is increasing pressure worldwide to improve health and social care services and growing interest in how manufacturing developments can be applied. The UK NHS is being driven to improve access through targets, with market competition driving the need to reduce length of stay within the system. One such approach is time buffer management which has been developed and implemented across many hospitals within

the UK and internationally. This paper has used case study evidence to explore how and why emergency and planned time buffer management applications have proved successful at improving patient flow. This commenced, by firstly understanding the logic and control elements present in the established manufacturing based applications of Drum Buffer Rope and Critical Chain Project Management. To support this evaluation, four common control functions were identified (prioritise, expedite, elevate and improve) to provide a framework for evaluation. All these control functions were present in both the health and social buffer management application designs. The implementation evidence also supported the importance of these functions in the effective use of the system. It is apparent that the success of time buffer management is due, in the main, to providing a simple and workable system that embraces these control functions.

The functional comparison with TPS kanban highlights the very different control needs associated with different levels of operations stability. Further research is proposed to investigate how the kanban and time buffer management functions can be combined in supporting the design and evaluation of control systems.

References

Baker, M., Taylor, I and Mitchell, A. (2009), *Making Hospitals Work,* Lean Enterprise Academy, Goodrich, UK.

Bhasin, S. and Burcher, P.(2006), "Lean viewed as a philosophy", *Journal of Manufacturing Technology Management,* Vol.17, pp. 56-72.

Deming, W.E. (1986), *Out of the Crisis,* MIT, Cambridge, MA.

Eisenhardt, K. M. (1989), "Building Theories from Case Study Research", *Academy of Management Review,* Vol.14 No 4, pp.532-550.

Fillingham, D., (2008), *Lean healthcare.* Kingsham Press, UK.

Goldratt, E.M. (1990), *Theory of Constraints.* North River Press, Great Barrington, MA.

Goldratt, E.M. (1997), *Critical Chain.* North River Press, Great Barrington, MA.

Goldratt, E.M. (2007), *The Goldratt Webcast Program on Project Management Viewer Notes.* V 4.71. GMG: Amsterdam.

Graban, M., (2009), *Lean Hospitals*, Productivity Press, New York,NY.

Herck, van P., Vanhaecht, K. And Sermeus, W, (2004) "Effects of clinical pathways: do they work?", *Journal of Integrated Care Pathways*, Vol.8, pp 95-105.

Jones, D., Mitchell, A. (2006), *Lean Thinking for the NHS*, NHS Confederation, London.

Jones, T.J. (2008), "Lean Enterprise Academy News Letter – 18 August", available at www.leanuk.org (accessed 13 November 2008).

Ohno, T. (1988), *The Toyota Production System; Beyond Large-Scale Production.* Productivity, Portland, OR.

Press Ganey Associates Inc (2007), *The Emergency Department Pulse Report: Patient Perspectives on American Health Care*, South Bend, IN.

Proudlove, N., Moxham, C. and Boaden, R. (2008), "Lessons for Lean in Healthcare from Using Six Sigma in the NHS", *Public Money and Management*, Feb, pp 27- 34.

Schmenner, R.W., and Swink, M.L. (1998), "On theory in operations management". *Journal of Operations Management*, 17, pp 97-113.

Schragenheim, E., and Detmer, W. (2001), *Manufacturing at Warp Speed.* CRC Press, Boca Raton, FL..

Shewart, W.A. (1931), *Economic Control of Quality of Manufactured Product*, Van Nostrand, New York, NY.

Shewart, W.A. (1939), *Statistical Method from the viewpoint of Quality Control*.

Graduate School of the Department of Agriculture, Washington, DC.

Steyn, H. (2000), "An investigation into the fundamentals of critical chain project scheduling", *International Journal of Project Management* Vol. 19, pp363-369.

Stratton, R., Robey D. and Allison I. (2008), "Utilising Buffer Management to Manage Uncertainty and Focus Improvement", *Proceedings of the International Annual Conference of EurOMA*, Groningen, the Netherlands (June).

Stratton, R. (2009), "Critical Chain Project Management – Theory and Practice", *Proceedings of the 20th POMS Conference, Orlando, May 1st-4th.*

Umble, M.M., and Srikanth, M.L.,(1997), *Synchronous Management: Profit based manufacturing for the 21st Century, Volume two: Implementation Issues and Case Studies.* Spectrum Publishing Company, Guilford, CT.

Umble, M.M. and Umble, E.J. (2006), "Utilising buffer management to improve performance in a healthcare environment", *European Journal of Operational Research* Vol. 174, pp. 1060-1075.

Walley, P., and Gowland, B. (2004), "Completing the cycle: from P.D. to P.D.S.A.", *International Journal of Health Care Quality Assurance*, Vol.17, No 6, pp. 349-358.

Yin, R.K. (1994), *Case Study Research – Design and Methods.* 2nd Ed., SAGE, London.

Zidel, T.G. (2006), *A Lean Guide To Transforming Healthcare,* ASQ Quality Press, Milwaukee, Wisconsin.